

Université de Montréal

**New statistical methods to assess the effect of time-dependent
exposures in case-control studies**

**par
Zhirong Cao**

Faculté de médecine

**Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de maîtrise
en sciences biomédicales**

Décembre, 2008

© Zhirong Cao, 2008

**Université de Montréal
Faculté des études supérieures**

Ce mémoire intitulé

**New statistical methods to assess the effect of time-dependent
exposures in case-control studies**

**présenté par
Zhirong Cao**

a été évalué par un jury composé des personnes suivantes :

Président-rapporteur :	Jean Lambert
Directeur de recherche :	Karen Leffondré
Membre du jury:	Andrea Benedetti

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor, Dr. Karen Leffondré, who gave me endless support, intense supervision, and patience. She has always been encouraging and helping me through my graduate study. I would not have completed my study without her help.

I would like to thank Willy Wynant. I thank you for your friendly advices, encouragement and reviewing my thesis.

Thanks are due to Dr. Jack Siemiatycki for his insightful discussion in epidemiology.

I would also like to thank Jerome Asselin for his consistently immediate assistance with technical details and data management issues.

I would like to acknowledge the financial support in the form of scholarships from the National Sciences and Engineering Research Council of Canada (NSERC), Canadian Institutes for Health Research (CIHR), and the Fonds de la Recherche en Santé du Québec (FRSQ).

Last, but not least, I would like to dedicate this thesis to my parents, my husband and my children, for their love, patience, and understanding—they allowed me to spend most of the time on this thesis.

Résumé et mots clés

Contexte. Les études cas-témoins sont très fréquemment utilisées par les épidémiologistes pour évaluer l'impact de certaines expositions sur une maladie particulière. Ces expositions peuvent être représentées par plusieurs variables dépendant du temps, et de nouvelles méthodes sont nécessaires pour estimer de manière précise leurs effets. En effet, la régression logistique qui est la méthode conventionnelle pour analyser les données cas-témoins ne tient pas directement compte des changements de valeurs des covariables au cours du temps. Par opposition, les méthodes d'analyse des données de survie telles que le modèle de Cox à risques instantanés proportionnels peuvent directement incorporer des covariables dépendant du temps représentant les histoires individuelles d'exposition. Cependant, cela nécessite de manipuler les ensembles de sujets à risque avec précaution à cause du sur-échantillonnage des cas, en comparaison avec les témoins, dans les études cas-témoins. Comme montré dans une étude de simulation précédente, la définition optimale des ensembles de sujets à risque pour l'analyse des données cas-témoins reste encore à être élucidée, et à être étudiée dans le cas des variables dépendant du temps.

Objectif: L'objectif général est de proposer et d'étudier de nouvelles versions du modèle de Cox pour estimer l'impact d'expositions variant dans le temps dans les études cas-témoins, et de les appliquer à des données réelles cas-témoins sur le cancer du poumon et le tabac.

Méthodes. J'ai identifié de nouvelles définitions d'ensemble de sujets à risque, potentiellement optimales (le *Weighted Cox model* and le *Simple weighted Cox model*), dans lesquelles différentes pondérations ont été affectées aux cas et aux témoins, afin de refléter les proportions de cas et de non cas dans la population source. Les propriétés des estimateurs des effets d'exposition ont été étudiées par simulation. Différents aspects d'exposition ont été générés (intensité, durée, valeur cumulée d'exposition). Les données cas-témoins générées ont été ensuite analysées avec différentes versions du modèle de Cox, incluant les définitions anciennes et nouvelles des ensembles de sujets à risque, ainsi qu'avec la régression logistique conventionnelle, à des fins de comparaison. Les différents modèles de régression ont ensuite été appliqués sur des données réelles cas-témoins sur le cancer du poumon. Les estimations des effets de différentes variables de tabac, obtenues avec les différentes méthodes, ont été comparées entre elles, et comparées aux résultats des simulations.

Résultats. Les résultats des simulations montrent que les estimations des nouveaux modèles de Cox pondérés proposés, surtout celles du *Weighted Cox model*, sont bien moins biaisées que les estimations des modèles de Cox existants qui incluent ou excluent simplement les futurs cas de chaque ensemble de sujets à risque. De plus, les estimations du *Weighted Cox model* étaient légèrement, mais systématiquement, moins biaisées que celles de la régression logistique. L'application aux données réelles montre de plus grandes différences entre les estimations de la régression

logistique et des modèles de Cox pondérés, pour quelques variables de tabac dépendant du temps.

Conclusions. Les résultats suggèrent que le nouveau modèle de Cox pondéré propose pourrait être une alternative intéressante au modèle de régression logistique, pour estimer les effets d'expositions dépendant du temps dans les études cas-témoins

Mots clés. Modèle de Cox pondéré, variables dépendant du temps, étude cas-témoins, régression logistique, exposition cumulée, intensité d'exposition, simulation, tabac, cancer.

Summary and keywords

Background: Case-control studies are very often used by epidemiologists to assess the impact of specific exposure(s) on a particular disease. These exposures may be represented by several time-dependent covariates and new methods are needed to accurately estimate their effects. Indeed, conventional logistic regression, which is the standard method to analyze case-control data, does not directly account for changes in covariate values over time. By contrast, survival analytic methods such as the Cox proportional hazards model can directly incorporate time-dependent covariates representing the individual entire exposure histories. However, it requires some careful manipulation of risk sets because of the over-sampling of cases, compared to controls, in case-control studies. As shown in a preliminary simulation study, the optimal definition of risk sets for the analysis of case-control data remains unclear and has to be investigated in the case of time-dependent variables.

Objective: The overall objective is to propose and to investigate new versions of the Cox model for assessing the impact of time-dependent exposures in case-control studies, and to apply them to a real case-control dataset on lung cancer and smoking.

Methods: I identified some potential new risk sets definitions (the *weighted Cox model* and the *simple weighted Cox model*), in which different weights were given to cases and controls, in order to reflect the proportions of cases and non cases in the source population. The properties of the estimates of the exposure effects that result from these new risk sets definitions were investigated through a simulation study.

Various aspects of exposure were generated (intensity, duration, cumulative exposure value). The simulated case-control data were then analysed using different versions of Cox's models corresponding to existing and new definitions of risk sets, as well as with standard logistic regression, for comparison purpose. The different regression models were then applied to real case-control data on lung cancer. The estimates of the effects of different smoking variables, obtained with the different methods, were compared to each other, as well as to simulation results.

Results: The simulation results show that the estimates from the new proposed weighted Cox models, especially those from the *weighted Cox model*, are much less biased than the estimates from the existing Cox models that simply include or exclude future cases. In addition, the *weighted Cox model* was slightly, but systematically, less biased than logistic regression. The real life application shows some greater discrepancies between the estimates of the proposed Cox models and logistic regression, for some smoking time-dependent covariates.

Conclusions: The results suggest that the new proposed weighted Cox models could be an interesting alternative to logistic regression for estimating the effects of time-dependent exposures in case-control studies.

Keywords: Weighted Cox model, time-dependent variables, case-control study, logistic regression, cumulative exposure, exposure intensity, simulation, smoking, cancer.

Table of contents

Acknowledgements.....	iii
Résumé et mots clés.....	iv
Summary and keywords.....	vii
Table of contents.....	ix
List of tables.....	xi
List of figures.....	xiii
List of acronyms and abbreviations	xiv
1 Introduction.....	1
2 Literature review	4
2.1 Overview of epidemiological study design.....	5
2.2 Overview of logistic regression	7
2.3 Overview of survival analysis.....	9
2.3.1 Survival and hazard functions.....	9
2.3.2 Parametric survival models.....	10
2.3.3 The Cox semi-parametric model.....	12
2.3.4 Adaptation of Cox's model to case-control study.....	15
3 Objectives	19
4 Methods.....	22
4.1 New proposed Cox models	23
4.2 Simulation study	27
4.2.1 Overview.....	27
4.2.2 Generation of the source population	27
4.2.3 Simulation of case-control studies	30
4.2.4 Summary of the different scenarios investigated.....	31

4.2.5	Data analytical models.....	33
4.2.6	Summary statistics to evaluate the performance of the different analytical models	34
4.3	Real data analysis:.....	35
4.3.1	Data source.....	35
4.3.2	Method to handle missing smoking data	36
4.3.3	Description of smoking intensity trajectories over time.....	39
4.3.4	Data analytical models.....	41
5	Results.....	49
5.1	Results from simulations.....	50
5.2	Results from real data analysis	54
5.2.1	Description of real data.....	54
5.2.2	Patterns of smoking intensity.....	56
5.2.3	Results from regression models.....	58
6	Conclusion and discussion.....	66
7	References.....	71
8	Appendix.....	75
8.1	Questionnaire on smoking	76
8.2	Simulation results from Leffondré 2003.....	77

List of tables

Table I. Characteristics of the Exponential, Weibull, and Gompertz distributions ...	11
Table II: definitions of weights in risk sets of models used for subject j at risk at age t_i :	26
Table III : Summary of simulation scenarios in Model 1, hazard depended on intensity and past duration separately.....	32
Table IV: Summary of simulation scenarios for Model 2, where the hazard depended on the value cumulative exposure*.....	32
Table V: Summary of smoking variables in the case-control study, Montréal, Quebec, Canada, 1996-2001.	36
Table VI: Comparison of the distribution of smoking intensities (n, mean, standard deviation) before and after imputation to handle missing smoking intensity at each of the four ages (25, 40, 50, 60 years) in the case-control study, Montréal, Quebec, Canada, 1996-2001.	38
Table VII : interpretation of logged Bayes factor ($2 \Delta BIC$) for model selection	41
Table VIII: Summary of sub-datasets from a case-control study, Montréal, Quebec, Canada, 1996-2001.	42
Table IX: Summary of the smoking models used to analyse the data from the case- control study, Montréal, Quebec, Canada, 1996-2001.	45
Table X: lifetime probability of developing lung cancer and probability of developing lung cancer within the next 10 years by age group, Canada.....	46
Table XI: Age-conditional probabilities of developing lung cancer in the future and the weights for each age categorical for weighted Cox model in the analysis of Montreal case-control study.....	47
Table XII: Results from simulations of the proposed Cox models and conditional logistic regression (CLR) for estimating the effects β of intensity $I(t)$ and duration $D(t)$ (Model 1), based on the 1000 simulations.	52
Table XIII: Results from simulations of all the Cox models and conditional logistic regression (CLR) for estimating the effect β of cumulative exposure† (Model 2), based on 1000 simulations	53

Table XIV: Demographic characteristics of subjects at the time of diagnosis/interview, Montréal, Quebec, Canada, 1996-2001.....	54
Table XV: Smoking-related characteristics of subjects in a case-control study of environmental exposure and cancer at the time of diagnosis/interview, Montréal, Quebec, Canada, 1996-2001.....	55
Table XVI: Percentages of patterns of intensity change over-time for current smokers in a case-control study, Montréal, Quebec, Canada, 1996-2001.....	58
Table XVII: Smoking effect estimates from the Cox models and standard unconditional logistic regression (LR), using Model 1 in current smokers, Montréal, Quebec, Canada, 1996-2001.....	63
Table XVIII: Smoking effect estimates from the Cox models and standard unconditional logistic regression (LR), using Model 2, Montréal, Quebec, Canada, 1996-2001.....	64
Table XIX: Smoking effect estimates from the Cox models and standard unconditional logistic regression (LR), using Models 3 and 4, Montréal, Quebec, Canada, 1996-2001.....	65
Table XX: Questions that were asked in the questionnaire regarding smoking history in the case-control study of lung cancer, Montréal, Quebec, Canada, 1996-2001.....	76
Table XXI. Mean of the 1,000 estimates, corresponding confidence interval, and relative bias for the adapted Cox model and conditional logistic regression.....	77

List of figures

Figure 1: Different patterns of smoking intensity change over time for three hypothetical subjects who have the same value of cigarette-years.....	3
Figure 2 : Exposure pattern of a hypothetical subject j diagnosed or selected at age t_j , with an increasing intensity over lifetime.....	28
Figure 3: The intensity at different age represented by the intensity at the reported age.....	36
Figure 4: make up of the intensity at some age with missing data.....	38
Figure 5: Trajectory of intensity for current male smokers in a case-control study, Montréal, Quebec, Canada, 1996-2001.	56
Figure 6: Trajectory of intensity for current female smokers in a case-control study, Montréal, Quebec, Canada, 1996-2001.	57

List of acronyms and abbreviations

OR	odds ratio
HR	hazards ratio
PH	proportional hazards
SE	standard errors
SD	standard deviation
CI	confidence interval
BIC	Bayesian information criterion
CSI	comprehensive smoking index
CCDPC	Centre for Chronic Disease Prevention and Control
CLR	Conditional logistic regression
LR	logistic regression

1 Introduction

Case-control studies, which consist in sampling subjects who have the disease of interest (the cases) and subjects free of disease (the controls) at the time of study, are very often used by epidemiologists to assess the impact of specific past exposure(s) on the disease of interest[1]. Many of these exposures such as smoking history may be represented by several time-dependent covariates [2], and new methods are needed to accurately assess their effects. Indeed, conventional logistic regression, the standard method to analyze case-control data, does not directly account for changes in covariate over time. For example, in studies investigating the impact of smoking, smoking history is often represented in the regression model by a cumulative exposure variable, cigarette-years, which is the product of 1) the average smoking intensity over lifetime, and 2) the total duration of smoking at the time of diagnosis for cases and at the time of interview for controls. Such a cumulative variable does not allow discrimination between subjects who indeed have cumulated the same amount of smoking, but who smoked with different patterns of intensity over lifetime, as illustrated with the three different hypothetical patterns in Figure 1. In this figure, the three subjects have cumulated 600 cigarette-years over their lifetime, but with very different patterns of intensity over time. Using cigarette-years in a standard logistic regression model to represent smoking history would assume that these three subjects have exactly the same risk of developing the disease at the age of 60 years, with respect to their past smoking consumption. However, since the effect of cigarettes smoked t years ago is likely to decrease as t increases [3], one could assume that subject B with increasing intensity has probably a higher risk at the age of 60,

compared to subject C who has decreasing intensity. By contrast, at ages earlier than 45 years, one could assume that subject B had a lower risk than subject C.

Such patterns over lifetime are difficult to model in a standard logistic regression model. One possibility could be using different variables to represent different periods of past exposures. However, one would have then to face with the problem of arbitrary choice of these periods, and potential multi-collinearity between the resulting covariates. By contrast, survival analytic methods such as the Cox proportional hazards (PH) model [4] can directly incorporate time-dependent covariates representing the individual entire exposure history. However, these methods were originally proposed for prospective data, and their application to retrospective designs requires some careful manipulation of risk sets [5, 6]. The optimal definition of risk sets for the analysis of case-control data remains unclear and has to be investigated in the case of time-dependent variables [7].

