Joint modeling of individual socio-professional trajectory and overall or cause-specific survival

Maryam Karimi

NNT : 2016SACLS120

THÈSE DE DOCTORAT
DE
L'UNIVERSITÉ PARIS-SACLAY
PRÉPARÉE À
L'UNIVERSITÉ PARIS-SUD XI

ÉCOLE DOCTORALE N ° 570
EDSP Santé Publique

Spécialité Santé publique - biostatistiques

Par

Mme Maryam Karimi

Modélisation conjointe de trajectoire socio-professionnelle individuelle et de la survie globale ou spécifique

❄❄❄❄❄❄❄❄❄

Joint modelling of individual socio-professional trajectory and overall or cause-specific survival

Composition du Jury:

Mme. Singh-Manoux, Archana	Directrice de Recherche, INSERM	Présidente
Mme. Rondeau, Virginie	Directrice de Recherche, INSERM	Rapporteur
M. Rizopoulos, Dimitris	Associate Professor, Rotterdam	Rapporteur
Mme. Cambois, Emmanuelle	Directrice de Recherche, INED	Examinatrice
M. Dauxois, Jean-Yves	Professeur des Universités, Toulouse	Examinateur
M. Latouche, Aurélien	Professeur des Universités, Cnam	Directeur de thèse
M. Rey, Grégoire	Ingénieur de Recherche, INSERM	Co-directeur de thèse

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique

Titre: Modélisation conjointe de trajectoire socio-professionnelle individuelle et de la survie globale ou spécifique
Mots clés: Statistique, Analyse de survie (biométrie), Méthode longitudinale, Modèles linéaires généralisés, Analyse de régression, Variables latentes , Sciences sociales – Méthodes statistiques, Mortalité, Modèles de hasards proportionnels, Cause de décès, Modèles conjoints, Données longitudinales, Modèles linéaires généralisés mixtes, Risques concurrents, Risque cause-spécifique, Modèle de Cox, Algorithme EM, Maximum de vraisemblance, Régression de Poisson, Effets aléatoires

Résumé: Appartenir à une catégorie socio-économique moins élevée est généralement associé à une mortalité plus élevée pour de nombreuses causes de décès. De précédentes études ont déjà montré l'importance de la prise en compte des différentes dimensions des trajectoires socio-économiques au cours de la vie. L'analyse des trajectoires professionnelles constitue une étape importante pour mieux comprendre ces phénomènes. L'enjeu pour mesurer l'association entre les parcours de vie des trajectoires socio-économiques et la mortalité est de décomposer la part respective de ces facteurs dans l'explication du niveau de survie des individus. La complexité de l'interprétation de cette association réside dans la causalité bidirectionnelle qui la sous-tend: Les différentiels de mortalité sont-ils dus à des différentiels d'état de santé initial influençant conjointement la situation professionnelle et la mortalité, ou l'évolution professionnelle influence-t-elle directement l'état de santé puis la mortalité?
Les méthodes usuelles ne tiennent pas compte de l'interdépendance des changements de situation professionnelle et de la bidirectionnalité de la causalité qui conduit à un biais important dans l'estimation du lien causale entre situation professionnelle et mortalité. Par conséquent, il est nécessaire de proposer des méthodes statistiques qui prennent en compte des mesures répétées (les professions) simultanément avec les variables de survie. Cette étude est motivée par la base de données Cosmop-DADS qui est un échantillon de la population salariée française.
Le premier objectif de cette thèse était d'examiner l'ensemble des trajectoires professionnelles avec une classification professionnelle précise, au lieu d'utiliser un nombre limité d'états dans un parcours professionnel qui a été considéré précédemment. A cet effet, nous avons défini des variables dépendantes du temps afin de prendre en compte différentes dimensions des trajectoires professionnelles, à travers des modèles dits de "life-course", à savoir *critical period*, *accumulation model* et *social mobility model*, et nous avons mis en évidence l'association entre les trajectoires professionnelles et la mortalité par cause en utilisant ces variables dans un modèle de Cox.
Le deuxième objectif a consisté à intégrer les épisodes professionnel comme un sous-modèle longitudinal dans le cadre des modèles conjoints pour réduire le biais issu de l'inclusion des covariables dépendantes du temps endogènes dans le modèle de Cox. Nous avons proposé un modèle conjoint pour les données longitudinales nominales et des données de risques concurrents dans une approche basée sur la vraisemblance. En outre, nous avons proposé une approche de type méta-analyse pour résoudre les problèmes liés au temps des calculs dans les modèles conjoints appliqués à l'analyse des grandes bases de données. Cette approche consiste à combiner les résultats issus d'analyses effectuées sur les échantillons stratifiés indépendants. Dans la même perspective de l'utilisation du modèle conjoint sur les grandes bases de données, nous avons proposé une procédure basée sur l'avantage computationnel de la régression de Poisson. Cette approche consiste à trouver les trajectoires types à travers les méthodes de la classification, et d'appliquer le modèle conjoint sur ces trajectoires types.

Title: Joint modelling of individual socio-professional trajectory and overall or cause-specific survival
Keywords: Statistics, Survival analysis, Longitudinal methods, Genelaized linear models, Regression analysis, Latent variables, Social science – Statistical methods, Mortality, Hazards proportional models, Causes of death, Joint models, Longitudinale data, Generalized linear mixed models, Competing risks, Cause-specific hazards, Cox model, EM algorithm, Maximum likelihood, Poisson regression, Random effects

Abstract: Being in low socioeconomic position is associated with increased mortality risk from various causes of death. Previous studies have already shown the importance of considering different dimensions of socioeconomic trajectories across the life-course. Analyses of professional trajectories constitute a crucial step in order to better understand the association between socio-economic position and mortality. The main challenge in measuring this association is then to decompose the respective share of these factors in explaining the survival level of individuals. The complexity lies in the bidirectional causality underlying the observed associations: Are mortality differentials due to differences in the initial health conditions that are jointly influencing employment status and mortality, or the professional trajectory influences directly health conditions and then mortality?
Standard methods do not consider the interdependence of changes in occupational status and the bidirectional causal effect underlying the observed association and that leads to substantial bias in estimating the causal link between professional trajectory and mortality. Therefore, it is necessary to propose statistical methods that consider simultaneously repeated measurements (careers) and survival variables. This study was motivated by the Cosmop-DADS database, which is a sample of the French salaried population.
The first aim of this dissertation was to consider the whole professional trajectories and an accurate occupational classification, instead of using limited number of stages during life course and a simple occupational classification that has been considered previously. For this purpose, we defined time-dependent variables to capture different life course dimensions, namely *critical period*, *accumulation model* and *social mobility model*, and we highlighted the association between professional trajectories and cause-specific mortality using the defined variables in a Cox proportional hazards model.
The second aim was to incorporate the employment episodes in a longitudinal sub-model within the joint model framework to reduce the bias resulting from the inclusion of internal time-dependent covariates in the Cox model. We proposed a joint model for longitudinal nominal outcomes and competing risks data in a likelihood-based approach. In addition, we proposed an approach mimicking meta-analysis to address the calculation problems in joint models and large datasets, by extracting independent stratified samples from the large dataset, applying the joint model on each sample and then combining the results. In the same objective, that is fitting joint model on large-scale data, we propose a procedure based on the appeal of the Poisson regression model. This approach consist of finding representative trajectories by means of clustering methods and then applying the joint model on these representative trajectories.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Acknowledgements

En premier lieu, je tiens à exprimer mes remerciements les plus sincères à Aurélien Latouche et Grégoire Rey, pour leur encadrements très attentifs, leur disponibilité, la confiance qu'ils m'ont accordés tout au long de ce doctorat, ainsi que leurs qualités humaines. Ils m'ont laissée une grande liberté d'initiative tout en me faisant profiter de leur expérience et de leur savoir. Ce fut un grand plaisir de travailler avec Aurélien et de bénéficier de son expérience dans le domaine des modèles de survie et plus spécifiquement des risques concurrents. Je suis ravi d'avoir travaillé avec Grégoire qui m'a proposé ce sujet de thèse à la suite de mon stage de master et avec qui j'ai pu enrichir mes connaissances en épidémiologie.

Je tiens à remercier chaleureusement Béatrice Geoffroy-Perez, avec qui nous avons eu l'occasion de travailler sur la base de données Cosmop-DADS. J'ai beacoup apprécié collaborer avec Béatrice et de bénéficier de ses commentaires constructifs durant la premère partie de cette thèse.

I would also thank all the members of the PhD committee, Professor Archana Singh-Manoux, Professor Dimitris Rizopoulos, Doctor Virginie Rondeau, Professor Jean-Yves Dauxois and Doctor Emmanuelle Cambois, for generously accepting to evaluate this dissertation. A special thanks to my referees, Professor Dimitris Rizopoulos and Doctor Virginie Rondeau for the time they dedicated to reading this dissertation and their constructive and kind comments.

Je voudrais remercier l'IReSP et l'InVS qui ont financé cette thèse, les congrès et les formations auxquels j'ai participé pendant ce doctorat.

Je souhaite également remercier l'ensemble de l'équipe du CépiDc pour leur accueil. Je remercie spécialement Laurence Camelin pour son aide, sa patience dans mes démarches administratives et pour ses cannelés! Mes remerciement s'adressent plus particulièrement à Cécile, Walid, Martin, Layla, Bruno, Nelly et Hajèr pour leur

bonne humeur quotidienne, leur soutien et les moments que nous avons partagés. Un grand merci à Karim pour m'avoir supporté dans le bureau ces derniers mois sans doute pas facile, pour ses conseils éclairés, son écoute et ses encouragements.

Mes remerciements s'adressent ensuite à l'ensemble de l'équipe du laboratoire Cédric du Cnam pour leur accueil chaleureux. Je remercie mes collègues de bureau, Julia et Théo avec qui j'ai pu partager des moments sympathiques.

Je n'oublierai pas mes chers amis de master avec qui j'ai partagé une bonne partie de ma vie à Paris. Kevin, Pierre, Dina, Mahmoud L., Adam et Judi: merci! Je suis heureuse de vous avoir dans ma vie.

جا دارد که از تمامی دوستان ایرانی ام، به خصوص دوستانم در پاریس، تشکر کنم برای تمامی خاطرات قشنگ و فراموش نشدنی ای که با هم ساختیم. گلاره، اسفند و امین بهترین سال های غربت رو با شما میشد تجربه کرد. شهاب، سارا، علی، فرشته، ماه منیر، مازیار و شایان ژوسیو همیشه از پر خاطره ترین ها در پاریس خواهد بود.

سپاس آخر را به بهترین های زندگیم، به پدر و مادرم، فواد و بهناز، و خواهر عزیزم، مژگان، تقدیم می کنم که بدون شک مهم ترین عامل رسیدن من به این مرحله هستند. قدردان پدر و مادرم هستم که با بزرگواری و از خودگذشتگی بهترین شرایط تحصیل را برایم در داخل و خارج از ایران مهیا کردند. از مژگان متشکرم که فقط یک خواهر نبود، بهترین دوست و دلسوز و صبور من بود به خصوص در این سال های آخر. هر آنچه که از این عزیزانم بنویسم کم است.

Scientific production

Published manuscript

Karimi M, Geoffroy-Perez B, Fouquet A, Latouche A, Rey G. Socioprofessional trajectories and mortality in France, 1976–2002: a longitudinal follow-up of administrative data. *J Epidemiol Community Health* 2015; 69(4): 339-346. DOI: 10.1136/jech-2014-204615.

Submitted manuscript

Karimi M, Rey G, Latouche A. Joint modelling of longitudinal nominal data and cause-specific hazards (*submitted*).

Unpublished manuscript

Karimi M, Rey G, Latouche A. Joint modelling for large-scale data using Poisson regression models (*in preparation*).

Conference presentations

Karimi M, Rey G, Latouche A. Joint modelling of socioprofessional trajectory and cause-specific mortality, *36th Annual Conference of the International Society for Clinical Biostatistics*, Utrecht, The Netherlands, August 2015.

Karimi M, Rey G, Latouche A. Joint modelling of socioprofessional trajectory and cause-specific mortality, *Journées 2015 de GDR "Statistique et Santé"*, Paris, France, June 2015.

Karimi M, Geoffroy-Perez B, Fouquet A, Latouche A, Rey G. Socio-professional trajectories and mortality in France, 1976-2002, *20th IEA World Congress of Epidemiology*, Anchorage, Alaska USA, August 2014.

Seminars

Karimi M, Rey G, Latouche A. Modélisation conjointe des trajectoires professionnelles et de la mortalité, *Séminaire scientifique d'Equipe de Recherche en Epidémiologie sociale (ERES), Institut Pierre Louis d'Epidémiologie et Santé Publique, Inserm UMRS 1136*, Paris, France, May 2016.

Rey G, Latouche A, Stéphane Rican, Karimi M, Ghosn W , Moreno-Betancur M. ILEM Inégalités sociales, lieux de résidence et mortalitépar causes: Analyses multi-niveaux de données longitudinales en population générale, *Séminaire clôture AAP 2011 "Santé mentale- prévention- prospective- thématiques générales de l'IReSP"*, Paris, France, November 2015.

Karimi M, Rey G, Latouche A. Utilisation des données DADS chaînées aux causes de décès pour étudier l'association entre trajectoires socioprofessionnelles et mortalité par cause, *Les Rencontres de Statistique Appliquée, Service des Méthodes Statistiques (SMS) de l'Ined et Société Française de Statistique (SFdS)*, Paris, France, November 2015.

Contents

List of Acronyms

CC Complete Case.

CI Confidence Interval.

CIF Cumulative Incidence Function.

CNAV National Old-Age Insurance Fund - Caisse National de l'Assurance Vieillesse.

CNIL National Commission on Informatics and Liberty - Commission Nationale de l'Informatique et des Libertés.

CSH Cause-Specific Hazard rate.

CSHR Cause-Specific Hazard Ratio.

DADS Annual Declarations of Social Data - Déclarations Annuelles des Données Sociales.

DERA Department of Employment and Activity Incomes - Département Emploi et Revenus d'Activité.

DST Departement of Occupational Health - Département Santé Travail.

EM Expectation-Maximization.

GEE Generalized Estimating Equations.

GLM Generalized Linear Model.

GLMM Generalized Linear Mixed Model.

HR Hazard Ratio.

ICD International Classification of Diseases.

INSEE French National Institute for Statistics and Economic Studies - Institut National de la Statistique et des Economics.

InVS Institute of Health Surveillance - Institut de Veille Sanitaire.

LMM Linear Mixed Models.

LOCF Last Observation Carried Forward.

MAR Missing at Random.

MCAR Missing Completely at Random.

MCMC Markov Chain Monte Carlo.

MI Multiple Imputation.

ML Maximum Likelihood.

MLE Maximum Likelihood Estimation.

MNAR Missing Not at Random.

MSM Marginal Structural Model.

OM Optimal Matching.

RNIPP National Identification Registery of Individuals - Répertoire National d'Identification des Personnes Physiques.

List of Figures

List of Tables

Synthèse

(extended summary in French)

Introduction

Contexte

Les inégalités de santé sont définies comme la différence d'état de santé ou comme la différence de répartition des déterminants de la santé entre les personnes ou entre les différents groupes de population en raison des facteurs sociaux, des facteurs biologiques ou d'autres facteurs. L'analyse de ces inégalités concerne l'analyse des différentiels en matière de comportements (nutrition, activité physique, tabagisme et etc.), d'exposition aux risques (pollutions, conditions de travail), d'attitudes face au système de soins (accessibilité, recours aux soins préventifs et curatifs) et aussi l'analyse des différentiels en matière de morbidité et de mortalité.

Les indicateurs de mortalité constituent des mesures synthétiques de l'état de santé d'une population. Les données de mortalité par cause de décès ont l'avantage d'être exhaustives et enregistrées sur l'ensemble du pays de manière homogène. Elles sont régulièrement employées en France, en particulier depuis la loi de santé publique de 2004 comme une information sanitaire de référence pour le pilotage des politiques de santé publique.

Les inégalités socio-économiques de mortalité, quantifiée par les écarts de mortalité entre les groupes sociaux, ont été étudiées dans de nombreux pays industriels [1–6]. En France, ces différences de mortalité entre les groupes sociaux ont été mis en

évidence depuis les années 1970, avec les études réalisées par l'Institut National de la Statistique Et des sciences Economiques (INSEE), l'Institut National de la Santé Et de la Recherche Médicale (INSERM), etc. [7, 8].

Malgré le faible niveau de la mortalité et sa diminution, les études menées au Royaume-Uni, aux états-Unis et en Europe ont montré que ces inégalités restent importantes dans de nombreux pays [9–11]. Plus précisément, en France, pour la plupart des causes de décès, on observe des écarts de mortalité importants (par profession, niveau d'étude, etc.) au niveau individuel [12]. Des études comparatives ont également montré qu'en France ces inégalités sont parmi les plus importantes en Europe [9–11]. En outre, ces études ont montré que les inégalités socio-économiques ont augmenté au cours du temps chez les hommes et chez les femmes [1, 13, 14], en particulier en France [12, 15, 16]. Par conséquent, l'analyse de ces inégalités est l'un des sujets les plus importants en sciences sociales et en santé publique.

Bien qu'il n'y a pas de mesure unique défini pour le statut socio-économique, plusieurs indicateurs, y compris le statut professionnel, l'éducation et les revenus ont été proposés dans la littérature [17]. Cependant, certaines études ont montré que le statut professionnel est plus prédictif de la mortalité que le niveau d'études et le revenu, car il est lié à la fois au niveau d'études et au revenu. Par ailleurs la profession est plus proche du moment du décès que le niveau d'études [18].

De nombreuses études ont constaté que les taux de mortalité sont plus élevés chez les individus ayant un "faible" niveau socio-économique [19, 20]; quel que soit l'indicateur socio-économique utilisé [17]. La plupart de ces études n'ont mesuré la position socio-économique qu'à un moment de la vie. Cette approche ne tient pas compte de l'impact des transitions entre les différents groupes socio-économiques. Ainsi, pour obtenir une meilleure compréhension de la relation entre la santé et la position socio-économique, il faut prendre en compte les diverses dimensions des trajectoires socio-économiques, telles que la position socio-économique dans l'enfance, l'évolution de la situation socio-économique et les modalités de transitions entre les groupes sociaux [21, 22].

Certaines études ont déjà démontré ce fait en prenant en compte le niveau socio-économique à travers la vie et en particulier l'influence des trajectoires socio-professionnelles [22–24] et la situation sociale dans l'enfance [25] sur les écarts de

mortalité. Ils ont considéré l'évolution de la profession au cours de la vie active en tant que marqueur pour différencier les facteurs liés à l'environnement familial dans l'enfance, et les facteurs liés à la personne. Plus précisément, en comparant deux personnes ayant le même niveau professionnel à un instant donné, l'individu le plus avantagé durant l'enfance aura commencé sa carrière dans une catégorie professionnelle supérieure et pourra avoir des caractéristiques intrinsèques différentes de caractéristiques de l'individu en progression professionnelle. Plusieurs hypothèses sont discutées le plus souvent pour caractériser l'influence de la trajectoire professionnelle sur la santé:

- Le niveau social réel d'un individu à un instant donné est peut être le reflet de sa situation sociale dans les différentes étapes de sa vie [26],
- La seconde est l'hypothèse d'accumulation, à savoir que plus le nombre d'années au cours desquelles la catégorie sociale d'un individu est défavorisée est élevé, plus l'effet délétère sur sa santé sera important,
- Certaines étapes ou des moments particuliers de la vie (enfance, études, entrée sur le marché du travail, vie professionnelle et retraite) sont considérés comme des périodes clés ayant un impact sur la santé.

Cependant, ces trois hypothèses ne sont pas exclusives, ni exhaustives pour expliquer la contribution des trajectoires de vie sur l'association entre le niveau socio-économique et la santé [27].

L'enjeu consiste ensuite à mesurer la part respective de ces différents facteurs dans l'explication du niveau de survie. La complexité de la mise en évidence de l'effet de ces facteurs réside dans les sens contradictoires de causalité sous-tendus par ce type d'association. Les différentiels de mortalité sont-ils dus à des différentiels d'état de santé initial influençant conjointement la situation professionnelle et la mortalité, ou, l'évolution professionnelle influence-t-elle directement l'état de santé puis la mortalité?

Matériel

La base de données Cosmop-DADS a été constituée dans le cadre du projet COSMOP [28] par l'Institut de Veille Sanitaire (InVS). Elle contient l'appariement des

causes de décès du CépiDc (Centre d'épidémiologie sur les causes médicales de Décès) avec un échantillon de salariés issu du Panel des Déclarations Annuelles de Données Sociales (DADS) de l'INSEE. Un taux d'appariement de 98% a été obtenu. Cette déclaration annuelle est une formalité administrative que doit accomplir toute entreprise employant des salariés, destinée aux administrations sociales et fiscales. Le panel DADS regroupe les déclarations, relatives aux épisodes salariés des individus nés en octobre d'une année paire. Dans le panel, une observation correspond à l'emploi d'un individu, dans une entreprise, pour un poste et une année donnée. Le champ de l'échantillon exploité recouvre les salariés hors agents de l'état, et hors des secteurs de l'agriculture, des services domestiques et des activités extraterritoriales, ayant eu une activité dans l'année, hors stagiaires et apprentis. Ces données forment donc un échantillon représentatif de la population salariée pour les années 1976 à 2002.

Objectif de la thèse

A notre connaissance, il n'existe aucune étude qui a examiné l'ensemble des trajectoires professionnelles, correspondant aux emplois successifs des individus. En effet, les études retrouvées dans la littérature ne prennent en compte que deux emplois (à l'entrée dans le marché du travail, un emploi en milieu de parcours professionnel) [21, 22] et une classification simple des catégories professionnelles (basse, moyenne, élevée). L'un des objectifs de nos travaux a consisté a prendre en compte l'ensemble de la trajectoire professionnelle.

Une première approche pour la mise en évidence d'une association entre les trajectoires professionnelles et les causes de décès a été l'utilisation des données administratives de la profession comme une covariable dépendante du temps dans un modèle à risques proportionnels dans le cadre des modèles de parcours de vie.

Cependant, les épisodes d'emploi sont recueillies uniquement pour les sujets vivants, et donc ce sont des variables dépendantes du temps endogènes, ce qui va avoir pour effet de biaiser les résultats de notre première approche [29]. Il est donc nécessaire de modéliser conjointement les données de la profession et de la survie en intégrant les épisodes d'emploi dans un sous-modèle longitudinal dans le cadre du modèle conjoint. Les extensions existantes sur les modèles conjoints dans la plupart

des cas sont concentrées sur des réponses continues, binaires et ordinales. Il y a eu moins d'attention aux données longitudinales nominales. Cependant ces types de données ne correspondent pas à nos données. Le second objectif de cette thèse était de proposer un modèle conjoint pour les données longitudinales nominales et les données de survie. La base de Cosmop-DADS étant très volumineuse, nous avons rencontrés des problèmes de temps de calcul, ce qui nous a amené à proposer une approche pertinente au regarde de la taille de cette base.

Mise en évidence de l'association entre trajectoires professionnelles et mortalité

Contexte

Il a déjà été démontré que le niveau social observé d'un individu à un instant donnée, peut refléter en partie sa position sociale à différentes étapes de sa vie passée [26]. Cependant, pour mieux décrire l'association entre la position sociale et la mortalité, les modèles de parcours de vie ont été introduits dans la littérature.

Cette approche admet à la fois que les expositions et les conditions de vie précoces et tardives agissent en tant que facteurs de risque ou de protection tout au long de la vie de l'individu [30]. L'objectif est d'examiner comment le niveau social pendant l'enfance, l'adolescence et la vie de jeune adulte influence le risque de maladie à l'âge d'adulte et la position socio-économique qui entraîne des inégalités sociales de santé et de mortalité. On peut citer des études démontrant qu'un niveau socio-économique faible durant la vie influence la mortalité par cause-spécifique et en particulier par maladies cardio-vasculaires [31, 32]. Les hypothèse les plus utilisées dans le cadre des modèles de parcours de vie sont: les modèles *critical periods*, *accumulation* et *social mobility*.

Une période critique (*critical period*) est une fenêtre temporelle dans laquelle une exposition peut avoir des effets indésirables ou protecteurs de longue durée sur le développement d'une maladie [30]. L'intérêt de ce modèle, qui est aussi parfois connu comme modèle latent, est de mettre l'accent sur le moment de l'exposition, il suppose qu'une exposition peut avoir des dommages irréversibles plus tard sur la santé [21]. Ce concept a été étendu à l'évolution sociale, de sorte que dans ce modèle

des étapes ou des moments précis dans la vie sont considérés comme des périodes clés qui affectent la santé.

Le modèle d'accumulation fait l'hypothèse que les différences de mortalité sont expliquées par l'accumulation de toutes les conditions actuelles et passées du travail, les modes de vie et les comportements. Les analyses à l'aide de ce modèle sont basées sur la durée de séjour cumulée dans le groupe social le plus défavorisé. Elles suggèrent que l'accumulation de l'exposition à un bas niveau socio-économique au cours de la vie augmente le risque de mortalité [21, 27, 30, 33].

Le modèle de mobilité sociale (*social mobility*) a été développé pour tenir compte de la modalité de transitions entre les groupes sociaux qui peuvent être divisés en mobilité intra-générationnelle et intergénérationnelle. La mobilité intergénérationnelle porte sur les changements dans un groupe social entre les générations, comme les changements entre la classe sociale des parents et la propre classe sociale de l'individu à l'âge adulte. La mobilité intra-générationnelle désigne les changements entre les classes sociales occupées par un individu à l'âge adulte. Différentes opinions quant à l'impact de la mobilité sociale sur la santé et la mortalité peuvent être trouvées dans la littérature. Certains auteurs montrent que les individus mobiles ont des niveaux de santé placés entre le niveau de santé de leur classe actuelle et le niveau de santé de leur classe d'origine, plus proche de la classe actuelle [34, 35].

Les modèles de parcours de vie contribuent à expliquer l'impact potentiel du statut socio-économique sur la santé. Toutefois, un biais pourrait être dû dans ces modèles à l'impact de la santé sur la situation socio-économique, ou une sélection liée à l'état de santé. Cesser son activité professionnelle en raison de problèmes de santé est un exemple de ce type de sélection, également connu sous le nom de causalité inverse. Cette causalité inverse, entre la santé et la position sociale devrait être pris en compte dans les analyses [36, 37].

Contributions

Dans un premier temps, nous avons défini une trajectoire professionnelle comme la séquence des positions professionnelles consécutives occupées par un individu. Pour tester les 3 hypothèses mentionnées précédemment, nous considérons les variables dépendantes du temps suivantes:

- La classe professionnelle à chaque année;
- Le temps cumulé dans les catégories socio-professionnelles défini comme la durée du séjour individuel dans chaque catégorie professionnelle. Cet indicateur a été calculé pour toutes les classes sauf la classe des cadres, de sorte que celle-ci ait constitué comme la référence;
- La mobilité sociale pour 10 ans, défini par les taux de transition entre les classes. Cet indicateur a été classé en trois groupes en utilisant les tertiles de sa distribution.

Ensuite, nous avons considéré un échantillon de la base Cosmop-DADS. Cet échantillon contient toutes les personnes nées dans les territoires français pour lesquelles une période salariée a été déclarée dans Cosmop-DADS entre 25 et 30 ans, à l'exclusion de celles qui travaillent en dehors du champ d'étude dans leur première année. Au total 337 706 hommes et 275 378 femmes sont incluses dans cette partie d'étude.

Le modèle de Cox a été utilisé pour estimer les risques relatifs toutes causes et cause-spécifique en prenant en compte la troncature à gauche induite par les entrées retardées. Ce modèle a été ajusté pour les 3 variables définies, la profession au début du suivi et la période d'observation. Pour limiter l'impact de la causalité inverse, ce qui est l'influence possible de la santé sur la situation sociale [36, 37], les catégories professionnelles ont été considérées avec un décalage de deux ans, soit au lieu d'utiliser la catégorie socio-professionnelle actuelle, celle de deux ans avant a été prise en compte.

Des études précédentes sur ce sujet ont généralement considéré la position socio-économique des individus à deux ou trois étapes de la vie, y compris l'enfance (position socio-économique du père), la position à l'entrée dans le marché du travail et celle du milieu de vie. Dans cette analyse, nous avons étudier l'association entre toute la trajectoire professionnelle et la mortalité toutes causes confondues et trois principales causes de décès: les maladies cardiovasculaires, le cancer et les causes externes. Dans ce cadre, nous avons démontré que l'exposition à long terme à une situation socio-économique faible est fortement associée à la mortalité chez les hommes et les femmes, en particulier pour les maladies cardiovasculaires. En outre, cette analyse a également mis en évidence la pertinence des modèles *critical period*

et *social mobility* (cf. Figure 1 et Figure 2). Les résultats de cette partie de la thèse ont fait l'objet d'une publication [38].