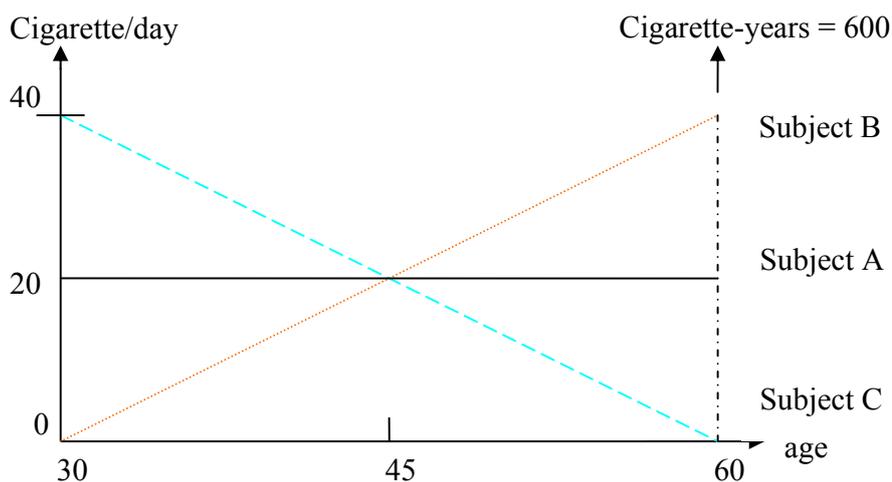


Figure 1: Different patterns of smoking intensity change over time for three hypothetical subjects who have the same value of cigarette-years.

2 Literature review

2.1 Overview of epidemiological study design

Two important types of epidemiological studies for generation hypotheses are cohort studies and case-control studies. The principles of these two epidemiological designs are briefly outlined below (for more details see e.g. [8] and [9]).

Cohort studies

In a cohort study, a group of individuals who are exposed to a certain condition are followed over time, and compared with another group unexposed to that condition. One measure of interest is the incidence rate, which is number of new disease cases per population at risk. The incidence rates for exposed and non-exposed subjects are calculated separately. The relative risk is the ratio of the incidence rate of exposed subjects to non-exposed subjects and is used to measure the association between exposure and disease.

Population based case-control studies

In a case-control study, subjects with the event/disease (cases) are selected, and then the history of exposure and/or other characteristics is recorded by interview or any other sources. A comparison group (control group) of subjects without the event/disease is assembled from the source population, and the history of controls is recorded as well as for cases. In these studies, individuals with the disease (the cases) are over-sampled. Thus, the percentage of people who has disease is greater in the study than in the source population. The percentage of subjects with the disease in the source population can not be estimated from a case-control study, so the relative risk can not be estimated. However, the odds ratio (OR) can always be estimated in a

case-control study. For a rare disease, the odds ratio (OR) approximates relative risk. Matching between cases and controls is sometimes performed in case-control studies. This ensures that the matching factors, such as age or sex, are equally distributed between cases and controls [10]. Both individual matching (one or more controls matched to each specified case) and frequency matching (groups of controls matched to groups of cases) are used.

Nested case-control studies

Since collecting covariate information for all individuals in a cohort may be very expensive, and time-consuming, Langholz [5] shows that the nested case-control design, which is cohort sampling design, is a useful alternative to a full cohort study. Nested case-control studies are case-control studies done in the population of an ongoing cohort study. The case-control study is thus said to be “nested” inside the cohort study. In the nested case-control design, each case is compared to a small sample of individuals (controls) selected from the risk set at the time of case’s event. The collection information includes all cases and only the sampled controls.

Advantages and disadvantages of cohort and case-control studies

Cohort studies allow complete information on the subject’s exposure(s), and estimate incidence rates as well as relative risk. Since cohort studies need a large number of subjects to follow up, they are expensive to carry out and are not suited for rare diseases study. Case-control studies are relatively inexpensive as compared with

cohort studies, so they permit the studies of rare diseases. However, the information on past exposure is usually based on interview, and may be prone to measurement errors because of recall bias. Nevertheless, many case-control studies collect huge information on past exposures, and appropriate statistical methods are needed to account for all this information.

While cohort studies are often analysed using survival analyses methods, case-control data are almost always analysed using logistic regression. In the following sections, I briefly present these two analytical approaches.

2.2 Overview of logistic regression

Logistic regression is a part of generalized linear models [11]. Logistic regression allows the investigation of the relationship between a binary response variable Y , such as presence/absence or success/failure, and a set of explanatory variables X . The outcome variable Y follows a Bernoulli distribution with a probability of “success” $P(Y=1|x) = \pi(x)$, which is given by :

$$\pi(x) = \frac{\exp(\alpha + \beta' X)}{1 + \exp(\alpha + \beta' X)} \quad (2.1)$$

where α is the intercept and β is the vector of regression coefficients. Equation 2.1 can also be written:

$$\text{logit}\pi(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \alpha + \beta' X = \alpha + \sum_{j=1}^p \beta_j x_j \quad (2.2)$$

which shows that the logistic regression model assumes that the effects of all X_j are linear on the logit of π . For each X_j , $\exp(\beta_j)$ corresponds to the OR for X_j . The latter result explains the popularity of logistic regression in epidemiology.

Let y_i be the indicator, taking the value 1 if the subject i has the event, and 0 otherwise. The regression coefficients are estimated by maximizing the likelihood function:

$$L(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (2.3)$$

Note that Prentice showed that we can analyze retrospective data as if they were prospective [12].

Conditional logistic regression

Conditional logistic regression is used to analyse individually matched case-control studies. Conditional logistic regression works in nearly the same way as regular logistic regression, except one needs to specify which individuals belong to which matched set (i.e. strata).

The theory is similar: one can derive a conditional likelihood and maximize it. From a practical perspective, the only difference is the need to specify the matched set to which each person belongs. When each matched set consists of a single case and a single control, the conditional likelihood is given by:

$$L(\beta) = \prod_i \frac{\exp[\beta'(x_{i1} - x_{i0})]}{1 + \exp[\beta'(x_{i1} - x_{i0})]}, \quad (2.4)$$

where x_{i1} and x_{i0} are the vectors of explanatory variables of the case and the control, respectively, of the i^{th} matched set.

2.3 Overview of survival analysis

Survival analysis is a popular data analytical approach for studies in which outcome variable of interest is time until an event occurs (referred to as time-to-event). Time means years, months, or days from the beginning of follow-up of a subject until an event occurs; alternatively, time can also refer to the age of a subject when an event occurs. Event means death, disease incidence, recovery (e.g. return to work) or any designated experience of interest that may happen to a subject.

Censoring is a key analytical problem in survival analysis, since one may ignore the survival time for some subjects. Censoring may occur when a person does not experience the event before the study ends, or a person is lost to follow-up during the study period. There are three types of possible censoring schemes, right censoring, interval censoring and left censoring. All along this thesis, I consider only the case of right censoring.

2.3.1 Survival and hazard functions

Let the random variable T denotes the survival time, and $f(t)$ its density probability function:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \quad (2.5)$$

$F(t)$ denotes the cumulative distribution function

$$F(t) = P(T \leq t). \quad (2.6)$$

The survival function $S(t)$ is therefore written

$$S(t) = P(T > t) = 1 - F(t). \quad (2.7)$$

The hazard function $h(t)$ is given by

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \quad (2.8)$$

$h(t)\Delta t$ is the probability to have the event of interest between t and $t + \Delta t$, conditionally on being still at risk at time t , i.e. not having experienced the event before t . The cumulative hazard function is finally given by

$$H(t) = \int_0^t h(u) du. \quad (2.9)$$

One can show the relationship

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d \log(S(t))}{dt}. \quad (2.10)$$

2.3.2 Parametric survival models

When using parametric survival models, one assumes that the survival time variable T follows a known distribution that depends on a finite number of unknown parameters. The most popular distributions are the Exponential distribution, the Weibull distribution and the Gompertz distribution. Table I gives their characteristics [13].

Maximum likelihood estimation is used to estimate the unknown parameters of the parametric distributions. Let t_i be the survival or censoring time of subject i , and d_i the indicator, taking the value 1 if the subject i has the event at time t_i , and 0 otherwise. The likelihood function under general non-informative censoring has the form

$$L(\theta) = \prod_{i=1}^n f(t_i|x_i)^{d_i} S(t_i|x_i)^{(1-d_i)} \quad (2.11)$$

and in general must be maximized numerically using a procedure such as Newton-Raphson.

Table I. Characteristics of the Exponential, Weibull, and Gompertz distributions [13].

Characteristic	Distribution		
	Exponential	Weibull	Gompertz
Parameter	Scale parameter $\lambda > 0$	Scale parameter $\lambda > 0$ Shape parameter $v > 0$	Scale parameter $\lambda > 0$ Shape parameter $\alpha \in (-\infty, \infty)$
Hazard function	$h_0(t) = \lambda$	$h_0(t) = \lambda v t^{v-1}$	$h_0(t) = \exp(\alpha t)$
Cumulative hazard function	$H_0(t) = \lambda t$	$H_0(t) = \lambda t^v$	$H_0(t) = \frac{\lambda}{\alpha} (\exp(\alpha t) - 1)$
Density function	$f_0(t) = \lambda \exp(-\lambda t)$	$f_0(t) = \lambda v t^{v-1} \exp(-\lambda t^v)$	$f_0(t) = \lambda \exp(\alpha t) \exp\left(-\frac{\lambda}{\alpha} (\exp(\alpha t) - 1)\right)$
Survival function	$S_0(t) = \exp(-\lambda t)$	$S_0(t) = \exp(-\lambda t^v)$	$S_0(t) = \exp\left(-\frac{\lambda}{\alpha} (\exp(\alpha t) - 1)\right)$

2.3.3 The Cox semi-parametric model

Since it is difficult to specify a priori correct assumption concerning the nature or the shape of the underlying survival distribution as required with parametric approaches, the Cox semi-parametric model is rather used in practice. This proportional hazards (PH) model, which was introduced by Cox [4], is the most popular regression model used for analyzing survival data. It consists in a product of a term depending on time and a term depending on the covariates

$$h(t|x) = h_0(t) \exp(\beta'x) \quad (2.12)$$

where $X = X_1, X_2, \dots, X_p$ is the vector of the p explanatory/predictor variables, and $h_0(t)$ the baseline hazard function, which does not have to be specified. The Cox model is a semi-parametric model because of the non parametric function $h_0(t)$ and the parametric function $\exp(\beta'x)$.

The most important feature of the model is the proportional hazards (PH) assumption. Let consider two subjects i and j who differ in their covariates value for x_l , but have the same value for other covariates. The hazard ratio for x_l between these two subjects will be

$$HR = \exp[\beta_1(x_{i,l} - x_{j,l})], \quad (2.13)$$

which is independent of time t . The hazards ratio remains therefore constant over time. That is why the Cox model is called the PH model. Note that the variables can be time-dependent as explained below. What can't be time-dependent with the PH assumption is their effect β .

Partial likelihood

The regression parameters in the Cox PH model are estimated by maximizing a partial likelihood, developed by Cox [4, 14]. The Cox partial likelihood is based on the observed order of events and is constructed by comparing the hazard of the subject i who fails at time t_i to the hazard of all subjects j at risk just before that time t_i . The Cox partial likelihood is given by

$$L(\beta) = \prod_{i=1}^k \frac{\exp(\beta' x_i)}{\sum_{j \in R_i} \exp(\beta' x_j)} \quad (2.14)$$

where R_i is the risk set at time t_i made of the group of subjects at risk for failure just before t_i . The above equation is correct if no ties occur at any of the failure times, i.e. each failure occurs at a distinct time. We can use either Breslow [15] or Efron [16] approximations if there are ties in the data set, since the exact method can be time-consuming.

Extension of the Cox PH model for time-dependent variables

The values of some covariates, such as intensity of exposure, may change over time for some subjects. Covariates can thus be either fixed or time-dependent. It is possible to incorporate time-dependent covariates in the Cox model:

$$h(t|x(t)) = h_0(t) \exp[\beta' x(t)] \quad (2.15)$$

Note that such a model no longer satisfies the “strict” PH assumption and is sometimes called extended Cox model, see Fisher [17]. However, even if a covariate is time-dependent, its effect β may be constant over time, which is in agreement with

the PH assumption. The regression coefficients can be estimated through the time-dependent partial likelihood [18].

$$L(\beta) = \prod_{i=1}^k \frac{\exp(\beta' x_i(t_i))}{\sum_{j \in R_i} \exp(\beta' x_j(t_i))} \quad (2.16)$$

Choice of time-scale

In Cox's model, the time axis (or time-scale) T , has to be defined when we build the risk sets. Many authors have frequently used time-on-study or calendar time as the time-scale, because Cox's model was developed originally for clinical applications, where individuals are generally followed up since the initiation of treatment until death. However, a previous simulation study [19] that simulated cohort data, showed positive bias for time-dependent covariates when using time-on-study as the time-scale, and when the baseline hazard was not an exponential function of age [19]. By contrast, no bias was observed when using age as the time-scale.

In many epidemiological studies, the effect of age needs to be tightly controlled because the incidence of the disease of interest depends strongly on age. This is the case for example in aging research. In such situations, using age as the time-scale allows direct adjustment for age, without having to model its effect with any linear function or arbitrary categorization. Note that in aging research, using age as the time-scale often requires left truncation [20]. For example, if the objective of the study is to investigate the risk of dementia as a function of age, the age at dementia

should be left-truncated if subjects enter the study at different ages (delayed entry) and are included in the study only if they are free of dementia at the age at entry into the study. The subjects are considered to be at risk of dementia only from the age at entry into the study. Another reason to prefer age as the time-scale in some studies is when the beginning of follow-up is not defined by any relevant meaningful specific event but just corresponds to the entry into the study. In such situations, the most natural time-scale is age rather than time-on-study. Similarly, in the context of case-control analyses, age is the most natural time scale.

2.3.4 Adaptation of Cox's model to case-control study

As I presented before, case-control designs are widely used in epidemiology, especially for rare disease studies. Logistic regression is the common tool to estimate the effect of covariates on the risk of disease in such studies but it does not efficiently use the information on covariates which varies over time. The Cox model with time-dependent covariate can handle such exposures that may vary over time. But the question is how to manipulate the risk sets when the Cox model is applied to such case-control data.

Dr. Leffondré's simulation study [7] investigated how the accuracy of the point estimates of Cox's model depends on the operational definition of risk sets and/or on some aspects of the time-varying exposure. This study simulated a hypothetical population-based matched case-control study that was conducted in years 1995-2000, among subjects aged 55-69 years. A population of size $N=3000$ subjects was first

generated, with 200 subjects born each year from 1931 to 1945. A permutation algorithm [21, 22] was used to generate survival times conditional on time-dependent covariates. Two different types of time-dependent covariates were investigated: one continuous covariate representing the duration of exposure, and one binary covariate representing the current status of exposure. In addition to these time-dependent covariates, each true Cox model included a binary fixed-in-time covariate, representing for example the sex of subjects. The cases included all subjects who had an event between years 1995 and 2000. For each case, a single control was selected among subjects who were still at risk at the age of case's event and were born the same year as the case.