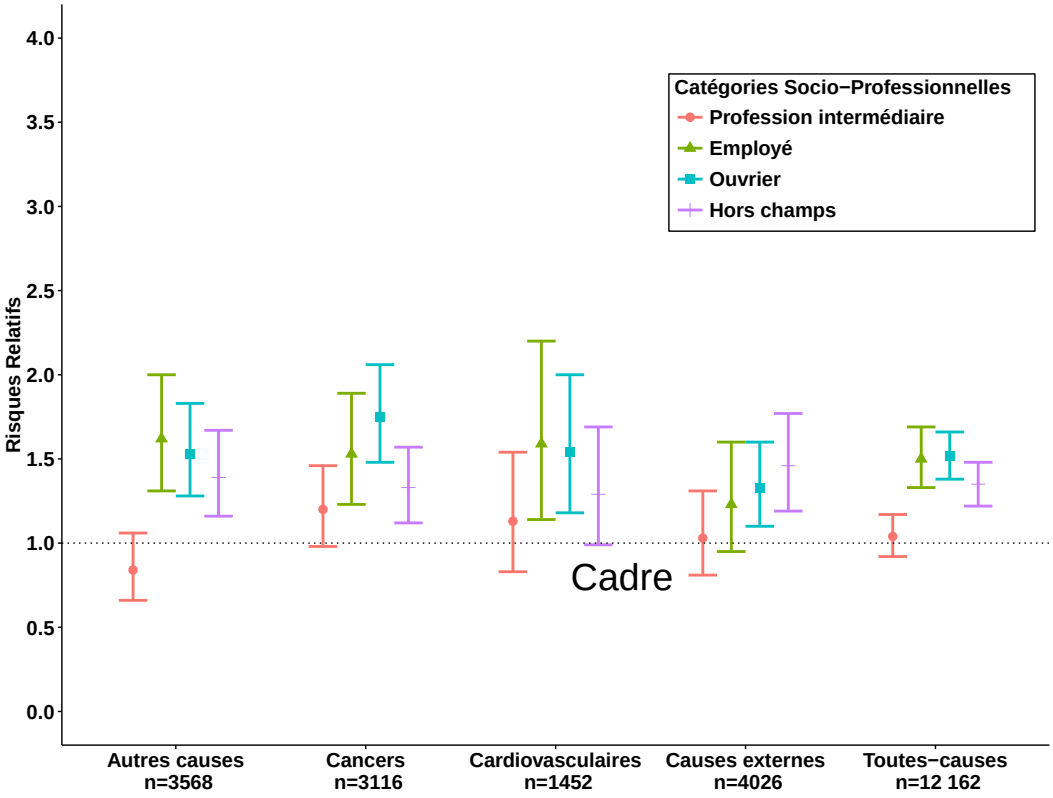


Figure 1 – RR toutes-causes et cause-spécifique chez les hommes par rapport les trajectoires professionnelles

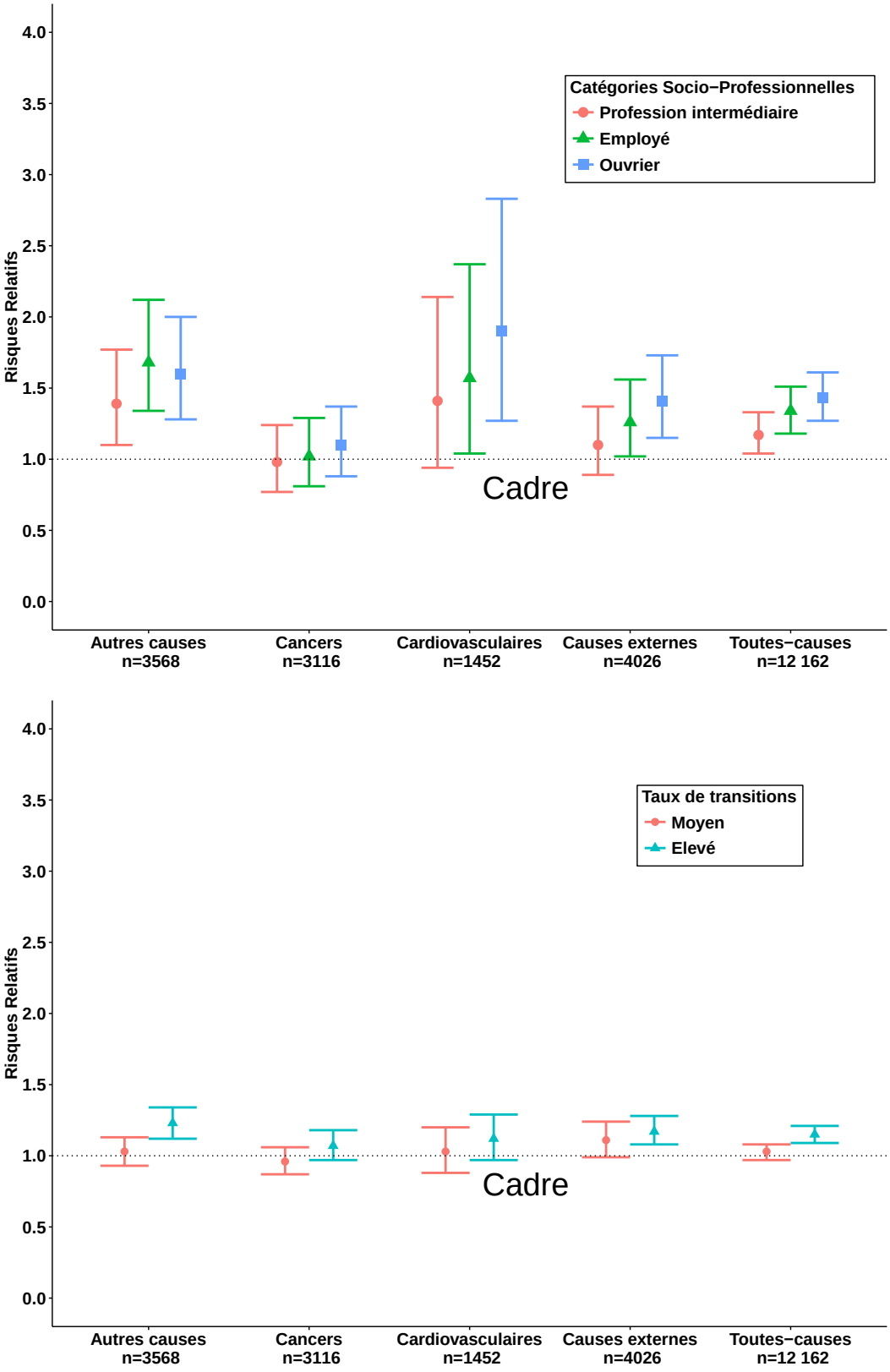


Figure 2 – RR toutes-causes et cause-spécifique chez les hommes par rapport les trajectoires professionnelles

Modélisation conjointe des données longitudinales nominales et des risques concurrents

Contexte

L'utilisation des covariables dépendantes du temps endogènes dans le modèle de Cox induit un biais dans les résultats. Par conséquent, il est nécessaire de modéliser conjointement le processus longitudinal et le processus de survie.

La modélisation conjointe des données longitudinales et de survie fait partie des *modèles de mélanges de profils*, ou des *modèles de sélection* et des *modèles à paramètres partagés*. Bien que mathématiquement tous ces modèles décrivent la distribution conjointe des données longitudinales et de survie, ils ont des interprétations statistiques différentes. Nous nous concentrons uniquement sur les modèles à paramètres partagés, dans la suite, nous appellerons ces modèles des modèles conjoints pour les données longitudinales et la survie.

Ces modèles ont été introduits pour l'étude de la relation entre le nombre de cellules CD4 et la date de survenue du diagnostic du SIDA ou du décès dans les essais cliniques sur le VIH. Elle visait également à déterminer si le nombre de cellules CD4 pouvait être considéré comme un marqueur de substitution utile dans l'évaluation de traitement [39, 40]. Dans ce genre d'études un modèle linéaire mixte a été utilisé pour décrire la trajectoire du nombre de cellules CD4 à l'échelle logarithmique. L'idée fondamentale des modèles conjoints est basée sur le lien entre le modèle de survie avec un modèle approprié pour les mesures longitudinales, généralement un modèle à effet aléatoire [41] dans lequel la corrélation entre les mesures répétées n'est pas ignorée. L'association entre le processus longitudinal et le processus de survie se fait par l'intermédiaire d'une structure latente.

Différentes approches ont été développées dans la littérature pour structurer l'association entre les deux processus longitudinal et survie dans les modèles conjoints. Un exemple est l'utilisation de modèle mixte défini pour lequel les données longitudinales sont une covariable du sous-modèle de survie [42, 43]. Une autre approche inclut directement les effets aléatoires dans les deux sous-modèles longitudinal et de survie avec une distribution conjointe supposée pour les effets aléatoires [44–46]. En pratique ces deux approches sont les plus utilisées dans la littérature, cepen-

dant il existe une démarche différente, appelé le modèle conjoint de classe latente. Elle consiste à diviser la population, supposée être hétérogène, en un nombre fini de classes homogènes où chaque classe est caractérisée par une trajectoire spécifique des données longitudinales et du risque spécifique de l'événement [47, 48].

Dans la littérature, une attention particulière a été portée sur la modélisation conjointe des données longitudinales et de survie au cours des dernières années. Plusieurs extensions ont été proposées dans la littérature pour les adapter à une plus grande variété de données et situations. Malgré toutes ces évolutions, la plupart de ces travaux ont porté sur des données continues [39, 49] ou des mesures de qualité de vie [50], sur des mesures binaires [51] ou sur des réponses ordinales [46, 52] et il y a eu moins d'attention aux données longitudinales catégorielles non ordinales. Récemment, Murawska et Rizopoulos [53] ont développé une extension de la modélisation conjointe de données longitudinales catégorielles et les données de survie en utilisant une approche bayésienne.

Contributions

Compte tenu de la structure de notre jeu de données, la base de données Cosmop-DADS, nous avons étendu le travail de Li et al. [46], en proposant une estimation des paramètres d'intérêt par maximisation d'une vraisemblance pour un modèle conjoint des données longitudinales nominales et des données de risques concurrents. Nous avons introduit les effets aléatoires dans chaque sous-modèle pour structurer l'association.

Soient n le nombre de sujets inclus dans l'étude, Y_{ij} la $j^{\text{ème}}$ observation nominale de l'individu i avec K modalité, $Y_{ij} = k \in \{1, \cdots, K\}$, X_{ij} les prédicteurs des effets fixes, W_{ij} les prédicteurs des effets aléatoires de la partie longitudinale et Z_i les covariables de la partie survie. Les b_ik représentent les effets aléatoires des mesures répétées et u_i représente l'effets aléatoires du modèle de survie. Le modèle conjoint proposé comprend trois composantes: Un modèle GLMM (pour les sigles en anglais de modèle linéaire mixte généralisé) pour les données nominales appelé sous-modèle longitudinal, et un modèle à risques concurrents avec effets aléatoires appelé sous-modèle de survie et la matrice de variance-covariance des effets aléatoires pour

décrire l'association conjointe des mesures répétées et des données de survie.

$$\begin{cases} g(\pi_{ijk}) = X_{ij}^{'}\beta_k + W_{ij}^{'}b_{ik} & k = 1, \cdots, K \\ \lambda_d(t|Z_i, u_i) = \lambda_{0d}(t)\exp\left(Z_i^{'}\gamma_d + \nu_d u_i\right) & d = 1, 2 \\ a_i = \begin{pmatrix} b_i \\ u_i \end{pmatrix} \sim N_{(K-1)q+1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_b & \Sigma_{bu}^{'} \\ \Sigma_{bu} & \sigma_u^2 \end{pmatrix}\right) \end{cases}$$

$$\Sigma_b = \begin{pmatrix} \sigma_{b_1}^2 & \sigma_{b_1 b_2} & \cdots & \sigma_{b_1 b_{K-1}} \\ \sigma_{b_2 b_1} & \sigma_{b_2}^2 & \cdots & \sigma_{b_2 b_{K-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{b_{K-1} b_1} & \sigma_{b_{K-1} b_2} & \cdots & \sigma_{b_{K-1}}^2 \end{pmatrix}$$

La performance de ce modèle a été évaluée dans une étude de simulation. Les paramètres du modèle ont été estimés par l'algorithme EM.

L'approche proposée a été appliquée à un sous-échantillon de la base Cosmop-DADS comprenant 20 000 personnes. L'estimation des paramètres nécessite des calculs complexes au niveau computationnel lorsque les données sont à grande échelle. A notre connaissance aucun travail n'a jusqu'à présent appliqué des modèles conjoint à des bases de données aussi volumineuse. Nous proposons donc une approche reproduisant le principe d'une méta-analyse. Elle consiste à échantillonner la base en S sous-ensemble de même taille présentant la même répartition des causes de décès que dans la grande base. Les paramètres du modèle conjoint sont estimés séparément sur chaque sous-échantillon stratifié, puis sont combinés. La combinaison d'une estimation est faite en prenant la moyenne des paramètres estimés dans chaque sous-échantillon. La présentation de cette méthode, sa validation et son application sont l'objet de la deuxième publication qui est en cours de relecture (*soumis*).

Dans la même perspective, nous avons proposé une utilisation de la régression de Poisson dans le modèle conjoint. Dans cette approche, le sous-modèle de survie est remplacé par une régression de Poisson. Cette régression est computationnellement avantageuse par rapport au modèle de Cox dans les grandes bases de données avec les covariables catégorielles, comme l'estimation des paramètres sont basées sur les données des tableaux de contingence. Donc une catégorisation des trajectoires professionnelles est nécessaires. Cette catégorisation est obtenue en trouvant les trajectoires types de toutes les trajectoires obtenues. Après la validation de ce modèle par une méthode de simulation, cette méthode sera appliquée au même échantillon

que la deuxième publication. La présentation de cette approche, sa validation et son application seront l'objet de la troisième publication.

Introduction

1 Context

Health inequalities are defined as the differences in health status or in the distribution of health determinants between people or different population groups due to social, biological or other factors. Analysis of health inequalities concerns the analysis of differentials in behaviours such as nutrition, physical activity and smoking, in exposure to risk as pollution and working conditions, in access to and attitudes toward health care system. In addition, these analysis concern also the analysis of differentials in morbidity and mortality.

Mortality indicators constitute synthetical measures of the health status of a population. Mortality data by cause of death has the advantage of being exhaustive and is recorded on the entire country uniformly. It is regularly used for international and historical comparisons, and especially in France, since the 2004 public health law, as a health information reference for the management of public health policies.

Socioeconomic inequalities in mortality, as quantified by mortality differentials between social groups, have been studied in many industrialized countries [1–6]. In France, these differences in mortality between social groups have been found since the 1970s, with the studies performed by demographers and social epidemiologists [7, 8].

Despite the low level in mortality and its continuous decrease, studies conducted in the UK, US and Europe have shown that these inequalities are still large in many countries [9–11]. More specifically, in France, for most causes of death, strong mortality differentials (by profession, educational level and etc.) are observed at

the individual level [12]. Comparative studies have also shown that in France these inequalities are among the largest in Europe [9–11]. Besides, these studies have found that socioeconomic inequalities have increased over time in both men and women [1, 13, 14], especially in France [12, 15, 16]. Therefore, the analysis of these inequalities is one of the most important topics in social sciences and public health.

Although, there is no unique defined measure for socioeconomic status, several indicators of socioeconomic status, including occupational status, education, and income have been proposed in the literature [17]. However, some studies have shown that occupational status is more predictive of mortality than educational level and income, as it is related to both education and income and it is also closer to time of death [18].

A large body of research has found that mortality rates are higher among those in lower socioeconomic positions [19, 20]; regardless of the socioeconomic indicator considered [17]. Most of these studies have measured socioeconomic positions only at one stage of life. This approach does not consider the impact of transitions between different socioeconomic groups. Thus, to obtain a better understanding of the relationship between health and socioeconomic position, various dimensions of socioeconomic trajectories, such as childhood's socioeconomic position, evolution of socioeconomic position and frequency and direction of transitions between social groups, need to be taken into account [21, 22].

Some studies have already investigated the impact of the socioeconomic level through life and in particular the influence of socio-professional trajectories [22–24] and childhood's social circumstances [25] on social mortality differentials. They considered the evolution of the profession during active life as a marker for differentiating the factors related to family environment in childhood, and the factors related to the person. More specifically, while comparing individuals in the same professional level at a given instant, those with more benefits in childhood who started their career in a higher professional category, would probably have some intrinsic characteristics different from those in professional progress. Several hypotheses are generally discussed to characterize the influence of professional trajectory on health:

- The actual social level of an individual in a given instant is the reflection of

his/her social situation in different stages of his/her life [26],

- The accumulation, i.e., as the number of years that an individual spends in a disadvantaged social category is higher, the deleterious effect on his/her health will be more important,
- Certain steps or specific moments of life (childhood, education, entry into the labour market, professional life and retirement) as key periods having an impact on health.

However, given the inherent correlation of observations for the same individual over time, these three hypotheses are not exclusive, nor exhaustive for explaining the contribution of the lifetime trajectories on the association between socioeconomic level and health [27].

The main challenge in measuring the association between life course socioeconomic trajectories and mortality is then to decompose the respective share of these factors in explaining the survival level of individuals. The complexity of this demonstration lies in the bidirectionnality of causality presented in Figure 3. Are the mortality differentials due to differences in the initial health conditions that are jointly influencing employment status and mortality, or the professional development influences directly health conditions and then mortality?

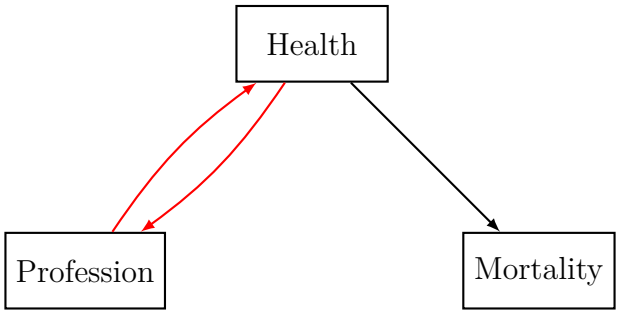


Figure 3 – Socio-professional trajectories and mortality

2 Motivating Database

2.1 Panel of DADS

The Panel of the Annual Declarations of Social Data - Déclarations Annuelles des Données Sociales (DADS) is managed by the Department of Employment and Activ-

ity Incomes - Département Emploi et Revenus d'Activité (DERA) of French National Institute for Statistics and Economic Studies - Institut National de la Statistique et des Economics (INSEE). These annual declarations are a mandatory administrative procedure that should be done by any company with employees pursuant to the article L133-5-4, R243-14 of the social security's code and the articles 87, 88, 240 and 241 of the tax's general code. Originally, these declarations are exploited primarily by the National Old-Age Insurance Fund - Caisse National de l'Assurance Viellesse (CNAV), and are used to calculate the retirement rights. They are also used by taxes administrations for control reasons.

From 1976, DERA collects and links, at the individual level, all professional episodes declared by employers, concerning individuals born in October of an even year. The scope of the exploitation of DADS by INSEE up to 2002 was covering the sectors semi-public and private non-agricultural. We note that owing to administrative reasons, professional episodes of the years 1981, 1983 and 1990 are missed in DADS. Episodes of careers declared as self-employed, employees of the state, employees in agriculture, domestic services, extra-territorial activities, interns and apprentices are excluded from its scope. The DADS is, therefore, a representative longitudinal sample of the French salaried population in this scope, that is, 80% of all paid occupations in France.

This panel contains the professional paths of about 2,9 million people. For each professional episode, the professional data consists of the dates of the beginning and the end of activity, the social category of the individual, the sector of activity and the place of activity.

The professional reference in DADS is the French classification of occupations and occupational classes [54]. This classification was created by INSEE regarding various social characteristics, such as type of work conditions (manual or non-manual), skills and employment status (self-employed, employed, etc.) and income level. The aim was to reflect both working conditions and social background, like the English classification of occupations in UK [55].

The most aggregated level of this classification is presented in Table 1 [56]. As described before, the Panel of DADS does not cover the farmers class, therefore, only the first five classes will be considered.

Table 1 – French classification of occupations

Short title	Examples
Upper class	Intellectual occupations, upper managerial staff and administrators, medical doctors, independent professionals, engineers
Intermediary occupations	Managerial staff, school teachers, skilled technicians, medical and social workers, intermediary managerial and administrators
Clerk class	Civil servants, police and army, company administrative staff, sales and direct personal services
Manual workers class	Skilled, unskilled and farm workers
Craftsmen and trade-related workers	Shop owners, fim managers, craft industry, independent workers (plumbers, electricians, etc.)
Farmers class	Various size farm buisness

2.2 Causes of Death Database

The underlying cause of death is defined as the disease or injury that initiated the morbid evolution leading directly to death, or the circumstances of the accident or violence which produced the fatal injury. In practice, the underlying cause of death is chosen between a number of conditions listed on the medical death certificate.

The underlying causes of death are coded from death certificates according to the International Classification of Diseases (ICD). In France, the French National Death Registry (INSERM-CépiDc) is in charge of this mission. From 1968 till now, three revisions of ICD were used, namely ICD-8 (1968 to 1978), ICD-9 (1979-1999) and ICD-10 (2000-2016).

Three broad categories of the underlying causes of death considered in this dissertation are presented in Table 2.

2.3 Cosmop-DADS database

The Cosmop-DADS database was constructed as part of the Cosmop project [28] by the Departement of Occupational Health - Département Santé Travail (DST), of

Table 2 – Causes of death according to the International Classification of Diseases (ICD)

Causes of death	**ICD-8**	**ICD-9**	**ICD-10**
Cardiovascular diseases	390–444.1, 444.3–458, 782.4	390–459	I00–I99
Cancers	140–239	140–239	C00–D48
Lung	155, 197.8	162	C33–C34
UADT[1]	140, 161	140, 161	C00–C14, C32
Breast	174	174–175	C50
External causes	E800–E999	E800–E999	V01–Y89

the Institute of Health Surveillance - Institut de Veille Sanitaire (InVS).

First for the Panel of DADS, the vital status of the subjects, their date and place of death up to the 1st of April 2006 were investigated by the Department of Demography with the National Identification Registery of Individuals - Répertoire National d'Identification des Personnes Physiques (RNIPP). Then a deterministic record linkage was used to match the occupational paths provided from the Panel of DADS with the causes of death database, reaching a matching rate of 98% using sex, date of birth, date of death and the commune of residence at the time of death as key identifiers.

In total, the Cosmop-DADS population is a sample of the French population (for whom the vital status and date of death are available), employed at least once as a salaried worker in the semi-public and private sectors between 1976 and 2002. This database contains 1 755 590 individuals (957 299 men and 798 291 women). A more complete description of the Cosmop-DADS database is shown in Appendix A.

These analysis were approved by the French data protection committee and institutional ethical review board: National Commission on Informatics and Liberty - Commission Nationale de l'Informatique et des Libertés (CNIL) (authorisation n° 904210v1).

3 Goals of the Thesis

Previous studies on such data have considered limited number of stages, either individual's position at entry into the labour market or his/her position at mid-

life age [21, 22] and used a simple classification for socioeconomic positions (low-medium-high). However, we would like to consider the whole professional trajectories, corresponding to the successive occupations of individuals and a more accurate classification of occupations.

An existing alternative approach is the use of the administrative employment episodes as a time-dependent covariate in a proportional hazards model. In Chapter 3, we highlight the association by analysing different characteristics of professional trajectories and their relationship with the cause of death. Based on life-course models, we define ancillary time-dependent covariates that characterize each professional trajectory.

However, the employment episodes that are collected only for the subjects under the study, are endogenous time-dependent covariates. It is thus natural to model the joint distribution of professional trajectory process and time-to-event process. In Chapter 4, we start by giving a brief literature review on the joint modelling of longitudinal and time-to-event data. Previous joint models mostly have focused on continuous, binary and ordinal responses. There has been less attention to non-ordinal categorical longitudinal outcomes. We therefore propose a joint model for nominal longitudinal data and competing risk data in a likelihood-based framework. We adopt a Generalized Linear Mixed Model (GLMM) for nominal responses to model the longitudinal trajectories and two cause-specific proportional hazards models for competing risk survival data.

Even in a reasonable sample size and moderate individual measurements, estimation of joint model parameters is computationally intensive [53, 57] and it becomes out of reach in the case of large datasets. So far, the existing joint models have been applied to sample size up to 2000 individuals. An approach mimicking a meta analysis is employed to address the calculation problems in joint models and large datasets (Chapter 4), by extracting independent stratified samples from the large dataset, applying the joint model on each sample and then combining the results. In Chapter 5, we propose a joint modelling approach for large-scale data by introducing a Poisson regression model in the survival sub-model.

Part I

Preliminaries

Chapter 1

Background on longitudinal nominal data

1.1 What is longitudinal data?

In epidemiological and medical studies, personal characteristics or environmental exposure, are often collected repeatedly over time. In the context of repeated measures, we refer to the so-called *longitudinal data* when the time itself is, at least in part, a subject of interest [58]. In longitudinal data the observed repeated measures for each subject are strongly correlated, tending to be more alike than the observed repeated measures for different subjects. The key feature of longitudinal data is that it is possible to evaluate the within-subject changes in the outcome of interest over time and to assess the association between covariates and these changes [59]. Standard statistical methods, used for cross-sectional data, that do not take into account this within-subject correlation and assume that observations are independent of each other, produce invalid standard errors [60].

1.2 Regression models for longitudinal outcomes

In longitudinal settings, observed data for each individual i consists of m_i repeated measures over time, $Y_{i1}, \cdots, Y_{im_i}$ for $i = 1, \cdots, n$. Observations of each subject i, $Y_{i1}, \cdots, Y_{im_i}$ are usually correlated and thus, their joint dependence $(Y_{i1}, \cdots, Y_{imi})$ should not be ignored. Two major approaches in the context of regression models for longitudinal data are *marginal models* [61] and *random effects models* [41]. Marginal

models are based on treating the joint dependence structure as a nuisance and aim to describe the population-averaged effects. The alternative approach incorporates unobserved *subject-specific* terms, namely *random effects*, into the model that remain constant within a subject, but changes across individuals.

A third approach, namely *transition models*, which is not of our interest, may also be found in the literature, in which each response is modelled conditional upon the past responses. This approach has been criticized by Diggle et al. [59] due to its difficulties in interpretations.

When the interest lies in the estimation of subject-specific effects, their variability and also in modelling the joint distribution of the repeated measures, the random effects approach is preferable [60, Chapter 13]. Since the random effects modelling implies the marginal model, one could recover marginal informations from the random effects modelling framework. Therefore, using random effects, not only is it possible to estimate the parameters that describe how the average response changes in the population, but also it is possible to analyse how individual response trajectories change over time. Thus, the random effects modelling methodology is more relevant in the context of joint modelling framework for longitudinal and time-to-event data, which will be discussed in Chapter 4.

1.2.1 Generalized linear mixed models

Different extensions of random effects models have been developed regarding type of the repeated measures, which is the key in choosing the appropriate statistical methods. For instance, Linear Mixed Models (LMM) are applicable only for normally distributed outcomes. However, the repeated measurements are not always continuous and normally distributed. As a result, using LMMs is not relevant in all cases. An alternative approach for analysis non-Gaussian longitudinal outcomes is the so-called Generalized Linear Mixed Model (GLMM).

The Generalized Linear Model (GLM)s described by McCullagh et al. [62] generalize linear regression models to allow for non-Gaussian variables. This generalization is done by using a link function relating the linear model to the non-Gaussian variable. The GLMM is an extension of the GLM, incorporating *random effects* as well as *fixed effects* in the linear predictor. It assumes that conditionally on random

effects, the repeated outcomes of a subject are independent, the so-called *assumption of conditional independence*. Adding random effects allows for multiple observations on each subject and though takes into account the correlation within the observations of each subject, by incorporating subject-specific random effects. The random effects represent the influence of an individual on his/her repeated outcomes and are usually assumed to be independent and normally distributed.

Regarding type of outcomes, extensions of the GLMMs such as mixed effects logit models for binary data [63], proportional odds model for ordinal data [64] and Poisson mixed models for count data [65] have been proposed in the literature. In this thesis, the professional careers are coded according to the French classification of occupations without a clear hierarchical order between the employment records. Therefore, the focus of this document will be on the GLMMs for categorical nominal data. Comprehensive overviews can be found for both Gaussian and non-Gaussian cases in Verbeke et al. [66] and Molenberghs et al. [58].

1.2.2 Baseline-Category Logit Random Effects Model

When the response variables are not ordered, an appropriate link function is the *baseline-category logit* [67, 68]. The baseline-category logit model with random effects, also known as *mixed-effects multinomial logistic regression model*, pairs each category with an arbitrary reference category [60, Chapter 13].

Let n be the number of subjects in the study and m_i the number of repeated values for each subject, $i = 1, \cdots, n$. Let Y_{ij} denotes the j-th value for subject i. We assume that the repeated values are nominal data with K modalities, $Y_{ij} = k \in \{1, \cdots, K\}$. Let X_{ij} be a $p \times 1$ vector of predictors for fixed effects and W_{ij} be a $q \times 1$ vector of predictors for the random effects. The linear predictor of the GLMM for nominal outcomes is defined as $\eta_{ijk} = \alpha_k + X_{ij}^{'}\beta_k + W_{ij}^{'}b_{ik}$, and the probability, π_{ijk}, that the modality k is observed for the j-th value of a given individual i, conditional on the random effects b_i, is given by:

$$\pi_{ijk} = P\Big(Y_{ij} = k | X_{ij}, W_{ij}, b_{ik}\Big) = \begin{cases} \frac{1}{1+\sum_{h=1}^{K-1} \exp\left(\alpha_h + X_{ij}^{'}\beta_h + W_{ij}^{'}b_{ih}\right)} & \text{if } k = K \\ \frac{\exp\left(\alpha_k + X_{ij}^{'}\beta_k + W_{ij}^{'}b_{ik}\right)}{1+\sum_{h=1}^{K-1} \exp\left(\alpha_h + X_{ij}^{'}\beta_h + W_{ij}^{'}b_{ih}\right)} & \text{if } k = 1, \cdots, K-1 \end{cases} \tag{1.1}$$

Let $\alpha = (\alpha_1, \cdots, \alpha_{K-1})'$, the vector of intercepts and $\alpha_K = 0$. $\beta_k = (\beta_{k1}, \cdots, \beta_{kp})'$ is a $p \times 1$ vector of the fixed effects parameters with $\beta_K = 0$. So $\exp(\beta_{ks})$, with β_{ks} the s-th element of β_k, can be interpreted as the increase in odds of falling into modality k versus modality K resulting from a one-unit increase in the s-th covariate, holding the other covariates constant. Let $\beta = (\beta_1', \cdots, \beta'_{K-1})'$. $b_{ik} = (b_{ik1}, \cdots, b_{ikq})'$ is a $q \times 1$ vector of the random effects for subject i in the k-th modality. The random effects b_{ik} are commonly assumed to follow a multivariate Gaussian distribution with the expectation vector zero and the covariance matrix Σ_{b_k}, $b_{ik} \sim \mathcal{N}_q(0, \Sigma_{b_k})$. Let $b_i = (b_{i1}', \cdots, b_{i,K-1}')'$ be the $(K-1)q \times 1$ vector of the random effects for subject i following a multivariate Gaussian distribution with the expectation vector zero and the covariance matrix Σ_b, $b_i \sim \mathcal{N}_{(K-1)q}(0, \Sigma_b)$, defined as:

$$\Sigma_b = \begin{pmatrix} \Sigma_{b_1} & \Sigma_{b_1 b_2} & \cdots & \Sigma_{b_1 b_{K-1}} \\ \Sigma_{b_2 b_1} & \Sigma_{b_2} & \cdots & \Sigma_{b_2 b_{K-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{b_{K-1} b_1} & \Sigma_{b_{K-1} b_2} & \cdots & \Sigma_{b_{K-1}} \end{pmatrix} \tag{1.2}$$

1.3 GLMM Model fitting and inference

As explained in section 1.2.1, the GLMM assumes that conditionally on random effects b_i, the response measures of a subject Y_{ij} are independent. The GLMM has two components, a linear predictor $\eta_{ij} = \alpha + X_{ij}'\beta + W_{ij}'b_i$ and a link function $g(.)$, satisfying the Equation (1.3)

$$\mu_{ij} = E[Y_{ij} \mid b_i] = g^{-1}(\eta_{ij}) \tag{1.3}$$

where random effects b_i are drawn independently from a multivariate normal distribution with mean vector 0 and covariance Σ_b, $\mathcal{N}(0, \Sigma_b)$. Let $f_{ij}(y_{ij} \mid b_i, \alpha, \beta)$ be the conditional density function corresponding to Y_{ij} given b_i, α and β. The likelihood contribution of subject i which is the marginal density for Y_i, is given by:

$$f_i(y_i \mid \alpha, \beta, \Sigma) = \int \prod_{j=1}^{m_i} f_{ij}(y_{ij} \mid b_i, \alpha, \beta) f(b_i \mid \Sigma_b) db_i \tag{1.4}$$

with $f(b_i \mid \Sigma_b)$ the density function of the random effects b_i. By assuming independence between subjects, the so-called *marginal likelihood* function is derived as

$$L(\alpha, \beta, \Sigma) = \prod_{i=1}^{n} f_i(y_i \mid \alpha, \beta, \Sigma) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} f_{ij}(y_{ij} \mid b_i, \alpha, \beta) f(b_i \mid \Sigma_b) db_i \tag{1.5}$$

Estimation of the parameters α, β and Σ is often based on the Maximum Likelihood (ML) method. When the response outcomes are normally distributed, the marginal likelihood function in (1.5) has a closed-form solution. However, when the marginal likelihood has no closed-form, numerical techniques is needed to approximate the integration over random effects. For this purpose different methods have been developed in the literature. As stated by Fahrmeir et al. [69, Chapter 7], two different strategies may be considered for the estimation, the *direct* approach and the *indirect* approach.

The direct approach uses directly integration techniques such as Gauss-Hermite or Monte Carlo to approximate the marginal likelihood function. Then, iterative algorithms are used in order to calculate ML estimators. The indirect approach applies an Expectation-Maximization (EM) algorithm, in which the conditional expectations in the Expectation step are calculated using Gauss-Hermite or Monte Carlo techniques. Fisher scoring is used in the maximization of the Maximization step. Although the second approach takes much more time than the first one, since the EM algorithm never decreases the log likelihood, the indirect approach is numerically more stable than the direct approach which does not have this property. Detailed accounts of these two strategies in the context of random effects can be found in Hedeker et al. [70], Pinheiro et al. [71], Fahrmeir et al. [69] and McCulloch et al. [72]. In this dissertation we focus on the indirect approach since it will be convenient for parameter estimation in the context of joint modelling presented in Chapter 4. Brief description of the EM algorithm and the so-called *Gauss-Hermite quadrature technique* are given followed by the indirect maximization approach.

1.3.1 The EM algorithm

The EM algorithm [73, 74] is a general iterative approach to obtain Maximum Likelihood Estimation (MLE) in the context of incomplete data. This algorithm can be applied to a remarkably broad family of estimation problems that are not usually considered to involve missing data. Detailed descriptions and applications of this algorithm can be found in Rubin [75].

The rational of the foundation of the EM algorithm is based on associating with the observed incomplete data problem, a complete data problem for which

MLE is much simpler. In order to start the algorithm, initial values are chosen for the parameters and then it continues by iteration between two steps, namely the Expectation (E) step and the Maximization (M) step, until convergence. The E-step calculates the conditional expectation of the complete data log-likelihood given the observed data and the parameter estimates. Then the M-step finds the parameter estimates which maximize the complete data log-likelihood from the E-step.

Let Y represents the data consisting of an observed part Y^o and a missing part Y^m. The EM algorithm aims to estimate the parameter vector α of the observed data Y^0 by iterating between E-step and M-step. Usually employing complete case estimate to choose the initial value for the parameter vector α. In the E-step, the expected value of the complete data log-likelihood given the current values $\alpha^{(t)}$ and the observed data is calculated as follows:

$$Q(\alpha \mid \alpha^{(t)}) = \int l(\alpha, Y) f(Y^m \mid Y^o, \alpha^{(t)}) dY^m = E\Big[l(\alpha \mid Y) \mid Y^o, \alpha^{(t)}\Big] \tag{1.6}$$

In the M-step, the updated parameters $\alpha^{(t+1)}$ are obtained satisfying

$$\alpha^{(t+1)} = argmax_{\alpha} Q(\alpha \mid \alpha^{(t)}) \tag{1.7}$$

As showed by Dempster et al. [74], in the EM algorithm, at each iteration, the observed data log-likelihood increases or stays constant, $\log f(Y^o \mid \alpha^{(t+1)}) \geq \log f(Y^o \mid \alpha^{(t)})$. Therefore, the convergence of the log-likelihood against a global or local maximum or stationary point is guaranteed. In general, if more than one maximum or stationary point exists, this convergence requires stronger regularity conditions, which are ensured for complete data densities of the exponential family. However, its convergence rate could be slow which is reflecting the relative size of the unobservable data.

1.3.1.1 Standard Errors Estimation

The EM algorithm does not provide directly an estimate of the covariance matrix of the MLE, contrary to other estimation methods. Methods for estimation of the covariance matrix in the context of the EM algorithm are usually based on the observed information matrix, $I(\alpha \mid Y) = [-\partial^2 l(\alpha \mid Y)/\partial\alpha\partial\alpha']$, the expected information matrix, $I(\alpha) = [E[-\partial^2 l(\alpha \mid Y)/\partial\alpha\partial\alpha']]$ or on resampling methods. For

the two first approaches, the covariance matrix could be estimated by inverting the observed or expected information matrices evaluated at the estimation parameter $\hat{\alpha}$ obtained by the MLE.

Louis [76] showed that the observed information matrix can be obtained in terms of the conditional moments of the gradient and curvature of the complete-data log-likelihood function, which are easier to handle than the corresponding derivatives of the log-likelihood function with random effects, proposed within the EM framework. An alternative approach is to obtain the Hessian by differentiating the likelihood function.

However, the estimation of the covariance matrix based on the observed or expected information matrices are guaranteed to be valid inferentially only asymptotically. For instance, in mixture models, to apply the asymptotic theory of maximum likelihood, the sample size n should be very large. To address this problem, the bootstrap approach was proposed in the literature as an alternative method for standard errors estimation [77]. More details on the EM algorithm and its extensions can be found in McLachlan et al. [78].

1.3.2 Gauss-Hermite Quadrature

A popular method for approximating normal integrals is the Gauss-Hermite quadrature [79].

Let

$$\omega_N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big)$$

be the normal density function with mean μ and variance σ^2. For every regular function $g(.)$, the Gauss-Hermite approximation gives

$$\int_{-\infty}^{+\infty} \exp(-x^2) g(x) dx \approx \sum_{i=1}^{k} \omega_i g(x_i) \tag{1.8}$$

where k is the number of sample points, the nodes x_i are the roots of the Hermite polynomial with degree k, $H_k(x)$, and the ω_i are the associated weights calculated by:

$$\omega_i = \frac{2^{k-1} k! \sqrt{\pi}}{k^2 [H_{k-1}(x_i)]^2} \tag{1.9}$$

More generally, if $f(x) = \omega_N(x \mid \mu, \sigma^2)g(x)$, the integral $\int_{-\infty}^{+\infty} f(x)dx$ is approximated by substituting $x = \sqrt{2}\sigma z + \mu$:

$$\int_{-\infty}^{+\infty} f(x)dx \approx \sum_{i=1}^{k} \nu_i g(\sqrt{2}\sigma x_i + \mu) \tag{1.10}$$

with $\nu_i = \pi^{-1/2}\omega_i$.

In the case of an m-dimensional $x = (x_1, \cdots, x_m)$, a multivariate integration is needed:

$$\int_{R^m} f(x)dx = \int_R \cdots \int_R \omega(x_1, \cdots, x_m)g(x_1, \cdots, x_m)dx_1 \cdots dx_m \tag{1.11}$$

with $\omega(x) = \exp(-x^{'}x)$ and $f(x) = \omega(x)g(x)$. By applying a Cartesian product rule and the univariate Gauss-Hermite rule on each component of x, the following approximation is obtained:

$$\int_{R^m} f(x)dx \approx \sum_{i_1}^{k1} \omega_{i_1}^{(1)} \cdots \sum_{im=1}^{km} \omega_{i_m}^{(m)} g\Big(x_{i_1}^{(1)}, \cdots, x_{i_m}^{(m)}\Big) \tag{1.12}$$

with $x_{i_r}^{(r)}$ being the i_r-th root of the Hermite polynomial of degree k_r and $w_{i_r}^{(r)}$ being the corresponding weight. The number of nodes increases exponentially with dimension and therefore, the Cartesian product rules are less appropriate for high-dimensional integrals. Likewise, in the general case, i.e., if $f(x) = \omega_N(x \mid \mu, \Sigma)g(x)$ with $\mu = (\mu_1, \cdots, \mu_m)$ and variance-covariance matrix Σ, the multivariate integrals are approximated by substituting $x = \sqrt{2}\Sigma^{1/2}z + \mu$, where $\Sigma^{1/2}$ is the left Cholesky square root.

The quadrature technique needs k points in each of m dimensions and thus the integrals are approximated with a summation over k^m quadrature points. This technique is computationally feasible for integral dimensions up to 6. Alternative approach that has been developed in the literature namely the Monte Carlo methods that uses k nodes randomly sampled. The issue with this approach is the choice of k since for small k, the method results to poor approximation and for big k computation time increases. To address this problem, automated Monte Carlo [67] has been proposed in which at each iteration if the Monte Carlo error exceed the change in the estimation of previous iteration, we increase k.

1.3.3 Newton-Raphson Method

The Newton-Raphson is a numerical method to solve equations numerically. Let x_0 be an estimation of $x = x_0 + h$, the true root of function $f(.)$. Since h is small, using the linear approximation we conclude that

$$0 = f(x) = f(x_0 + h) \approx f(x_0) + hf^{'}(x_0) \tag{1.13}$$

And therefore,

$$x = x_0 + h \approx x_0 - \frac{f(x_0)}{f^{'}(x_0)} \tag{1.14}$$

where the right side of (1.14) is the new estimation of x. The process can then be repeated until convergence to a fixed point,

$$x_{n+1} = x_n - \frac{f(x_n)}{f^{'}(x_n)} \tag{1.15}$$

In the EM algorithm, at each iteration, the parameter estimations are updated via a one-step Newton-Raphson update.

1.3.4 Indirect Maximization Based on the EM Algorithm

One of the approaches for maximizing the marginal likelihood function (1.5) is based on an EM algorithm. Let $Y = (y_1, \cdots, y_n)$ be the incomplete data, $B = (b_1, \cdots, b_n)$ be the unobserved data and Ψ be the parameter vector. As explained in Section 1.3.1, EM algorithm uses the complete-data log-likelihood defined by

$$\log f(Y, B \mid \Psi) = \sum_{i=1}^{n} \log f(y_i \mid \Psi, b_i) + \sum_{i=1}^{n} \log f(b_i) \tag{1.16}$$

where $f(y_i)$ denotes the density function of the incomplete data Y and $f(b_i)$ denotes the density function of the unobserved random effects b_i.

In the E-step, the expectation of (1.16) conditional on the observed data and parameter vector from the previous step is determined:

$$Q(\Psi \mid \Psi^{(t)}) = E\Big[log(Y, B \mid Y, \Psi)\Big] = \int log\Big(f(Y, B \mid \Psi)\Big) f(B \mid Y, \Psi^{(t)}) dB \tag{1.17}$$

Using Bayes' theorem and the conditional independence assumption explained in Section 1.2.1, the posterior function $f(B \mid Y, \Psi^{(t)})$ is obtained by

$$f(B \mid Y, \Psi^{(t)}) = \frac{\prod_{i=1}^{n} f(y_i \mid b_i, \Psi^{(t)}) \prod_{i=1}^{n} f(b_i)}{\prod_{i=1}^{n} \int f(y_i \mid b_i, \Psi^{(t)}) f(b_i) db_i} \tag{1.18}$$

Hence,

$$Q(\Psi \mid \Psi^{(t)}) = \sum_{i=1}^{n} \frac{\int \Big[\log f(y_i \mid b_i, \Psi) + \log f(b_i)\Big] f(y_i \mid b_i, \Psi^{(t)}) f(b_i) db_i}{\int f(y_i \mid b_i, \Psi^{(t)}) f(b_i) db_i} \tag{1.19}$$

Calculating these integrals might be challenging, however, employing the Gauss-Hermite quadrature techniques enable to approximate the integrals in (1.19) providing that the random effects b_i are gaussian.

In the M-step, the obtained function $Q(\Psi \mid \Psi^{(t)})$ should be maximized with respect to the parameter vector Ψ. If $Q^{GH}(\Psi \mid \Psi^{(t)})$ be the approximation of $Q(\Psi \mid \Psi^{(t)})$ using the Gauss-Hermite rule, in the M-step we should solve the following equation,

$$S(\Psi \mid \Psi^{(t)}) = \frac{\partial Q^{GH}(\Psi \mid \Psi^{(t)})}{\partial \Psi} \tag{1.20}$$

With the GLMM formulation, a closed-form solution cannot be obtained for the fixed effects. Hence, a one-step Newton-Raphson method is applied to update these parameters in each iteration:

$$\hat{\Psi}^{(t+1)} = \hat{\Psi}^{(t)} - \frac{S(\hat{\Psi}^{(t)})}{\partial S(\hat{\Psi}^{(t)})/\partial \Psi} \tag{1.21}$$

where $\hat{\Psi}^{(t)}$ denotes the value of the parameter vector in the t-th iteration, and $\partial S(\hat{\Psi}^{(t)})/\partial \Psi$ denotes the Hessian matrix evaluated at $\hat{\Psi}^{(t)}$.

1.4 Missing data in longitudinal studies

In longitudinal studies, for each individual, data is collected at specific follow-up times. However, it is possible that some individuals miss some of their planned measurements. An important challenge in analysing longitudinal outcomes is the problem of these missing data. Depending on the missing data patterns, two type of missingness can be distinguished, namely *monotone* and *non-monotone*. Monotone missingness covers the cases where all values of individual are not observed after a scheduled time-point and the individual is said to have dropped-out of the study. The reason might be events such as death or country moving. On the other hand, non-monotone missingness, also called *intermittent* missingness, covers the cases where the responses of an individual are observed following some missing values for that individual.

Suppose that for individual i, it is designed to measure the outcome of interest Y at m_i time-points which means that $Y_i = (Y_{i1}, \cdots, Y_{im_i})'$ is the expected vector of the outcome for individual i. The *missing data indicator*, R_{ij} is defined as

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases} \tag{1.22}$$

Then, the vector Y_i is factorized into two subvectors Y_i^o and Y_i^m, namely the *observed* subvector and the *missing* subvector. These subvectors are representing the vector containing Y_{ij} for which $R_{ij} = 1$ and the vector containing Y_{ij} for which $R_{ij} = 0$, respectively. Therefore, the *full data* (Y_i, R_i) consists of the *complete data*, which refers to the vector of outcome that would have been recorded if there were no missing data, and the vector of missing data indicators $R_i = (R_{i1}, \cdots, R_{im_i})'$.

In a time-to-event setting, which is the concern of this work, one could consider monotone missingness as an event, identified as the time that terminates the repeated measurements sequence.

If missingness process is associated with longitudinal measurements, unobserved data can introduce bias in the results, which is the main concern of longitudinal analysis with missing data. Consequently, it is important to distinguish between different missing data mechanism. This mechanism can be seen as the probability model that describes the relation between the response data y_i and the missing data r_i processes. Rubin's taxonomy of missing data mechanism has been developed based on the conditional density of the missing data process r_i given the complete data y_i [80, 81]:

$$f(r_i \mid y_i^o, y_i^m; \alpha_r) \tag{1.23}$$

with α_r being the parameter vector of missingness process. Rubin's classification distinguishes three types of missing data mechanism [58], namely Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR).

1.4.1 Missing Completely at Random

Under the MCAR mechanism, the probability of an observation being missing is independent of the responses:

$$f(r_i \mid y_i^o, y_i^m; \alpha_r) = f(r_i; \alpha_r) \tag{1.24}$$

This means that the distribution of the observed data y_i^o is the same as the distribution of the complete data y_i. In this situation, the data can be analysed supposedly that the missing data process was predetermined. Therefore, under MCAR it is possible to ignore the missing data process and to obtain valid inferences, whether using likelihood-based or Bayesian approaches.

Methods that are usually used to analyse longitudinal data with missing values under MCAR are Complete Case (CC) analysis and Last Observation Carried Forward (LOCF). In CC analysis, all individuals with missing values are excluded from statistical analysis. However, using this approach leads to loss of information. The other approach, LOCF [82], that can be regarded as an imputation strategy, consists of substituting the last observed value whenever a value is missing. This approach is based on a strong assumption, that is the subject's measurements values do not change during the period they are unobserved. When this assumption is violated, the magnitude and direction of the produced bias depend on the true unknown regression coefficients [58, Chapter 27].

1.4.2 Missing at Random (MAR)

The MAR mechanism supposes that the probability of missingness is conditionally independent of the unobserved data given the observed values:

$$f(r_i \mid y_i^o, y_i^m; \alpha_r) = f(r_i \mid y_i^o; \alpha_r) \tag{1.25}$$

This class of missingness is also known as the *random missingness*. Under MAR, the missingness process depends on the observed value of y_i^o, so the distribution of y_i does not match the distribution of y_i^o and therefore, the observed data is not a

random sample of the original population. However, the conditional distribution of missing values given the observed data can be written as:

$$\begin{aligned} f(y_i^m \mid y_i^o, r_i; \alpha) &= \frac{f(y_i^m, y_i^o, r_i; \alpha)}{f(y_i^o, r_i; \alpha)} = \frac{f(r_i \mid y_i^o, y_i^m; \alpha_r) f(y_i^o, y_i^m; \alpha_y)}{f(r_i \mid y_i^o; \alpha_r) f(y_i^o; \alpha_y)} \\ &= \frac{f(r_i \mid y_i^o; \alpha_r) f(y_i^o, y_i^m; \alpha_y)}{f(r_i \mid y_i^o; \alpha_r) f(y_i^o; \alpha_y)} = \frac{f(y_i^o, y_i^m; \alpha_y)}{f(y_i^o; \alpha_y)} \\ &= f(y_i^m \mid y_i^o; \alpha_y) \end{aligned} \tag{1.26}$$

where α is the parameter vector of the joint distribution of the measurements and missingness processes and α_y is the parameter vector of the measurements model. The Equation (1.26) shows that under MAR, missing values can be predicted using the observed data assuming a model for the joint distribution (y_i^o, y_i^m).

Under MAR the likelihood of the complete data (y_i^o, y_i^m, r_i) for the i-th subject factors into two components as follows:

$$\begin{aligned} L_i(\alpha) &= \int f(y_i, r_i; \alpha) dy_i^m \\ &= \int f(y_i^o, y_i^m; \alpha_y) f(r_i \mid y_i^o, y_i^m; \alpha_r) dy_i^m \\ &= \int f(y_i^o, y_i^m; \alpha_y) f(r_i \mid y_i^o; \alpha_r) dy_i^m \\ &= f(y_i^o; \alpha_y) f(r_i \mid y_i^o; \alpha_r) \\ &= L_i(\alpha_y) \times L_i(\alpha_r) \end{aligned} \tag{1.27}$$

In addition if the two parameter vectors α_y and α_r are *disjoint*, i.e. if the parameter space of the full vector $\alpha = (\alpha_y^{'}, \alpha_r^{'})^{'}$ is the product of the parameter spaces of vectors α_y and α_r, then inference for α_y can be based on the direct likelihood inference using all observed data, ignoring the likelihood of the missing values [81]. This important property under the MAR is known as *ignorability*.

An alternative approach under MAR is the Multiple Imputation (MI) approach of Rubin [75]. The idea is based on replacing missing values with a set of M values drawn from the distribution of the missing data given the observed values. Standard statistical procedures for complete data are then applied on each imputed dataset and the results are combined.

1.4.3 Missing Not at Random (MNAR)

In this situation, the probability that a measurement is not observed depends on the unobserved values. The MNAR missingness is also called *nonrandom missingness*. Similarly to MAR, under MNAR, observed data is not a random sample of the original population. However, contrary to the MAR, under MNAR the predictive distribution of y_i^m conditional on the observed data depends on both observed values y_i^o and $f(r_i \mid y_i)$. In this case, the MNAR mechanism is *nonignorable* and thus, the model for the missingness process should be included in the analysis.

Under MNAR, valid inferences based on the likelihood require specification of the joint distribution of the measurement and missingness processes. The specification of this joint distribution can be classified into three type of model families [81, 83]: *pattern mixture models*, *selection models* and *shared-parameter models*.

The pattern-mixture approach [84] models the distribution of data conditional on the missingness mechanism which correspond to the following fatorization:

$$f(y_i^o, y_i^m, r_i; \alpha) = f(y_i^o, y_i^m \mid r_i; \alpha_y) f(r_i; \alpha_r) \tag{1.28}$$

In this factorization, the joint distribution is written as the product of a conditional model for the longitudinal data given the missingness process and a marginal model for the missingness process. In this factorization, the interest lies in estimating the longitudinal trajectory conditional on the missingness process. Therefore, first the samples are stratified according to the missingness process and then, different models can be postulated for the longitudinal data [85, 86].

The selection models factorize the joint distribution as follows:

$$f(y_i^o, y_i^m, r_i; \alpha) = f(y_i^o, y_i^m; \alpha_y) f(r_i \mid y_i^o, y_i^m; \alpha_r) \tag{1.29}$$

This approach models the complete data together with the missingness process conditional on the complete data. In this class of models, a marginal density for the longitudinal data and a model for the missingness process conditional on the longitudinal outcomes are chosen. The focus is therefore on estimating the missingness process given the repeated outcomes.

Extensions of these models, namely *random pattern-mixture models* and *random selection models*, have been developed in the literature by incorporating random

effects, u, into the models. Omitting parameters, factorization of these two extensions are as follows, respectively for random pattern-mixture models and random selection models:

$$f(y_i^o, y_i^m, r_i, u_i) = f(y_i^o, y_i^m \mid r_i) f(r_i \mid u_i) f(u_i) \tag{1.30}$$

$$f(y_i^o, y_i^m, r_i, u_i) = f(r_i \mid y_i^o, y_i^m) f(y_i^o, y_i^m \mid u_i) f(u_i) \tag{1.31}$$

Diggle [87] defined a third class of models, *random effects models*, which assumes that both longitudinal data and missingness process depend on an unobserved random effect, with a specified bivariate distribution for the random effects [85]:

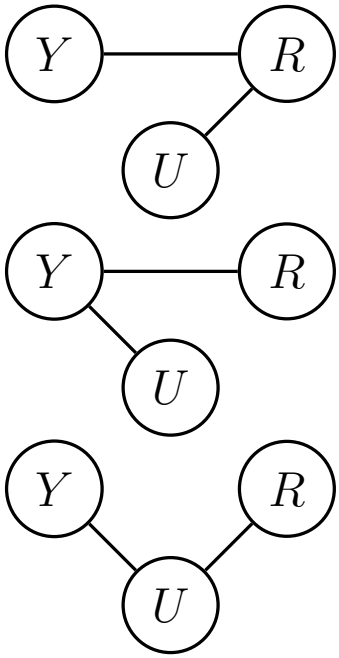
$$f(y_i^o, y_i^m, r_i, u_i) = f(y_i^o, yi^m \mid u_{i1}) f(r_i \mid u_{i2}) f(u_i) \tag{1.32}$$

where $u_i = (u_{i1}, u_{i2})'$. This class of models are also known as *shared-parameter models* as the measurements process and missingness mechanism are modelled by sharing random effects. In this class of models, the two measurement and missingness processes are assumed to be independent given random effects.

These three classes are shown visually in Figure 1.1 by diagrams presented in Diggle [87], where Y, R and U are representing longitudinal outcomes process, missingness process and an unobserved process, respectively. The absence of an edge between two nodes indicates conditional independence between the two nodes given the third one.

A final remark regarding Rubin's Taxonomy is that as shown by Molenberghs et al. [88], in practice, it is not possible to distinguish between MAR and MNAR. Besides studies with missingness by design, other missingness mechanism are not verifiable. Often, primary analysis are based on the MAR assumption unless the cases where the obvious MAR model does not fit the observed data. In this situation, it is attractive to fit a model under MNAR and then use the MAR model for sensitivity analysis.

Comprehensive overviews of models for the joint analysis of longitudinal outcomes with missing data can be found in Diggle et al. [59] and Hogan et al. [86, 89]. *Random effects models* will be the focus of Chapter 4 and Chapter 5.



(a) Pattern-mixture models: The individual random effects determine the missingness process, which after being predefined develops the longitudinal measurements.

(b) Selection models: Longitudinal data are influenced by individual random effects. The realisation of the longitudinal process influences the missingness process.

(c) Random effects models: Both longitudinal and missingness processes are a joint response to an unobserved individual process. The measurement and missingness processes are independent conditional on this unobserved process.

Figure 1.1 – Graphical model for different strategies in joint analysis of longitudinal data with *nonignorable* missing values

1.4.4 Missing data and professional scope in Cosmop-DADS

Since DADS declarations are mandatory for employers, there were theoretically no missing occupational episodes for employees working in companies within the DADS scope. However, professional trajectories were not fully observed for several individuals. We note that only around 17% of subjects have complete data (141733 women and 159149 men).

The first set of missing episodes concerned the years 1981, 1983 and 1990 which was mentioned in Section 2.1. We completed these episodes with information from the previous years. However, for other years, some occupations could not be classified in the five occupational classes, called miscoded occupations in Appendix A. We decided to impute these occupations using a multivariable multinomial logistic regression [90], incorporating sex, age and type of employment in the imputation model.

Regional and local authorities were not fully covered by DADS declarations before 1987. Therefore, any occupation of this type was excluded from our professional scope. The same decision was taken for occupations declared in the craftsmen and trade-related workers class, as those in DADS are not representative of this class

in the general population. In summary, the professional scope in this dissertation contains the DADS scope mentioned in Section 2.1 excluding regional and local authorities, and craftsmen and trade related workers class.

The Cosmop-DADS database is also containing individuals that left the follow-up. Some of them reapear in the database after some missing years and some of them not. We refer to these type of missing data as temporary exit and permanent exit, respectively. These exits may represent those professional episodes practiced by an individual in careers not covered by the professional scope of this study or an inactivity or retirement, as they are not covered in the DADS panel.

Missing data is a common issue in longitudinal study. The temporary and permanent exits of the motivating example, Cosmop-DADS, contain inactive individuals. In the literature it is already well established that inactivity is associated with an increased mortality risk [5, 91], consequently, these exits should not be ignored. As an alternative approach, imputation methods were introduced in the literature which need making assumptions on the missing data mechanism [75]. However, in the Cosmop-DADS database, missing professional episodes, the temporary and permanent exits, resulted from different scenarios. Considering that these missing professional episodes are a mixture of working outside the study scope, being inactive or retired, in the absence of other complementary database, building a sound imputation model is not feasible. We decided to add an additional *outside the scope* category to the four remaining categories. In other words, the *outside the scope* category, will gather all professional episodes that are not covered by DADS scope, careers in regional and local authorities class and in craftsmen and trade related workers class, in addition to the inactive and retired episodes. An overview of the extent of all these cases of missingness is shown in Appendix A.

Chapter 2

Background on survival analysis and competing risks

Survival analysis focuses on the study of time-to-event data defined as the time until the occurrence of an event of interest. In epidemiological and clinical studies, the event of interest may be death, onset or recurrence of a disease, while in demography, this event could be marriage or divorce. The term *failure* is also utilized in survival analysis signifying the event of interest. The occurrence of an event may be modelled as a transition from one state to another one, which indicates the possibility of analysing the time-to-event data as a multi-state model, as shown in Figure 2.1. If all subjects are present at the beginning of the study, each individual is at the "Event-free" or transient state until the occurrence of the event, if the event happens the subject moves to the absorbing state "Failure".

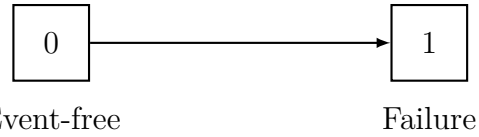


Figure 2.1 – Multi-state model representation for survival analysis

However, in many contexts, there is more than one event of interest such that the occurrence of one event prevent the occurrence of other events [92–94]. For example, in cardiovascular studies, death from other causes should be taken into account in addition to death from cardiovascular diseases. Thus, in this framework, the observed time is the time until the occurrence of any first event. As shown in

Figure 2.2, a multi-state formulation for the competing risks problems can also be adopted, with a transient "Event-free" state occupied by all subjects at the beginning and g absorbing states representing competing failure types $d \in \{1, \cdots, g\}$. In such formulation, the occurrence of an event may be modelled as a transition into any absorbing state. An alternative existing approach for competing risks is based on latent failure times. However, this formulation appears to cause some interpretational confusions and identifiability problems [95–97]. Therefore, due to the lack of plausibility of this later, the multi-state model formulation for competing risks will be considered in this dissertation.

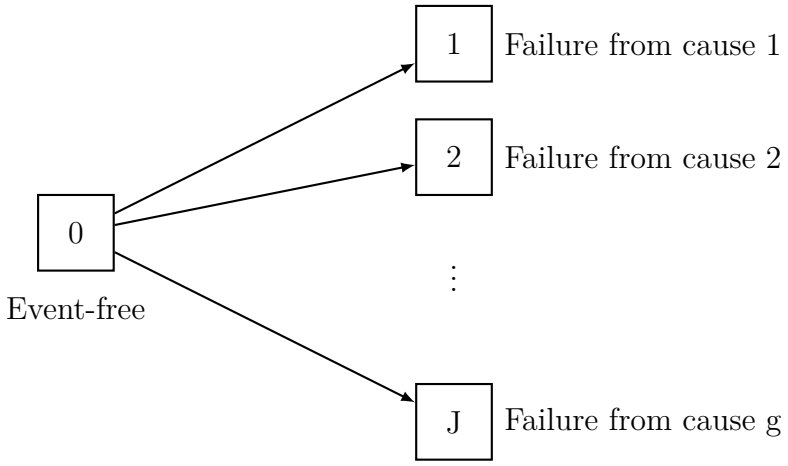


Figure 2.2 – Multi-state model representation for competing risks problem

In other words, competing risks models generalize survival analysis from a single event to multiple competing events. Using this approach is necessarily in many domains, particularly in clinical studies and in epidemiology. Different regression models have been proposed in the literature in order to summarize the effect of explanatory covariates in the competing risks setting. We start by introducing some notations and basic functionals in the competing risks framework.

2.1 Notations and basic definitions

In this section, we consider the competing risks framework in which more than one event may occur. Let T, be the response variable representing the *failure time* or the waiting time until the occurrence of the first event. As in this thesis the focus is on mortality, the failure events can be considered as death from different causes,

indexed by $d \in \{1, 2, \cdots, g\}$. Let D be a random variable representing the *cause of failure* and let Z be a vector of covariates.

The important characteristics of survival analysis that distinguishes this domain from other statistical analysis are *censoring* and *truncation*. In the presence of censoring and truncation, the survival data is not fully collected, i.e., the failure time on all subjects is not observed. Censoring and truncation are various disturbances, independent or not from the multivariate failure time. In this document we focus on *right-censored* and *left-truncated* data that we detail in the following.

In some studies, for a subset of individuals under study, the event is only known to occur after a certain time-point C. For instance, some individuals may be dropped out of the study due to relocation or the study may be closed while there are still event-free individuals at the endpoint. In such cases, the only available information for surviving individuals is that their failure time is greater than the value C, named *censoring time*. This mechanism leads to incomplete data known as *right-censoring*. Defining T_i as the observed event time or censoring of subject i, given that T_i^* is the survival time and C_i is the right-censoring time of subject i, by definition $T_i = min(T_i^*, C_i)$. Let $\delta_i = 1\{T_i^* \leq C_i\}$ be the indicator of censorship which indicates if a failure occurred or not and $\epsilon_i = \delta_i \times D_i \in \{0, 1, 2, \cdots, g\}$ be the status indicator. We note that $\epsilon_i = 0$ if the failure time is censored.