Three alternative definitions of modified risk sets (\tilde{R}) were considered in the Cox partial likelihood (2.15). In the first risk sets definition, \tilde{R}_i included the case failed at time t_i and only his matched control who was randomly selected at that time t_i . The partial likelihood resulting from that definition of risk sets is actually equivalent to the usual conditional logistic likelihood for 1-1 matched case-control data [6]. In the second risk sets definition, \tilde{R}_i included the case failed at time t_i , his matched control at t_i , and all future cases and their controls failed or selected after t_i . The model resulting from this risk set definition was referred to the 'naïve Cox model'. Since this model was likely to induce an under-estimation bias because of over-representation of future cases in each risk set, another risk set definition was considered. The third definition of risk set \tilde{R}_i , which was referred to the 'adapted

Cox model', excluded all future cases from the 'naïve Cox model'. Thus, while the naïve Cox model used entire covariate history for all cases and controls, the adapted Cox model used the entire covariate history for controls only. By contrast, in the conditional logistic analysis, only one value per covariate (assessed at the time of event for cases and at the time of selection for controls) contributed to the analysis.

In this simulation study, several scenarios were investigated. As expected, the results showed that the naïve Cox model induced a systematic serious under-estimation of all the regression parameters by 40%-50%, depending on the scenarios. By contrast, the adapted Cox model induced a systematic over-estimation of all effects, but the amount of relative bias was much smaller than the naïve Cox model. Indeed, for the fixed-in-time covariate, the amount of relative bias varied within 3%-17%, depending on the scenarios. (See Table XXI in the Appendix). For the time-dependent binary covariate representing the current exposure status, the relative bias was equal to 7%-13%, while for the continuous covariate representing the duration of exposure, it was equal to 12%-22%.

Logistic regression yielded quite accurate results for the effect of the fixed in-time covariate (sex), and for the effects of the two time-dependent covariates when only one of them was included in the model (Scenarios 1-4, Table XXI in the Appendix). However it over-estimated the effect of the time-dependent covariates when both of them were included in the model and when at least one of them had a weak effect (Scenarios 5, 7, 8, Table XXI in the Appendix). For example, it under-estimated the

weak effect of duration by as much as 27%-38% in the models that adjust for current exposure. This bias was much stronger than that from the adapted Cox model. It seems, therefore, that logistic regression has difficulties in separating the impact of current exposure from that of exposure duration, over-estimating the former and under-estimating the latter. The superiority of the adapted Cox model over logistic regression might be due to the fact that the adapted Cox model used additional information on past values of both variables among controls.

Overall, the results of this simulation study suggested that a better definition of risk set for the Cox analysis of case-control data should be intermediate between those from 1) the Naïve Cox model which included all future cases and systematically under-estimated the effects of exposure, and 2) the adapted Cox model which excluded future cases and systematically over-estimated the effects.

3 Objectives

The literature review, presented in Chapter 2, indicates that the conventional logistic model, which is widely used in case-control analyses, does not directly account for changes in the covariate values over time. On the other hand, some adaptations of Cox's model for case-control studies still needed to be improved and evaluated using simulated and real data. Thus, in this thesis, I attempted to propose and validate new versions of Cox models for case-control data. I expected that this approach would yield better results than the conventional logistic regression model and the recently proposed "adapted Cox model" [7] (see section 2.3.4), for estimating the effect of time-dependent variables in case-control studies.

To this aim, the following specific objectives were addressed:

- To propose new weighted Cox models for analysing case-control data with time-dependent exposures;
- To compare their point estimates to that from conventional logistic regression, for estimating the effects of different time-dependent aspects of smoking history on the risk of lung cancer, using data from a case-control study undertaken in Montreal.
- To interpret the results of the real data analysis in light of those obtained in a simulation study.

The simulation study was conducted in parallel to my thesis, by Willy Wynant, under the supervision of Dr. Karen Leffondré. Although I did not do the programming of

this simulation study, I contributed to all the discussions about its development. Moreover, since the results of this simulation study are essential to understand and interpret my own results on real data, I decided to incorporate them in the main body of my thesis.

4 Methods

I first identified some potential new risk sets definitions, which were adapted from the Naïve and the adapted Cox models proposed in Leffondré et al (2003) [7]. The properties of the estimators of the exposure effects that resulted from these new risk sets definitions were then investigated using data from a real case-control study on lung cancer. The exposure of interest all along this thesis was smoking history, which may be represented by various time-dependent covariates (e.g. intensity, duration, cumulative exposure). The case-control data were analysed using different versions of Cox's models corresponding to existing and new definitions of risk sets, as well as with logistic regression model, for comparison purpose. The results from the real data analysis were compared to those from a simulation study.

The method section is organized as follows. Section 4.1 presents the new proposed Cox models, section 4.2 presents the methods used to generate the data for the simulation study, and Section 4.3 presents the methods used to analyze the real data. The simulation study is presented before the real data analysis in order to further help the interpretation of the real data results.

4.1 New proposed Cox models

The Cox models that I propose in this thesis are based on the results of the two previous versions investigated in Leffondré et al. [7], i.e. the naïve Cox model and the adapted Cox model. As mentioned in Section 2.3.4, the naïve version, which included at each failure time all future cases and all future controls, systematically under-

estimated the effect of all covariates. By contrast, the adapted version, which excluded future cases as opposed to the Naïve one, induced a systematic over-estimation.

I propose two new definitions of risk sets $R(t_i)$ for the Cox models, which are intermediate between the naïve version and the adapted version. The general principle is to apply different weights to cases and controls in each risk set, such that the new risk sets reflect the actual composition of the full (unknown) risk set of the source population.

Suppose N_i (unknown) subjects are at risk of developing the disease at age t_i in the source population and p_i is the age-conditional probability to develop that disease at age t_i or at a later age in that population. Among those N_i subjects at age t_i , $N_i p_i$ subjects will develop the disease at age t_i or later (the population cases), and $N_i(1-p_i)$ subjects will never develop the disease at age t_i or later (the *pure* population controls). The case:pure control ratio at age t_i in the source population is therefore given by $p_i : (1-p_i)$ or

$$1 : \frac{1-p_i}{p_i} \quad (4.1)$$

Denote $n_{cases}(t_i)$ and $n_{controls}(t_i)$ the number of cases and controls in the case-control risk set at age t_i . Specifically, $n_{cases}(t_i)$ is the number of subjects who have been diagnosed at age t_i or at a later age (future cases), while $n_{controls}(t_i)$ is the number of

subjects who have been randomly selected as a control at age t_i or at a later age (future controls). The case:control ratio in the case-control risk set at age t_i is therefore given by $n_{cases}(t_i) : n_{controls}(t_i)$ or

$$1 : \frac{n_{controls}(t_i)}{n_{cases}(t_i)} \quad (4.2)$$

In order to get in each case-control risk set at age t_i , a case:control ratio similar to the population ratio (4.1), I propose to weight each case and controls as follows:

$$\omega_j(t_i) = \begin{cases} 1 & \text{if the subject } j \text{ is a case diagnosed at age } t_j \geq t_i \\ \frac{1-p_i}{p_i} \times \frac{n_{cases}(t_i)}{n_{controls}(t_i)} & \text{if the subject } j \text{ is a selected as a control at age } t_j \geq t_i \end{cases} \quad (4.3)$$

Since the age-conditional probability of developing the disease p_i might be difficult to estimate in some applications, I also proposed a *simple weighted Cox model*. In that *simple weighted Cox model*, I used lifetime probability p to develop the disease of interest instead of the age-conditional probability p_i . The weights used in the *simple weighted Cox model* are therefore not time-dependent, and are given by

$$\omega_j(t_i) = \begin{cases} 1 & \text{if the subject } j \text{ is a case diagnosed at age } t_j \geq t_i \\ \frac{1-p}{p} \times \frac{n_{cases}}{n_{controls}} & \text{if the subject } j \text{ is a selected as a control at age } t_j \geq t_i \end{cases} \quad (4.4)$$

where n_{cases} and $n_{controls}$ are the total number of cases and controls in the case-control study, respectively.

In simulation studies, the true age-conditional probability p_i and the true lifetime probability p can be directly calculated from the population data (for details, see section 4.2.5). For a real population based case-control study, the population data is unknown, but the probabilities of developing the disease can be estimated from relevant national health statistics (for details on lung cancer, see section 4.3.4). The results of standard logistic regression which is the conventional method for case-control study are present in this thesis for comparison purposes.

The details of definitions of weights in risk sets of Cox models are summarized in Table II.

Table II: definitions of weights in risk sets of models used for subject j at risk at age t_i :

Weights	Naïve Cox model	Adapted Cox model	Weighted Cox model	Simple Weighted Cox model	Conditional Logistic Regression (CLR)
Current case ($t_j = t_i$)	1	1	1	1	1
Current control ($t_j = t_i$)	1	1	$\frac{1 - p_i}{p_i} \times \frac{n_{cases}(t_i)}{n_{controls}(t_i)}$	$\frac{1 - p}{p} \times \frac{n_{cases}}{n_{controls}}$	1
Future case ($t_j > t_i$)	1	0	1	1	0
Future control ($t_j > t_i$)	1	1	$\frac{1 - p_i}{p_i} \times \frac{n_{cases}(t_i)}{n_{controls}(t_i)}$	$\frac{1 - p}{p} \times \frac{n_{cases}}{n_{controls}}$	0

p : the life time probability of developing the disease of interest in the source population.

p_i : the age-conditional probability of developing the disease of interest at age t_i or at a later age in the source population.

All the weighted Cox models were implemented by using the *coxph* function in the R statistical software. This function can directly handle such weights and relies on the robust sandwich variance estimator proposed by Binder [23].

4.2 Simulation study

4.2.1 Overview

To assess the performance of the Cox model with the new risk sets definitions, we carried out a series of simulations. First a hypothetical population was generated, and then a case-control study within that population was simulated. Several scenarios were investigated, and for each scenario the generating of population and case-control data was repeated 1000 times. The coding of the simulations was elaborated by Willy Wynant, who was Dr. Leffondré's research assistant.

4.2.2 Generation of the source population

Source populations of size $N=1000$ subjects were generated. For each subject, several covariates were generated based on the empirical distributions of smoking variables observed in our real data. These data came from a large population-based case-control study of lung cancer undertaken in Montréal, Quebec, Canada, in 1996-2001 (for details see section 4.3.1). In all our simulation scenarios, all subjects were currently exposed at failure (i.e. diagnosis in real data) or censoring and we focussed our attention to three aspects of exposure: the cumulative value of exposure, the intensity and the duration of exposure. Note that the first aspect of exposure is a combination of the two others. We did not generate any non-time dependent exposure in our

simulation study since logistic regression performs well for estimating the effect of such exposures [7].

1. The age at exposure initiation, A_j , $j=1,\dots,N$, was generated from lognormal distribution such that the age at exposure initiation had a mean $\mu = \log(16.1)$ and a standard deviation $\sigma = \log(4.1)$.
2. The intensity at exposure initiation, X_{0j} , $j=1,\dots,N$, was generated from lognormal distribution such that the intensity initiation had a mean $\mu = \log(37.5)$ and a standard deviation $\sigma = \log(20.0)$.
3. To reflect our real data on smoking intensity, where all subjects were asked to report their average number of cigarettes smoked per day at age 25, 40, 50, and 60 years, we generated an intensity of exposure according to a step function with predefined age intervals, as illustrated in Figure 2.

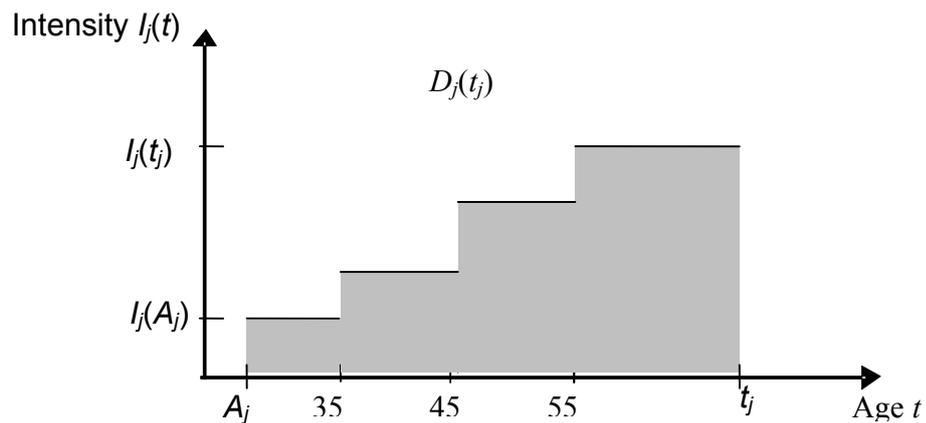


Figure 2 : Exposure pattern of a hypothetical subject j diagnosed or selected at age t_j , with an increasing intensity over lifetime. The value of the cumulative intensity at age t_j , $E_j(t_j)$, equals to the gray area under the curve. Where A_j is the smoking initial age, $I_j(A_j)$ is smoking intensity at the initial age, $I_j(t_j)$ is the intensity at age t_j , $D_j(t_j)$ is the smoking duration at age t_j which is equal to t_j minus A_j .

Therefore, we assumed that the intensity of exposure was a monotone increasing or decreasing step function. Thus, we first generated a discrete variable to define the pattern of intensity over time: constant, increasing, or decreasing intensity, with probabilities q_1 , q_2 , or q_3 , respectively. These percentages varied depending on the scenario. Then, we assumed that the rate of increase or decrease at each step was constant over time within subject. For each subject j , its rate τ_j of increase or decrease was generated from a lognormal distribution with mean $\mu=0.4$ and standard deviation $\sigma=0.085$ for increasing intensity, and mean $\mu=0.1$ and standard deviation $\sigma =0.075$ for decreasing intensity.

4. Use a permutational algorithm to assign each subject j defined by the set of variables (X_{0j}, A_j, τ_j) a survival time t_j . See details below.

Permutational algorithm to generate survival times conditional on time dependent covariates

To generate survival time conditional on time-dependent covariates, we used a permutational algorithm, which was first proposed by Abrahamowicz [21] and validated by MacKenzie [22]. Our algorithm consisted of the following steps:

- Simulation of survival times T_i^* , $i=1,\dots,N$, assuming Gompertz distribution with mean $\mu = 67.1$ and standard deviation $\sigma = 6.0$.
- Right censoring time C_i , $i=1,\dots,N$, was generated from uniform distribution $U[35, \text{upper}]$, where upper was defined so that 10% of survival ages T_i^* were finally uncensored.

- Sort the N tuples (t_i, δ_i) , where $\delta_i = I\{T_i^* \leq C_i\}$ is the indicator of non-censoring and $t_i = \min(T_i^*, C_i)$ is the survival time.
- Create the vector of current covariates values $X_j(t_i)$ for each subject j at each failure time t_i , i.e. past duration $D_j(t_i)$, current intensity $I_j(t_i)$ and cumulative exposure $E_j(t_i)$, as illustrated in Figure 2.
- Randomly pair the vector of current covariates values $X_j(t_i)$ and (t_i, δ_i) ($j=1, \dots, N; i=1, \dots, N$), according to probabilities based on the partial likelihood, from the earliest observed time t_i to the last time. If $\delta_i = 1$, then randomly select an individual from the risk set $R(t_i)$, with probability equal to his/her contribution to the partial likelihood at that time t_i . The subject who will be selected in such a way will then be considered as the subject who fails at that time. Otherwise, if $\delta_i = 0$ and the censoring is independent of covariates (as we assume here), then select an individual by simple random sampling from the risk set $R(t_i)$. This subject will then be considered as the subject who is censored at that time.