In addition, regarding the study design, sometimes individuals enter the study at a time L later than time 0. In this situation, the failure time T is observed with *delayed entry*, $T > L$, and the data is said to be *left-truncated*. This implies that only the survival of subjects surviving to the date of inclusion in the study may be examined. In short, in the presence of right-censoring and left-truncation, the observed data can be summarized as $(L_i, T_i, \epsilon_i, Z_i)$.

Even though censoring and truncation are two phenomenon representing a particular type of missing data, they should not detract the attention from the main objective which is making inferences about the joint distribution of (T, D), if there were neither censoring nor truncation in the study [97]. To address valid inference on the joint distribution of (T, D), it is convenient to make *random censoring* and *random truncation* assumptions, which means that (T, D) is independent of (L, C) given the covariates Z. However, weaker assumptions, namely *independent cen-*

soring and *independent truncation*, suffice for application of the martingale theory and counting processes underlying most of the main results in competing risks [92]. These assumptions suggest that if a subject is still alive at time t, the additional information that the individual is uncensored and not delayed entry will not change his/her instantaneous probability of failing from cause d [97].

Recalling from the multi-state formulation of the competing risks problem (Figure 2.2), every subject is initially in the "event-free" state. Each subject stays in the initial state until the occurrence of any first event, at time T. Occurrence of the failure type d is modelled by the transition from state 0 to state d at this time. Therefore, at the event time two components of the competing risks, (T, D), are observed with D representing the event type. As argued in Andersen et al. [92, Chapter II.6], the stochastic behaviour of a competing risks process is completely determined through the Cause-Specific Hazard rate (CSH), $\lambda_d(t), d = 1, \cdots, g$ or the transition intensities in a multi-state formulation, describing the instantaneous risk of failure from cause d,

$$\lambda_d(t, z) := \lim_{dt \to 0} \frac{P(t \leq T < t + dt, D = d \mid T \geq t, Z = z)}{dt}, \quad d = 1, \cdots, g \tag{2.1}$$

Under the random censoring assumption, the likelihood function can be introduced as a function of the CSHs

$$L = \prod_{i=1}^{n} \lambda_{d_i}(t_i, z_i)^{\delta_i} S(t_i, z_i) = \prod_{i=1}^{n} \prod_{d=1}^{g} \lambda_d(t_i, z_i)^{1(\epsilon_i = d)} \exp\Big(- \Lambda_d(t_i, z_i)\Big) \tag{2.2}$$

where $\Lambda_d(t, z) = \int_0^t \lambda_d(u, z)du$ is the cumulative hazard for cause d and $S(t, z) = \prod_{d=1}^{g} \exp(-\Lambda_d(t, z))$ is the overall survival function. Regression modelling based on the CSHs allows for a "direct" formulation of the covariates effect on the instantaneous forces that drive the patients remaining at risk at each time point t, i.e., those without any prior event.

Alternatively, the other key concept of competing risks in order to describe the joint distribution of (T, D) is the so-called Cumulative Incidence Function (CIF). The CIF are defined as the occurrence probability of the event d before time t:

$$F_d(t) := P(T \leq t, D = d), \quad d = 1, \cdots, g \tag{2.3}$$

The CIF describes the absolute risk of failing from cause d until time t. As a

result, regression models based on the CIFs may be useful when the prognosis is of interest [98, 99].

In the multi-state formulation of Figure 2.2, the CIF for failure cause d is interpreted as the probability of having transitioned to state d by time t given that the subject was in state 0 at time 0. The CIFs may be estimated using the CSHs, as following:

$$S(t) = P(T \geq t) = exp\Big(- \int_0^t (\lambda_1(u) + \cdots + \lambda_g(u))du\Big) \tag{2.4}$$

and

$$F_d(t) = \int_0^t S(u)\lambda_d(u)du, \quad d = 1, \cdots, g \tag{2.5}$$

It is important to note that in the standard survival analysis, given that $F(t) = 1 - \exp(-\int_0^t \lambda(u)du)$, there is a one-to-one correspondence between the rate $\lambda(.)$ and the risk $F(.)$. This implies that analysing survival data based on the hazard function leads to the same conclusions obtained by the analysis of the risk function. For instance, if a hazard-based regression shows an association between a certain factor and higher hazard function then the presence of this factor is also associated with a higher risk. However, the Equation 2.5 shows that CIF for cause d depends on all CSHs $\lambda_d, d = 1, \cdots, g$ through $S(u)$. Therefore, an increase in one of the CSHs will not necessarily leads to an increase in the corresponding CIF, as it depends also on the behaviour of other CSHs. This is the key feature of competing risks, that the one-to-one correspondence between CIF and CSH is no longer valid in competing risks context. Both rates and risks measures are useful in order to have a complete understanding of competing risks mechanism, as they tend to complement each other[100, 101]. We further (cf. Section 2.3) explain that the prediction of the CIFs is not possible if the model includes interval time-dependent covariates. Since the focus of this thesis is on this type of covariates, only regression methods based on the CSHs are considered here.

2.2 Regression models for the cause-specific hazard

The main objective is to assess the effect of a covariates vector Z on CSH. Figure 2.3 shows the association considered in regression models for the CSHs where Z is the observed covariates and T is the survival event process.

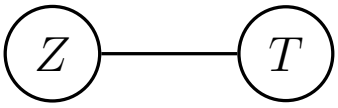


Figure 2.3 – Graphical representation of the regression models for the CSH

Let n be the number of individuals in the study, and Z_i be the $l \times 1$ vector of the observed covariates for individual i at baseline. Z_i may also be a time-dependent covariate that we will explain in more detail in Section 2.3.

2.2.1 Cox model

Among the existing methods for regression modelling of the CSH, the most widely used regression models are the *proportional hazards models*. The advantages of these regression modellings is that they are easy to fit and simple to interpret [100]. The popular *semi-parametric Cox model* [102], assumes a multiplicative effect of covariates on CSHs:

$$\lambda_d(t \mid Z) = \lambda_{0d}(t) \exp(\gamma_d^{'} Z), \qquad d = 1, \cdots, g \tag{2.6}$$

where γ_d is a l-vector of regression coefficients, $\gamma_d^{'} = (\gamma_{d1}, \cdots, \gamma_{dl})$, and $\lambda_{0d}(t)$ is an unspecified, non-negative *baseline hazard function* for cause d. Since a parametric form is only assumed for the covariate effect, this class of models are called semi-parametric. Cox model is a proportional hazards model which refers to its special property that the ratio of the CSHs of any two individuals i and j with covariates Z_i and Z_j is constant over time:

$$\frac{\lambda_d(t \mid Z_i)}{\lambda_d(t \mid Z_j)} = \exp\left(\gamma_d^{'}(Z_i - Z_j)\right) \tag{2.7}$$

The proportional hazards property can be checked by a graphical method using Schoenfeld residuals [103]. The quantity (2.7) is called the Cause-Specific Hazard

Ratio (CSHR), Hazard Ratio (HR) if $g = 1$, or the relative risk for the event d conditional on the covariate Z. The $\exp(\gamma_d)$ represents the relative change in the CSH for cause d for a one unit change in the covariate Z:

$$\exp(\gamma_d) = \frac{\lambda_d(t \mid Z = z + 1)}{\lambda_d(T \mid Z = z)} \tag{2.8}$$

Estimation of regression coefficients in (2.6) is based on the partial likelihood function in which specification of the baseline CSHs is not necessary [102],

$$pL(\gamma) = \prod_{d=1}^{g} \prod_{i=1}^{q_d} \frac{\exp(\gamma_d' Z_{(d)i(d)})}{\sum_{j \in R(t_{di})} \exp(\gamma_d' Z_{dj})} \tag{2.9}$$

where q_d denotes the number of distinct failure times due to cause d, $t_{d1} < \cdots < t_{dq_d}$, t_{di} corresponds to the i-th such time, $R(t_{di})$ is the set of individuals at risk just prior to time t_{di} and $i(d)$ is the index of the subject that died at t_{di}. The estimation of regression coefficients is then calculated by maximizing the partial likelihood 2.9. Note that the partial likelihood is a product over all observed failure times, all individuals and all failure causes. This partial likelihood can be factorized into g components, and the d-th component is algebraically identical to the partial likelihood that may be obtained by treating observed competing failure causes $\tilde{d}, \tilde{d} \in \{1, \cdots, g\} \setminus \{d\}$[1], as censoring [95, 100, 104]. In this case, one could fit the Cox model using standard software packages for the classical Cox regression by censoring the subjects who failed from other causes.

2.2.2 Poisson regression

In Cox regression models, the baseline hazard function is unspecified. However, another possibility is to choose a parametric form for the baseline hazard function. As an example we can mention the exponential model in which the baseline hazard function is constant $\lambda_0(t \mid \theta) = \theta$. A well-known example is the *Poisson regression*. Poisson regression is based on choosing time-intervals in which the baseline hazard rate is assumed to be constant and thus, compared to the Cox model, the baseline hazard function is approximated by a piece-wise constant function [105, 106].

This approach estimates covariates effects on event rates and is particularly interesting when data consists of much observations and less covariates, since estimation

[1] $A \setminus B = A \cap B^c$

of this model can be performed with much less computation. As in the Cosmop-DADS database, we are encountering large-scale data with 'large n and small p', we will focus on this model also known as piece-wise exponential regression.

Let $0 = t_0 < t_1 < t_2 < \cdots < t_K = \tau$ be a partitioning of the study time interval $[0, \tau]$, then define the baseline hazard for cause d to be a step function with a constant value in each interval, i.e.,

$$\lambda_{0d}(t) = \sum_{k=1}^{K} \theta_{kd} 1\{t \in (t_{k-1}, t_k]\} \tag{2.10}$$

with $1\{t \in (t_{k-1}, t_k]\}$ being the indicator of the k-th interval, $\theta_d = (\theta_{1d}, \cdots, \theta_{Kd})'$ and $\theta = (\theta_1', \cdots, \theta_g')'$.

The likelihood function presented in Equation (2.2) in this formulation is written as

$$L(\theta, \gamma) = \prod_{k=1}^{K} \prod_{i=1}^{n} \prod_{d=1}^{g} \Big\{ \{\theta_{kd} \exp(\gamma_d' Z_i)\}^{O_{ikd}} \exp\Big(- \theta_{kd} \exp(\gamma_d' Z_i) R_{ik} \Big) \Big\} \tag{2.11}$$

where O_{ikd} is the death indicator for cause d in the k-th interval and R_{ik} is the individual's exposure time in the k-th interval.

One of the advantages of this modelling appears for large-scale data with categorical covariates. In this case, the likelihood could be simplified to

$$L(\theta, \gamma) = \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{d=1}^{g} \Big\{ \{\theta_{kd} \exp(\gamma_d' Z^{(l)})\}^{O_{kd}^{(l)}} \exp\Big(- \theta_{kd} \exp(\gamma_d' Z^{(l)}) R_k^{(l)} \Big) \Big\} \tag{2.12}$$

where L is the number of distinct values of the covariate Z, $Z^{(1)}, \cdots, Z^{(L)}$ and

$$O_{kd}^{(l)} = \sum_{i: Z_i = Z^{(l)}} O_{ikd}, \qquad R_k^{(l)} = \sum_{i: z_i = Z^{(l)}} R_{ik}$$

Consequently, without loss of information, when the covariates are categorical or categorized, estimation can be based on aggregated quantities $O_{kd}^{(l)}$ and $R_k^{(l)}$ which is much less computational when n is much larger than L.

In this approach, the likelihood function from the piece-wise Exponential model is proportional to the likelihood one would obtain if the number of death from cause d in the k-th interval $O_{kd}^{(l)}$ were treated as independent and Poisson distributed random variable with a mean that is the product of the $R_k^{(l)}$ and the hazard rate. Accordingly, the term *Poisson regression models* is also utilized in the literature representing the piece-wise constant hazard models. This means that one could use statistical software for the Poisson distribution by including $log(R_k^{(l)})$ as an *offset*.

2.3 Time-dependent covariates

In section 2.2, we assumed that the observed covariates are time-independent, such as sex or age at baseline. However, it is possible that some components of Z be time-dependent, i.e., $Z = Z(t)$, when successive measurements are collected on study subjects as they are followed over time [107]. Such covariates could include environmental factors or clinical measurements collected during the follow-up.

For instance, in a multi-state formulation, Figure 2.2, a binary time-dependent covariate may be modelled by adding an additional transient state $\tilde{0}$:

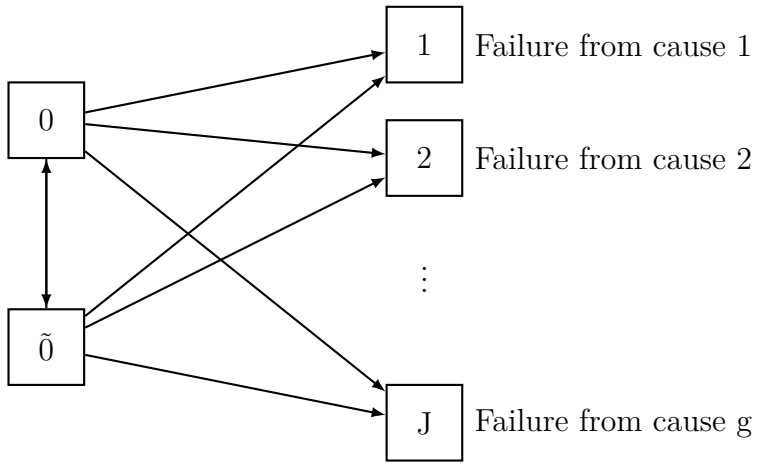


Figure 2.4 – Multi-state model representation for competing risks problem with a binary time-dependent covariate

Transitions between two transient states 0 and $\tilde{0}$ reflect the changes in binary time-dependent covariate over time. A regression model for the CSH of failure d, compares the hazard of transition $\tilde{0} \longrightarrow d$ with the hazard of transition $0 \longrightarrow d$ [104].

As explained in Kalbfleisch et al. [29, Chapter 6], two different categories of time-dependent covariates can be distinguished, *internal* or *endogenous* covariates and *external* or *exogenous* covariates. A covariate is said to be external if it satisfies the condition:

$$P(s \leq T_i < s + ds \mid T_i \geq s, \mathcal{Z}_i(s)) = P(s \leq T_i < s + ds \mid T_i \geq s, \mathcal{Z}_i(t)) \qquad (2.13)$$

for all $s, t, 0 < s \leq t$, and $ds \to 0$ with $\mathcal{Z}_i(t) = \{z_i(s), 0 \leq s < t\}$ representing the covariate's history observed for subject i up to t [29, Chapter 6]. It means that the hazard function at time u depends on the observed history of the covariate up to u,

but the occurrence of a failure in the time interval $[u, u + du)$ is independent of the future observations of the covariate. On the contrary, an internal time-dependent covariate, is the one that does not satisfy Equation (2.13) which means that the history of the covariate until t has an impact on the occurrence of a failure before t. In other words, external covariates' path is external to the individuals under study and is not directly generated by individuals behaviour in time. For instance, environmental temperature, air pollution levels and individual's age are examples of external time-dependent covariates and clinical characteristics such as blood pressure and size of tumour are internal time-dependent covariates.

External covariates may vary in a predetermined way, namely *defined time-dependent covariate*, such as individual's age [29, Chapter 6.3]. Other external covariates such as environmental temperature, are based on a stochastic process with a distribution that does not contain the parameters of the regression model of survival time, namely *ancillary time-dependent covariates*. The defined covariates path is fixed in advance and therefore, inference can be based on the partial likelihood conditional on the covariates. For an ancillary time-dependent covariate, as it is completely external to the individuals, modelling of this covariate does not include the parameter of interest and is not necessary to be specified. As a result, the survival function conditional on the observed covariate path does not change and thus, inference based on the partial likelihood can still be performed [92, Chapter III.5]. To handle these covariates, Cox model have been extended using the counting process formulation, known as *extended Cox model* or the *Andersen-Gill model* [92]. Interpretation of regression coefficients is exactly the same as it was in the standard Cox model. However, the proportionality assumption is no longer valid as the covariate is time-dependent.

Internal time-dependent covariates complicate statistical analysis, as the survival function is a function of both the hazard rate and the development process of the covariates. When the interest lies in estimation of hazards functions, it is still possible to use the partial likelihood conditionally on the observed covariates up to the time just before t [29]. But since the extended Cox model is based on the assumption that the covariates path is predictable, analysis based on the extended Cox model for internal covariates is not optimal and might be involved by a potential

bias.

In the presence of internal time-dependent covariates, estimating the survival probability and the CIFs is no longer possible based on the obtained cause-specific hazards [107]. This could be explained by the fact that the probability $S(t \mid Z(t)) = P(T \geq t \mid Z)$ is equal to one since observing the covariate at time t denotes the survival of the individual at this time. Therefore, prediction of the CIF which depends on the survival probability (2.5) is not possible if the model includes an internal time-dependent covariates.

2.4 Survival models with random effects or frailty models

In the classical survival analysis, it is assumed that the survival of individuals with the same values of the covariates is the same. However, there might be extra heterogeneities that are not included in the model. The survival modelling with random effects or the so-called *frailty models* gives the possibility to introduce random effects in the survival model in order to take into account the association and the unobserved heterogeneity. The term *frailty* appeared for the first time in a study by Vaupel et al. [108] in which a univariate survival model was considered. Clayton [109] was the first to apply this concept to a multivariate situation.

This model in its simplest way is based on adding a random effect that has a multiplicative effect on the hazard function of an individual or a cluster of individuals. Several extensions of the classical survival regressions incorporating random effects exist in the literature, including mixed effects Cox model [110–112] and Poisson mixed effects models [113, 114].

2.4.1 Cox Model with Random Effects

This model is obtained by incorporating the random effect in the classical Cox model [102] in which the random effect has as a multiplicative effect on the hazard rates:

$$\lambda(t \mid Z, u) = \lambda_0(t) \exp(Z^{'}\gamma + R^{'}u) \tag{2.14}$$

where Z and R are the fixed and random effects, γ is the vector of fixed-effects coefficients and u is the vector of random effects. We can assume that the random effects are normally distributed with mean 0 and a variance-covariance matrix Σ_u.

Klein [110] proposed to estimate the frailty and covariates effects by an EM algorithm to extend the partial likelihood techniques. In the E-step, the expectation of the full likelihood with respect to the observable data is computed. Then in the M-step, a partial likelihood is constructed to estimate the covariate effects using a profile likelihood technique. In this approach a nonparametric estimate of the baseline hazard function is necessary at each iteration. Other estimation techniques can also be found in the literature[111, 112].

Part II

Highlighting the association between socioprofessional trajectories and mortality

Chapter 3

Socio-professional trajectories and mortality

3.1 Background on life-course models

The aim of this chapter is to highlight the association between life course professional trajectory and adult mortality. Previous studies based on life course models have considered two or three stages in professional life and used a simple classification for socioeconomic positions (low-medium-high). Here, we go further by considering the whole professional trajectories and all-cause and cause-specific mortality. For this purpose, we use life course models on a representative sample of the French salaried population in the semi-public and private sectors from 1976 to 2002 to investigate the possible ways in which professional trajectories may be associated with adult mortality.

It has already been shown that an observed individual's social level at a given time, partially, reflects his/her social position at different stages of his/her past life [26]. However, to better describe the association, life course epidemiology models have been introduced in the literature, which was defined as "the study of long term effects on later health or disease risk of physical or social exposures during gestation, childhood, adolescence, young adulthood and later adult life" [30, 115].

This approach admits that both early and later life exposures and conditions are acting as risk or protective factors throughout individual's life [30]. The objective is to examine how the social level during childhood, adolescence and early adult life

influence the disease risk in adulthood and socioeconomic position that causes social inequalities in adult's health and mortality. We can mention studies showing that being in a low socio-economic level through life influences cause-specific mortality and cardiovascular diseases [31, 32]. Popular used hypothesis in the life course field are: *critical periods*, *accumulation*, and *social mobility* models.

A *critical period* is a time window in which an exposure can have long-lasting adverse or protective effects on development and subsequent disease outcome [30]. The attention of this model, which is also sometimes known as *latent* model, is more on the timing of an exposure and it assumes that an exposure can have irreversible damages for later health [21]. This concept has been extended to social developments, so that in this model some stages or specific moments in life are considered as key periods affecting health.

The *accumulation* model hypothesizes that mortality differentials are explained by the accumulation of all present and past working conditions, lifestyles and behaviours. Analyses using this model are based on the life-cumulative length of stay in the most disadvantaged social group. They suggest that the accumulation of poor socioeconomic exposure in life increases the risk of mortality [21, 27, 30, 33].

The *social mobility* model was developed to take into account the modality of transitions between social groups which can be divided into *intra-generational* and *inter-generational* mobilities. The *inter-generational* mobility addresses the changes in social group between generations, such as the changes between parental social class and own social class in adulthood. The *intra-generational* mobility is the changes between occupied social classes by an individual in adulthood. Different opinions regarding the impact of social mobility on health and mortality can be found in the literature. Some authors state that mobile individuals are placed in health levels between those of their current class and their original class, closest to the current class [34, 35].

Other models have also been proposed in the literature, such as *pathway* model, which assumes that the influence of childhood social level is attenuated after adjusting for other later conditions. Complete overviews of the life course models can be found in Galobardes et al. [25], Kuh et al. [30], Mishra et al. [116], and Niedzwiedz et al. [117].

The life course models help to explain the potential impact of socioeconomic status on health. However, a bias might be involved in the results obtained by this framework due to the impact of health on socioeconomic position, or health related selection. Selection out of the labour marker into an unemployment position due to health problems is an example of this kind of selection, also known as reverse causation. This reverse causation, between health and social position, is another issue that should be taken into account [36, 37].

3.2 Professional trajectory

A professional trajectory may be defined as the sequence of consecutive professional positions occupied by an individual. Figure 3.1[1] shows an example of 5 fictional trajectories. For instance, the second individual was working in the manual workers class from 1978 until 1985. No information on his professional category was available between 1986 and 2001. Finally he worked in an intermediary occupation in 2002. The fourth individual is an example of those individuals that do not stay in a single occupational category. The fourth horizontal bar is showing an individual who starts in a manual workers class and after experiencing some transitions between professional categories, he ends up in an upper class occupation. The third example shows an individual working in the manual workers class and clerk class. No further information is available for his occupational category from 1994 until his death in 2000.

As we can observe in Figure 3.1, a professional trajectory may be characterized by the occupied categories at each year, by the transitions between social classes and by the length of stay at each social class. To mimic the accumulation and social mobility hypothesis, we consider the following time-dependent covariates:

- *Occupational class at each year*;
- *Cumulative social class indicator*, defined as individuals length of stay in each occupational class. This indicator was calculated for all classes except the upper class, so the latter served as reference;

[1] Plotted with the R package, TraMineR [118]

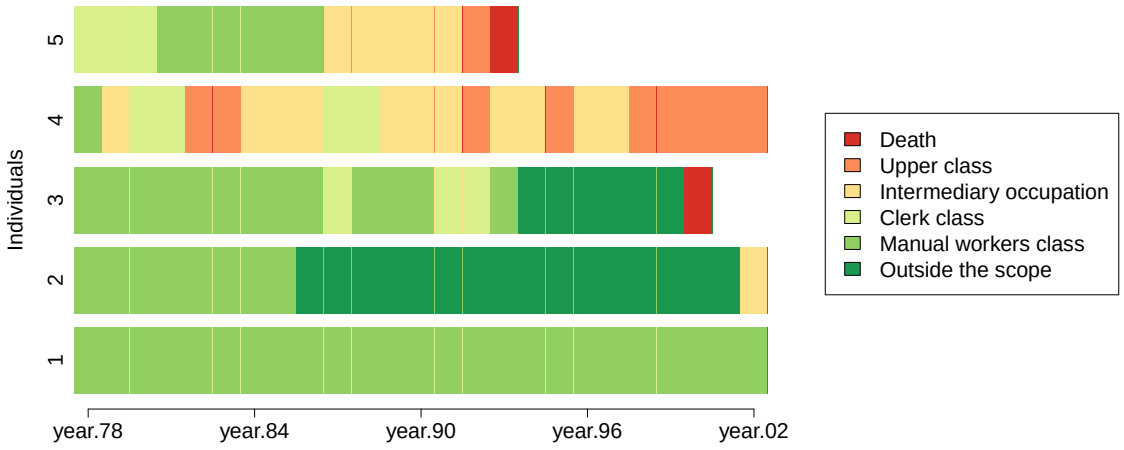


Figure 3.1 – Examples of fictional trajectories

- *10-year social mobility indicator*, defined by the transition rates between classes, excluding the *outside the scope* category and calculated as follows:

$$\frac{\text{number of transitions between occupational classes}}{\text{duration of follow-up}} \times 10$$

  This indicator was categorised into three groups using tertiles, separately for men and women.

To capture the critical period, third hypothesis, we consider the occupational class at the beginning of follow-up that is the most representative position of the end of childhood, given the fact that no information on childhood's socio-economic position was available in the motivating database.

3.3 Analysis of the Cosmop-DADS database

3.3.1 Study population

We consider the Cosmop-DADS database described in the Introduction. As explained previously, this database is obtained by linking the occupational life-course provided from the panel of DADS with the causes of death recorded by INSERM–CépiDc. In Section 1.4.4 of the Introduction, we explained the professional scope considered in this study, containing five categories: *Upper class*, *Intermediary occupations*, *Clerk class*, *Manual workers class* and the additional *Outside the scope*

class. The decision regarding this additional category will induce a bias but given the structure of the data, building a sound imputation model would require additional assumptions for which no auxiliary data, such as data on employees of the public sector, were available.

In this analysis, all individuals born in the French territories for whom a salaried period was declared in Cosmop-DADS between ages 25 and 30, excluding those working outside the study scope in their first year were considered. We excluded individuals born outside France due to the uncertainty of their vital status. In total 337 706 men and 275 378 women are included in the study. Less than 1% of occupations were imputed (corresponding to the so-called miscoded professions), and in total, 22% and 30% of follow-up years were outside the study scope for men and women, respectively. 52% of men and 61% of women were outside the study scope for at least one year of their follow-up.

Owing to the non-negligible number of episodes outside the study scope and the lack of available information for making more hypotheses about these episodes, a replicated analysis was carried out on a sub-sample of the analysed population for whom the first five years of their follow-up was covered by the study scope in order to ensure that an observed trajectory was complete (in the first five years) for the analysis (198 381 males and 134 784 females, with fewer than 14% of follow-up years outside the study scope in total).

3.3.2 Mortality

The Cosmop-DADS database is a sample of the French population for whom the vital status and date of death are available. All individuals of this sample were followed up to 2002 and the administrative censoring date was set at 31st December 2002. The underlying causes of death, recorded by INSERM-CépiDc, were coded according to the International Classification of Diseases, 8th, 9th and 10th revisions (ICD-8, ICD-9 and ICD-10), presented in Section 2.2 of the Introduction. We considered three broad categories of causes in this part: cardiovascular diseases, cancer and external causes (Table 2).

3.3.3 Statistical analysis

The Cox proportional hazards model, presented in Section 2.2.1, were used to estimate all-cause hazard ratios (HRs), cause-specific hazard ratios (CSHRs) and their 95% Confidence Intervals (CIs) while accounting for left truncation induced by the delayed entries. Age was used as the time-scale [119]. In the presence of competing risks, for each cause of death, we can fit the classical Cox model by censoring the participants who failed from competing causes of death [95, 100, 104].

The 3 indicators, *occupational class at each year*, *cumulative social class indicator* and *10-year social mobility indicator*, defined in Section 3.2 were calculated for the considered sample. To limit the impact of reverse causation, which is the possible influence of health on social position [36, 37], occupational classes were considered with a two-year time lag, i.e. instead of using the current occupational class, that of two years before, was taken into account. Adjustment for the covariates, occupational class at the beginning of the follow-up as a baseline covariate and the three indicators of professional trajectory as time-dependent covariates, was done by performing univariable analysis in the first step. After calculating the sample correlation between these covariates and finding no strong correlation between them, all these covariates were used in a multivariable analysis. Considering the decrease in mortality rates over time in France, all-cause and CSH models were adjusted for observation periods.

The *occupational class at each year* and the *10-year social mobility indicator* were introduced into the models as categorical variables, and the upper class and those without any mobility between classes were considered as the reference categories. For the cumulative social class indicator, HRs were interpreted as the hazard corresponding to an increase in the time spent in an occupational class versus that in the upper class. These HRs were calculated for a 10-year increase. No violation of the proportional hazards assumptions was found according to Schoenfeld residuals.

Proportional hazards models were conducted separately for men and women using the Survival package of the R software [120] and the imputation was carried out by the IVEware software [121].

3.3.4 Results

The average number of transitions between occupational classes differed between the age categories. Transitions were more numerous between the ages of 25 and 44 in women and between the ages of 25 and 34 in men. At the beginning of the follow-up, the largest class was the clerk class (about 54%) in women and manual workers (about 60%) in men. For young men (25-34 years), 49.3% of the cumulated time spent was in the manual workers class and much less in the upper class (6.5%). The same magnitude was observed in young women for the clerk and the upper class (25-34 years) (Table 3.1).

Table 3.1 – Characteristics of study population according to occupational trajectories

		Average number of transitions/10 years follow-up	Proportion of time spent in occupational classes					
			Upper class	Intermediary occupations	Clerk class	Manual workers class	Outside the scope	Total
Men	**At the beginning**	0	5.5	17.3	17.7	59.5	0	100
	25-34	1.0	6.5	17.0	12.4	49.3	14.8	100
	35-44	0.9	9.6	17.8	7.4	38.4	26.8	100
	45-54	0.6	12.9	17.9	5.8	31.8	31.6	100
	55-56	0.6	15.5	18.5	5.2	28.4	32.4	100
	All ages	0.9	8.8	17.4	9.4	42.1	22.3	100
Women	**At the beginning**	0	4.2	19.4	53.5	22.9	0	100
	25-34	0.8	4.3	17.0	41.0	16.0	21.7	100
	35-44	0.8	4.5	15.8	30.9	12.3	36.5	100
	45-54	0.6	5.5	16.6	28.1	11.4	38.4	100
	55-56	0.6	7.0	17.8	25.4	9.7	40.1	100
	All ages	0.8	4.6	16.5	35.0	13.8	30.1	100

During the follow-up, 12 162 (3.6%) men and 3551 (1.3%) women died. Most deaths occurred between the ages of 35 and 44. 48.7% of deaths among women and 39.8% of deaths among men occurred while individuals were outside the study scope two years before death. Most other deaths in men and women occurred while they were in the manual workers class and the clerk class, respectively (Table 3.2).

Table 3.2 – Distribution of study population according to occupational trajectories

		Number of death (%)		**Person-year (%)**	
		Men	**Women**	**Men**	**Women**
Observation period	**Beginning of follow-up**				
	Upper class	344 (2.8)	104 (3.0)		
	Intermediary occupations	1371 (11.3)	568 (16.0)		
	Clerk class	2042 (16.8)	1699 (47.8)		
	Manual workers class	8405 (69.1)	1180 (33.2)		
	Outside the scope	0 (0)	0 (0)		
	End of follow-up				
	Upper class	525 (4.3)	118 (3.3)		
	Intermediary occupations	1299 (10.7)	417 (11.7)		
	Clerk class	941 (7.7)	868 (24.5)		
	Manual workers class	4558 (37.5)	419 (11.8)		
	Outside the scope	4839 (39.8)	1729 (48.7)		
	1976-1980	306 (2.5)	66 (1.9)	4.52	3.87
	1981-1985	970 (8.0)	268 (7.5)	12.41	11.65
	1986-1990	1739 (14.3)	464 (13.1)	17.49	17.05
	1991-1995	2999 (24.6)	856 (24.1)	23.38	23.50
	1996-2002	6148 (50.6)	1897 (53.4)	42.20	43.93
Age category	**25-34**	2930 (24.1)	831 (23.4)	44.18	45.53
	35-44	4637 (38.1)	1396 (39.3)	38.77	38.49
	45-54	4329 (35.6)	1251 (35.2)	16.40	15.48
	55-56	266 (2.2)	73 (2.1)	0.65	0.50
Total		12 162	3551	337 706	275 378

The results of the univariable and multivariable analysis are subsequently presented in Table 3.3 – Table 3.6. Overall, the same magnitude was found for the results of the univariable and multivariable analysis, except for the estimated hazard ratios for the social mobility indicator, although, adjusting for all indicators led to some attenuation in the increased risk of death in association to professional trajectory indicators.

Table 3.3 – All-cause and cause-specific mortality hazard ratios among men according to socio-professional trajectories (univariable analysis)

	All-cause (n=12 162)	**Cardiovascular (n=1452)**	**Cancer (n=3116)**	**External causes (n=4026)**	**Other causes (n=3568)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.23 [1.09, 1.38]***	1.56 [1.06, 2.30]*	1.10 [0.88, 1.37]	1.20 [0.97, 1.47]	1.29 [1.04, 1.62]*
Clerk class	1.79 [1.60, 2.01]***	2.21 [1.52, 3.22]***	1.43 [1.16, 1.77]***	1.62 [1.33, 1.98]***	2.23[1.81, 2.76]***
Manual workers class	2.05 [1.84, 2.28]***	2.84 [1.98, 4.05]***	1.85 [1.52, 2.26]***	2.03 [1.69, 2.45]***	2.01 [1.64, 2.46]***
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.44 [1.30, 1.60]***	1.76 [1.31, 2.35]***	1.49 [1.24, 1.80]***	1.44 [1.20, 1.72]***	1.26 [1.04, 1.54]*
Clerk class	2.33 [2.09, 2.60]***	2.73 [1.99, 3.74]***	2.41 [1.96, 2.97]***	1.88 [1.56, 2.28]***	2.73 [2.24, 3.33]***
Manual workers class	2.34 [2.14, 2.56]***	2.74 [2.10, 3.58]***	2.52 [2.13, 2.98]***	2.46 [2.10, 2.90]***	1.83 [1.54, 2.18]***
Outside the scope	3.67 [3.35, 4.01]***	3.68 [2.83, 4.79]***	3.20 [2.71, 3.78]***	3.05 [2.59, 3.60]***	4.80 [4.06, 5.68]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.94 [0.86, 1.03]	1.00 [0.81, 1.25]	1.04 [0.90, 1.21]	0.91 [0.75, 1.10]	0.80 [0.67, 0.95]
Clerk class	1.68 [1.54, 1.83]***	1.66 [1.33, 2.06]**	1.57 [1.36, 1.83]***	1.35 [1.12, 1.63]	2.09 [1.80, 2.42]***
Manual workers class	1.66 [1.56, 1.76]***	1.69 [1.45, 1.96]***	1.74 [1.57, 1.93]***	1.68 [1.48, 1.91]***	1.51[1.34, 1.70]***
Outside the scope	2.09 [1.95, 2.23]***	1.84 [1.55, 2.18]***	1.79 [1.60, 2.00]***	2.10 [1.82, 2.42]***	2.61 [2.30, 2.96]***
Social mobility indicator[b]					
Low (= 0)	1	1	1	1	1
Medium	0.84 [0.79, 0.88]***	0.83 [0.72, 0.95]***	0.77 [0.70, 0.84]***	0.87 [0.78, 0.97]*	0.89 [0.80, 0.98]*
High (> 1.11)	0.83 [0.79, 0.86]***	0.80 [0.71, 0.90]***	0.75 [0.69, 0.81]***	0.85 [0.79, 0.92]***	0.88 [0.81, 0.95]***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted separately for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator

†: age as the time-scale in Cox proportional hazards model

Table 3.4 – All-cause and cause-specific mortality hazard ratios among women according to socio-professional trajectories (univariable analysis)

	All-cause (n=3551)	**Cardiovascular (n=304)**	**Cancer (n=1388)**	**External causes (n=894)**	**Other causes (n=965)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	0.95 [0.77, 1.17]	1.20 [0.54, 2.63]	0.98 [0.70, 1.37]	0.93 [0.63, 1.38]	0.87 [0.58, 1.31]
Clerk class	1.06 [0.87, 1.30]	1.23 [0.58, 2.64]	1.05 [0.76, 1.45]	0.97 [0.67, 1.40]	1.15 [0.79, 1.69]
Manual workers class	1.26 [1.03, 1.54]*	1.73 [0.81, 3.69]	1.28 [0.92, 1.77]	1.18 [0.81, 1.71]	1.21 [0.82, 1.79]
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.05 [0.85, 1.29]	2.92 [1.04, 8.18]	0.84 [0.63, 1.11]	1.27 [0.83, 1.97]	1.09 [0.68, 1.75]
Clerk class	1.12 [0.92, 1.35]	2.66 [0.97, 7.34]	0.86 [0.66, 1.12]	1.40 [0.93, 2.11]	1.29 [0.83, 2.01]
Manual workers class	1.35 [1.10, 1.66]**	4.50 [1.61,12.56]**	0.91 [0.69, 1.22]	1.77 [1.15, 2.72]**	1.60 [1.01, 2.55]*
Outside the scope	2.11 [1.75, 2.54]***	5.12 [1.90,13.75]**	1.34 [1.04, 1.72]*	2.22 [1.48, 3.32]***	3.82 [2.50, 5.85]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.83 [0.70, 0.97]*	1.38 [0.74, 2.60]	0.88 [0.70, 1.11]	0.79 [0.56, 1.10]	0.59 [0.40, 0.86]**
Clerk class	0.95 [0.83, 1.09]	1.71 [0.98, 2.99]	0.93 [0.76, 1.13]	0.80 [0.60, 1.08]	0.95 [0.73, 1.27]
Manual workers class	1.11 [0.96, 1.28]	2.15 [1.24, 3.73]**	1.03 [0.84, 1.26]	1.14 [0.85, 1.53]	1.02 [0.76, 1.36]
Outside the scope	1.47 [1.29, 1.68]***	2.88 [1.67, 4.96]***	1.17 [0.96, 1.41]	1.31 [0.99, 1.73]	1.95 [1.50, 2.54]***
Social mobility indicator[b]					
Low (= 0)	1	1	1	1	1
Medium	0.93 [0.84, 1.03]	0.99 [0.72, 1.38]	0.83 [0.71, 0.97]*	1.15 [0.93, 1.44]	0.93 [0.76, 1.14]*
High (> 0.91)	0.93 [0.86, 1.00]	0.84 [0.65, 1.10]	0.95 [0.85, 1.07]	0.90 [0.77, 1.05]	0.94 [0.81, 1.09]

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted separately for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator

†: age as the time-scale in Cox proportional hazards model

Table 3.5 – All-cause and cause-specific mortality hazard ratios among men according to socio-professional trajectories (multivariable analysis)

	All-cause (n=12 162)	**Cardiovascular (n=1452)**	**Cancer (n=3116)**	**External causes (n=4026)**	**Other causes (n=3568)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.17 [1.04, 1.33]*	1.41 [0.94, 2.14]	0.98 [0.77, 1.24]	1.10 [0.89, 1.37]	1.39 [1.10, 1.77]**
Clerk class	1.34 [1.18, 1.51]***	1.57 [1.04, 2.37]*	1.02 [0.81, 1.29]	1.26 [1.02, 1.56]*	1.68 [1.34, 2.12]***
Manual workers class	1.43 [1.27, 1.61]***	1.90 [1.27, 2.83]**	1.10 [0.88, 1.37]	1.41 [1.15, 1.73]**	1.60 [1.28, 2.00]***
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.16 [1.03, 1.30]*	1.26 [0.90, 1.76]	1.10 [0.88, 1.37]	1.23 [1.01, 1.51]*	1.07 [0.86, 1.33]
Clerk class	1.49 [1.31, 1.69]***	1.58 [1.09, 2.30]*	1.50 [1.16, 1.93]**	1.43 [1.14, 1.79]**	1.58 [1.26, 1.98]***
Manual workers class	1.39 [1.25, 1.56]***	1.43 [1.03, 1.99]*	1.26 [1.02, 1.56]*	1.73 [1.42, 2.12]***	1.09 [0.89, 1.33]
Outside the scope	2.57 [2.31, 2.85]***	2.45 [1.80, 2.34]***	2.21 [1.81, 2.71]***	2.20 [1.81, 2.68]***	3.25 [2.69, 3.94]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	1.04 [0.92, 1.17]	1.13 [0.83, 1.54]	1.20 [0.98, 1.46]	1.03 [0.81, 1.31]	0.84 [0.66, 1.06]
Clerk class	1.50 [1.33, 1.69]***	1.59 [1.14, 2.20]**	1.53 [1.23, 1.89]***	1.23 [0.95, 1.60]	1.62 [1.31, 2.00]***
Manual workers class	1.52 [1.38, 1.66]***	1.54 [1.18, 2.00]**	1.75 [1.48, 2.06]***	1.33 [1.10, 1.60]**	1.53 [1.28, 1.83]***
Outside the scope	1.35 [1.22, 1.48]***	1.29 [0.99, 1.69]	1.33 [1.12, 1.57]***	1.46 [1.19, 1.77]***	1.39 [1.16, 1.67]***
Social mobility indicator[b]					
Low (= 0)	1	1	1	1	1
Medium	1.03 [0.97, 1.08]	1.03 [0.88, 1.20]	0.96 [0.87, 1.06]	1.11 [0.99, 1.24]	1.03 [0.93, 1.13]
High (> 1.11)	1.15 [1.09, 1.21]***	1.12 [0.97, 1.29]	1.07 [0.97, 1.18]	1.17 [1.08, 1.28]***	1.23 [1.12, 1.34]***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Table 3.6 – All-cause and cause-specific mortality hazard ratios among women according to socio-professional trajectories (multivariable analysis)

	All-cause (n=3551)	**Cardiovascular (n=304)**	**Cancer (n=1388)**	**External causes (n=894)**	**Other causes (n=965)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	0.99 [0.79, 1.23]	0.93 [0.39, 2.24]	1.02 [0.71, 1.47]	0.90 [0.60, 1.35]	1.03 [0.69, 1.55]
Clerk class	1.05 [0.85, 1.29]	0.81 [0.35, 1.86]	1.07 [0.76, 1.52]	0.92 [0.62, 1.36]	1.15 [0.78, 1.70]
Manual workers class	1.15 [0.93, 1.43]	1.04 [0.45, 2.43]	1.35 [0.94, 1.94]	0.91 [0.60, 1.37]	1.09 [0.73, 1.63]
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.04 [0.82, 1.32]	2.18 [0.76, 6.23]	0.77 [0.55, 1.09]	1.40 [0.86, 2.27]	1.27 [0.74, 2.19]
Clerk class	1.00 [0.80, 1.26]	1.49 [0.52, 4.26]	0.75 [0.54, 1.04]	1.58 [0.98, 2.53]	1.12 [0.66, 1.89]
Manual workers class	1.13 [0.88, 1.45]	2.63 [0.88, 7.85]	0.71 [0.49, 1.04]	1.65 [0.99, 2.74]	1.47 [0.84, 2.58]
Outside the scope	1.81 [1.45, 2.27]***	2.48 [0.89, 6.87]	1.20 [0.86, 1.66]	2.18 [1.38, 3.47]***	3.16 [1.90, 5.26]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.98 [0.78, 1.23]	1.61 [0.63, 4.10]	1.12 [0.82, 1.53]	0.86 [0.55, 1.36]	0.66 [0.39, 1.12]
Clerk class	1.12 [0.92, 1.36]	2.65 [1.14, 6.13]*	1.16 [0.88, 1.54]	0.79 [0.52, 1.19]	1.14 [0.77, 1.69]
Manual workers class	1.12 [0.91, 1.38]	2.05 [0.86, 4.89]	1.10 [0.81, 1.49]	1.08 [0.71, 1.64]	1.07 [0.70, 1.64]
Outside the scope	1.21 [1.01, 1.47]*	3.16 [1.41, 7.05]**	1.08 [0.82, 1.43]	1.01 [0.69, 1.49]	1.37 [0.93, 2.01]
Social mobility indicator[b]					
Low (= 0)	1	1	1	1	1
Medium	1.00 [0.90, 1.11]	1.05 [0.76, 1.47]	0.85 [0.73, 0.99]*	1.26 [1.01, 1.58]*	1.09 [0.89, 1.33]
High (> 0.91)	1.13 [1.04, 1.22]**	1.10 [0.84, 1.46]	1.04 [0.91, 1.18]	1.05 [0.88, 1.24]	1.40 [1.19, 1.64]***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Occupation at beginning of follow-up

As shown in Table 3.5, men in the manual workers class at the beginning had a higher mortality risk compared to those who were in the upper class (except for cancer mortality) but to a different degree depending on the causes of death (HRs:1.43 [1.27, 1.61], 1.90 [1.27, 2.83], 1.41 [1.15, 1.73] and 1.60 [1.28, 2.00] respectively for mortality from all causes, cardiovascular diseases, external causes and other causes). Also, being in the clerk class at the beginning of follow-up increased the mortality risk among men compared to being in the upper class at the beginning (HRs: 1.34 [1.18, 1.51], 1.57 [1.04, 2.37], 1.26 [1.02, 1.29] and 1.68 [1.34, 2.12] respectively for mortality from all causes, cardiovascular diseases, external causes and other causes). In women, this association was not statistically significant (Table 3.6).

Current occupational class

Among men, being in the clerk class increased the mortality risk compared to being in the upper class (HRs: 1.49 [1.31, 1.69], 1.58 [1.09, 2.30], 1.50 [1.16, 1.93], 1.43 [1.14, 1.79] and 1.58 [1.26, 1.98] respectively for mortality from all causes, cardiovascular diseases, cancer, external causes and other causes. Among men, those in the manual workers class had an increased mortality risk compared to those in the upper class (HRs: 1.39 [1.25, 1.56], 1.43 [1.03, 1.99], 1.26 [1.02, 1.56] and 1.73 [1.42, 2.12] respectively for all-cause, cardiovascular, cancer and external-cause mortality). Those outside the study scope had the highest mortality risk except for cardiovascular and cancer mortality among women, i.e. about two to three-fold higher than the mortality risk in the upper class (Table 3.5 and Table 3.6).

Cumulative time spent in occupational classes

The cumulative time spent in occupational classes was strongly associated with men's all-cause and cause-specific mortality (Table 3.5) and women's all-cause and cardiovascular mortality (Table 3.6), with less pronounced associations for men's external-cause mortality. Among men, more time spent in an occupational class increased the mortality risk compared to that in the upper class. This increase in manual workers was associated with a 1.8-fold higher cancer mortality risk (HR: 1.75 [1.48, 2.06]) and that outside the study scope was associated with a 1.5-fold

higher external-cause mortality risk (HR: 1.46 [1.19, 1.77]) compared to that in the upper class. Among women, more time spent in the clerk class was associated with a 2.7-fold higher cardiovascular mortality risk compared to that in the upper class (HR: 2.65 [1.14, 6.13]).

Social mobility indicator

In the univariable analysis (Table 3.3 and Table 3.6), an inverse association between the social mobility indicator and mortality was systematically found among men, and only for cancer mortality among women. Adjusting for other indicators changed the direction of the results, except for women's cancer mortality.

In multivariable analysis (Table 3.5 and Table 3.6), the same magnitude was observed for this indicator among men and women except for women's external-cause mortality, with significant results for men and women's all-cause, external-cause and other causes mortality, and women's cancer mortality. Having a high social mobility indicator increased the all-cause mortality risk (HRs: 1.15 [1.09, 1.21] and 1.13 [1.04, 1.22] respectively for men and women), the other causes mortality risk (HRs: 1.23 [1.12, 1.34] and 1.40 [1.19, 1.64] respectively for men and women) and the external-cause mortality risk (HR: 1.17 [1.08, 1.28] for men) compared to not experiencing any mobility during professional life (Table 3.5 and Table 3.6).

3.3.5 Ad-hoc sensitivity analysis

When replicated analyses were performed on the sub-sample, including individuals working in the study scope during their first five years of follow-up, the estimated all-cause and cause-specific hazard ratios did not change much for any of the indicators except for men's cardiovascular mortality (Table 3.7 and Table 3.8).

Table 3.7 – All-cause and cause-specific mortality hazard ratios according to socio-professional trajectories among men working in the scope of study on their first five years of follow-up

	All-cause (n=6884)	**Cardiovascular (n=949)**	**Cancer (n=2067)**	**External causes (n=1979)**	**Other causes (n=1889)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.25 [1.04, 1.50]*	2.04 [1.10, 3.79]*	0.88 [0.64, 1.21]	1.04 [0.75, 1.46]	1.84 [1.27, 2.66]**
Clerk class	1.38 [1.15, 1.66]***	2.18 [1.16, 4.08]*	0.96 [0.71, 1.32]	1.30 [0.93, 1.82]	1.86 [1.29, 2.69]***
Manual workers class	1.31 [1.10, 1.57]**	2.57 [1.39, 4.76]**	0.90 [0.66, 1.22]	1.22 [0.88, 1.69]	1.61 [1.13, 2.31]**
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.09 [0.94, 1.25]	1.09 [0.75, 1.59]	1.01 [0.77, 1.31]	1.20 [0.91, 1.56]	1.04 [0.79, 1.37]
Clerk class	1.40 [1.19, 1.66]***	1.29 [0.83, 2.01]	1.29 [0.94, 1.76]	1.33 [0.97, 1.82]	1.70 [1.26, 2.31]***
Manual workers class	1.28 [1.12, 1.48]***	1.09 [0.74, 1.58]	1.12 [0.87, 1.45]	1.78 [1.37, 2.32]***	1.08 [0.83, 1.41]
Outside the scope	2.50 [2.18, 2.86]***	2.05 [1.42, 2.96]***	2.01 [1.58, 2.56]***	2.22 [1.70, 2.89]***	3.54 [2.77, 4.54]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	1.07 [0.93, 1.23]	1.05 [0.73, 1.49]	1.26 [0.99, 1.60]	1.10 [0.82, 1.47]	0.87 [0.65, 1.15]
Clerk class	1.48 [1.27, 1.71]***	1.58 [1.08, 2.33]*	1.57 [1.21, 2.04]***	1.22 [0.88, 1.68]	1.55 [1.18, 2.03]**
Manual workers class	1.60 [1.43, 1.80]***	1.58 [1.08, 2.33]*	1.57 [1.21, 2.04]***	1.22 [0.88, 1.68]	1.65 [1.32, 2.06]***
Outside the scope	1.57 [1.38, 1.78]***	1.31 [0.92, 1.86]	1.67 [1.34, 2.08]***	1.73 [1.31, 2.29]***	1.61 [1.26, 2.05]***
Social mobility indicator[b]					
Low (= 0)	1	1	1	1	1
Medium	0.98 [0.92, 1.06]	0.97 [0.81, 1.17]	0.94 [0.83, 1.06]	1.06 [0.92, 1.22]	0.99 [0.87, 1.13]
High (> 1.18)	1.08 [1.02, 1.16]*	1.07 [0.89, 1.27]	1.01 [0.89, 1.14]	1.09 [0.97, 1.22]	1.18 [1.05, 1.33]**

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Table 3.8 – All-cause and cause-specific mortality hazard ratios according to socio-professional trajectories among women working in the scope of study on their first five years of follow-up

	All-cause (n=1544)	**Cardiovascular (n=136)**	**Cancer (n=723)**	**External causes (n=316)**	**Other causes (n=369)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.04 [0.72, 1.51]	2.10 [0.25,17.80]	1.12 [0.64, 1.95]	1.15 [0.52, 2.52]	0.78 [0.42, 1.47]
Clerk class	0.98 [0.68, 1.39]	2.02 [0.24,16.62]	1.09 [0.63, 1.87]	0.86 [0.40, 1.86]	0.80 [0.43, 1.49]
Manual workers class	0.99 [0.68, 1.44]	2.35 [0.28,19.74]	1.24 [0.71, 2.19]	0.90 [0.41, 2.01]	0.58 [0.30, 1.14]
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.16 [0.82, 1.65]	2.21 [0.56, 8.71]	0.97 [0.61, 1.56]	0.85 [0.41, 1.77]	2.60 [0.98, 6.72]
Clerk class	1.12 [0.80, 1.56]	1.65 [0.41, 6.67]	0.94 [0.60, 1.48]	1.24 [0.60, 2.55]	1.92 [0.77, 4.81]
Manual workers class	1.16 [0.80, 1.68]	4.06 [0.94,17.59]	0.84 [0.50, 1.40]	0.98 [0.44, 2.16]	2.10 [0.77, 5.69]
Outside the scope	2.08 [1.49, 2.91]***	3.03 [0.76,12.11]	1.45 [0.92, 2.29]	1.52 [0.73, 3.15]***	6.73 [2.73,16.60]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.94 [0.69, 1.28]	1.55 [0.45, 5.28]	0.91 [0.60, 1.39]	1.14 [0.58, 2.24]	0.69 [0.33, 1.45]
Clerk class	1.05 [0.80, 1.39]	2.38 [0.78, 7.29]*	0.95 [0.65, 1.39]	0.86 [0.45, 1.65]	1.15 [0.64, 2.09]
Manual workers class	1.12 [0.83, 1.50]	1.44 [0.43, 4.78]	0.98 [0.65, 1.48]	1.38 [0.72, 2.64]	1.26 [0.65, 2.43]
Outside the scope	1.21 [0.90, 1.61]*	2.56 [0.82, 8.01]**	1.10 [0.74, 1.64]	1.12 [0.55, 2.26]	1.27 [0.69, 2.32]
Social mobility indicator[b]					
Low (= 0)	1	1	1	1	1
Medium	0.94 [0.81, 1.08]	1.18 [0.73, 1.90]	0.84 [0.68, 1.03]	1.03 [0.74, 1.43]	1.01 [0.75, 1.36]
High (> 1)	1.03 [0.91, 1.15]	1.17 [0.79, 1.72]	0.97 [0.81, 1.15]	0.86 [0.66, 1.12]	1.27 [0.99, 1.61]***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

3.4 Discussion

Previous studies on this topic have generally considered individuals' socioeconomic position at two or three stages of life including childhood (father's socioeconomic position), entry into the labour market and mid-life position. To our knowledge, the present study is the first to investigate the association between the whole professional trajectory and all-cause mortality and within that, three major causes of death: cardiovascular disease, cancer and external causes. Overall, our results add to the existing evidence of the strong relationship between professional trajectory and all-cause mortality among men, with less pronounced associations among women [20–22, 32, 122–125].

Compared to previous studies, a new aspect of our study is the use of the duration of time spent in occupational classes as a measure of socioeconomic exposure and the transition rates between occupational classes as a measure for capturing the social mobility dimension. The three most commonly used life-course models, namely the critical period, cumulative and social mobility models were taken into account. Our results suggest that all three dimensions are associated to men's all-cause mortality. For women, only the cumulative and the social mobility models were confirmed by this analysis.

3.4.1 Interpretations and comparisons with other studies

As shown in previous studies, strong associations between professional trajectories and men's and women's mortality was found [20–22, 32, 122–125]. However, a direct comparison with other studies cannot be easily made given the different occupational classifications in each country, and the fact that we used whole professional trajectories.

The present study only focused on professional trajectories with no information on childhood circumstances. The individual's first occupation is likely to be the most representative dimension of the end of childhood. We found that the association between the first occupation and mortality was strong for men's cardiovascular and external-cause mortality. Previously, strong associations have also been reported between socioeconomic circumstances in childhood and mortality from some causes

of death, such as cardiovascular diseases [21, 122–124].

On the other hand, for some other causes of death such as external causes and lung cancer [124], stronger associations were found between socioeconomic circumstances in adulthood and adult mortality than those in childhood. Our results are in accordance with the literature, since in other studies, for some causes of death such as external causes and cancer, occupational classes were found to be strongly linked with men's mortality. Supplementary analysis on different cancers also reported the same associations or even stronger ones, for deaths by UADT cancers (See Appendix B, Table B.1 and Table B.2). For women, the results were not statistically significant.

Another hypothesis in the literature is the putative association between the accumulation of exposure to different socioeconomic conditions and mortality. However, the use of only three stages of life limited the number of possible trajectories, so the different trajectories could be compared. By investigating the duration of time spent in each occupational class instead of comparing different trajectories, we found a strong relationship between the duration of exposure to low professional position and mortality. This association was stronger for cardiovascular and cancer mortality in men but was significant only for all-cause and cardiovascular mortality in women. This is consistent with the results of previous studies [21, 22, 123, 126]. The large mortality risk of those who stay longer in the low occupational categories can be explained by exposure to poor working conditions and by the fact that the least skilled are less likely to move upward. Furthermore, staying a long time in the same professional conditions could reflect a greater adherence to a professional class and its specific lifestyle.

The changes between occupational categories and their dynamics were also pointed out in previous studies. Some studies have shown that within classes, male movers have a mortality risk situated between that of non-movers in their class of origin and that of their destination [23, 127]. Here, we investigated the association between the frequency of changes between occupational classes and mortality. Instability in professional life may be interpreted in two ways. If instability is chosen, it could be the reflection of high dynamism with the ability to change and adapt to several professional environments. Conversely, if instability is forced, it could be

due to difficulties in finding one's place, to a high dependence on the work market or to personal events. We found an inverse association for this indicator in the univariable analysis, as it does not take into account the occupational classes before and after the transitions. Our results of the multivariable analysis show that subjects with high transition rates have an increased risk of all-cause and external-cause mortality. These results suggest that the instability measured is more forced than chosen, with a deleterious association on mortality. In a very explorative approach to disentangle the chosen and forced instability, we considered the following naive order of occupations from high to low level: *upper class*, *intermediary occupations*, *clerk class*, and *manual workers*. Although this order is not strictly hierarchical, upward and downward changes were studied as separate variables. The risk of mortality was positively associated with downward changes, for instance, going from the *upper class* to the *clerk class*, and negatively with upward changes, for example, going from the *manual workers class* to the *intermediary occupations class* (Results shown in Appendix B, Table B.3 and Table B.4).

3.4.2 Limitations

The main limitation in this investigation is the high percentage of follow-up years outside the scope of the study. The decision to consider all these data in the *outside the scope* category could induce a bias. However, we examined a wide range of occupational sectors and the occupational stages are sufficiently reliable as they were collected within the context of administrative procedures. Furthermore, the replicated analysis on the subsample with sufficient follow-up provided almost the same results, which strengthens the findings.

All participants had worked at least once between the ages of 25 and 30 and were likely to be healthier than the general population, so the sample should not be interpreted as representative of the French population.

Finally, taking into account the individual's occupation with a two-year time lag could reduce the reverse causation bias. Moreover, for some causes of death such as transport accidents, the problem of reverse causation is less likely to be a source of bias.

Despite these drawbacks, the large size of the sample, the annual nature of the

information collected and the causes of death coded with high precision are the major strength of this study. Using repeated measures of occupational category over the follow-up could provide insight into changes that may have occurred during a person's professional life. However, using endogenous time-dependent covariates in a Cox proportional hazards model results in bias. Consequently, to reduce the bias and to better address the bidirectional association between professional trajectories and mortality, in the next part we will focus on models that take into account simultaneously professional trajectory and mortality, namely the *joint modelling of longitudinal data and cause-specific mortality*.

Part III

Joint modelling of professional trajectory and mortality

Chapter 4

A joint modelling of longitudinal nominal data and cause-specific hazards

4.1 Background on joint models

In Chapter 3, we highlighted the association between socio-professional trajectories and mortality using the administrative employment records as time-dependent covariates in a Cox proportional hazards model. However, the employment records that are collected only for the subjects under the study are endogenous (internal) time-dependent covariates. As mentioned in Section 2.3, the extended Cox model, which is the extension of the Cox model to handle time-dependent covariates, is based on the assumption that the covariates path is predictable and thus, is not appropriate for internal time-dependent covariates. Consequently, it is of interest to model jointly the longitudinal process and time-to-event process.

Joint analysis of longitudinal outcomes and survival data can be categorized into *pattern-mixture models*, *selection models* and *random effects models* (cf. Section 1.4.3). Although mathematically all these models describe the joint distribution of longitudinal outcomes and survival data, they have different statistical interpretations. We focus on the *random effects models*, also known as *shared-parameter models* and refer to this class of models as *joint models for longitudinal and time-to-event data*.

A motivating example for the field of joint models was the study of the relationship between the CD4 cell counts and the time to AIDS diagnosis or death in HIV clinical trials. It aimed also to determine whether the CD4 cell counts could be considered as a useful surrogate marker in the treatment evaluation [39, 40]. In these kind of studies a simple linear mixed model was used to describe the log of CD4 cell counts trajectories. The fundamental idea of the so-called joint models is based on linking the survival model with a suitable model for the longitudinal measurements, usually a random effect model [41], in which the correlation between the repeated measures is not ignored, via a common unobserved structure, to capture the correlation between the two longitudinal and survival processes.

Different approaches have been developed in the literature for the association structure in joint models. One may include the mixed model defined for the longitudinal outcomes as a covariate in the survival sub-model [42, 43]. An alternative approach would be including directly the random effects in both longitudinal and survival sub-models with an assumed joint distribution for the random effects [44–46]. These are the most used approaches in the literature, however, a different approach has also been proposed, namely the *joint latent class model*. The idea is based on dividing the population, which is assumed to be heterogeneous, into a finite number of homogeneous classes where each class is characterized by a specific trajectory of the longitudinal outcome and a specific event risk [47, 48].

Considerable attention has been paid to the joint modelling of longitudinal outcome and survival data in recent years and since its appearance, several extensions have been proposed in the literature to adapt them to a wider variety of outcomes and situations. These extensions, with a preference for the association structure via the random effects, include joint modelling for multiple longitudinal outcomes [49, 128–130], for multiple events in a cause-specific context for competing risks [45, 131], for multiple correlated events [132], for survival data in the presence of recurrent events [47, 133], or in the presence of cure fraction [134–137] and in the presence of censored and missing time-varying covariates [138], using either likelihood-based approaches or Bayesian ones. Some good overviews of this class of models can be found in [43, 139–143].

Despite all these progress in the joint modelling of longitudinal and time-to-

event, most previous works have focused on continuous measurements [39, 49] or the quality of life measurements [50], on binary measurements [51] or on ordinal responses [46, 52] and there has been less attention to non-ordinal categorical longitudinal outcomes. Recently, Murawska and Rizopoulos [53] developed an extension of the joint modelling of categorical longitudinal data and time-to-event data using a Bayesian approach.