4.2.3 Simulation of case-control studies

In each generated source population, we simulated a hypothetical population-based case-control study that was 1:1 age-matched case-control study. The cases were all the subjects who had event in the source population. Since we generated source populations of 1000 subjects, and that in average about 10% of these subjects had an

event, each case-control dataset was made of about 100 cases. For each case, a single control among subjects who were still at risk at the age of case's event was randomly selected. This resulted in about 100 cases and 100 controls in each case-control data set. The set of potential controls for a given case included all future cases and controls, as well as past controls provided they were still at risk at the age of case's event. Thus, if a subject was selected as a control, for example at age of 45 years, he could also be selected as a control again, e.g. at age 50, and had an event at age 55, for example. However in this case, in order to consider these subjects as different subjects, they were assigned a distinct ID number.

4.2.4 Summary of the different scenarios investigated

Several different scenarios were considered, each based on the proportional hazards model for the true effects of covariates on the hazard. With respect to exposure, we focus on distinguishing the effects of two aspects of exposure which may make distinct etiologic contributions: intensity and duration of exposure. In model 1 (see Table III), the hazard depended on current intensity and past duration, which were represented by two separate covariates. The distribution of the patterns of change in intensity over time varied across scenarios: in scenario 1, all subjects had decreasing intensity over time ($q_1 = 0\%$, $q_2 = 0\%$, and $q_3 = 100\%$); in scenario 2, all subjects had increasing intensity over time ($q_1 = 0\%$, $q_2 = 100\%$, and $q_3 = 0\%$); and in scenarios 3, 60% of the subjects had constant intensity, 25% had increasing intensity, and 15% had decreasing intensity over time ($q_1 = 60\%$, $q_2 = 25\%$, and $q_3 = 15\%$). For the

effects of intensity (β_I) and duration (β_D) on the hazard, we assumed $\beta_I = 0.02$ and $\beta_D = 0.05$ or $\beta_I = 0.03$ and $\beta_D = 0.08$ in each scenario.

Table III : Summary of simulation scenarios in Model 1, hazard depended on intensity and past duration separately.

Scenario No	Intensity patterns of subjects¶ (%)			True effect β	
	q_1	q_2	q_3	Intensity β_I	Duration β_D
1	0	0	100	0.02	0.05
				0.03	0.08
2	0	100	0	0.02	0.05
				0.03	0.08
3	60	25	15	0.02	0.05
				0.03	0.08

¶ q_1 , q_2 and q_3 are the proportions of subjects who have a constant, increasing and decreasing intensity over lifetime in the source population.

In Model 2 (scenarios 4-8, see Table IV), the hazard depended on the value of cumulative exposure, which was calculated as the area under the curve of intensity over time (see Figure 2). As in scenarios 1-3 for Model 1, the distribution of the patterns of change in intensity over time (q_1 , q_2 , and q_3) varied across scenarios 4-8, and the true effect (β_E) was equal to 0.005 or 0.010.

Table IV: Summary of simulation scenarios for Model 2, where the hazard depended on the value cumulative exposure*.

Scenario No	Intensity patterns of subjects¶ (%)			True effect β_E
	q_1	q_2	q_3	
4	0	0	100	0.0005
				0.0010
5	0	100	0	0.0005
				0.0010
6	0	50	50	0.0005
				0.0010
7	33	33	33	0.0005
				0.0010
8	60	25	15	0.0005
				0.0010

¶ q_1 , q_2 and q_3 are the proportions of subjects who have a constant, increasing and decreasing intensity over lifetime in the course population.

* Cumulative exposure was calculated as the area under the curve of intensity over time, as illustrated in Figure 2.

4.2.5 Data analytical models

Logistic regression models

For the aim of comparison, logistic regression was used since it is the conventional method for case-control studies. Because individual matching was used in the simulation study, matched-pair analysis (e.g., conditional logistic regression) was used, and the matching variable itself cannot be analysed, e.g. age. As illustrated in the last column of Table II in the section 4.1, a standard conditional logistic regression procedure can be equivalent to a weighted Cox model, in which a weight of one is assigned to the two subjects (case and its control) at t_i , while a weight of zero is assigned to the rest subjects.

Existing Cox models

The two existing Cox models, i.e. the naïve version and the adapted version (See section 2.3.4 and Table II) were used in the simulation study, for the aim of comparison.

New proposed weighted Cox models

Since the source population was known in this simulation context, the weights of the risk sets of the two new models introduced in Section 4.1 were estimated from the generated population data. The age-conditional probabilities p_i were the proportions of cases who had the events at age t_i or at a later age $t > t_i$ in the source population; and the lifetime probability p was the proportion of cases in the source population. Since it was a 1:1 age-matched case-control study, the ratio $n_{\text{cases}} / n_{\text{controls}}$ was

systematically equal to one. Similarly, the case-control ratio in each risk set at any age t_i , $n_{cases}(t_i)/n_{controls}(t_i)$, was systematically equal to one. Thus, the weights in equations (4.3) and (4.4) become:

$$\omega_j(t_i) = \begin{cases} 1 & \text{if the subject } j \text{ was a case failed at age } t_j \geq t_i \\ \frac{1-p_i}{p_i} & \text{if the subject } j \text{ was selected as a control at age } t_j \geq t_i \end{cases} \quad (4.5)$$

and

$$\omega_j(t_i) = \begin{cases} 1 & \text{if the subject } j \text{ was a case failed at age } t_j \geq t_i \\ \frac{1-p}{p} & \text{if the subject } j \text{ was selected as a control at age } t_j \geq t_i \end{cases} \quad (4.6)$$

Since the ages at event were precise enough in the simulation study, there were no ties among the survival times to handle in all the Cox models.

4.2.6 Summary statistics to evaluate the performance of the different analytical models

As mentioned in Section 4.2.1, all the scenarios were repeated 1000 times. The mean $\bar{\hat{\beta}}$ of the 1000 estimated regression coefficients $\hat{\beta}$, which is the log hazard ratio for Cox's models and the log odds ratio for logistic models, and the relative bias $(\bar{\hat{\beta}} - \beta)/\beta$ were calculated for each scenario investigated. The ratio between the mean of the 1000 standard errors (SE) and the empirical standard deviation (SD) of the 1000 estimates was calculated to assess the accuracy of the variance estimators.

The coverage rate of the 95% confidence interval (CI) of the estimates, $\hat{\beta} \pm 1.96 \times SE$, was calculated. The power of the Wald test was also calculated.

4.3 Real data analysis:

4.3.1 Data source

Study design

The real data that I used to investigate the performance of the new versions of the Cox models came from a large population-based case-control study of lung cancer undertaken in Montréal, Quebec, Canada, 1996-2001. The objective of this case-control study was to investigate the association between lung cancer and environmental and occupational exposures. It included both males and females, aged 35-75. Controls were frequency matched to cases on sex and age [24]. For the present analysis, our exposure of interest is smoking history, which was represented by different variables as explained below.

Smoking information

The information on smoking history available in the real case-control data are described in Table V. Each smoker was required to report average number of cigarettes smoked per day in average in the whole smoking period, and also at aged 25, 40, 50, and 60 years. To investigate the effect of smoking history in the Cox's models, I assumed that those smoking intensities were constant around these ages 25, 40, 50, and 60 years. The smoking intensity was then represented as a time-dependent variable. For example: the average number of cigarettes smoked per day reported for

aged 40 represented the intensity from ages 35 to 45, the average number of cigarettes smoked per day reported for aged 50 was used to represent the intensity from ages 45 to 55, and the reported smoking intensity for aged 60 was used for the intensity after ages 55, as illustrated in Figure 3. Since this case-control study only included subjects aged from 35 to 75, the intensity at aged 25 was never used.

Table V: Summary of smoking variables in the case-control study, Montréal, Quebec, Canada, 1996-2001.

Variable name	Type of variable
Smoking status	Categorical (never smoker, current smoker, ex-smoker)
Age at initiation	continue (years)
Total duration	continue (years)
Time since cessation	continue (years)
Average number of cigarettes smoked per day over lifetime	
Average number of cigarettes smoked per day at age 25	continue (cig/day)
Average number of cigarettes smoked per day at age 40	
Average number of cigarettes smoked per day at age 50	
Average number of cigarettes smoked per day at age 60	
Average number of cigarettes smoked per day at age 75	

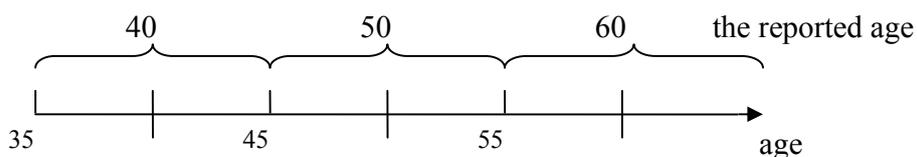


Figure 3: The intensity at different age represented by the intensity at the reported age.

4.3.2 Method to handle missing smoking data

In all the analyses, I removed the subjects who were smokers but had not reported the average amount smoked cigarettes per day (11 subjects out off total 2190 subjects).

Therefore, all smokers had the average smoking intensity over lifetime in the dataset of my analysis, however not all smokers had all intensities at the four reported ages. I did some imputation to handle these missing data. The method of imputation is described below.

One reason of the missing data is that some subjects started smoking later, or stopped smoking earlier than some reported ages. For example, if a subject started smoking at age 30, and quitted smoking (or was diagnosed with lung cancer) at age 55, he did not have reported intensities at age 25 and 60. Since in such a situation, the ages with missing value were out off the smoking period of that subject, the missing data did not influence the analyses when I represented intensity as a time dependent variable in the Cox models.

However, the situation was different if the missing intensity was inside the smoking period. For example, if a subject started smoking at age 30, and was diagnosed/interviewed at age 59, then the subject reported the average number of cigarettes smoked per day only at age 40 and 50. Since these two average numbers of cigarettes smoked per day can only represent the smoking intensity from age 35 to 55, there was a gap of the intensity from age 55 to 59. This kind of missing intensity at some age was caused by the definition of the range around each reported age, and could not be considered as missing at random.

To handle this kind of missing data, I used the intensity reported at the closest earlier age to represent the missing intensity at the later age, as illustrated in Figure 4, where

the intensity from age 55 to 59 was presented by the intensity at age 50. If no intensity was reported at a younger age, the average number of cigarettes smoked per day during the whole smoking period was used.

Table VI: Comparison of the distribution of smoking intensities (n, mean, standard deviation) before and after imputation to handle missing smoking intensity at each of the four ages (25, 40, 50, 60 years) in the case-control study, Montréal, Quebec, Canada, 1996-2001.

Age	Males†				Females†			
	Cases		Controls		Cases		Controls	
	Before	After	Before	After	Before	After	Before	After
25 yrs	670	687	672	703	383	424	238	281
	31.3	31.6	26.1	26.1	23.6	23.6	17.4	17.1
	(14.5)	(14.7)	(14.8)	(14.9)	(11.7)	(11.7)	(11.0)	(11.5)
40 yrs	669	683	585	702	412	421	227	279
	36.4	36.4	30.3	29.4	27.1	27.0	20.1	19.4
	(17.3)	(17.4)	(17.4)	(17.3)	(12.7)	(12.6)	(12.2)	(12.3)
50 yrs	591	650	417	672	342	371	160	248
	36.9	36.9	29.6	29.6	27.4	27.3	19.4	19.2
	(17.5)	(17.8)	(17.6)	(17.8)	(12.6)	(12.4)	(11.8)	(13.3)
60 yrs	381	509	230	544	206	241	80	157
	34.1	35.1	27.2	29.5	25.3	25.9	18.1	18.5
	(17.0)	(18.2)	(17.3)	(18.1)	(12.3)	(12.5)	(11.3)	(13.5)

† Among current and ex-smokers.

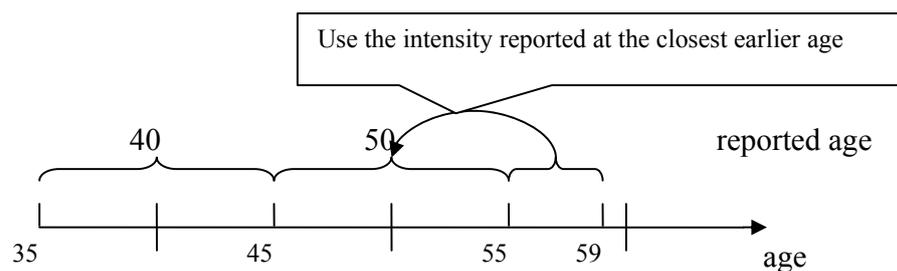


Figure 4: make up of the intensity at some age with missing data

The impact of the imputation on the distribution of the four reported smoking intensities is illustrated in Table VI. The data shown in this table suggest that the

imputation changes only very slightly the distribution of smoking intensities over the ages. Actually, the imputation did even not affect the general patterns of results when we compared the estimates from the different regression models (data not shown). Indeed, we fitted all the regression models (presented in Section 4.3.4) using either the data with missing information, or the data with imputed values for intensities, and the differences between the estimates from the different regression models were similar. Thus, we decided to present only one set of results throughout this thesis, i.e. those based on imputed values for intensity. The results on real data are shown in Section 5.2.

4.3.3 Description of smoking intensity trajectories over time

In order to study the different patterns of change in intensity among subjects, I used two different methods as described below.

The simple heuristic method

First I calculated the absolute difference of the average amount smoked cigarettes daily between the four ages, as the maximum minus the minimum values of the four reported intensities. Then if this difference in intensity was lower than 5 cig/day, the intensity of smoking was considered as stable; otherwise the intensity was considered as changing over time. I finally considered a priori three further different subgroups within the changing group: one group with consistently increasing intensity over time, one with consistently decreasing intensity, and the last group of subjects with non consistent changes over time.

The group based trajectories approach

The second method that I used to identify distinct groups of individual smoking intensity trajectories within the population is implemented in the SAS PROC TRAJ. This procedure estimates group-based trajectory models [25], and was originally proposed to describe individual longitudinal sequences of behavioural measurements.

In order to use SAS PROC TRAJ, I organized my data in multivariate format, where there was only one row of data for each subject and multiple observations included in one line of data. The four variables of smoking intensity correspond to the four repeated measurement taken at four different reported ages. The data sets that I used to do this analysis were data set A (current male smokers) and data set B (current female smokers), details see chapter 4.3.4.

Unlike the simple method in which the number of groups was fixed a priori, the SAS PROC TRAJ uses a model selection procedure to determine the optimal number of groups to compose the mixture. The Bayesian information criterion (BIC) is used for selecting the optimal model [26]. Specifically, the comparisons are completed in a step-wise manner, such that the BIC of the two-group model is compared to the BIC of the one-group model, and the three-group model to the two-group model, and so on. For each increasingly complex model that is tested, the BIC of the more complex (larger number of groups) less the BIC of the less complex model is used to select the model that better fits the data. The change in BIC is given by

$$\Delta BIC = BIC_{(complex)} - BIC_{(null)},$$

and its interpretation in terms of model preference is show in Table VII.