Given the structure of our motivation dataset, the Cosmop-DADS database, in this chapter, we extend the work of Li et al. [46], by proposing a joint model for nominal longitudinal data and competing risk data in a likelihood-based framework. The association structure was modelled by introducing the random effects in each sub-model.

However, even in the case of a reasonable sample size and moderate individual measurements, the joint modelling of longitudinal outcomes and survival is computationally intensive [53, 57] and it becomes out of reach in the case of large datasets. So far, the existing joint models have been applied to sample size up to 2000 individuals. To address this issue we propose an approach mimicking a meta analysis.

4.2 Joint modelling framework

The proposed joint model comprises three components: the nominal longitudinal sub-model, the cause-specific sub-models and the variance-covariance matrix of random effects to describe the joint association of repeated values and competing risks data.

4.2.1 Nominal longitudinal sub-model

Let n be the number of subjects in the study and let Y_{ij} denotes the j-th observed value for subject i, $j = 1, \cdots, m_i$ with $Y_{ij} = k \in \{1, \cdots, K\}$. We postulate a baseline-category logit model for Y_{ij} with random effects incorporated into the model. Recall that the probability that the modality k is observed for the j-th value of individual i, conditional on the random effects b_i, is given by:

$$\pi_{ijk} = P\Big(Y_{ij} = k|X_{ij}, W_{ij}, b_{ik}\Big) = \begin{cases} \frac{1}{1+\sum_{h=1}^{K-1}\exp\Big(\alpha_h + X_{ij}^{'}\beta_h + W_{ij}^{'}b_{ih}\Big)} & \text{if } k = K \\ \frac{\exp\Big(\alpha_k + X_{ij}^{'}\beta_k + W_{ij}^{'}b_{ik}\Big)}{1+\sum_{h=1}^{K-1}\exp\Big(\alpha_h + X_{ij}^{'}\beta_h + W_{ij}^{'}b_{ih}\Big)} & \text{if } k = 1, \cdots, K-1 \end{cases} \tag{4.1}$$

where $\eta_{ijk} = \alpha_k + X_{ij}^{'}\beta_k + W_{ij}^{'}b_{ik}$, X_{ij} vector of predictors for the fixed effects and W_{ij} vector of predictors for the random effects.

Similarly to Section 1.2.2, $\alpha = (\alpha_1, \cdots, \alpha_{K-1})^{'}$, the vector of intercepts with $\alpha_K = 0$. $\beta_k = (\beta_{k1}, \cdots, \beta_{kp})^{'}$ a $p \times 1$ fixed effects parameters vector with $\beta_K = 0$, $\beta = (\beta_1^{'}, \cdots, \beta'_{K-1})^{'}$, $b_{ik} = (b_{ik1}, \cdots, b_{ikq})^{'}$ a $q \times 1$ vector of the random effects for subject i in the k-th modality. We also assume that the random effects b_{ik} follow a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix Σ_{b_k}, $b_{ik} \sim \mathcal{N}_q(0, \Sigma_{b_k})$. Then, $b_i = (b_{i1}^{'}, \cdots, b_{i,K-1}^{'})^{'}$ the $(K-1)q \times 1$ vector of random effects for subject i follows a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix Σ_b, $b_i \sim \mathcal{N}_{(K-1)q}(0, \Sigma_b)$, defined as:

$$\Sigma_b = \begin{pmatrix} \Sigma_{b_1} & \Sigma_{b_1 b_2} & \cdots & \Sigma_{b_1 b_{K-1}} \\ \Sigma_{b_2 b_1} & \Sigma_{b_2} & \cdots & \Sigma_{b_2 b_{K-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{b_{K-1} b_1} & \Sigma_{b_{K-1} b_2} & \cdots & \Sigma_{b_{K-1}} \end{pmatrix} \tag{4.2}$$

4.2.2 Cause-specific hazards sub-model

The survival sub-model is defined as a proportional hazards model for each CSH [95] incorporated a subject-specific random effect.

Let Z_i be a $l \times 1$ vector of covariates, $T_i = min(T_i^*, C_i)$ the right-censored event time with C_i the censoring time of subject i and T_i^* the survival time of subject i, and $\epsilon_i = \delta_i \times D_i$ with $\delta_i = 1\{T_i^* \leq C_i\}$ the indicator of censorship and $D_i \in \{1, \cdots, g\}$ indicating the failure type of subject i. The sub-model for event d, $d = 1, \cdots, g$, is specified as:

$$\begin{aligned} \lambda_d(t|Z_i, u_i) &= \lim_{h \to 0} h^{-1} P\Big(t \leq T_i < t + h, \epsilon_i = d | T_i \geq t, Z_i, u_i\Big) \\ &= \lambda_{0d}(t) \exp\Big(Z_i^{'}\gamma_d + \nu_d u_i\Big) \end{aligned} \tag{4.3}$$

where $\lambda_d(t|Z_i, u_i)$ is the instantaneous risk of failure from cause d at time t given the vector of covariates Z_i and the frailty u_i, $\lambda_{0d}(t)$ is an unspecified baseline hazard function for event d and $\gamma = (\gamma_1^{'}, \cdots, \gamma_g^{'})^{'}$ is a vector of fixed regression coefficients. In this formulation, the heterogeneities, that are not observed through the covariates Z_i, are accounted for by the random effect u_i and the parameter ν_d. The ν_d represents the effect of the random effect u_i, and $\nu = (\nu_1, \cdots, \nu_g)^{'}$ is therefore, the vector of coefficients of the random effects u_i with ν_1 set to 1 to ensure identifiability [131].

Let $Y_i = (Y_{i1}, \cdots, Y_{im_i})^{'}$, $Y = (Y_1^{'}, \cdots, Y_n^{'})^{'}$, $\widetilde{T_i} = (T_i, \epsilon_i)^{'}$ and $\widetilde{T} = (T_1, \epsilon_1, \cdots, T_n, \epsilon_n)^{'}$. The association between longitudinal data Y and competing risks data $\widetilde{T}$ is modelled by assuming that the joint distribution of the random effects of the two sub-models, b_i and u_i follows a multivariate Gaussian distribution:

$$a_i = \begin{pmatrix} b_i \\ u_i \end{pmatrix} \sim \mathcal{N}_{(K-1)q+1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} , \begin{pmatrix} \Sigma_b & \Sigma_{bu}^{'} \\ \Sigma_{bu} & \sigma_u^2 \end{pmatrix} \right) \tag{4.4}$$

4.3 Likelihood function, Estimation and Inference

Initial works in the joint models class have focused on a two-stage approach for the estimation of the model parameters [40]. In this approach, in the first step, missing values for all subjects at each time point are imputed using the assumed model for the longitudinal data and all the information available at that time point. The second step consists of fitting a Cox proportional hazards model treating the imputed values as the true values of longitudinal outcome. However, since in the imputation step, survival data is not employed, this approach may result in bias. Moreover, the simulation studies have also shown the loss of efficiency of this approach [42, 144].

As a result, to eliminate this bias, a second approach based on the joint likelihood from the two sub-models of longitudinal outcomes and survival data has been developed. In a likelihood-based framework, the maximum likelihood method was proposed for the parameter estimation [44, 145, 146]. An alternative method in the literature is the Bayesian approach using Markov Chain Monte Carlo (MCMC) techniques for estimating the parameters [42, 53, 128]. Both of these two approaches, Bayesian and maximum likelihood, are based on specification of an appropriate joint

likelihood of longitudinal outcomes and survival data.

A completely different approach was proposed by Tsiatis et al. [144] in which the random effects are treated as nuisance parameters. In this work, we follow the maximum likelihood estimation method for estimating the joint model of longitudinal nominal outcomes and competing risks data, where each individual can fail from one out of two or more possible event type and only the time to the first of these events can be observed.

4.3.1 Likelihood formulation

Let $\Psi = (\alpha, \beta, \gamma, \nu, \Sigma, \lambda_{01}(t), \cdots, \lambda_{0g}(t))$ be the vector of all parameters in (4.1) and (4.3), where Σ is the variance-covariance matrix of a_i defined in (5.4). We assume that the longitudinal outcomes are independent of the competing risks survival data conditional on covariates and random effects. The joint distribution of $(Y, \widetilde{T})$ is completely determined by $f(Y \mid a, \Psi)$, $f(\widetilde{T} \mid a, \Psi)$ and $f(a \mid \Psi)$ where $f(.)$ stands for the probability density function and $a = (b, u)^{'}$ represents the vector of random effects of the two sub-model. The observed-data likelihood function for Ψ, conditional on the observed data $(Y_i, \widetilde{T}_i)$ for $i = 1, \cdots, n$, is

$$\begin{aligned} L(\Psi|Y, \tilde{T}) &\propto \prod_{i=1}^{n} f(Y_i, \widetilde{T}_i|\Psi) \\ &= \prod_{i=1}^{n} \int_a f(Y_i|\widetilde{T}_i, a, \Psi) f(\widetilde{T}_i|a, \Psi) f(a|\Psi) da \end{aligned} \tag{4.5}$$

Since Y and $\widetilde{T}$ are independent given the covariates, the Equation (4.5) can be formulated as:

$$\begin{aligned} L(\Psi|Y, \tilde{T}) &\propto \prod_{i=1}^{n} \int_a f(Y_i|a, \Psi) f(\widetilde{T}_i|a, \Psi) f(a|\Psi) da \\ &= \prod_{i=1}^{n} \int_a \Big[\prod_{j=1}^{m_i} \prod_{h=1}^{K} \{\pi_{ijh}\}^{I(Y_{ij}=h)} \Big] \Big\{ \prod_{d=1}^{g} \lambda_d(T_i|Z_i, u, \gamma_d, \nu_d)^{I(\epsilon_i=d)} \Big\} \\ &\times \exp\Big[- \int_0^{T_i} \{ \sum_{d=1}^{g} \lambda_d(t|Z_i, u_i, \gamma_d, \nu_d) \} dt \Big] \\ &\times \frac{1}{\sqrt{(2\pi)^{(K-1)q+1}|\Sigma|}} \exp\Big(- \frac{1}{2} a^{'} \Sigma^{-1} a \Big) da \end{aligned} \tag{4.6}$$

4.3.2 Estimation

We used the Maximum likelihood estimation approach to estimate the parameters. Maximizing the observed-data likelihood, Equation (4.5), in the presence of integration over random effects a_i is difficult. As a result and for simplification, the complete-data likelihood conditional on the random effects will be considered.

$$\begin{aligned} L(\Psi|Y,\widetilde{T},a) \propto & \prod_{i=1}^{n}\Big[\prod_{j=1}^{m_i}\prod_{k=1}^{K}\{\pi_{ijk}\}^{I(Y_{ij}=k)}\Big] \\ & \times\Big\{\prod_{d=1}^{g}\lambda_d(T_i|Z_i,u)^{I(\epsilon_i=d)}\Big\} \\ & \times\exp\Big[-\int_0^{T_i}\{\sum_{d=1}^{g}\lambda_d(t|Z_i,u_i)\}dt\Big] \\ & \times\frac{1}{\sqrt{(2\pi)^{(K-1)q+1}|\Sigma|}}\exp\Big(-\frac{1}{2}a_i^{'}\Sigma^{-1}a_i\Big) \end{aligned} \tag{4.7}$$

The EM algorithm is used to obtain the maximum likelihood estimates of Ψ. This algorithm iterates between E-steps and M-steps. The E-step computes the expected logarithm of the complete-data likelihood conditional on the observed data and the current estimates of the parameters. This means that in each iteration, the conditional expectations of all functions of a_i that appears in the log-likelihood must be evaluated. We write the complete-data log-likelihood, $l(\Psi \mid Y,\widetilde{T},a)$ as:

$$\begin{aligned} l(\Psi|Y,\widetilde{T},a) &= \log L(\Psi|Y,\widetilde{T},a) \\ &= \sum_{i=1}^{n}\Big[\Big(\sum_{j=1}^{m_i}\sum_{k=1}^{K}I(Y_{ij}=k)log(\pi_{ijk})\Big) \\ &+\Big(\sum_{d=1}^{g}I(\epsilon_i=d)log(\lambda_d(T_i|Z_i,u))\Big) \\ &+\Big(-\int_0^{T_i}\{\sum_{d=1}^{g}\lambda_d(t|Z_i,u_i)\}dt\Big) \\ &+\Big(log(\frac{1}{\sqrt{(2\pi)^{(K-1)q+1}|\Sigma|}})\Big(-\frac{1}{2}a_i^{'}\Sigma^{-1}a_i)\Big)\Big] \end{aligned} \tag{4.8}$$

where

$$log(\pi_{ijk}) = \begin{cases} 0 - log\Big(1+\sum_{h=1}^{K-1}\exp(X_{ij}^{'}\beta_h + W_{ij}^{'}b_{ih})\Big) & \text{if } k=K \\ (X_{ij}^{'}\beta_k + W_{ij}^{'}b_{ik}) - log\Big(1+\sum_{h=1}^{K-1}\exp(X_{ij}^{'}\beta_h + W_{ij}^{'}b_{ih})\Big) & \text{if } k\neq K \end{cases} \tag{4.9}$$

In the $(m+1)$-th iteration of the E-step, we evaluate:

$$\begin{aligned} E_{a_i|Y_i,\widetilde{T}_i,\Psi^{(m)}}\Big[h(a_i)\Big] &= \int h(a_i) f(a_i|Y_i,\widetilde{T}_i,\Psi^{(m)})da_i \\ &= \frac{\int h(a_i) f(Y_i,\widetilde{T}_i,a_i|\Psi^{(m)})da_i}{f(Y_i,\widetilde{T}_i|\Psi^{(m)})} \\ &= \frac{\int h(a_i) f(Y_i|a_i,\Psi^{(m)}) f(\widetilde{T}_i|a_i,\Psi^{(m)}) f(a_i|\Psi^{(m)})da_i}{\int f(Y_i|a_i,\Psi^{(m)}) f(\widetilde{T}_i|a_i,\Psi^{(m)}) f(a_i|\Psi^{(m)})da_i} \end{aligned} \tag{4.10}$$

for $h(a)$ being a function of $l(\Psi|Y,\widetilde{T},a)$. Each integral (of dimension $K+1$) of Equation (4.10) can be approximated using Gauss-Hermite quadrature method. The Gauss-Hermite method becomes intractable if the number of modalities of longitudinal outcome exceeds 5. Thus, one could aggregate the modalities, if possible, to reduce the dimension of integrals or use Monte Carlo techniques. Given the motivating data in this study, we focus on the first approach.

The M-step estimates the new parameter by maximizing the expected log-likelihood mentioned in Equation (4.10):

$$\Psi^{(m+1)} = \arg\max_{\Psi} E_{a|Y,\widetilde{T},\Psi^{(m)}}\Big[l(\Psi|Y,\widetilde{T},a)\Big] \tag{4.11}$$

In this step, each cumulative baseline hazard function for cause d, $H_{0d}(t)$, is assumed to be a step function with jumps at observed event times due to cause d, $d=1,\cdots,g$:

$$\begin{aligned} H_{0d}^{(m+1)}(t_{dq}) &= \sum_{j=1}^{q} \lambda_{0d}^{(m+1)}(t_{dj}) \\ &= \sum_{j=1}^{q} \frac{n_{dj}}{\sum_{r\in R\left(t_{dj}\right)} \exp\left(Z_r'\gamma_d^{(m)}\right) E\Big(\exp\left(\nu_d^{(m)}u_r\right)\Big)} \end{aligned} \tag{4.12}$$

where q_d is the number of distinct failure times due to the dth cause, $t_{d1} \leq \cdots \leq t_{dq_d}$ for $d=1,\cdots,g$ and $R(t_{dj})$ is the risk set at time t_{dj} and n_{dj} is the number of failures due to cause d at time t_{dj}.

The variance-covariance matrix Σ is updated as following:

$$\begin{aligned}
\Sigma_{b_k}^{(m+1)} &= \tfrac{1}{n}\sum_{i=1}^{n} E\Big(b_{ik}b_{ik}'\Big),\\
\Sigma_{b_k b_j}^{(m+1)} &= \tfrac{1}{n}\sum_{i=1}^{n} E\Big(b_{ik}b_{ij}'\Big),\\
\sigma_u^{2(m+1)} &= \tfrac{1}{n}\sum_{i=1}^{n} E\Big(u_i^2\Big),\\
\Sigma_{bu}^{(m+1)} &= \tfrac{1}{n}\sum_{i=1}^{n} E\Big(b_i u_i\Big)
\end{aligned} \tag{4.13}$$

where E stands for $E_{a_i|Y_i,\widetilde{T}_i,\Psi^{(m)}}$.

Since no closed-form solution exists for score equations of α, β, γ and ν, a one-step Newton-Raphson method is required to update the parameter estimations at each iteration:

$$\Psi^{(m+1)} = \Psi^{(m)} - \frac{S_\Psi^{(m)}}{I_\Psi^{(m)}} \tag{4.14}$$

More details are given in Appendix C, Equations (C.1)–(C.10). The updated parameter estimation, $\Psi^{(m+1)}$ is then considered as the input of the E-step in the next iteration. These two steps are iterated until the convergence criteria is met.

4.3.3 Standard Error Estimation

The baseline hazard function λ_{0d}, being unspecified, the dimension of the maximum likelihood estimates increases as the sample size increases and thus, the method proposed by Louis [76] becomes time-consuming and computationally unattractive in the calculation and inversion of this matrix. As a result, estimation of the standard errors is based on a profile likelihood approach [43]. This approach aims to eliminate a parameter of the likelihood function by replacing it with its maximum likelihood estimator as a function of the remaining parameters.

We followed the approach applied in Elashoff et al. [45]. The parameter vector $\Psi = \Big(\alpha, \beta, \gamma, \nu, \Sigma, \lambda_{01}(t), \cdots, \lambda_{0g}(t)\Big)$ was splitted into two components, $\Omega = \Big(\alpha, \beta, \gamma, \nu, \Sigma\Big)$ and $\Lambda = \Big(\lambda_{01}(t), \cdots, \lambda_{0g}(t)\Big)$. The variance-covariance matrix of Ω is approximated by inverting the empirical Fisher information obtained from the profile likelihood where the baseline hazards function have been profiled out. The observed information matrix of Ω is approximated by

$$\sum_{i=1}^{n} l^{(i)}(\hat{\Omega}|Y,\widetilde{T})l^{(i)}(\hat{\Omega}|Y,\widetilde{T})'$$

given that $l^{(i)}(\hat{\Omega}|Y,\widetilde{T})$ is the observed score vector from the profile likelihood on the i-th subject evaluated at $\hat{\Omega}$.

4.3.4 Estimation of marginal membership probabilities in the longitudinal sub-sample

In a mixed-effects multinomial logistic model with quantitative predictors, it is of interest to plot the estimated marginal probabilities. As explained in Hedeker [68], the marginal probabilities, $\pi_k^m, k = 1, \cdots, K$ can be estimated in two steps. In the first step, the *subject-specific* probabilities are calculated by replacing specific values of covariates and estimated parameters, $\widehat{\Psi}$, in the Equation (4.1). These probabilities are functions of the subject-specific random effects b_i, $\widehat{\pi_k^{ss}(b)}$. Then the marginal probabilities can be obtained by integrating over the random effects distribution of these subject-specific probabilities:

$$\widehat{\pi_k^m} = \int_b \widehat{\pi_k^{ss}(b)} f(b) db, \qquad \text{for } k = 1, \cdots, K \tag{4.15}$$

where $f(b)$ denotes the probability density function of the random effects. Since we assume that the random effects are normally distributed, we can resolve the integration by the numerical quadrature techniques, such as Gauss-Hermite quadrature.

Confidence intervals of these marginal probabilities can be estimated employing delta method as follows. Let $\Psi^{(l)} = (\Psi_1, \cdots, \Psi_P)$ be the parameter vector of the longitudinal sub-model. For the ease of notation we use Ψ instead of $\Psi^{(l)}$. Let Ψ^0 be the true value of Ψ and $\widehat{\Psi}$ an estimation of Ψ obtained by the MLE. We have the asymptotic property of the maximum likelihood estimators [147] as $n \longrightarrow \infty$:

$$I(\Psi^0)^{1/2}(\widehat{\Psi} - \Psi^0) \xrightarrow{D} \mathcal{N}_P(0, I_P) \tag{4.16}$$

where $I(\Psi)$ is the Fisher information matrix, I_P is the identity matrix of rank P and $\xrightarrow{D}$ denotes convergence in distribution. Thus for large n, $\widehat{\Psi} \overset{D}{\sim} \Psi^0 + I(\Psi^0)^{-1}\mathcal{N}_P(0, I_p)$, where $\overset{D}{\sim}$ is the shorthand for *is approximately distributed as*. The marginal probability π_k^m, $k = 1, \cdots, K$ is approximately distributed as:

$$\begin{aligned}\pi_k^m[\widehat{\Psi}] - \pi_k^m[\Psi^0] &\overset{D}{\sim} \pi_k^m[\Psi^0 + \mathcal{N}_P(0, I^{-1}(\Psi^0))] - \pi_k^m[\Psi^0] \\ &\overset{D}{\sim} \partial_\Psi \pi_k^m(\Psi^0)\mathcal{N}_P(0, I^{-1}(\Psi^0)) \\ &\overset{D}{\sim} \mathcal{N}_P\Big(0, \partial_\Psi \pi_k^m(\Psi^0) I^{-1}(\Psi^0)(\partial_\Psi \pi_k^m(\Psi^0))'\Big)\end{aligned} \tag{4.17}$$

with $\partial_\Psi \pi_k^m$ being the derivative of the function π_k^m with respect to the parameter vector Ψ. This result is often called the *delta method*.

We can apply this result to the multivariate case where

$$\Pi(\Psi) = \begin{pmatrix} \pi_1^m(\Psi) \\ \pi_2^m(\Psi) \\ \vdots \\ \pi_K^m(\Psi) \end{pmatrix} \tag{4.18}$$

Then

$$\Pi(\widehat{\Psi}) - \Pi(\Psi^0) \overset{D}{\sim} \mathcal{N}\Big(0, \partial_\Psi \Pi(\Psi^0) I^{-1}(\Psi^0)(\partial_\Psi \Pi(\Psi^0))'\Big) \tag{4.19}$$

with

$$\partial_\Psi \Pi(\Psi) = \begin{pmatrix} (\nabla \pi_1^m(\Psi))' \\ (\nabla \pi_2^m(\Psi))' \\ \vdots \\ (\nabla \pi_K^m(\Psi))' \end{pmatrix} \tag{4.20}$$

and

$$\nabla \pi_k^m(\Psi) = \begin{pmatrix} \partial \pi_k^m(\Psi)/\partial \Psi_1 \\ \partial \pi_k^m(\Psi)/\partial \Psi_2 \\ \vdots \\ \partial \pi_k^m(\Psi)/\partial \Psi_P \end{pmatrix} \tag{4.21}$$

Given the Equation (4.15) and by the so-called *differentiation under the integral sign*, the gradient of the k-th marginal probability is

$$\frac{\partial \pi_k^m(\Psi)}{\partial \Psi_p} = \frac{\partial \int_b \widehat{\pi_k^{ss}(b)} f(b) db}{\partial \Psi_p} = \int_b \frac{\partial}{\partial \Psi_p} \Big(\widehat{\pi_k^{ss}(b)} f(b)\Big) db, \text{ for } p = 1, \cdots, P \tag{4.22}$$

In Section 4.5, we consider a particular case in which $K = 3$. We suppose a random intercept b_i and the variance-covariance matrix Σ_b defined as:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Let $M(\alpha, \beta, b) = \exp(\alpha + X^{'}\beta + b)$, the probabilities defined in Equation (4.1) are as follows:

$$\begin{aligned} \pi_1(b_1, b_2) &= \frac{M(\alpha_1, \beta_1, b_1)}{1 + M(\alpha_1, \beta_1, b_1) + M(\alpha_2, \beta_2, b_2)} \\ \pi_2(b_1, b_2) &= \frac{M(\alpha_2, \beta_2, b_2)}{1 + M(\alpha_1, \beta_1, b_1) + M(\alpha_2, \beta_2, b_2)} \\ \pi_3(b_1, b_2) &= \frac{1}{1 + M(\alpha_1, \beta_1, b_1) + M(\alpha_2, \beta_2, b_2)} \end{aligned} \tag{4.23}$$

and

$$f(b_1, b_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \exp\Big(\frac{-1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)}(\sigma_2^2 b_1^2 - 2\sigma_{12} b_1 b_2 + \sigma_2^2 b_2^2)\Big) \tag{4.24}$$

$$= (2\pi)^{-1}(\det)^{-1/2} \exp\Big(-\frac{1}{2}(\det)^{-1}(\sigma_2^2 b_1^2 - 2\sigma_{12} b_1 b_2 + \sigma_2^2 b_2^2))\Big) \tag{4.25}$$

with $\det = \det(\Sigma_b) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$. Each parameter is replaced by its estimation obtained by MLE. The derivative term presented in Equation (4.22) is calculated and the Gauss-Hermite technique is used to approximate integration over the random effects. Using Equation (4.19), we obtain the variance-covariance matrix and hence, the 95% CI of the marginal probabilities.

As an alternative method we can use a parametric bootstrap. This could be done by simulating B samples drawn from the estimated parameter, $\widehat{\Psi}$. Separated analysis should be performed on each sample to obtain $\widehat{\Psi_b}$, $b = 1, \cdots, B$. For each sample, marginal probabilities $\widehat{\pi_k^{m(b)}}$, $k = 1, \cdots, K$ and $b = 1, \cdots, B$, are then estimated as explained earlier. The variance of marginal probabilities can be approximated by the empirical variance

$$var_k = \frac{1}{B}\sum_{b=1}^{B}\Big(\widehat{\pi_k^{m(b)}} - \bar{\widehat{\pi_k^{m}}}\Big)^2 \tag{4.26}$$

with

$$\bar{\widehat{\pi_k^{m}}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\pi_k^{m(b)}} \tag{4.27}$$

Given the fact that the latter approach requires much more time, in this dissertation, the first approach is used to calculat confidence intervals.

4.4 Simulation study

4.4.1 Design

We performed a simulation study in order to assess the effects of various censoring rate in our proposed joint model. In particular, different scenarios were conducted to investigate the performance of the proposed joint model for different censoring rates.

The data were generated from two sub-models 4.1 and 4.3, while an administrative censoring rate of 50% (scenario I) and 80% (scenario II) was superimposed.

The longitudinal data were simulated with 3 nominal classes, just as in the motivating example. The covariate $X_{ij} = (x_i, t_{ij})'$, where $t_{ij} \in [0, 0.15]$ in scenario I and $t_{ij} \in [0, 0.70]$ in scenario II, up to 9 observation times, and $x_i \sim Bernoulli(0.5)$. We set $W_{ij} = 1$, so that b_{ik} is the random intercept for subject i and category k. Random effects b_i and u_i were simulated from a multivariate Gaussian distribution using the *mvrnorm* function of the `R` software, setting the mean vector to 0 and the covariance matrix Σ. For the longitudinal part, all $\pi_{ijk} = P(Y_{ij} = k)$ were calculated to be used in a multinomial experiment in order to generate longitudinal categories (using the *rmultinom* function).

The competing risk data were simulated with scheme proposed in [148] with constant baseline hazards $\lambda_{01} = 0.15$, $\lambda_{02} = 0.25$, $Z_i = (z_i, x_i)'$ with $z_i \sim N(2, 1)$. Table 4.1 shows the rate of each event type in our simulation.

Table 4.1 – Description of simulation scenarios

	Scenario I		Scenario II	
	Mean	**Proportion**	**Mean**	**Proportion**
Repeated values	8.2		6.4	
Categories ($K = 3$)				
1		21.47		21.99
2		43.50		45.71
3		35.03		32.20
Event Type				
0		82.73		50.85
1		10.33		27.99
2		6.94		21.16

4.4.2 Results

Table 4.3 and Table 4.2 summarize the simulation results on 500 replications with the sample size $n = 500$. The parameters of the proposed joint model and the separate analysis, consisting of a nominal GLMM for longitudinal data and a Gaussian frailty model for each cause-specific hazard ratio, were estimated respectively. The estimated parameters, simulated bias, standard errors (SE), 95% confidence interval coverage probability (CP), root-mean-square error (RMSE) and the comparison of the mean square errors (MSE) of both joint and separate analysis are presented in Table 4.3 and Table 4.2.

The joint model provided unbiased estimates for all the parameters in both scenarios. The coverage probabilities reach their nominal value of 95%. On the contrary, the separate analysis produces much larger bias in the time trend β_{12} and β_{22} for the longitudinal sub-model. The non-ignorable missing values after death cannot be taken into account in the nominal GLMM alone and therefore, produces bias in the estimated time trend (β_{12} and β_{22}). The standard error of the random effect coefficient, ν_2, is poorly estimated by the separate analysis. We observe that the joint model provides us much more accurate estimates of the coefficient of the random effect in the survival sub-model, ν_2. The ratio of the MSE of the separate analysis (MSE_S)/the MSE of the joint model (MSE_J) for the parameter ν_2 is much larger than 1. This indicates that combining longitudinal data and survival outcomes improves the estimations of the survival sub-model.

Overall, the joint model performs better than the separate analysis as the MSE_J is smaller than or equal to the MSE_S, or the ratio of $\text{MSE}_S/\text{MSE}_J$ is larger than or close to 1. Finally, we observe that in both joint model and separate analysis, the estimation of the variance of the random effect in the survival sub-model, σ_u^2, is biased. Estimation of the variance of the survival random effect, σ_u^2, requires a larger sample size.

Table 4.2 – Comparison of the joint model and the separate analyses (sample size = 500, 50% administrative censoring)

		Separate Analysis					Joint Analysis					
Parameter	**True**	**Estimate**	**Bias**	**SE**	**CP**	**RMSE**	**Estimate**	**Bias**	**SE**	**CP**	**RMSE**	**MSE$_S$/MSE$_J$**
Longitudinal												
Fixed effects												
θ_1	-1.000	-0.975	-0.025	0.135	0.960	0.138	-1.000	-0.000	0.138	0.957	0.138	1.000
θ_2	0.000	0.016	-0.016	0.111	0.968	0.112	0.000	-0.000	0.112	0.975	0.112	1.000
β_{11}	0.200	0.159	0.041	0.151	0.962	0.156	0.191	0.009	0.153	0.953	0.153	1.043
β_{12}	0.300	0.410	-0.110	0.243	0.939	0.267	0.308	-0.008	0.252	0.957	0.252	1.120
β_{21}	0.100	0.081	0.019	0.133	0.960	0.134	0.099	0.001	0.134	0.961	0.134	1.001
β_{22}	0.500	0.600	-0.100	0.205	0.941	0.228	0.516	-0.016	0.213	0.955	0.214	1.143
Random effects												
Σ_{b_1}	1.000	0.986	0.014	0.195	0.954	0.195	1.003	-0.003	0.198	0.959	0.198	0.969
Σ_{b_2}	1.000	0.979	0.021	0.159	0.960	0.161	1.004	-0.004	0.161	0.967	0.161	0.992
Survival												
Fixed effects												
γ_{11}	0.800	0.833	-0.033	0.109	0.876	0.114	0.815	-0.015	0.110	0.885	0.111	1.058
γ_{12}	-1.000	-1.028	0.028	0.198	0.884	0.200	-0.997	-0.003	0.198	0.908	0.198	1.017
γ_{21}	0.500	0.522	-0.022	0.123	0.899	0.125	0.510	-0.010	0.119	0.736	0.119	1.108
γ_{22}	-1.000	-1.076	0.076	0.236	0.926	0.248	-1.038	0.038	0.226	0.926	0.230	1.171
Random effects												
ν_2	0.500	0.679	-0.179	1.296	0.867	1.308	0.498	0.002	0.494	0.818	0.494	7.008
σ_u^2	0.500	0.770	-0.270	0.313	0.644	0.414	0.543	-0.043	0.324	0.721	0.326	1.605
Covariance												
$\Sigma_{b_1 b_2 \sigma_{22}}$	-0.500	-0.505	0.005	0.125	0.979	0.125	-0.503	0.003	0.127	0.973	0.127	0.974
$\Sigma_{b_1 u}$	-0.200	-	-	-	-	-	-0.196	-0.004	0.174	0.883	0.174	-
$\Sigma_{b_2 u}$	-0.200	-	-	-	-	-	-0.202	0.002	0.151	0.867	0.151	-

Table 4.3 – Comparison of the joint model and the separate analyses (sample size = 500, 80% administrative censoring)

		Separate Analysis					Joint Analysis					
Parameter	**True**	**Estimate**	**Bias**	**SE**	**CP**	**RMSE**	**Estimate**	**Bias**	**SE**	**CP**	**RMSE**	**MSE$_S$/MSE$_J$**
Longitudinal												
Fixed effects												
θ_1	-1.000	-0.997	-0.003	0.124	0.969	0.124	-1.003	0.003	0.125	0.975	0.125	0.984
θ_2	0.000	0.000	-0.000	0.104	0.981	0.104	-0.005	0.005	0.104	0.981	0.105	0.981
β_{11}	0.200	0.191	0.009	0.136	0.935	0.137	0.203	-0.003	0.137	0.944	0.137	1.000
β_{12}	0.300	0.456	-0.156	0.963	0.956	0.975	0.309	-0.009	0.974	0.952	0.974	1.002
β_{21}	0.100	0.095	0.005	0.121	0.979	0.121	0.106	-0.006	0.122	0.979	0.122	0.984
β_{22}	0.500	0.629	-0.129	0.806	0.948	0.817	0.512	-0.012	0.815	0.944	0.815	1.005
Random effects												
Σ_{b_1}	1.000	0.993	0.007	0.166	0.965	0.166	1.000	0.000	0.168	0.971	0.168	0.976
Σ_{b_2}	1.000	0.993	0.007	0.137	0.973	0.137	0.999	0.001	0.137	0.975	0.137	1.000
Survival												
Fixed effects												
γ_{11}	0.800	0.873	-0.073	0.165	0.919	0.181	0.849	-0.049	0.163	0.919	0.170	1.133
γ_{12}	-1.000	-1.077	0.077	0.331	0.939	0.340	-1.061	0.061	0.326	0.948	0.332	1.049
γ_{21}	0.500	0.532	-0.032	0.244	0.969	0.247	0.509	-0.009	0.207	0.954	0.207	1.424
γ_{22}	-1.000	-1.109	0.109	0.476	0.985	0.488	-1.087	0.087	0.418	0.979	0.427	1.306
Random effects												
ν_2	0.500	0.443	0.057	2.866	1.000	2.866	0.399	0.101	0.964	0.971	0.969	8.748
σ_u^2	0.500	0.954	-0.454	0.416	0.687	0.615	0.714	-0.214	0.380	0.644	0.437	1.980
Covariance												
$\Sigma_{b_1 b_2}$	-0.500	-0.501	0.001	0.108	0.979	0.108	-0.498	-0.002	0.109	0.975	0.109	0.982
$\Sigma_{b_1 u}$	-0.200	-	-	-	-	-	-0.207	0.007	0.259	0.886	0.259	-
$\Sigma_{b_2 u}$	-0.200	-	-	-	-	-	-0.197	-0.003	0.224	0.877	0.224	-

4.5 Application to the Cosmop-DADS database

We return to the analysis of the Cosmop-DADS database, our motivating database. This database is the result of a record linkage to match the panel of DADS with the causes of death database, using sex, date of birth, date of death and the commune of residence at the time of death as key identifiers. The panel of DADS contains approximately 80% of all paid occupations in France. The Cosmop-DADS population is a sample of the French population, composed of people for whom the vital status and the date of death are available, employed at least once as salaried workers in the semi-public and private sectors. All individuals in this database, 957 299 men and 798 291 women, were followed-up to 2002 and the administrative censoring date was set at 31 December 2002.

As explained in Section 1.4.4 of the Introduction, professional categories were regrouped as *upper class*, *intermediary occupations*, *clerk class*, *manual workers class* and the *outside the scope* class.

We illustrate an application of our proposed joint model using a large sample size of the Cosmop-DADS database[1]. This subset contains all individuals without missing professional episodes, followed-up at least for 10 years, at the age of 34 or more at the entrance and with a transition rate higher than 0.1. Here, the transition rate is defined as the ratio of the number of transitions between professional categories and the total number of follow-up years. For instance, if an individual was working in total for 20 years, first in the manual workers class for 5 years and then in the clerk class for 15 years, his transition rate is $1/20 = 0.05$. Finally, there are 11 852 men and 9827 women included in this application, with 94% right censoring rate.

4.5.1 Joint modelling on large-scale data

The joint modelling of longitudinal data and survival outcomes becomes almost out of reach in large-scale data. The considered sample in this section is very large compared to the existing applications of joint models on real databases. We

[1]It is worthy of note that the results obtained here are not comparable with those in Karimi et al. [38](Chapter 3) as the considered populations in these two studies have not the same characteristics.

therefore, propose an approach mimicking a meta analysis.

In this approach, the initial sample is stratified randomly into S equal size sub-samples. The parameters of the joint model are estimated separately on each stratified sub-sample, $\widehat{\Psi_1}, \cdots, \widehat{\Psi_S}$, and then were combined. As the sub-samples are chosen randomly, the combination of estimations is done by taking the mean of the estimated parameters and therefore, the pooled estimation of the parameter is obtained as follows:

$$\widehat{\Psi}^* = \frac{1}{S} \sum_{s=1}^{S} \widehat{\Psi_s} \tag{4.28}$$

As a consequence, their variance is estimated by:

$$var(\widehat{\Psi}^*) = \frac{1}{S^2} \sum_{s=1}^{S} var\widehat{\Psi_s} \tag{4.29}$$

4.5.2 Results

We divide the initial sample into 10 stratified sub-samples. The sub-samples were stratified by cause of death to preserve the proportion of different causes. Table 4.4 gives a description of these sub-samples. Two competing events, cancer mortality

Table 4.4 – Description of the meta-analysis sub-samples

	Sample	Total
Men	1184	11852
Death	101	1020
Cancer	48	488
Other causes	53	532
Women	982	9827
Death	24	242
Cancer	14	141
Other causes	10	101

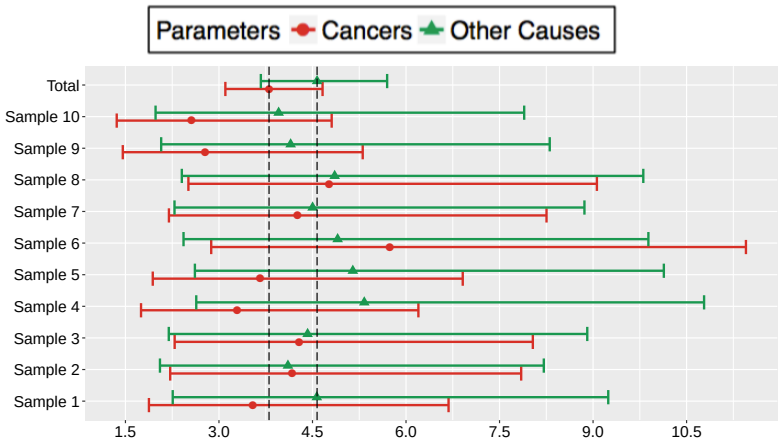
and death by other causes were considered in this part. By using profession as time-dependent variable in cause-specific Cox proportional hazards model, close estimates were found for the upper class and the intermediary occupations [38], thus, we combined these two classes. Therefore, three broad categories were considered: upper class, clerk class and manual workers class, where the upper class were used as the reference category. The longitudinal sub-model was adjusted for sex and age

(mean-centered). The Cox model was adjusted for the variable sex and age was used as the time-scale in this sub-model. To account for left truncation induced by the delayed enteries, we include the age at the entry as the second variable of adjustment.

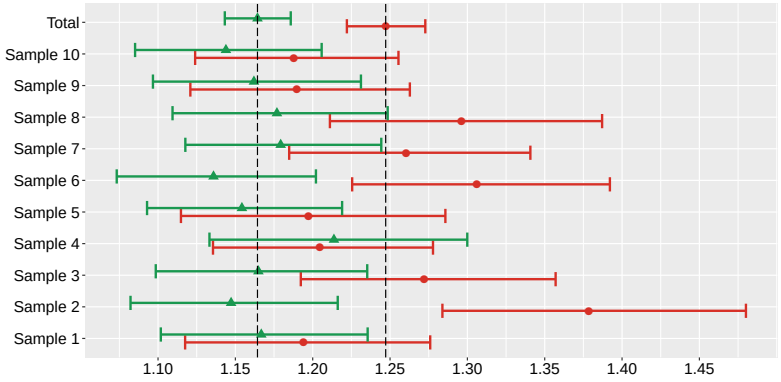
The parameter estimations which resulted from fitting a joint model on each sub-sample are presented in Figure 4.2 and Figure 4.1 for each covariate and the pooled estimates of these 10 stratified sub-samples are summarized in Table 4.5.

In the longitudinal sub-model, the estimated intercept, θ_2, has the following interpretation: The estimated log-odds of the clerk class versus the upper class in the reference group, i.e., men at the mean age, 48 years, is 1.044 (95% CI = [1.006, 1.082]). As highlighted earlier, the exponential of a fixed-effect coefficient in the longitudinal sub-model is interpreted as, for those with the same random effect, the increase in odds of falling into the modality of interest versus the reference modality resulting from a one-unit increase in that covariate, holding the other covariates constant. Therefore, the estimated sex effect in the clerk class versus the upper class, -1.861, is interpreted as follows: Holding age constant, for those with the same random effect, among men the odds of working in the clerk class rather than in the upper class are $0.155 = \exp(-1.861)$ (95% CI = [0.147, 0.163]) times (84.5% lower than) the odds among women. This means that Men are less likely to work in the clerk class versus the upper class than women. No difference was found between men and women for working in the manual workers class versus the upper class.

For women, with the same random effect, a one-year increase in age multiplies the odds of working in the manual workers class rather than working in the upper class by $0.914 = \exp(-0.090)$ (95% CI = [0.912, 0.915]), i.e. decreases this odds by 8.6%. On the other hand, among men, with the same random effect, a one-year increase in age multiplies the odds of working in the manual workers class rather than working in the upper class by $0.926 = \exp(-0.090 + 0.013)$ (95% CI = [0.922, 0.930]), i.e. decreases this odds by 7.4%. The relatively large $\sigma^2_{11} = 5.879$ and $\sigma^2_{22} = 3.325$, random effects of the longitudinal sub-model, reflect strong positive associations between the repeated values.

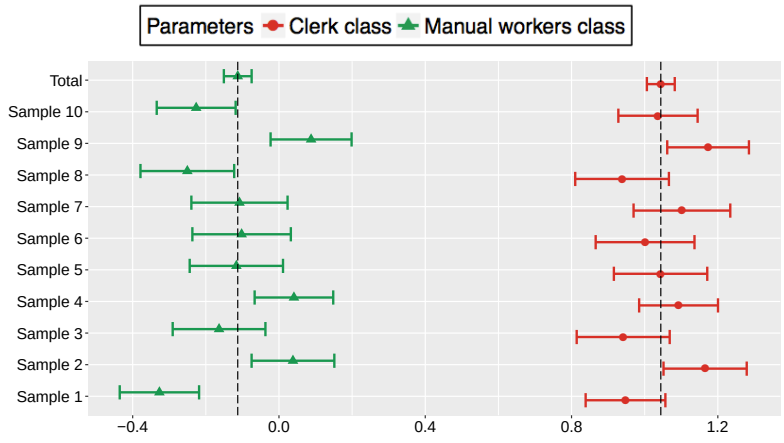


(a) Sex Effect

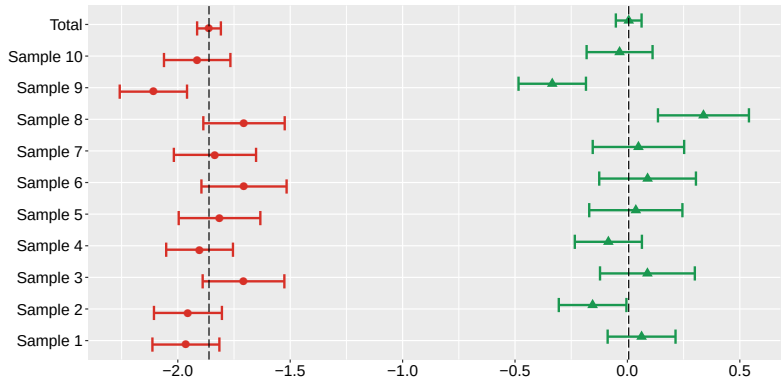


(b) Age at Entry Effect

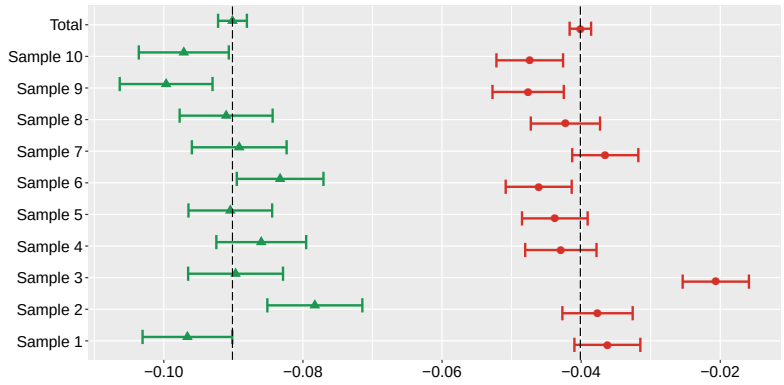
Figure 4.1 – Estimation of covariates effects: Survival sub-model



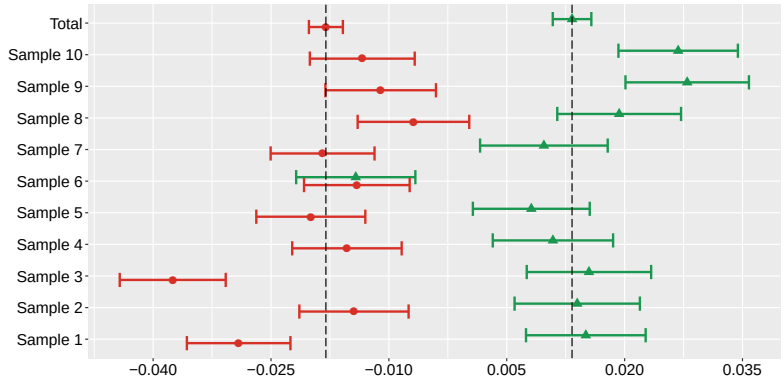
(a) Intercept



(b) Sex Effect



(c) Age Effect



(d) Sex by Age Effect

Figure 4.2 – Estimation of covariates effects: Longitudinal sub-model

For illustrating the difference between men and women over time, it is of interest to plot the estimated membership probabilities of each professional category. As explained in Section 4.3.4, each marginal probability was calculated by integrating the subject-specific probabilities over the random effects distribution. Figure 4.3 plots the estimated probabilities of working in the upper class, working in the clerk class or working in the manual workers class, for men and women, as a function of age[2].

As shown in Figure 4.3, the probability of working in the clerk class decreases strongly for men over time, but it stays almost constant for women over time. The probability of working in the clerk class is strongly higher among women than among men. For working in the manual workers class, there is a more pronounced effect of age among men than among women.

In the survival sub-model, significant effects of both covariates sex and age at entry were observed for mortality by cancer and by other causes. The results show that men had an increased cause-specific mortality compared to women, CSHRs = 3.803 [3.103, 4.662] and 4.574 [3.672, 5.697], respectively, for mortality from cancer and other causes.

Even though, the estimates $\widehat{\sigma_{b_1u}}$, $\widehat{\sigma_{b_2u}}$ or $\widehat{\nu_2}$ are not significant, it's worth detailing the interpretation of their signs. Notably a negative sign of the estimated covariance between the random intercept b_{i2} in the longitudinal sub-model and the random effect u_i in the survival sub-model, $\widehat{\sigma_{b_2u}} = -0.050$, together with the positive $\widehat{\nu_2} = 0.210$, coefficient of the random effect in the survival sub-model, highlights that there is a lower cause-specific hazards for both death by cancers and death by other causes for individuals working in the clerk class as compared to the upper class. On the contrary, a positive sign of the estimated covariance $\widehat{\sigma_{b_1u}} = 0.172$ and the positive $\widehat{\nu_2}$ highlights that, there is a higher cause-specific hazards for both death from cancer and death from other causes for individuals working in the manual workers class as compared to the upper class. By using the professional occupations as time-dependent covariate in the Cox's proportional hazards model, we find the same conclusion, which is the higher cause-specific hazards of individuals in the manual workers class compared to those in the upper class (Table 4.6).

[2]At each age, sum of the estimated marginal membership probabilities is equal to 1.

Table 4.5 – Pooled estimates of 10 stratified sub-samples

(a) Longitudinal sub-model

	Estimate	[95% CI]
Upper class	Reference category	
Manual workers class		
Intercept	**-0.113**	[-0.151,-0.075]
Sex (Men=1, Women=0)	0.006	[-0.051, 0.063]
Age	**-0.090**	[-0.092,-0.088]
Sex by Age	**0.013**	[0.011, 0.016]
Clerk class		
Intercept	**1.044**	[1.006, 1.082]
Sex (Men=1, Women=0)	**-1.861**	[-1.914,-1.808]
Age	-0.040	[-0.042,-0.038]
Sex by Age	**-0.018**	[-0.020,-0.016]

(b) Survival sub-model

	CSHR	[95% CI]
Fixed effects		
Cancer		
Sex	**3.803**	[3.103, 4.662]
Age at entry	**1.247**	[1.222, 1.273]
Other causes		
Sex	**4.574**	[3.672, 5.697]
Age at entry	**1.164**	[1.143, 1.186]
	Estimate	[95% CI]
Random effect		
ν_2	0.210	[-0.027, 0.448]

(c) Estimated variance-covariance matrix of random effects

$$\widehat{\Sigma} = \begin{pmatrix} \mathbf{5.879} & [\ 5.669, 6.089] & \mathbf{2.817} & [\ 2.694, 2.941] & 0.172 & [-0.124, 0.468] \\ \mathbf{2.817} & [\ 2.694, 2.941] & \mathbf{3.325} & [\ 3.208, 3.442] & -0.050 & [-0.278, 0.178] \\ 0.172 & [-0.124, 0.468] & -0.050 & [-0.278, 0.178] & \mathbf{4.936} & [\ 4.552, 5.320] \end{pmatrix}$$

bold indicates p-value < 0.05

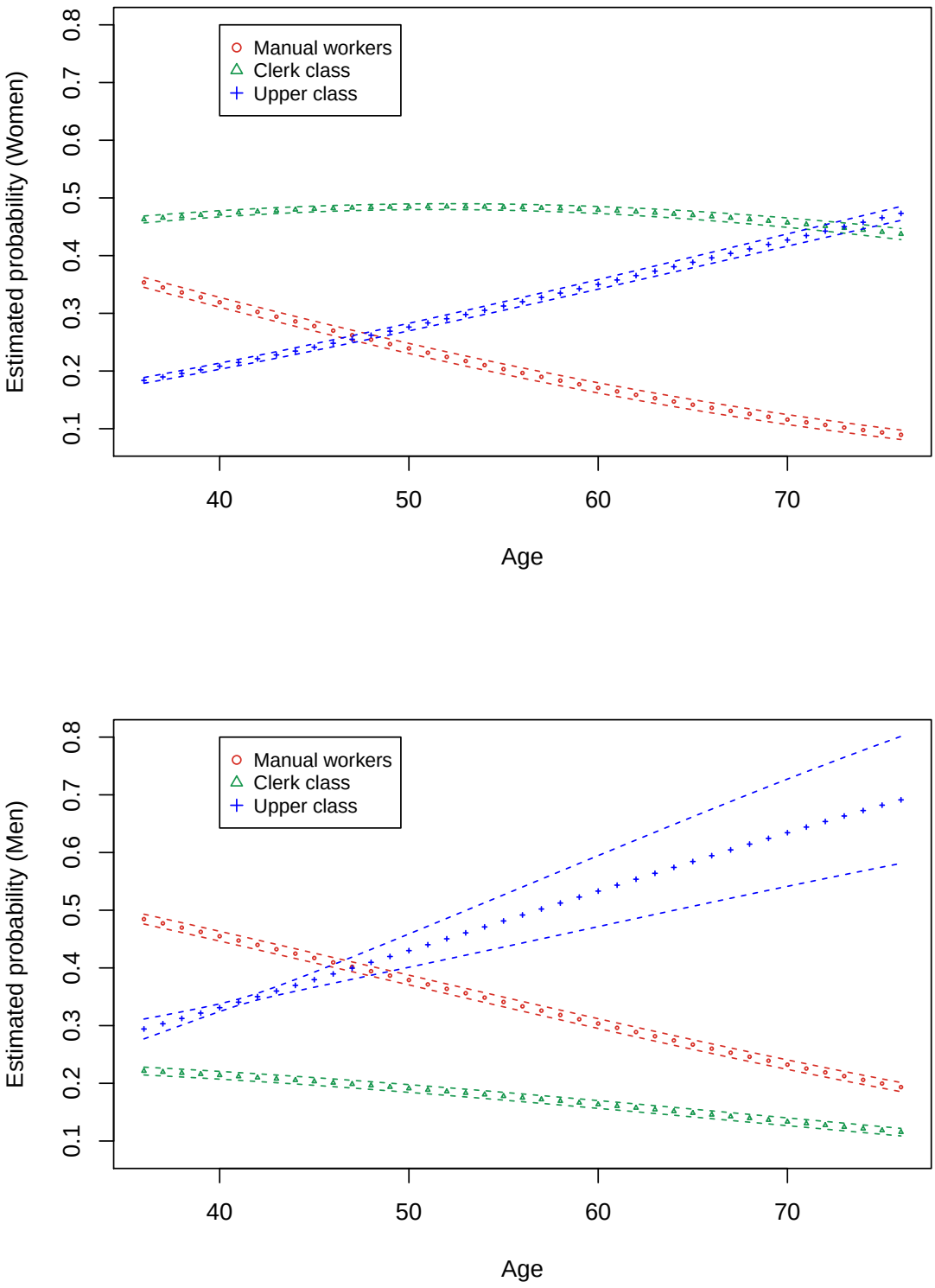


Figure 4.3 – Estimated membership probabilities in professional category

Table 4.6 – Cox analysis on the same sample

	Cancer (n=620)	**Other causes (n=630)**
	$CSHR_{\dagger}$ [95% CI]	$CSHR_{\dagger}$ [95% CI]
Sex		
Women	1	1
Men	**3.18** [2.61, 3.87]	**5.10** [4.10, 6.35]
Age at entry	1.00 [0.98, 1.01]	**1.02** [1.00, 1.03]
Current occupational class		
Upper class	1	1
Manual workers class	**1.42** [1.19, 1.69]	**1.55** [1.30, 1.85]
Clerk class	0.87 [0.70, 1.08]	1.21 [0.98, 1.50]

bold indicates p-value < 0.05

†: age as the time-scale in Cox proportional hazards model

4.6 Discussion

In this Chapter, we proposed an extension of the joint modelling of longitudinal outcomes and competing risk data that handles longitudinal nominal data, such as professional category, and a competing risk model, like time to death by cancer or by other causes. In order to model the repeated observations collected in longitudinal studies, by taking into account their correlation, both GLMMs and marginal models as Generalized Estimating Equations (GEE) [61] have been proposed in the literature. However, these analyses are limited in that they do not consider simultaneously longitudinal outcomes and time-to-event data. Conducting a joint analysis of longitudinal outcomes and time-to-event data allows modelling these two types of data together. A nominal GLMM was used to model the evolution of longitudinal nominal observations over time. Longitudinal outcomes and survival data are then linked by assuming a multivariate Gaussian distribution for the random effects of the two aspects. For the sake of simplicity and given the Cosmop-DADS set also, we considered time-independent covariates in the cause-specific hazards model, however, introducing time-dependent covariates in the survival model is feasible. This can be achieved by evaluating the value of the time-dependent covariate at each time point for the risk set at that time.

Parameter estimation was performed by MLE through an EM algorithm and a one-step Newton-Raphson method. For numerical integrations in the expectation step, Gauss-Hermite rules were applied. Simulation scenarios were carried out to show that the proposed joint model provides less biased estimates than the separate approach, even under high censoring rates.

A large sample size of administrative data on professional trajectories has motivated this work. In the case of large datasets, the calculation of parameter estimates of the joint model is out of reach, therefore, we adopted a meta-analysis strategy. We applied the proposed joint model on stratified sub-samples of the large dataset and then, combined the results by taking the average of the estimations. This approach provides a practical application of the joint models for very large datasets.

In the current parameterization of our joint model, the value of the longitudinal outcome is associated with the cause-specific hazard of an event of interest.

However, as explained in Chapter 3, a professional trajectory can be characterized by the cumulative time spent in each professional category and the transition rates between professional trajectories as well as the current professional category itself. Furthermore, we would like to investigate whether other characteristics of individual's professional trajectory may be associated with the cause-specific hazards. Besides, recently some efforts have been made for situations with 'large n and small p', i.e., many observations and small number of explanatory variables, for Poisson regression models [106]. Future research will focus on including these two aspects in the joint modelling framework.

4.7 Software

To estimate the parameters and the standard errors, a program was developed in C programming language available in `https://www.dropbox.com/sh/ye8su77fa6wjffo/AAC2ClZWNinnsFWns9DjrG5Ca?dl=0`. In order to reduce the calculation time, OpenMP [149] was used. OpenMP is a specification for a set of compiler directives, library routines, and environment variables that can be used to specify high-level parallelism in Fortran and C/C++ programs.

Chapter 5

Joint modelling for large-scale data using Poisson regression models

There is an increasing attention to the analysis of large-scale data. The joint modelling approaches are computationally intensive even in reasonable sample sizes [53, 57]. The parameter estimation in this class of models becomes almost out of reach in large-scale datasets. In Chapter 4 we proposed an approach mimicking meta-analysis to address the estimation problems in large data sets which was based on stratifying the large database and estimating the parameter of interest in each stratified sample.

In the proposed joint model in Chapter 4, a cause-specific proportional hazards model was considered for the competing risks data. The Poisson regression model is equivalent to the Cox model in which the baseline hazard function is approximated by a piecewise constant function. The appeal of the Poisson regression model compared to Cox model appears in the case where all covariates are categorical. That is in large data sets, the hazard rate $\lambda(t)$ of a Poisson regression can be estimated using aggregated data when covariates are categorical or categorized [106]. Using this aggregated data saves a large amount of computation time. Based on this feature, we derived an extension of the joint model for large-scale data.

5.1 Poisson regression model

Let $0 = t_0 < t_1 < t_2 < \cdots < t_L = \tau$ be a partitioning of the study time interval $[0, \tau]$, a Poisson regression models assumes the baseline hazard for cause d to be a step function with a constant value in each interval, i.e.,

$$\lambda_{0d}(t) = \sum_{l=1}^{L} \theta_{ld} 1\{t \in (t_{l-1}, t_l]\} \tag{5.1}$$

with $1\{t \in (t_{l-1}, t_l]\}$ being the indicator of the l-th interval, $\theta_d = (\theta_{1d}, \cdots, \theta_{Ld})'$ and $\theta = (\theta_1', \cdots, \theta_g')'$.

The likelihood of the Poisson regression when all covariates are categorized is then:

$$L(\theta, \gamma) = \prod_{c=1}^{C} \prod_{l=1}^{L} \prod_{d=1}^{g} \Big\{ \{\theta_{ld} \exp(\gamma_d' Z^{(c)})\}^{O_{ld}^{(c)}} \exp\Big(- \theta_{ld} \exp(\gamma_d' Z^{(c)}) R_l^{(c)} \Big) \Big\} \tag{5.2}$$

where C is the number of distinct values of the covariate Z, $Z^{(1)}, \cdots, Z^{(C)}$ and

$$O_{ld}^{(l)} = \sum_{i: Z_i = Z^{(c)}} O_{ild}, \qquad R_l^{(c)} = \sum_{i: z_i = Z^{(c)}} R_{il}$$

Estimation of the parameters $\Phi = (\theta, \gamma)$ is then based on aggregated quantities $O_{ld}^{(c)}$ and $R_c^{(c)}$. When n is much larger than C the parameter estimation in Poisson regression model is much less computational. This can be illustrated by the fact that in a Poisson regression for the cause d, $l \times C$ tables containing the number of deaths by each cause and the number of person-years at risk according to the categorical covariate, are sufficient for the parameter estimation instead of working with all observed data.

5.2 Large-scale joint modelling approach

We consider the same notation as previous chapter. We propose fitting the joint model as follows:

- First, a set of representative trajectories of all observed trajectories, R, should be obtained. This set should have the maximum coverage possible of all trajectories. Let $C = |R|$
- The joint model, combined of three components

– A baseline-category logit model for the c-th representative trajectory, as the longitudinal sub-model (cf. 4.2.1), with

$$\pi_{jk}^{(c)} = P(Y_j^{(c)} = k | X_j^{(c)}, W_j^{(c)}, b_k^{(c)}) = \begin{cases} \frac{1}{1+\sum_{h=1}^{K-1} \exp\left(\alpha_h + {X_j^{(c)}}' \beta_h + {W_j^{(c)}}' b_h^{(c)}\right)} & \text{if } k = K \\ \frac{\exp\left(\alpha_k + {X_j^{(c)}}' \beta_k + {W_j^{(c)}}' b_k^{(c)}\right)}{1+\sum_{h=1}^{K-1} \exp\left(\alpha_h + {X_j^{(c)}}' \beta_h + {W_j^{(c)}}' b_h^{(c)}\right)} & \text{if } k = 1, \cdots, K-1 \end{cases} \tag{5.3}$$

– A Poisson regression model with random effects as the cause-specific hazards sub-model,

– and the variance-covariance matrix of random effects to describe the joint association of longitudinal values and competing risks data,

$$a^{(c)} = \begin{pmatrix} b^{(c)} \\ u^{(c)} \end{pmatrix} \sim \mathcal{N}_{(K-1)q+1} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} , \begin{pmatrix} \Sigma_{b^{(c)}} & \Sigma'_{b^{(c)}u^{(c)}} \\ \Sigma_{b^{(c)}u^{(c)}} & \sigma^2_{u^{(c)}} \end{pmatrix} \right) \tag{5.4}$$

is then performed based on the following likelihood:

$$\begin{aligned} L(\Psi | Y, \widetilde{T}, a) \propto & \prod_{c=1}^{C} \left[\prod_{j=1}^{m^{(c)}} \prod_{k=1}^{K} \{\pi_{jk}^{(c)}\}^{I(Y_j^c = k)} \right]^{\omega_{(c)}} \\ & \times \left\{ \prod_{l=1}^{L} \prod_{d=1}^{g} \left\{ \{\theta_{ld} \exp(\gamma_d' Z^{(c)} + \nu_d u^{(c)})\}^{O_{ld}^{(c)}} \exp\left(- \theta_{ld} \exp(\gamma_d' Z^{(c)} + \nu_d u^{(c)}) R_l^{(c)} \right) \right\} \right\} \\ & \times \frac{1}{\sqrt{(2\pi)^{(K-1)q+1} |\Sigma|}} \exp\left(- \frac{1}{2} {a^{(c)}}' \Sigma^{-1} a^{(c)} \right) \end{aligned} \tag{5.5}$$

with ω_c the percentage of trajectories represented by the c-th element of S and $m^{(c)}$ the number of observed consecutive for the c-th trajectory.