Table VII : Interpretation of the ΔBIC for model selection

$2 \Delta BIC$	Evidence against H_0 (i.e. against simple model)
0 to 2	Not worth mentioning
2 to 6	positive
6 to 10	strong
>10	Very strong

Other criterion for model selection could be considered, but the change in BIC has been shown to be valid for testing the number of components in a mixture [25].

4.3.4 Data analytical models

Smoking confounders

There were three potential confounders that I systematically included in all analytical models: occupational exposure to asbestos (ever/never), ethnic group (represented by 6 dummy variables, in which French group was the reference group), and annual income which was represented by a continuous log-transformed variable.

Datasets and smoking covariates

Smoking history has many aspects, and there is considerable variation in the way to model this multidimensional phenomenon in the analyses. The choice of the modelling strategy depends on the objective of the study, which varies from one study to the other. While in some applications, one may want to estimate the effects of smoking intensity and smoking duration, some others just want to adjust adequately

for smoking and then rather used an overall indicator of smoking. In some other applications, it may be of interest to investigate accurately the impact of time since cessation. Leffondré et al. [2] illustrated the impact of several decisions that must be made when modeling smoking variables. The objective of my thesis is not to revisit all these issues dealing with the best representation of smoking history in the regression models. Rather, it is to compare the estimates of the proposed weighted Cox models and of the existing models (Naïve and adapted Cox, and logistic regression), for estimating these different aspects of smoking. Therefore, I used several models using different representations of smoking, as well as different sub-datasets, which are all described below.

First, I did separate analyses for males and females, since some previous studies have suggested that female smokers may not have the same risk of developing lung cancer as male smokers[27]. Second, I performed some analyses using current smokers only, some others using all smokers only, and some others using all subjects. Table VIII shows the summary of the different sub-datasets that I used to conduct all the analyses.

Table VIII: Summary of sub-datasets from a case-control study, Montréal, Quebec, Canada, 1996-2001.

Name of datasets	Sex of subjects	Smoking status	No. of cases	No. of controls
A	male	current smokers	256	230
B	female	current smokers	174	107
C	male	All smokers (current + ex-smokers)	687	703
D	female	All smokers (current + ex-smokers)	424	281
E	male	All subjects (current + ex- + never smokers)	706	866
F	female	All subjects (current + ex- + never smokers)	455	592

To compare the estimation of the effects of smoking intensity and duration obtained with the different models, I conducted the analyses in current smokers only (Dataset A and B), in order to avoid adjustment for time since cessation, as suggested in Leffondré et al [2]. The model using intensity and duration as separate variables is later referred as Model 1.

Datasets A and B were also used to estimate the effect of the cumulative smoking exposure, measured by the cigarettes-years variable (Model 2), which is the area under the curve of intensity over time as shown in Figure 2. Cigarettes-years, which implies that intensity and duration have the same impact, is the most commonly used variable to model smoking history, although using intensity and duration separately may lead a better fit to data [2]. However, when estimating the effects of time since cessation, cigarettes-years is still useful because it reduces multicollinearity [2]. For the analyses that do not only include current smokers but also ex-smokers (Datasets C-F), there is a need to adjust for time since smoking cessation. In some further analyses, cigarettes-years was log-transformed (Model 3) because of potential violation of the linearity assumption [2] [3].

When the analyses included never smokers (Datasets E and F), the effect of smoking status (never/ever-smoker) was estimated. I centered cigarettes-years by subtracting the mean cigarettes-years value from the original value for all smokers, while keeping zero for never smokers. Such a linear transformation of cigarettes-years does not change its estimated effect [28], but it allows the effect of ever smoking to compare average smokers with never smokers [2].

Using Datasets E and F, I also estimated the effect of a single aggregate measure of smoking exposure (Model 4), named the comprehensive smoking index (CSI) [2] [3].

The CSI was originally proposed to reduce multicollinearity problems that arise when modelling simultaneously several time-related smoking components, since it incorporates intensity, duration, and time since cessation. This smoking variable depend on a half-life (τ) and a lag time (δ) parameters that have to be fixed a priori, or estimated by maximizing the fit. The new version of CSI for lung cancer [3] is:

$$CSI = (1 - 0.5^{dur^*/\tau})(0.5^{tsc^*/\tau}) \ln(int+1),$$

where dur is the duration of smoking, $dur^* = \max(dur - \delta, 0)$, tsc is the time since cessation, and $tsc^* = \max(tsc - \delta, 0)$. The parameters of the half-life (τ) and lag (δ) for this case-control data set were estimated [3]. For males, the estimated half-life (τ) was equal to 26 years and the estimated lag (δ) was equal to 1 year; for females, the estimated half-life (τ) was 26 years and the estimated lag (δ) was 0.7 year. Note that the CSI implies some non linear effects of duration, intensity, and time since cessation (for more details, see [3]).

Table IX summarizes the different smoking models that I estimated.

Table IX: Summary of the smoking models used to analyse the data from the case-control study, Montréal, Quebec, Canada, 1996-2001.

Dataset	Smoking model	variables
A, B (Current smokers)	1	Intensity Duration
	2	Cigarettes-years
	3	Log Cigarettes-years
C, D (All smokers)	2	Cigarettes-years Time since cessation
E, F (All subjects)	2	Indicator of ever-smoking Cigarettes-years (centered) Time since cessation
	4	CSI

Regression models

Logistic regression models

Since the real case-control study was not individually matched, I did not use conditional logistic regression. However, I adjusted for the matching variables (age and sex) in all the unconditional logistic regression analyses.

Existing Cox models

For the comparison purpose, the two existing Cox models (the naïve version and the adapted version, see Section 2.3.4) were used in this real data analysis.

New proposed Cox models

I used the new proposed weighted Cox models (see Table II) to investigate their estimates in a real case-control data analysis. Since the true population source is unknown, the weights for these new risk sets can not be calculated as in the

simulation study. Indeed, they rely on probabilities of developing lung cancer in the source population, which is not available in a real case-control data analysis. These probabilities are the long term age-conditional and lifetime probabilities of developing lung cancer in the population source, p and p_i , respectively. To estimate these probabilities, I relied on the 2006 Canadian cancer statistics monograph. This report provided estimated lifetime probabilities p of developing lung cancer and estimated short-term age-conditional probabilities π_i of developing lung cancer for each decade of ages from 30 to 89 years, for male and female separately, as shown in Table X.

Table X: lifetime probability of developing lung cancer and probability of developing lung cancer within the next 10 years by age group, Canada.

	Lifetime probability of developing lung cancer		Probability (%) π_i of developing cancer in next 10 years by age group					
	%	One in:	30-39	40-49	50-59	60-69	70-79	80-89
Males	8.8	1.4	<0.05	0.2	0.9	2.7	4.3	3.7
Females	5.9	16.8	<0.05	0.2	0.8	1.7	2.3	1.8

Note: the probability of developing cancer is calculated based on age- and sex-specific cancer incidence and mortality rates for Canada in 2001. Source: Surveillance division, CCDPC, Public health agency of Canada.

For the *simple weighted Cox model*, I used therefore a lifetime probability p of 8.8% for males and 5.9% for females. For the *weighted Cox model*, p_i is the long-term age-conditional probability of developing lung cancer at age t_i or later. However, Table X provides only the short-term age-conditional probability of developing lung cancer in the next ten years. Thus, I used the age-conditional probability (π_i) of developing lung cancer within the next 10 years to estimate the long-term age-conditional probability p_i . For each subject belonging to age group i , the probability

of not developing lung cancer within the next 10 years is $(1 - \pi_i)$. The long-term age-conditional probability of not developing lung cancer in the future for subjects belonging to that decade of age i , can then be approximated by the product of short-term probabilities of not developing lung cancer in each next decades:

$$\prod_{l=i}^{\text{\# of age groups in the future}} (1 - \pi_l)$$

Thus, the probability p_i of developing lung cancer in the future is given by:

$$\left(1 - \prod_{l=i}^{\text{\# of age groups in the future}} (1 - \pi_l) \right).$$

Table XI provides the resulting estimates of p_i .

Table XI: Age-conditional probabilities of developing lung cancer in the future and the weights for each age categorical for weighted Cox model in the analysis of Montreal case-control study.

Age t_i	Male j		Female j	
	π_i^*	$p_i^\#$	π_i^*	$p_i^\#$
30-39	0.000	0.113	0.000	0.066
40-49	0.002	0.113	0.002	0.066
50-59	0.009	0.111	0.008	0.064
60-69	0.027	0.103	0.017	0.057
70-79	0.043	0.078	0.023	0.041
80-89	0.037	-	0.018	-

* Estimated age-conditional probability to develop lung cancer in the next ten years following age t_i . These estimates were provided in [29], and based on age- and sex-specific cancer incidence and mortality rates for Canada in 2001 and on life tables based on 1999-2001 all cause mortality rates.

Estimated age-conditional probability to develop lung cancer after age t_i calculated as

$$p_i = \left(1 - \prod_{l=i}^{\text{\# of age groups in the future}} (1 - \pi_l) \right).$$

It should be noted that the probabilities p_i and p are estimated from the general population data, so these probabilities may be more appropriate for the average individual in the source population than for the smokers who have higher risk to

develop lung cancer. However, in order to simplify the analysis, all datasets were analysed using the same probabilities p_i and p .

Moreover, there were some ties in age at diagnosis/interview the real data. I used Efron's approximation [16], which is more intensive computationally but also more precise than the Breslow method, to handle these ties.

5 Results

5.1 Results from simulations

In all scenarios that we investigated (see Table III and Table IV), as we expected, the naïve Cox model systematically under-estimated all the effects, while the adapted version systematically over-estimated them. Moreover, these two existing Cox models had more bias than the new proposed weighted Cox models. More details are described below.

Table XII shows the results for Model 1 which investigated the effects of both intensity and duration of exposure. These exposure components were represented as separate time-dependent covariates in the Cox models and they were fixed at their values at the age of event/selection in conditional logistic regression (CLR). The proportions of subjects with constant (q_1), increasing (q_2), and decreasing (q_3) intensity varied across scenarios, with the last scenario ($q_1 = 0.6$, $q_2 = 0.25$, $q_3 = 0.15$) close to the distribution of intensity over time in our real data (see Section 5.).

From Table XII, the relative bias ($\frac{\bar{\hat{\beta}} - \beta}{\beta}$) of the estimates of the effects of both intensity and duration from the new proposed weighted Cox models was systematically lower than that from the two other existing Cox models. Indeed, the naïve Cox model under-estimated all the effects by about 30%, while the adapted Cox model over-estimated them by about 45%. The proposed *weighted Cox model* estimates were biased by generally less than 10%, and had thus a better performance than the *simple weighted Cox model* which over-estimated all effects by about 20%. The later result suggests that the *simple weight Cox model*, which used fixed weights

for subjects, only partly corrects the bias observed with the Naïve and the Adapted Cox models. The *weighted Cox model* estimates were slightly less biased than those from CLR. Interestingly, the new proposed *weighted Cox model* and CLR systematically over-estimated the effects of intensity in all scenarios, but tended to under-estimate the weaker effect of duration. It seems all Cox models had better power than CLR in all scenarios. However, the coverage rate of CLR was systematically closer to the nominal level of 95% than that from all Cox models, which suggests that CLR has actually a better control of the type I error. Further scenarios implying no effect of the covariates would be necessary to confirm that CLR has a better control of the type I error. The under coverage of the Cox models estimates might be partly due to the fact that the robust sandwich variance estimator systematically underestimates the true variance, as shown with the ratio (mean SE/SD) in Table XII.

Table XIII compared the results of CLR and all Cox models when the hazard only depended on the time-dependent cumulative exposure in the true model. In each scenario, a specific value for the effect of the cumulative exposure was defined, as well as specific proportions of subjects with constant (q_1), increasing (q_2), and decreasing (q_3) intensity. Similarly to Model 1, the new proposed weighted Cox models had a better performance than the existing Cox models, and tended to systematically over-estimate the effects of the cumulative exposure. The relative bias of the *weighted Cox model* tended to be lower than that from CLR and was reduced when the cumulative exposure effect increased.

Table XII: Results from simulations of the proposed Cox models and conditional logistic regression (CLR) for estimating the effects β of intensity $I(t)$ and duration $D(t)$ (Model 1), based on the 1000 simulations.

Intensity patterns (q_1, q_2, q_3)*	Variable†	β	Method	Relative bias (%)	Mean SE/SD	Coverage	Power
0,0,100	Intensity	0.02	Naïve Cox	-26.3	0.83	82.5	57.8
			Adapted Cox	+43.6	0.77	84.2	54.1
			Simple weighted Cox	+24.8	0.80	83.8	57.0
			Weighted Cox	+08.3	0.81	87.9	57.6
			CLR	+07.8	0.98	95.7	36.6
	Duration	0.05	Naïve Cox	-33.5	0.78	84.5	26.8
			Adapted Cox	+27.3	0.27	91.7	29.5
			Simple weighted Cox	+10.0	0.82	91.0	30.2
			Weighted Cox	-03.0	0.78	90.4	30.9
			CLR	-01.2	0.82	95.0	23.8
	Intensity	0.03	Naïve Cox	-26.4	0.81	69.0	86.9
			Adapted Cox	+44.6	0.72	76.6	83.1
			Simple weighted Cox	+22.2	0.76	78.6	85.4
			Weighted Cox	+05.7	0.78	85.8	84.8
			CLR	+07.1	0.97	95.3	75.1
	Duration	0.08	Naïve Cox	-25.1	0.33	81.9	55.7
			Adapted Cox	+32.2	0.24	88.3	56.1
			Weighted Cox	+17.7	0.44	88.7	56.4
			Simple weighted Cox	+05.9	0.32	89.4	57.0
			CLR	+06.3	0.60	94.4	53.6
0,100,0	Intensity	0.02	Naïve Cox	-25.4	0.83	29.0	100.0
			Adapted Cox	+44.8	0.65	48.6	100.0
			Simple weighted Cox	+19.4	0.76	59.4	100.0
			Weighted Cox	+02.5	0.81	87.8	100.0
			CLR	+04.2	0.99	95.6	100.0
	Duration	0.05	Naïve Cox	-32.3	0.34	87.1	32.5
			Adapted Cox	+11.8	0.21	85.3	29.9
			Simple weighted Cox	+6.5	0.29	87.5	30.6
			Weighted Cox	-5.2	0.30	88.0	31.8
			CLR	-5.7	0.33	94.3	22.3
	Intensity	0.03	Naïve Cox	-22.3	0.83	20.0	100.0
			Adapted Cox	+44.5	0.62	28.8	100.0
			Simple weighted Cox	+17.2	0.78	53.3	100.0
			Weighted Cox	+02.1	0.83	89.3	100.0
			CLR	+04.3	0.96	95.5	100.0
	Duration	0.08	Naïve Cox	-21.3	0.84	85.8	64.5
			Adapted Cox	+27.9	0.74	85.8	48.5
			Simple weighted Cox	+14.5	0.79	88.3	58.5
			Weighted Cox	+02.1	0.81	89.5	63.8
			CLR	+02.7	0.99	96.8	34.0
60,25,15	Intensity	0.02	Naïve Cox	-25.3	0.80	42.8	99.6
			Adapted Cox	+54.3	0.60	55.7	99.3
			Simple weighted Cox	+22.8	0.73	62.9	99.4
			Weighted Cox	+04.5	0.78	85.5	99.4
			CLR	+09.7	0.95	95.6	98.7
	Duration	0.05	Naïve Cox	-31.5	0.48	85.8	29.6
			Adapted Cox	+20.0	0.31	89.0	30.2
			Simple weighted Cox	+09.9	0.43	90.1	31.2
			Weighted Cox	-02.9	0.44	89.5	32.5
			CLR	+00.6	0.46	95.6	23.9
	Intensity	0.03	Naïve Cox	-23.2	0.86	25.0	100.0
			Adapted Cox	+52.2	0.47	37.5	100.0
			Simple weighted Cox	+19.5	0.76	54.8	100.0
			Weighted Cox	+03.3	0.82	90.0	100.0
			CLR	+07.0	0.93	96.5	100.0
	Duration	0.08	Naïve Cox	-25.5	0.74	85.7	59.3
			Adapted Cox	+33.6	0.61	86.0	51.3
			Simple weighted Cox	+17.0	0.84	88.0	58.5
			Weighted Cox	+02.3	0.78	90.7	60.7
			CLR	+06.7	0.76	96.0	44.3

*The percentages of subjects in the population source that had a constant (q_1), increasing (q_2), and decreasing (q_3) intensity over lifetime, respectively.

† Intensity and duration were represented by time-dependent covariates in all Cox models. In CLR, we used the values of intensity and duration at the time of event/selection.

Table XIII: Results from simulations of all the Cox models and conditional logistic regression (CLR) for estimating the effect β of cumulative exposure† (Model 2), based on 1000 simulations

Intensity patterns (a) (q_1, q_2, q_3)*	β	Method	Relative bias (%)	Mean SE/SD (d)	Coverage (e)	Power (f)
0,0,100	0.0005	Naïve Cox	-28.3	0.80	74.8	71.3
		Adapted Cox	+46.9	0.73	79.5	67.6
		Simple weighted Cox	+26.5	0.77	79.7	70.4
		Weighted Cox	+7.3	0.79	87.2	72.4
		CLR	+8.7	0.99	96.2	58.3
	0.0010	Naïve Cox	-27.8	0.80	39.1	100.0
		Adapted Cox	+45.9	0.68	63.2	99.2
		Simple weighted Cox	+21.8	0.76	69.4	99.1
		Weighted Cox	+3.9	0.79	86.9	99.6
		CLR	+4.8	0.96	95.7	98.8
0,100,0	0.0005	Naïve Cox	-30.6	0.81	44.9	97.2
		Adapted Cox	+46.4	0.68	70.2	96.5
		Simple weighted Cox	+23.9	0.77	72.6	97.0
		Weighted Cox	+3.5	0.80	87.7	97.6
		CLR	+4.5	0.98	96.0	94.1
	0.0010	Naïve Cox	-26.1	0.84	17.2	100.0
		Adapted Cox	+46.3	0.62	44.5	100.0
		Simple weighted Cox	+19.9	0.77	57.2	100.0
		Weighted Cox	+2.3	0.83	90.0	100.0
		CLR	+5.1	0.96	95.1	100.0
0,50,50	0.0005	Naïve Cox	-30.5	0.80	53.6	92.3
		Adapted Cox	+47.9	0.68	69.5	91.5
		Simple weighted Cox	+25.1	0.73	71.9	91.8
		Weighted Cox	+4.3	0.77	86.1	93.5
		CLR	+7.1	0.96	95.7	90.0
	0.0010	Naïve Cox	-26.8	0.87	21.3	100.0
		Adapted Cox	+47.5	0.64	46.3	100.0
		Simple weighted Cox	+21.0	0.78	58.2	100.0
		Weighted Cox	+2.8	0.84	90.1	100.0
		CLR	+3.6	0.97	95.3	99.9
33,33,33	0.0005	Naïve Cox	-30.2	0.76	56.9	88.3
		Adapted Cox	+47.6	0.67	74.4	87.8
		Simple weighted Cox	+25.1	0.72	75.3	88.9
		Weighted Cox	+4.9	0.73	84.7	90.5
		CLR	+6.3	0.95	95.9	83.5
	0.0010	Naïve Cox	-27.0	0.81	27.1	100.0
		Adapted Cox	+47.0	0.64	48.6	100.0
		Simple weighted Cox	+21.2	0.75	60.3	100.0
		Weighted Cox	+2.9	0.80	87.8	100.0
		CLR	+3.8	0.98	94.7	99.9
60,25,15	0.0005	Naïve Cox	-29.8	0.81	61.9	89.4
		Adapted Cox	+49.5	0.71	73.9	87.5
		Simple weighted Cox	+26.3	0.77	76.0	89.0
		Weighted Cox	+05.7	0.78	87.7	90.6
		CLR	+06.7	1.01	95.2	82.2
	0.0010	Naïve Cox	-27.0	0.88	25.1	100.0
		Adapted Cox	+47.5	0.65	51.3	100.0
		Simple weighted Cox	+21.8	0.78	60.4	100.0
		Weighted Cox	+03.6	0.85	90.2	100.0
		CLR	+04.7	0.93	94.6	100.0

*The percentages of subjects in the population source that had a constant (q_1), increasing (q_2), and decreasing (q_3) intensity over lifetime, respectively.

† Cumulative exposure was calculated as the area under the curve as illustrated in Figure 2. It was represented by a time-dependent variable in all the Cox models, and was fixed to the final value observed at the age of event/selection in CLR.

5.2 Results from real data analysis

5.2.1 Description of real data

Demographic and smoking characteristics of study subjects in the Montréal lung cancer case-control study at the time of diagnosis/interview, are shown in Table XIV and Table XV, respectively. Since this was a frequency matched case-control study on age and sex, cases and controls had similar distributions of age and sex. As expected, cases had higher average smoking intensity, and longer average smoking duration than controls. For the ex-smokers, cases had shorter time since cessation than controls.

Table XIV: Demographic characteristics of subjects at the time of diagnosis/interview, Montréal, Quebec, Canada, 1996-2001.

Variables	Males				Females			
	Cases (706)		Controls (866)		Cases (455)		Controls (592)	
	%	Mean(S D)	%	Mean(S D)	%	Mean (SD)	%	Mean (SD)
Age (yrs) *		64.2 (7.6)		65.1 (7.5)		61.6 (9.1)		61.9 (9.2)
Occup. exposure†								
Never exposed	95.2		77.7		99.6		99.2	
Ever exposed	4.8		22.3		0.4		0.8	
Duration (yrs)		39.4 (11.2)		16.4 (10.3)		21.0 (18.9)		11.6 (10.3)
Age at init. (yrs)		24.5 (8.3)		22.5 (6.9)		32.6 (22.8)		37.2 (13.4)
Ethnic group								
Francophone	77.5		64.3		78.2		68.1	
Anglophone	4.8		6.4		9.7		4.2	
Italian	7.2		11.1		2.6		7.3	
European	6.4		11.0		4.8		7.8	
Jewish	0.7		1.4		1.5		1.0	
Other	3.4		5.9		3.1		11.7	
Annual income (CAD)		33,023 (14,992)		35,164 (14,067)		33,722 (19,969)		38,473 (14,586)

* Controls were age-stratified to match the age and sex distribution of cases.

† Occupational exposure to asbestos.

Table XV: Smoking-related characteristics of subjects in a case-control study of environmental exposure and cancer at the time of diagnosis/interview, Montréal, Quebec, Canada, 1996-2001.

Variables	Males				Females			
	Cases (706)		Controls (866)		Cases (455)		Controls (592)	
	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)
Smoking status								
Never smoker	2.7		18.8		6.8		52.5	
Ex-smoker*	61.0		54.6		54.9		29.4	
Current smoker	36.3		26.6		38.3		18.1	
Intensity† (cigarettes/day)								
Average		35.6 (16.8)		28.7 (16.3)		27.2 (12.0)		20.0 (12.9)
25 yrs		31.6 (14.7)		26.1 (14.9)		23.6 (11.7)		17.1 (11.5)
40 yrs		36.4 (17.4)		29.4 (17.3)		27.0 (12.6)		19.4 (12.3)
50 yrs		36.9 (17.8)		29.6 (17.8)		27.3 (12.4)		19.2 (13.3)
60 yrs		35.1 (18.2)		29.5 (18.1)		25.9 (12.5)		18.5 (13.5)
Duration† (yrs)		43.8 (10.4)		35.4 (13.2)		40.5 (10.3)		31.5 (13.1)
Total cigarette-years (cigarettes)		1536.5 (900.6)		849.0 (796.8)		1023.9 (601.4)		307.7 (466.6)
Time since cessation‡ (yrs)		5.9 (8.4)		16.4 (10.3)		3.2 (6.1)		11.6 (10.3)
Age at initial† (yrs)		15.6 (3.5)		16.7 (4.0)		18.1 (5.5)		20.2 (7.0)

* Subjects who had stopped smoking at least 1 day before the interview/diagnosis.

‡ Mean values and standard deviation among ex-smokers.

† Mean values and standard deviation among current and ex-smokers after imputation.

5.2.2 Patterns of smoking intensity

Results from the group based trajectories approach:

First, I used the SAS Proc Traj to investigate the patterns of smoking intensity over time. I tested six models (one group up to six groups), and obtained six BIC values to review. The comparisons were completed in a step-wise manner so that the two-group model was compared to the one-group model and the three-group model to the two-group model and so on.

The best model for both the male current smokers (Dataset A) and the female current smokers (Dataset B) was the five-group model. Figure 5 and Figure 6 show the trajectories of intensity of the best models for these datasets. From Figure 5 and Figure 6 we can see that the intensities are quite constant over time for most current smokers: about 68% male smokers have consistent intensity over time, and about 85.7% for female smokers.

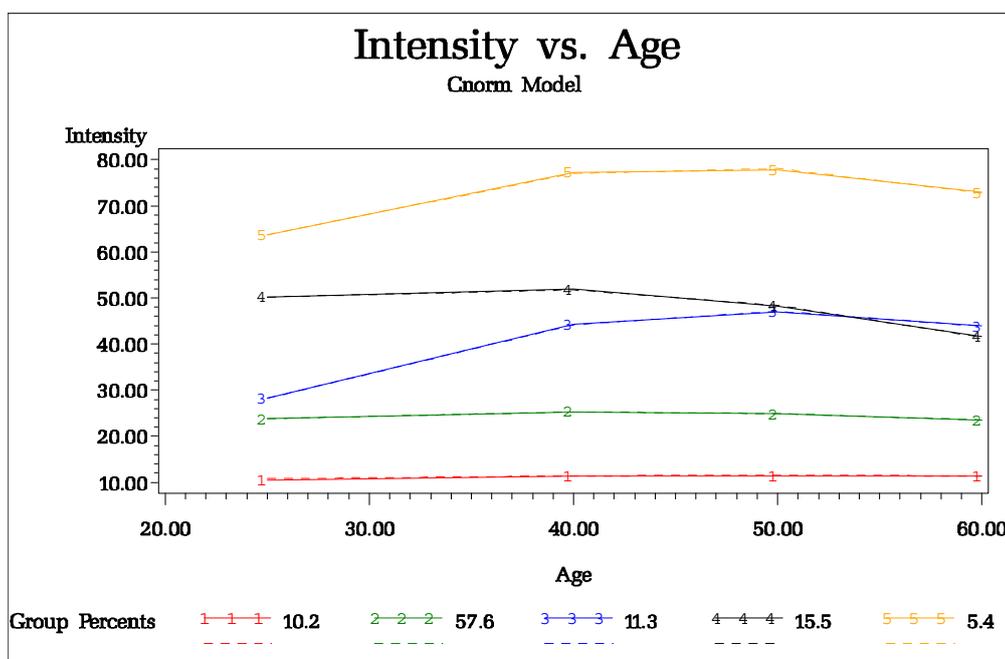


Figure 5: Trajectory of intensity for current male smokers in a case-control study, Montréal, Quebec, Canada, 1996-2001.

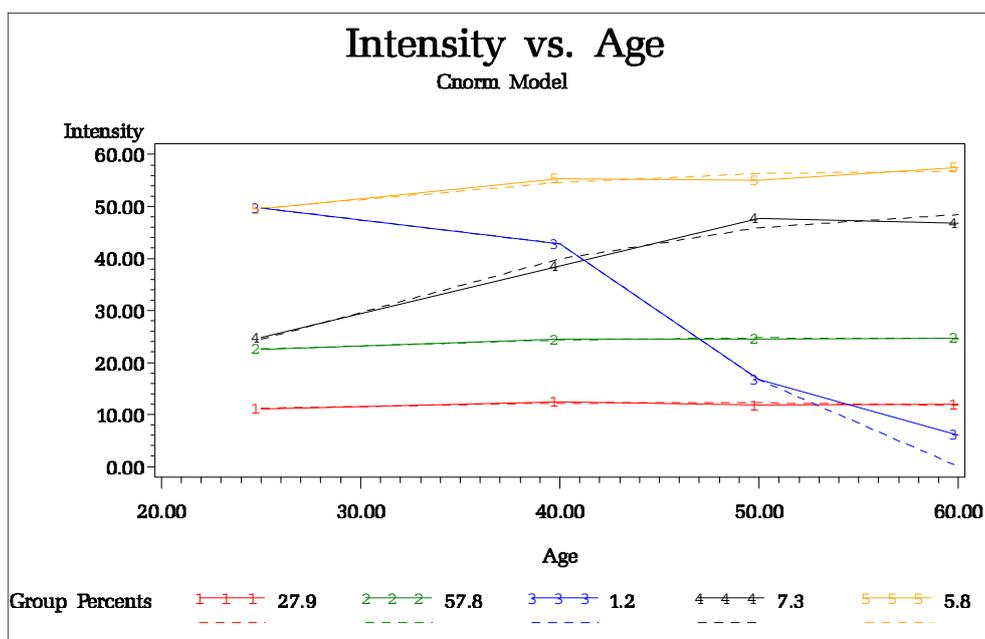


Figure 6: Trajectory of intensity for current female smokers in a case-control study, Montréal, Quebec, Canada, 1996-2001.

Result from the simple heuristic method

In the simple method, if the maximum absolute difference between any two age-specific intensities (reported at age 25, 40, 50, and 60 years) was less than 5 cigarettes per day, the subject was considered with constant intensity over life time. Otherwise, the intensity was classified as increasing, decreasing, or unstable, depending on the direction of change between all consecutive reported intensities. Table XVI shows the percentages of the different patterns of change in intensity over lifetime among subjects who were still smoking at diagnosis/interview (current smokers, Datasets A and B). More than two third of these subjects had a nearly constant smoking intensity over lifetime, and this proportion was higher in controls than in cases. More than one half of the subjects who had non constant intensity over lifetime had a monotone increasing intensity over lifetime.

From the results of both simple method and SAS Proc Traj, I thus found that about 70% of subjects had a constant smoking intensity over lifetime, half of the remaining subjects had an increasing intensity, and a smaller proportion of subjects had either a

decreasing or unstable smoking intensity over lifetime, similarly to the last scenario of the simulation study ($q_1 = 0.6$, $q_2 = 0.25$, $q_3 = 0.15$, Table XII and Table XIII).

Table XVI: Percentages of patterns of intensity change over-time for current smokers in a case-control study, Montréal, Quebec, Canada, 1996-2001.

Characteristics	Dataset A†		Dataset B†	
	Males (n = 486)		Females (n = 281)	
	Cases (n = 256)	Controls (n = 230)	Cases (n = 174)	Controls (n = 107)
Pattern of change in intensity (%) #				
Constant	67.4	77.5	66	74.0
Increasing	21.5	16.5	25.2	14.6
Decreasing	6.0	3.3	4.2	6.0
Unstable	5.1	2.7	4.5	5.3

† Dataset A included male current smokers; dataset B included female smokers, Details see section 4.3.4.

Intensity was considered constant over lifetime if the maximum difference between any two age-specific intensities (reported at age 25, 40, 50, or 60 years) was lower than 5 cigarettes per day. Otherwise, intensity was classified as increasing, decreasing, or unstable, depending on the direction of change between all consecutive reported intensities.

5.2.3 Results from regression models

Table XVII to Table XIX show the results of the estimated effects $\hat{\beta}$ and their robust standard error (SE) from the different real datasets, based on different regression models. I also calculated the hazard ratios (HR) for Cox models and odds ratios (OR) for logistic regression, and their 95% confidence intervals. For continuous covariates, in order to better assess the difference between the estimates, I computed the HR and OR corresponding to approximately one standard deviation increase in the covariate, i.e. 10 units in intensity or duration, 800 units in cigarettes-years, and one year in time since cessation.

From all the results, one can see that there is no dramatic difference between the estimates from the different models. As expected from the simulation study, the estimates of the new proposed weighted Cox models were systematically intermediate between those of the naïve Cox model and of the adapted Cox model. However, the

differences in estimates between the *weighted Cox model* and the *simple weighted Cox model* were not as big as those in the simulation study. This may be because the age-conditional probabilities of developing lung cancer p_i used in the *weighted Cox model* and life-time probabilities p used in the *simple weighted Cox model* were very close for each age (see Table X and Table XI), which was not necessarily the case in the simulation study.

The differences between the estimates of logistic regression and the *weighted Cox model* were more pronounced in this real data analysis than in the simulation study, in which those estimates were very close. Table XVII (Model 1) shows that logistic regression always had stronger estimates of intensity effects than any Cox model, even when compared to the adapted Cox model which systematically over-estimated all the effects in the simulation study. On the other hand, logistic regression had weaker estimates of duration effects than any Cox models, even when compared to the Naïve Cox model which systematically under-estimated all the effects. This seems to confirm the results of the simulation study which suggested that CLR tended to over-estimate any effect of intensity, and to under-estimate weaker effects of duration (Table XVII).

Similarly to intensity, the estimates of cigarettes-years (Model 2, Table XVIII) of logistic regression were systematically farther away from zero than the estimates of all Cox models, even when compared to the Adapted Cox model, which systematically over-estimated the effect of cumulative exposure in the simulation study. Table XVIII also shows the estimated effect of time since cessation. As expected, all the results from the different models indicate that the risk of lung cancer significantly decreases with increasing time since cessation. The only exception is

with the Naïve Cox model which shows a non significant HR in females, but we know from the simulation study that this models seriously under-estimates all the effects. As for the other covariates, the estimates from the proposed weighted Cox models were systematically intermediate between the Naïve and the Adapted Cox models. Interestingly, the direction of the differences between the logistic estimates and the Cox estimates for time since cessation depended on the Datasets. While for males the logistic estimates were close to those of the weighted Cox estimates, for females, the logistic estimates were stronger than any Cox estimates, even when compared to the Adapted Cox model. For ever smoking, the logistic estimates were stronger than any Cox model in both males and females.

Table XIX shows the estimates of the log transformed cigarettes-years (Model 3) and of the comprehensive smoking index (CSI, Model 4). Similarly as for the first two models, the estimates of new proposed Cox models were systematically intermediate between the naïve Cox model and the adapted Cox model. One interesting thing is that, for males, the logistic estimate was stronger than any Cox estimates, even than the adapted Cox estimate, but for females, the logistic estimate was close to that of the weighted Cox models. Note that the estimated effect of the CSI from any model was stronger in females than in males, suggesting that the overall impact of smoking is stronger in females than in males. The 95% CI of the HR from the *simple weighted Cox model* do even not overlap (2.67-3.87 versus 4.07-6.67). This confirms the findings from some previous studies[27] .

It is hard to say which model has the most accurate estimates in the real data analysis, since the true model and the true effects of the parameters are unknown. However, when comparing logistic regression to the naïve Cox model which systematically

under-estimated all the effects in the simulation study, and to the adapted Cox model which always over-estimated all the effects, it seems that logistic regression likely over-estimates the effects of cigarette-years and of smoking intensity, and under-estimate the effect of smoking duration, in both males and females. The smoking status, the log of cigarette-years, the CSI, and time since cessation were not considered in the simulation study. However, because of the systematic under/over-estimation of the Naïve/Adapted Cox models observed in the current simulation study and in the previous one [7], one may argue that logistic regression seems to over-estimate the effects of smoking status in both males and females, of time since cessation in females, and of the log of cigarette-years and CSI in males. However, further simulation studies involving such aspects of exposure history are needed to confirm these points, as well as to explain them. Indeed, there is a need to understand why for some covariates the patterns of results differ for males or females. The differences may be partly due to the fact that the strength of association between these different covariates and the risk of lung cancer is not the same for males and females. Indeed, as shown in the previous simulation study [7], logistic regression may have some difficulties in separating the effects of two inter-correlated variables when at least one of them as a weak effete.

Moreover, further studies are needed to explore the linearity assumptions of the effects of these continuous covariates, and to investigate whether some violation of this assumption could partly explain the differences between the estimates from logistic regression and from the *weighted Cox model*, observed between males and females. Indeed, all our analyses assumed that all these covariates have a linear effect on the log hazard for the Cox models, and on the logit for logistic regression. Yet, this assumption might be violated for some covariates in males or females. For example,

in a previous study using the same datasets, it has been found that cigarettes-years had a non-linear effect on the logit of lung cancer in both males and females [3]. The suggested over-estimation bias of logistic regression for the effect of cigarette-years might be partly due to this violation of the linearity assumption. Indeed, after the log-transformation of cigarette-years, the estimated effect of cigarette-years from logistic regression was much closer to the weighted Cox models in females. Similarly, the logistic estimates for the effect of the CSI, which supposes non linear effects of duration, intensity, and time since cessation, was closer to the weighted Cox estimates in females than in males. However, further studies involving simulations and real data analyses are needed to explore this issue of non linearity which was beyond the scope of my thesis.

Table XVII: Smoking effect estimates from the Cox models and standard unconditional logistic regression (LR), using Model 1 in current smokers, Montréal, Quebec, Canada, 1996-2001.

Dataset †	Smoking Variables‡	Method	$\hat{\beta}^*$	SE	HR¶	95% CI¶	
A	Intensity	Naïve Cox	0.0088	0.0034	1.09	1.02	1.17
		Adapted Cox	0.0202	0.0060	1.22	1.09	1.38
		Simple weighted Cox	0.0204	0.0060	1.23	1.09	1.38
		Weighted Cox	0.0202	0.0060	1.22	1.09	1.38
		LR	0.0344	0.0072	1.41	1.23	1.62
	Duration	Naïve Cox	0.0293	0.0153	1.34	0.99	1.81
		Adapted Cox	0.0513	0.0300	1.67	0.93	3.01
		Simple weighted Cox	0.0461	0.0271	1.59	0.93	2.70
		Weighted Cox	0.0473	0.0274	1.60	0.94	2.74
		LR	0.0215	0.0069	1.24	1.08	1.42
B	Intensity	Naïve Cox	0.0182	0.0064	1.20	1.06	1.36
		Adapted Cox	0.0557	0.0146	1.75	1.31	2.32
		Simple weighted Cox	0.0493	0.0117	1.64	1.30	2.06
		Weighted Cox	0.0486	0.0116	1.63	1.29	2.04
		LR	0.0625	0.0146	1.87	1.40	2.48
	Duration	Naïve Cox	0.1027	0.0201	2.79	1.88	4.15
		Adapted Cox	0.1288	0.0400	3.62	1.66	7.93
		Simple weighted Cox	0.1185	0.0368	3.27	1.59	6.73
		Weighted Cox	0.1200	0.0364	3.32	1.63	6.77
		LR	0.0930	0.0276	2.53	1.47	4.35

† Dataset **A** included male current smokers; **B** female current smokers.

* Estimated regression coefficient, which is the log of hazard ratio for Cox models and log odds ratio for logistic regression, adjusted for ethnic group, occupational exposure, and annual income. Logistic model also adjusted for age.

¶ Hazard ratio (or odds ratio for logistic regression) calculated for an increase of 10 cigarettes per day for intensity and 10 years for duration, which all correspond approximately to one standard deviation of these variables.

‡ All the smoking covariates were time-dependent in all the Cox models. For logistic regression (LR), these variables were fixed, for each subject, at their value at the subject's age at diagnosis/interview.

Table XVIII: Smoking effect estimates from the Cox models and standard unconditional logistic regression (LR), using Model 2, Montréal, Quebec, Canada, 1996-2001.

Dataset†	Smoking Variables‡	Method	$\hat{\beta}^*$	SE	HR¶	95% CI¶	
A	Cigarette-years	Naïve Cox	0.0002	0.0001	1.20	1.08	1.34
		Adapted Cox	0.0005	0.0001	1.51	1.20	1.88
		Simple weighted Cox	0.0005	0.0001	1.48	1.22	1.79
		Weighted Cox	0.0005	0.0001	1.47	1.21	1.78
		LR	0.0008	0.0001	1.92	1.52	2.41
B	Cigarette-years	Naïve Cox	0.0006	0.0002	1.66	1.31	2.10
		Adapted Cox	0.0016	0.0003	3.52	2.04	6.09
		Simple weighted Cox	0.0014	0.0003	3.10	2.05	4.68
		Weighted Cox	0.0014	0.0003	2.99	1.99	4.49
		LR	0.0021	0.0004	5.44	2.93	10.08
C	Cigarette-years	Naïve Cox	0.0002	0.0000	1.16	1.08	1.24
		Adapted Cox	0.0004	0.0001	1.34	1.16	1.54
		Simple weighted Cox	0.0003	0.0001	1.30	1.15	1.46
		Weighted Cox	0.0003	0.0001	1.30	1.16	1.47
		LR	0.0005	0.0001	1.55	1.36	1.78
	Time since cessation	Naïve Cox	-0.0352	0.0055	0.97	0.96	0.98
		Adapted Cox	-0.0682	0.0090	0.93	0.92	0.95
		Simple weighted Cox	-0.0644	0.0084	0.94	0.92	0.95
		Weighted Cox	-0.0645	0.0084	0.94	0.92	0.95
		LR	-0.0622	0.0070	0.94	0.93	0.95
D	Cigarette-years	Naïve Cox	0.0006	0.0001	1.58	1.35	1.85
		Adapted Cox	0.0012	0.0002	2.57	1.83	3.59
		Simple weighted Cox	0.0011	0.0002	2.46	1.84	3.30
		Weighted Cox	0.0011	0.0002	2.38	1.79	3.18
		LR	0.0017	0.0002	4.00	2.77	5.77
	Time since cessation	Naïve Cox	-0.0117	0.0087	0.99	0.97	1.01
		Adapted Cox	-0.0485	0.0146	0.95	0.93	0.98
		Simple weighted Cox	-0.0468	0.0146	0.95	0.93	0.98
		Weighted Cox	-0.0454	0.0140	0.96	0.93	0.98
		LR	-0.0636	0.0132	0.94	0.91	0.96
E	Cigarette-years	Naïve Cox	0.0002	0.0000	1.16	1.09	1.24
		Adapted Cox	0.0004	0.0001	1.34	1.16	1.54
		Simple weighted Cox	0.0003	0.0001	1.29	1.15	1.45
		Weighted Cox	0.0003	0.0001	1.30	1.16	1.46
		LR	0.0006	0.0001	1.57	1.37	1.80
	Time since cessation	Naïve Cox	-0.0346	0.0055	0.97	0.96	0.98
		Adapted Cox	-0.0676	0.0089	0.93	0.92	0.95
		Simple weighted Cox	-0.0623	0.0081	0.94	0.92	0.95
		Weighted Cox	-0.0625	0.0081	0.94	0.92	0.95
		LR	-0.0610	0.0070	0.94	0.93	0.95
	Ever smoking	Naïve Cox	1.6658	0.2425	5.29	3.29	8.51
		Adapted Cox	2.5129	0.2728	12.34	7.23	21.07
		Simple weighted Cox	2.3493	0.2666	10.48	6.21	17.67
		Weighted Cox	2.3741	0.2671	10.74	6.36	18.13
		LR	2.5557	0.2556	12.88	7.80	21.26
F	Cigarette-years	Naïve Cox	0.0006	0.0001	1.57	1.33	1.84
		Adapted Cox	0.0012	0.0002	2.56	1.83	3.58
		Simple weighted Cox	0.0011	0.0002	2.32	1.80	2.99
		Weighted Cox	0.0010	0.0002	2.22	1.72	2.86
		LR	0.0017	0.0002	3.99	2.78	5.74
	Time since cessation	Naïve Cox	-0.0123	0.0086	0.99	0.97	1.00
		Adapted Cox	-0.0486	0.0145	0.95	0.93	0.98
		Simple weighted Cox	-0.0442	0.0136	0.96	0.93	0.98
		Weighted Cox	-0.0423	0.0129	0.96	0.93	0.98
		LR	-0.0627	0.0130	0.94	0.92	0.96
	Ever smoking	Naïve Cox	1.8618	0.1884	6.44	4.45	9.31
		Adapted Cox	2.8615	0.2368	17.49	10.99	27.82
		Simple weighted Cox	2.7503	0.2331	15.65	9.91	24.71
		Weighted Cox	2.7293	0.2278	15.32	9.80	23.95
		LR	3.0501	0.2288	21.12	13.49	33.06

† Dataset **A** included male current smokers; **B** female current smokers; **C** all male smokers; **D** all female smokers; **E** all male subjects; **F** all female subjects.

* Estimated regression coefficient, which is the log of hazard ratio for Cox models and log odds ratio for logistic regression, adjusted for ethnic group, occupational exposure, and annual income. Logistic model also adjusted for age.

¶ Hazard ratio (or odds ratio for logistic regression) calculated for an increase of 800 units in cigarettes-years or one year in time since cessation. For ever smoking, never smokers was reference group

‡ All the smoking covariates were time-dependent in all the Cox models. At each age, the number of cigarettes-years was calculated as the corresponding area under the curve as in Figure 2. For logistic regression (LR), these variables were fixed, for each subject, at their value at the subject's age at diagnosis/interview.

Table XIX: Smoking effect estimates from the Cox models and standard unconditional logistic regression (LR), using Models 3 and 4, Montréal, Quebec, Canada, 1996-2001.

Dataset†	Smoking Variables‡	Method	$\hat{\beta}^*$	SE	HR	95% CI	
A	Log cigar.-years	Naïve Cox	0.5094	0.1169	1.66	1.32	2.09
		Adapted Cox	1.1835	0.2541	3.27	1.98	5.37
		Weighted Cox	1.0809	0.2079	2.95	1.96	4.43
		Simple weighted Cox	1.0935	0.2035	2.98	2.00	4.45
		LR	1.4513	0.2224	4.27	2.76	6.60
B	Log cigar.-years	Naïve Cox	0.8805	0.1545	2.41	1.78	3.27
		Adapted Cox	1.8547	0.4043	6.39	2.89	14.1
		Weighted Cox	1.7004	0.3344	5.48	2.84	10.5
		Simple weighted Cox	1.7366	0.3420	5.68	2.90	11.1
		LR	1.7024	0.3048	5.49	3.02	9.97
		LR	1.7208	0.0596	2.06	1.83	2.31
E	CSI	Adapted Cox	1.2914	0.1191	3.64	2.88	4.59
		Simple weighted Cox	1.1681	0.0948	3.22	2.67	3.87
		Weighted Cox	1.1836	0.0974	3.27	2.70	3.95
		LR	1.3769	0.0905	3.96	3.32	4.73
		LR	1.0145	0.0821	2.76	2.35	3.24
F	CSI	Adapted Cox	1.7570	0.1536	5.80	4.29	7.83
		Simple weighted Cox	1.6512	0.1259	5.21	4.07	6.67
		Weighted Cox	1.6078	0.1238	4.99	3.92	6.36
		LR	1.6780	0.1086	5.35	4.33	6.63
		LR	1.6780	0.1086	5.35	4.33	6.63

† Dataset **A** included male current smokers; **B** female current smokers; **E** all male subjects; **F** female subjects.

* Estimated regression coefficient, which is the log of hazard ratio for Cox models and log odds ratio for logistic regression, adjusted for ethnic group, occupational exposure, and annual income. Logistic model also adjusted for age.

‡ Log of cigar.-years (Model 3); CSI, Comprehensive smoking index (Model 4). See details in Section 4.3.4. All the smoking covariates were time-dependent in all the Cox models. For logistic regression (LR), these variables were fixed, for each subject, at their values at the subject's age at diagnosis/interview.

6 Conclusion and discussion

In this thesis, I proposed two new weighted Cox models to accurately estimate the effects of different time-dependent aspects of exposure in case-control data. Indeed, standard logistic regression does not directly account for temporal changes in covariate values. The proposed weighted estimators are based on weights that depend either on the lifetime probability to develop the disease of interest in the source population (for the *Simple Weighted Cox model*) or on the age-conditional probabilities (for the *Weighted Cox model*). These probabilities can be estimated using relevant health national statistics. The *Simple Weighted Cox model* can be useful in studies in which it is difficult to obtain the age-conditional probabilities of developing the disease. The performances of the new proposed Cox models were compared to standard logistic regression and to two earlier versions of Cox models for case-control data: the *Naïve* and the *Adapted Cox models* that were investigated in Leffondré et al. [7]. These previous models simply consisted in including or excluding future cases from each risk set, which led to serious under or over-estimation bias of any effect, respectively [7]. The performances of all the models (the two new weighted and the two earlier Cox models, as well as standard logistic regression) were compared through simulations. They were then applied to estimate the effects of different smoking components on lung cancer, using data from a case-control study undertaken in Montréal, in 1996-2001. The results from the real case-control data on lung cancer show some differences between the estimated effects from the different methods, and the simulation results help us to interpret these differences.

The simulation results show that the estimates from the two proposed Weighted Cox models are only moderately biased: around +20% for the *Simple weighted Cox model*,

and around +5% for the *Weighted Cox model*. These estimates are intermediate between those from the *Naïve* and the *Adapted Cox models*. When applied to the real case-control data on lung cancer, the differences of the estimates between the two weighted Cox models were very small. One reason of this is likely that the age-conditional probabilities of developing lung cancer did not change greatly among the different age groups. This result suggests that the *Simple weighted Cox model* can be applied in studies in which the incidence rate does not change rapidly from different age groups. However, further simulation studies are needed to investigate the impact of misspecification of the proposed weights, for both the *Simple weighted Cox model* and the *Weighted Cox model*. This issue is of particular interest for analyses focussing on a specific subpopulation at a higher (or lower) risk than the general population, such as for example the subpopulation of smokers in our real life example. Moreover, it should be noted that the simulation results suggest the robust sandwich variance estimator used for all Cox models tends to under-estimate the variance of the estimators, which implies under-coverage of the 95% CI of the regression parameters. Further studies are needed to investigate this issue.

The simulation study also indicates that the *Weighted Cox model* estimates are slightly less biased than those from conventional logistic regression. However, these estimates were relatively close in all scenarios of the simulation study. The results from the real case-control data on lung cancer generally show stronger differences between these estimates. For smoking intensity, cigarette-years, and smoking status, the logistic regression estimates were stronger than the estimates from the *Adapted Cox model*. Since the latter model over-estimated all effects in the simulation study, it seems that most of the logistic estimates were over-estimated. By contrast, for duration, the logistic regression seems to under-estimate the effect since its point

estimate was lower than that of the *Naïve Cox model* which under-estimated all effects in the simulation study. For time since cessation, the log of cigarette-years, and the comprehensive smoking index (CSI), the difference between the Weighted Cox and the logistic estimates depended on sex. This result might be due to some differences in the strength of their effect on the risk of lung cancer between males and females, as well as on some violation of the linearity assumption in males or females. Further studies involving further simulations and real data investigation are needed to investigate these issues. Further studies are also needed to investigate the proportional hazards assumption. Indeed, for some of our smoking covariates, the effect might depend on age, which could potentially bias the estimates from the Cox models. Further studies are needed to explore this assumption in our real data, as well as further simulations to investigate the impact of the violation of this assumption on Cox estimates. Indeed, in our simulation study, we assumed constant effects over time of all time-dependent covariates.

In the real case-control data analysis, I used the value of current intensity in the models using intensity and duration as separate variables. However, the average number of cigarettes smoked per day over lifetime is more often used in applications, including those on lung cancer. Thus, I performed some additional analyses using the average intensity instead of the current intensity in the real case-control data analyses. I obtained very similar effect estimates. This is likely due to the fact that most subjects (about 70%) in the real data had an approximately constant intensity over lifetime. It would be interesting to compare the estimates using data with greater within subject variability of intensity over lifetime.

Overall, the results suggest that the new proposed weighted Cox model could be an interesting alternative to logistic regression for estimating the effects of time-dependent exposures in case-control studies. However, further studies are needed to propose a better variance estimator for the weighted Cox models, as well as to investigate the impact of non-linear and time-dependent effects on the estimates.

7 References

1. Breslow, N.E., *Statistics in epidemiology: The case-control study*. Journal of the American Statistical Association, 1996. **91**(14-28).
2. Leffondré, K., Abrahamowicz, M., Siemiatycki, J., Rachet, B., *Modeling smoking history: A comparison of different approaches*. American Journal of Epidemiology, 2002. **156**: p. 813-23.
3. Leffondré, K., Abrahamowicz, M., Xiao, Y. and Siemiatycki, J. , *Modeling smoking history using a comprehensive smoking index: Application to lung cancer*. Statistics in Medicine, 2006. **25**: p. 4132-46.
4. Cox, D.R., *Regression models and life tables (with discussion)*. Journal of the Royal Statistical Society, 1972. **34**: p. 187-220.
5. Langholz, B., and Thomas, DC., *Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison*. American Journal of Epidemiology, 1990. **131**(1): p. 169-176.
6. Prentice, R., and Breslow, NE., *Retrospective studies and failure time models*. Biometrika, 1978. **65**: p. 153-8.
7. Leffondré, K., Abrahamowicz, M., and Siemiatycki, J., *Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: A simulation study*. Statistics in Medicine, 2003. **22**: p. 3781-94.
8. Kleinbaum, D., Kupper, LL., and Morgenstern, H. , *Epidemiologic Research. Principles and Quantitative Methods*. 1982: Wadsworth, Belmont.
9. Rothman KJ, G.S., Lash TL. , *Modern Epidemiology*. (3rd edn). 2008: Philadelphia, PA: Lippincott Williams & Wilkins
10. Breslow, N.E., *Case-Control Studies*. 2005: Springer Berlin Heidelberg.

11. Hosmer, D., and Lemeshow, S., *Applied Logistic Regression*. 1989: John Wiley & Sons (Sd).
12. Prentice, R., and Pyke, R., *Logistic disease incidence models and case-control studies*. *Biometrika*, 1979. **66**(3): p. 403-411.
13. Bender, R., Augustin, T. and Blettner, M. , *Generating survival times to simulate Cox proportional hazards models*. *Statistics in Medicine*, 2005. **24**: p. 1713–1723.
14. Cox, D.R., *Partial likelihood*. *Biometrika*, 1975. **62**: p. 269-75.
15. Breslow, N.E., *Covariance analysis of censored survival data*. *Biometrics*, 1974. **30**: p. 89-99.
16. Efron, B., *Efficiency of Cox's likelihood function for censored data*. *Journal of the American Statistical Association*, 1977. **72**: p. 557–65.
17. Fisher, L., and Lin, D.Y. , *Time-dependent covariates in the Cox proportional-hazards regression model*. *Annu. Rev. Public Health.*, 1999. **20**: p. 145–57.
18. Klein, J., and Moeschberger, M.L., *Survival Analysis: Techniques for Censored and Truncated Data*. 1997, New York: Springer-Verlag.
19. Thiebaut, A.a.B., J., *Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study*. *Statistics in medicine*, 2004. **23**(24): p. 3803-20.
20. Lamarca, R., Alonso, J., Gomez, G., and Munoz, A., *Left-truncated data with age as time scale: an alternative for survival analysis in the elderly population*. *Journals of Gerontology Series A*, 1998. **53**(5 M337-M343).
21. Abrahamowicz, M., MacKenzie, T. and Esdaile, J.M., *Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis*. *Journal of the American Statistical Association* 1996. **91**: p. 1432-39.

22. MacKenzie, T., and M. Abrahamowicz, *Marginal and hazard ratio specific random data generation: Applications to semi-parametric bootstrapping*. *Statistics and Computing* 2002. **12**: p. 245-52.
23. Binder, D.A., *Fitting Cox's proportional hazards models from survey data*. *Biometrika*, 1992. **79**: p. 139-147.
24. Benedetti, A., Parent, ME. and Siemiatycki, J., *Consumption of alcoholic beverages and risk of lung cancer: results from two case-control studies in Montreal, Canada*. *Cancer Causes Control*, 2006. **17**(4): p. 469-80.
25. Jones, B., Nagin, DS. and Roeder, K., *A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories*. *Sociological Methods & Research*, 2001. **29**(3): p. 374-393.
26. D'unger, A., Land, K., McCall, P., and Nagin, D., *How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses of the London, Philadelphia, and Racine cohorts studies*. *American Journal of Sociology*, 1998. **103**: p. 1593-1630.
27. Risch, H., Howe, GR. , Jain, M., Burch, JD, Holowaty, EJ., and Miller, AB. , *Are Female Smokers at Higher Risk for Lung Cancer than Male Smokers? A Case-Control Analysis by Histologic Type*. *American Journal of Epidemiology* 1993. **138**(5): p. p. 281-293.
28. Kleinbaum, D., Kupper, LL., Muller, KE., and Nizam, A., *Applied Regression Analysis and Multivariable Methods*, ed. 2nd. 1988, Boston: MA:PWS-Kent Publishing Company. .
29. Canadian Cancer Society/National Cancer Institute of, C., *Canadian Cancer Statistics 2006*. 2006, Toronto, Canada

8 Appendix

8.1 Questionnaire on smoking

The questions that were asked in the questionnaire of the Montreal case-control study regarding smoking history are shown in Table XX.

Table XX: Questions that were asked in the questionnaire regarding smoking history in the case-control study of lung cancer, Montréal, Quebec, Canada, 1996-2001.

Question	Answer
“Have you ever smoked at least 100 cigarettes in your entire life?”	yes, no, or DK*
“Has there ever been a period when you smoked cigarettes regularly (at least once a year)?”	yes, no, or DK*
“About how old were you when you first started smoking cigarettes regularly?”(years)
“Do you still smoke cigarettes now?”	yes, no, or DK*
“During all the years that you smoked cigarettes how many did you smoke per day on average?” (cigarettes per day)
“Were there ever any periods when you gave up smoking for at least 12 months and then took it again?”	yes, no, or DK*
If answer is “yes”:	From age...to age ...
“We would like to have an idea of how much you smoked at different times in your life, and what type of cigarettes they were. We would like you to think of four different years in your life. Please try to recall your smoking habits at these times (if applicable). ”	Number of cigarettes per day Favourite brand Type(filter, non-filter, rolled, DK*)
Age at 25, 40, 50, and 60 years old	

*: DK means don't know

8.2 Simulation results from Leffondré 2003

Table XXI. Mean of the 1,000 estimates, corresponding confidence interval, and relative bias for the adapted Cox model and conditional logistic regression.

Scenario	True effect	Adapted Cox model				Conditional logistic regression			
		β	$\hat{\beta}$	95% CI	$\frac{\hat{\beta} - \beta}{\beta}$	$\hat{\beta}$	95% CI	$\frac{\hat{\beta} - \beta}{\beta}$	
1	Sex	.4	.428	.407, .448*	.070	.406	.388, .424	.016	
	Current E	.4	.427	.403, .450*	.067	.488	.467, .509	.221	
2	Sex	.4	.441	.416, .465*	.102	.422	.401, .442*	.054	
	Current E	1.4	1.507	1.483, 1.531*	.076	1.509	1.486, 1.532*	.078	
3	Sex	.4	.412	.392, .432	.030	.398	.381, .416	-.004	
	Duration 6	.00	.0073	.0065, .0081*	.217	.0065	.0059, .0072	.083	
4	Sex	.4	.435	.413, .458*	.088	.412	.393, .430	.029	
	Duration	.03	.0337	.0329, .0346*	.123	.0311	.0304, .0317*	.037	
5	Sex	.4	.453	.432, .475*	.134	.430	.412, .448*	.074	
	Current E	.4	.451	.421, .481*	.127	.534	.508, .561*	.336	
	Duration 6	.00	.0067	.0057, .0078	.117	.0037	.0029, .0045*	-.383	
6	Sex	.4	.469	.439, .500*	.173	.415	.391, .438	.037	
	Current E	1.4	1.544	1.511, 1.576*	.103	1.633	1.587, 1.678*	.166	
	Duration	.03	.0365	.0352, .0379*	.217	.0304	.0294, .0315	.013	
7	Sex	.4	.443	.418, .467*	.107	.411	.391, .431	.028	
	Current E	.4	.444	.412, .476*	.109	.543	.515, .570*	.356	
	Duration	.03	.0355	.0344, .0366*	.183	.0297	.0288, .0306	-.010	
8	Sex	.4	.454	.428, .480*	.135	.416	.394, .437	.039	
	Current E	1.4	1.547	1.515, 1.580*	.105	.575	1.545, 1.605*	.125	
	Duration 6	.00	.0072	.0060, .0084	.200	.0044	.0035, .0053*	-.267	

Abbreviations: CI: confidence interval of $\hat{\beta}$, Current E: Indicator of current exposure.

* Indicate confidence interval that does not include the true value of β .