5.3 Classification of longitudinal trajectories

Different clustering approaches can be found in the literature regarding type of data, such as clustering approaches based on a mixture of regression models [58, 150]. Given the main objective of this part, that is searching for typical professional trajectories, we focus on classification methods for longitudinal nominal trajectories. The analysis of categorical sequences, also known as *sequence analysis*, is one of the most discussed approaches for this purpose.

Sequence analysis has been initialized in social science since the work of Abbott et al. [151], the so-called Optimal Matching (OM) analysis. The idea of sequence

analysis is based on comparing life course trajectories according to different dissimilarity metrics and clustering these trajectories based on the calculated distances between them. One can identify typical trajectories through visual inspection [152]. However, when the number of trajectories in a cluster increases, finding visually the typical trajectory may be difficult. To address this issue two approaches can be considered: One could search for representative trajectories among the observed trajectories, or creating an artificial trajectory that verify a supposed criteria. The latter approach can produces a sequence that is not plausible in social context and thus, we focus on the first approach.

5.3.1 Dissimilarity metrics

A dissimilarity metric is a method to evaluate the level of difference between two trajectories. These metrics can be classified into three categories:

- Measuring the distance between the distributions of two sequences;
- Measuring the number of similar (common) patterns between two trajectories;
- Measuring the cost of operations transforming one trajectory to another, also called as *edit* distances. These operations are substitutions, deletions and insertions (*indels*), compression and expansions, and swaps.

A complete overview of these dissimilarity metrics can be found in Studer et al. [153]. In this chapter, we focus on the most common approach in social science which is OM.

OM which is an *edit* metric, aims to evaluate the distance between two trajectories x and y, $d(x, y)$ as the minimum cost of transforming trajectory x to trajectory y by operations *substitution*, *insertion* and *deletion* of states. The main drawback of this approach is that these operations and their costs are not meaningful in sociological terms [154, 155].

Let $S = \{s_1, \cdots, s_N\}$ be the list of the possible states. The dissimilarity between two sequences x and y is then defined as:

$$d_{OM}(x, y) = min_j \sum_{i=1}^{l_j} \gamma(O_i^j) \tag{5.6}$$

where O_i^j are operations that transform trajectory x into y, l_j is the necessary number of operations and $\gamma(O_i^j)$ is the cost of operation O_i^j. These operations could be substituting s_m with s_n ($s_m \rightarrow s_n$), deleting s_m ($s_m \rightarrow$) and inserting s_m ($\rightarrow s_m$).

Different strategies can be considered to determine the substitution costs. They can be set based on a priori knowledge. For instance, this can be done when a hierarchy order exists between the states. Even if by doing this, an order is set but the costs of the operations are chosen arbitrary. Another option is to attribute a value at each state and derive operation costs. A third solution would be using the data.

For choosing the *indel* cost, most applications are based on a single indel cost without giving any importance to the inserted or deleted state. As an alternative, one could choose state-dependent indel costs, by giving a higher cost to rare states. By choosing different costs, variants of the OM, such as *dynamic Hamming distance* [155], *localized optimal matching* [156], *optimal matching sensitive to spell length* [157], *optimal matching between sequences of spells* [153] and *optimal matching between sequences of transitions* [158] can be obtained.

5.3.2 Typical trajectories

Based on obtained distances between sequences, finding the sequences that represent the trajectories is then possible. Given that the defined distances does not represent the distribution of trajectories, typical trajectories can not be obtained directly using these distances.

Gabadinho et al. [159] proposed the following approach for searching a typical trajectory:

1. First, a representativeness score (*frequency*, *neighbourhood density* or *centrality*) for each distinct trajectory should be computed;

2. Sorting distinct trajectories according to their scores;

3. Starting from the most representative trajectory, keep from the list each trajectory for which the distance with any already retained trajectory is greater than a given threshold. Repeat until the expected overall coverage is attained.

It is then necessary to evaluate obtained representative trajectories. The contribution of the i-th representative trajectory r_i to the overall absolute coverage is defined as the number of trajectories among those assigned to r_i that are in its neighbourhood:

$$c_i = \sum_{j \in R_i} \Big(d(x_j, r_i) < \delta \Big) \tag{5.7}$$

where δ is the neighbourhood radius for the overall coverage. Then the absolute coverage c is defined as the sum of the c_i and the overall coverage for n trajectories is c/n.

5.4 Typical trajectories in Cosmop-DADS

We illustrate this approach by considering the sample of the previous chapter. This sample contains 21 660 individuals with professional episodes in *Upper class* and *Intermediary occupations* (UP), *Clerk class* (CL) and *Manual workers class* (MA). Between these 21 660 professional trajectories, there are 16 625 distinct trajectories. For calculating the dissimilarities between these trajectories, we consider OM with a unit *indel* cost and the substitution cost matrix based on the transition rates.

$$\begin{array}{c|ccc} & UP \rightarrow & CL \rightarrow & MA \rightarrow \\ UP \rightarrow & 0.000 & 1.708 & 1.746 \\ CL \rightarrow & 1.708 & 0.000 & 1.712 \\ MA \rightarrow & 1.746 & 1.712 & 0.000 \end{array} \tag{5.8}$$

Figure 5.1 shows the representative trajectory between all observed trajectories. We used the medoid [160], which is the criteria to find the most central sequence. This typical trajectory represents the indivdual who was followed for 7 years, starting in the Manusal workers class at first, moving to the Clerk class and working in this class for 2 years, and then going back to the Manual workers class. This individual worked in the Upper class for the last 3 years of his follow-up. It is clear that the professional trajectory of this individual is not representative of all professional trajectories. The overall coverage of this trajectory, obtained for a neighbourhood radius of 10% of the maximal possible distance is about 2% which is not satisfactory. Thus, we should search for a set of representative trajectory with more satisfied coverage.

Figures 5.2 and 5.3 show representative trajectories for 25%, 50%, 80% and 90% of coverage, respectively. As it is shown in these Figures, 5, 22, 208 and 469

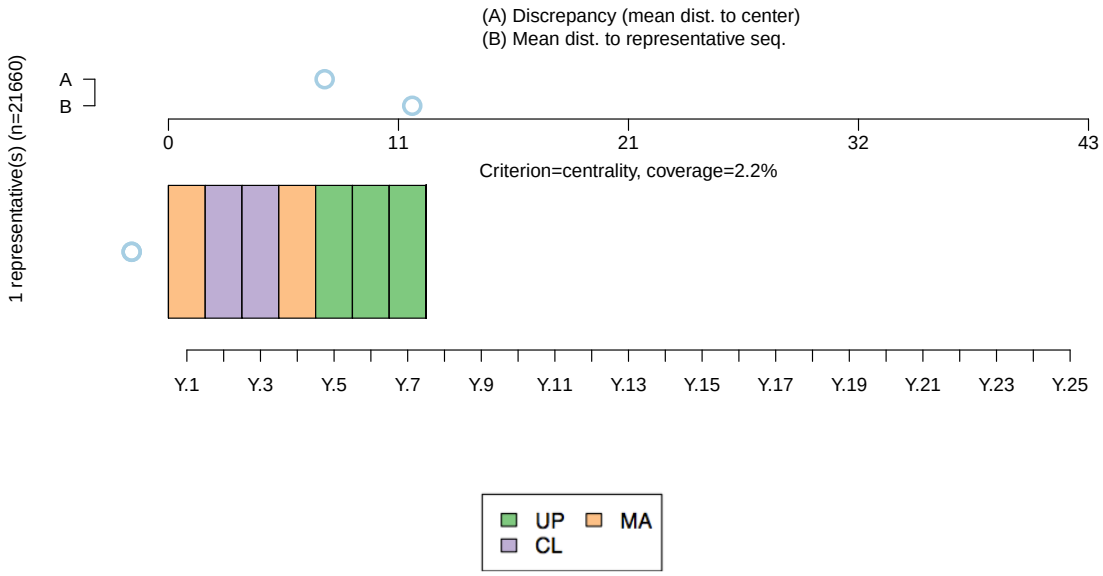
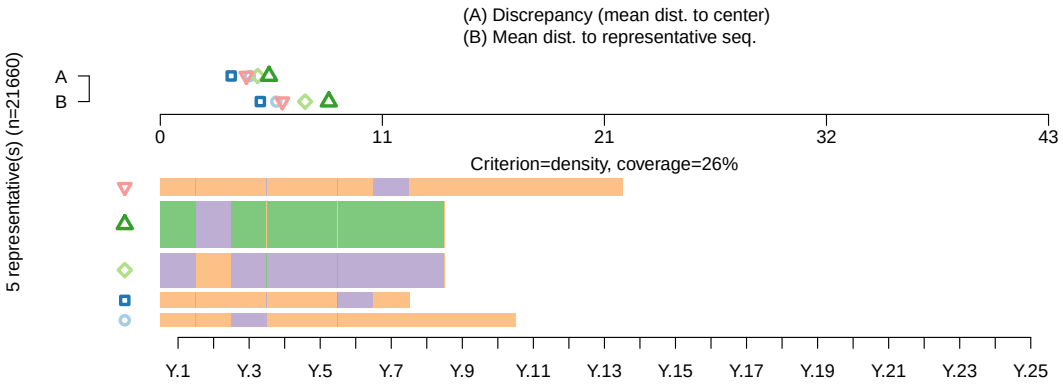
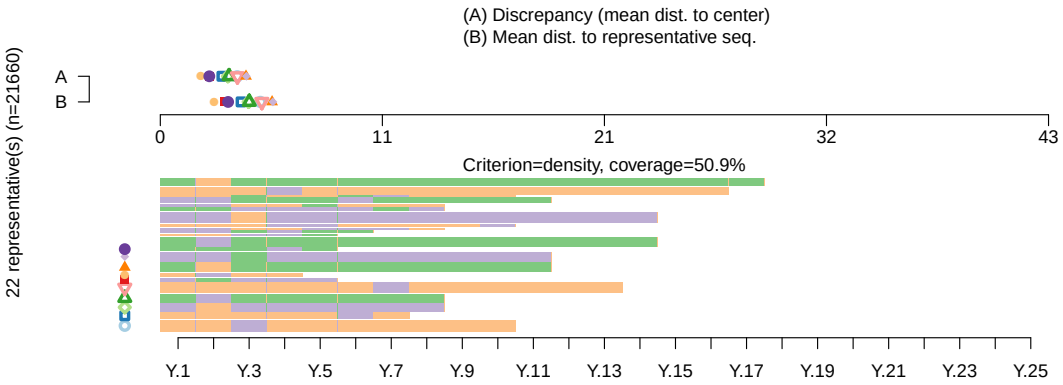


Figure 5.1 – Representation of typical trajectory

trajectories could represent 21660 professional trajectories for having respectively, 25%, 50%, 80% and 90% coverages. Based on these results, having 469 representative professional trajectories coveraging almost all trajectories, is an advantage as the proposed joint model in this chapter can be applied. It is obvious that running through 469 individuals is much easier and needs less computational time.



(a) Coverage 25%



(b) Coverage 50%

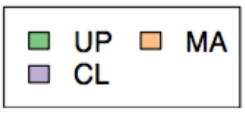
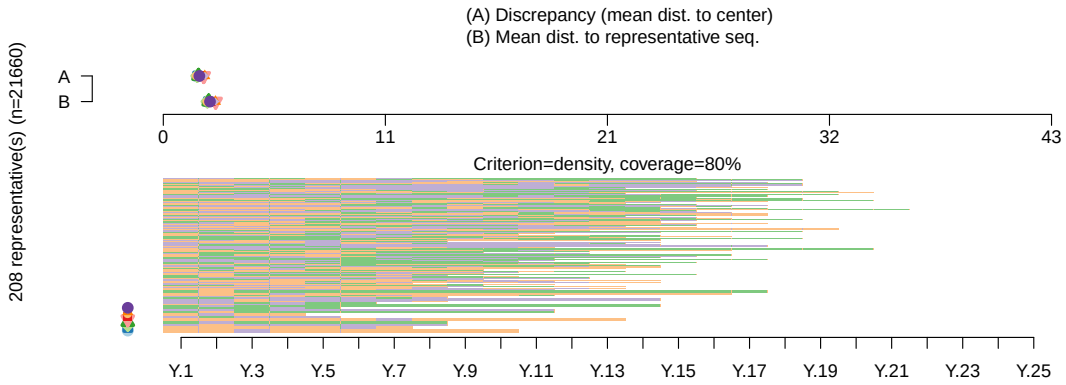
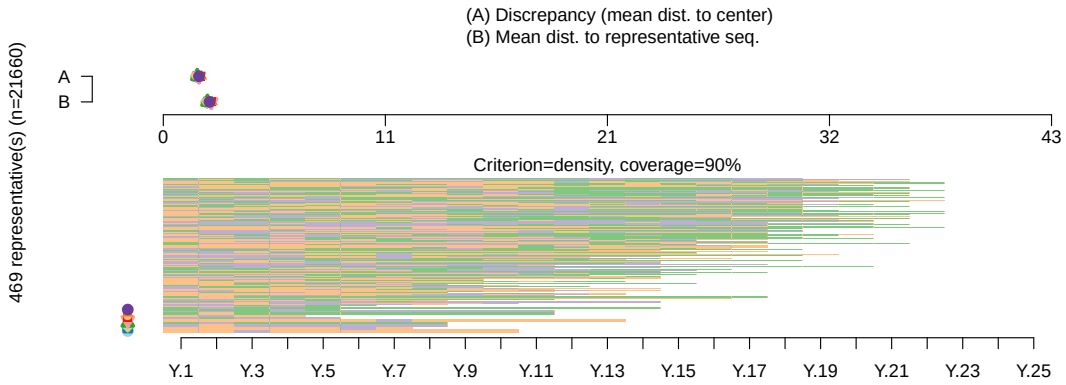


Figure 5.2 – Representative trajectories for different coverage percentage



(a) Coverage 80%



(b) Coverage 90%

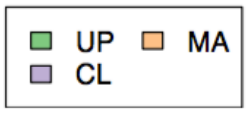


Figure 5.3 – Representative trajectories for different coverage percentage

5.5 Discussion

In this Chapter, we proposed an approach to address the calculation problem of joint models in large-scale data. Since we are dealing with professional trajectories with less than 5 occupational category, and due to the fact that some transitions between occupational categories are more likely to happen compared to the others, the trajectories are likely to repeat. Therefore, we considered an OM approach to find representative trajectories of all observed trajectories and apply the presented joint model on the obtained set of representative trajectories. The practical implementation and evaluation of the joint model are the object of ongoing research. A future research should focus on a comparison of different approaches to classify professional trajectories and their impact on the proposed joint modelling framework.

General discussion and future research

In this dissertation we have presented and discussed different approaches to address an epidemiological question that is measuring the association between life-course socio-professional trajectories and cause-specific mortality. From an epidemiological point of view, this has already been studied in different countries using limited number of stages in professional trajectories, such as professional position at labour market and professional position at midlife, while in this study we add the opportunity to consider the whole professional trajectory defined as the consecutive occupational positions of an individual during his/her life course.

Some of previous studies on this subject were based on life course models, namely *critical period*, *accumulation model* and *social mobility model*. In this context, we proposed defining variables to capture these life-course hypothesis (Chapter 3). The duration of time spent in occupational classes was considered as a measure of socioeconomic exposure and the transition rates between occupational classes was defined to take into account the social mobility dimension. Through our analysis we showed that long-time exposure to poor socioeconomic position is strongly associated with adult mortality, especially for cardiovascular diseases. In addition, our analysis also provided support for the critical period and social mobility model. Additional analysis showed the necessity of separating downward and upward analysis by defining an order between professional categories if possible.

Regarding the bias due to using internal time-dependent variables in Cox proportional hazards model, we focused on joint modelling for longitudinal nominal

outcomes and competing risks data. Occupational episodes were incorporated in a longitudinal sub-model to model professional trajectory. Longitudinal sub-model is then linked with competing risks model through unobserved random effects (Chapter 4). Furthermore, this approach provides additional summary measures, such as estimation of membership probabilities of each professional category taking into account competing risks data. In the parameterization of the proposed joint model, the value of the longitudinal outcome was associated with the hazard of an event of interest. Further research should focus on associating different characteristics of longitudinal trajectories, such as the accumulation of time spent in each category, with cause-specific mortality. An alternative approach would be the use of dynamic landmarking [161], that aims to estimate the effect of a time-dependent exposure covariate on survival. This approach is interesting, however building a super-model and landmark data set in such large database could require data preprocessing.

One drawback of joint modelling is the computational burden required by multiple integrations. We therefore, proposed in Chapter 4 an approach mimicking meta analysis to handle large-scale data in the joint modelling framework. This approach gives us the opportunity to apply the joint model on a sample of 20 000 individuals. In the same objective, that is fitting joint model on large-scale data, we propose a procedure based on the technical advantage of the poisson regression model (Chapter 5). First a set of representative trajectories of all observed longitudinal trajectories should be obtained by means of clustering methods following by a joint model on these representative trajectories.

We should note that an important condition in using any statistical method is the availability of software. Implemented `R` codes are available to perform the procedures proposed in Chapter 3. In the joint modelling framework, different packages have already been implemented, but none of them covers the proposed model in Chapter 4 for longitudinal nominal data. presented simulation results and application on real data in Chapter 4 are obtained using a C code implemented for this purpose. Also, software implementation for the proposed model in Chapter 5 is part of ongoing work. Once finished, different simulation scenarios as well as application on the same dataset as the dataset used in Chapter 4 will be performed to evaluate the approach.

It is also of interest to mention the life histories framework, such as multi-state models [94, 162]. The idea is based on considering life course trajectories as a sequence of states and transitions between these states. In a multi-state framework, a separate state should be considered for each professional category and for each competing risk. The competing risks states are *absorbing* states and professional categories are *transient*. The parameters of interest in this model are transition rates between states that are estimated from data by counting the number of events and individuals at risk. One of the interesting `R` packages in life history data analysis is *Biograph*. In addition to the calculation of the parameter of interest in multi-state model, *Biograph* provides functions for visualisation of life histories data using functions of the *TraMineR* pacakge [118]. *TraMineR* also provides the model-free data mining method of *sequence analysis*. In an exploratory analysis, the visualisation and classification of professional trajectories in the Cosmop-DADS was studies previously [163]. As stated by Eerola et al. [164], these approaches complement each other, as one focuses on finding typical patterns and the other focuses on transition rates between states.

It is necessary to emphasize that the utility of approaches discussed in this dissertation depends on the quality and characteristics of the data. The main limitation of our study was the structure of the motivating database. The Cosmop-DADS database does not cover all professional categories and thus, contains about 40% missing data that could not be completed by standard methods such as imputation approaches. In order to better estimate the association between professional trajectories and cause-specific mortality, there is need for improving the quality of database. Linking this database with datasets on educational level, on salaries and on quality of life would improve the precision of socioeconomic categories and thus more precise results on the association between socio-professional trajectories and adults mortality can be obtained. The new Permanent Demographic Sample (EDP++) that has been constructed linking census and civil state data with DADS could be the alternative dataset for future studies. Moreover, in the presence of professional episodes (time-dependent exposure) and salaries (confounding time-dependent covariates), methods on causal inference such as Marginal Structural Model (MSM) [165] could go further in the analysis of this association. In this

context, some approaches has also been proposed for competing risks problem [166].

It is also of interest to mention the mortality data used in this study registered by INSERM-CépiDc. We focused on these data and especially on the underlying cause of death as an indicator of mortality. Being exhaustive and the fact that these data are recorded on the entire country uniformly are their main advantages. The mortality data gives the possibility for international comparisons of epidemiological and demographical studies. However, It is often difficult to define a single underlying cause of death, especially in elderly populations where they suffer from multiple pathologies [167]. In this context, the analyses of multiple causes of death should be considered. Recently, Moreno-Betancur et al. [168] discussed this issue based on an empirical approach. Employing multiple causes of death in joint modelling framework could be the object of study in the future.

Bibliography

[1] Marang-van de Mheen, P. J., Davey Smith, G., Hart, C. L., and Gunning-Schepers, L. J. (1998). Socioeconomic differentials in mortality among men within Great Britain: time trends and contributory causes. *J Epidemiol Community Health*, 52(4): 214–218.

[2] Martikainen, P., Valkonen, T., and Martelin, T. (2001). Change in male and female life expectancy by social class: decomposition by age and cause of death in Finland 1971-95. *J Epidemiol Community Health*, 55(7): 494–499.

[3] Davey Smith, G., Bartley, M., and Blane, D. (1990). The Black report on socioeconomic inequalities in health 10 years on. *BMJ*, 301(6748): 373–377.

[4] Desplanques, G. (1984). L'inégalité sociale devant la mort. *estat*, 162(1): 29–50.

[5] Menvielle, G., Luce, D., Geoffroy-Perez, B., Chastang, J.-F., and Leclerc, A. (2005). Social inequalities and cancer mortality in France, 1975–1990. *Cancer Causes Control*, 16(5): 501–513.

[6] Lerlerc, A., Lert, F., and Goldberg, M. (1984). Les inégalités socials devant la mort en grande-bretagne et en France. *Soc Sci Med*, 19(5): 479–487.

[7] Michel, E., Jougla, E., and Hatton, F. (1996). Mourir avant de vieillir. *Insee première*, (429).

[8] Leclerc, A., Fassin, D., Granjean, H., Kaminski, M., and Lang, T. (2000). Les inégalités sociales de santé. La Découverte.

[9] Kunst, A. E. and Mackenbach, J. P. (1994). The size of mortality differences associated with educational level in nine industrialized countries. *Am J Public Health*, 84(6): 932–937.

[10] Leclerc, A., Lert, F., and Fabien, C. (1990). Differential mortality: some comparisons between England and Wales, Finland and France, based on inequality measures. *Int J Epidemiol*, 19(4): 1001–1010.

[11] Mackenbach, J. P., Kunst, A. E., Cavelaars, A. E., Groenhof, F., and Geurts, J. J. (1997). Socioeconomic inequalities in morbidity and mortality in western Europe. *The Lancet*, 349(9066): 1655–1659.

[12] Leclerc, A., Chastang, J.-F., Menvielle, G., and Luce, D. (2006). Socioeconomic inequalities in premature mortality in France: have they widened in recent decades? *Soc Sci Med*, 62(8): 2035–2045.

[13] Mackenbach, J. P., Bos, V., Andersen, O., Cardano, M., Costa, G., Harding, S., Reid, A., Hemstrom, O., Valkonen, T., and Kunst, A. E. (2003). Widening

socioeconomic inequalities in mortality in six Western European countries. *Int J Epidemiol*, 32(5): 830–837.

[14] Pappas, G., Queen, S., Hadden, W., and Fisher, G. (1993). The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986. *N Engl J Med*, 329(2): 103–109.

[15] Menvielle, G., Chastang, J. F., Luce, D., Leclerc, A., and Edisc Group (2007). Changing social disparities and mortality in France (1968-1996): cause of death analysis by educational level. *Rev Epidemiol Sante Publique*, 55(2): 97–105.

[16] Monteil, C. and Robert-Bobée, I. (2005). Les différences sociales de mortalité: en augmentation chez les hommes, stables chez les femmes. *Insee première*, (1025).

[17] Galobardes, B., Shaw, M., Lawlor, D. A., and Lynch, J. W. (2006). Indicators of socioeconomic position (part 1). *J Epidemiol Community Health*, 60(1): 7–12.

[18] Stringhini, S., Dugravot, A., Shipley, M., Goldberg, M., Zins, M., Kivimëki, M., Marmot, M., Sabia, S., and Singh-Manoux, A. (2011a). Health behaviours, socioeconomic status, and mortality: further analyses of the British Whithehall II and the French GAZEL prospective cohorts. *PLoS Med*, 8(2): e1000419.

[19] Steenland, K., Henley, J., and Thun, M. (2002). All-Cause and Cause-specific Death Rates by Educational Status for Two Million People in Two American Cancer Society Cohorts, 1959–1996. *Am J Epidemiol*, 156(1): 11–21.

[20] Kunst, A. E., Groenhof, F., Mackenbach, J. P., and Leon, D. A. (1998). Occupational class and cause specific mortality in middle aged men in 11 European countries: comparison of population based studies: EU Working Group on Socioeconomic Inequalities in Health. *BMJ*, 316(7145): 1636–1642.

[21] Davey Smith, G., Hart, C., Blane, D., Gillis, C., and Hawthorne, V. (1997). Lifetime socioeconomic position and mortality: prospective observational study. *BMJ*, 314(7080): 547–552.

[22] Melchior, M., Berkman, L. F., Kawachi, I., Krieger, N., Zins, M., Bonenfant, S., and Goldberg, M. (2006). Lifelong socioeconomic trajectory and premature mortality (35-65 years) in France: findings from the GAZEL Cohort Study. *J Epidemiol Community Health*, 60(11): 937–944.

[23] Blane, D., Harding, S., and Rosato, M. (1999a). Does social mobility affect the size of the socioeconomic mortality differential?: evidence from the Office for National Statistics Longitudinal Study. *J R Stat Soc Ser A Stat Soc*, 162(1): 59–70.

[24] Lynch, J. W., Kaplan, G. A., Cohen, R. D., Kauhanen, J., Wilson, T. W., Smith, N. L., and Salonen, J. T. (1994). Childhood and adult socioeconomic status as predictors of mortality. *The Lancet*, 343(8896): 524–527.

[25] Galobardes, B., Lynch, J. W., and Davey Smith, G. (2008). Is the association between childhood socioeconomic circumstances and cause-specific mortality established? Update of a systematic review. *J Epidemiol Community Health*, 62(5): 387–390.

[26] Stringhini, S., Dugravot, A., Kivimaki, M., Shipley, M., Zins, M., Goldberg, M., Ferrie, J. E., and Singh-Manoux, A. (2011b). Do different measures

of early life socioeconomic circumstances predict adult mortality? Evidence from the British Whitehall II and French GAZEL studies. *J Epidemiol Community Health*, 65(12): 1097–1103.

[27] Singh-Manoux, A., Ferrie, J. E., Chandola, T., and Marmot, M. (2004). Socioeconomic trajectories across the life course and health outcomes in midlife: evidence for the accumulation hypothesis? *Int J Epidemiol*, 33(5): 1072–1079.

[28] Geoffroy-Perez, B., Imbernon, E., and Goldberg, M. (2005). "Projet Cosmop : cohorte pour la surveillance de la mortalité par profession. Premiers résultats de l'étude de faisabilité á partir de l'Échantillon démographique permanent". Département santé-travail, InVS.

[29] Kalbfleisch, J. and Prentice, R. (2002). The Statistical Analysis of Failure Time Data. 2nd ed. Wiley, New York.

[30] Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J., and Power, C. (2003). Life course epidemiology. *J Epidemiol Community Health*, 57(10): 778–783.

[31] Galobardes, B., Lynch, J. W., and Davey Smith, G. (2004). Childhood Socioeconomic Circumstances and Cause-specific Mortality in Adulthood: Systematic Review and Interpretation. *Epidemiol Rev*, 26(1): 7–21.

[32] Pollitt, R. A., Rose, K. M., and Kaufman, J. S. (2005). Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. *BMC Public Health*, 5(1): 1–13.

[33] Heslop, P., Davey Smith, G., Macleod, J., and Hart, C. (2001). The socioeconomic position of employed women, risk factors and mortality. *Soc Sci Med*, 53(4): 477–485.

[34] Blau, P. M. (1956). Social Mobility and Interpersonal Relations. *Am Social Rev*, 21(3): 290–295.

[35] Blane, D., Davey Smith, G., and Hart, C. (1999b). Some Social and Physical Correlates of Intergenerational Social Mobility: Evidence from the Western of Scotland Collaborative Study. *Sociology*, 33(1): 169–183.

[36] Fox, A. J., Goldblatt, P. O., and Jones, D. R. (1985). Social class mortality differentials: artefact, selection or life circumstances? *J Epidemiol Community Health*, 39(1): 1–8.

[37] Smith, J. P. (1999). Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status. *J Econ Perspect*, 13(2): 144–166.

[38] Karimi, M., Geoffroy-Perez, B., Fouquet, A., Latouche, A., and Rey, G. (2015). Socioprofessional trajectories and mortality in France, 1976-2002: a longitudinal follow-up of administrative data. *J Epidemiol Community Health*, 69(4): 339–346.

[39] DeGruttola, V. and Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4): 1003–1014.

[40] Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *J Am Stat Assoc*, 90(429): 157–169.

[41] Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for longitudinal Data. *Biometrics*, 38(4): 963–974.

[42] Faucett, C. L. and Thomas, D. C. (1996). Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: a Gibbs Sampling Approach. *Stat Med*, 15(15): 1663–1685.

[43] Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. 1st ed. Chapman & Hall/CRC.

[44] Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4): 465–480.

[45] Elashoff, R. M., Li, G., and Li, N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Stat Med*, 26(14): 2813–2835.

[46] Li, N., Elashoff, R. M., Li, G., and Saver, J. (2010). Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. *Stat Med*, 29(5): 546–557.

[47] Han, J., Slate, E. H., and Peña, E. A. (2007). Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Stat Med*, 26(29): 5285–5302.

[48] Proust-Lima, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2015). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Stat Med*, 35(3): 382–398.

[49] Brown, E. R., Ibrahim, J. G., and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1): 64–73.

[50] Zeng, D. and Cai, J. (2005). Simultaneous Modelling of Survival and Longitudinal Data with an Application to Repeated Quality of Life Measures. *Lifetime Data Anal*, 11(2): 151–174.

[51] Faucett, C. L., Schenker, N., and Elashoff, R. M. (1998). Analysis of Censored Survival Data with Intermittently Observed Time-Dependent Binary Covariates. *J Ame Stat Assoc*, 93(442): 427–437.

[52] Li, N., Elashoff, R. M., Li, G., and Tseng, C.-H. (2012). Joint analysis of bivariate longitudinal ordinal outcomes and competing risks survival times with nonparametric distributions for random effects. *Stat Med*, 31(16): 1707–1721.

[53] Murawska, M. and Rizopoulos, D. (2013). "Extensions in Joint Modeling of Survival and Longitudinal Outcomes". PhD thesis. Erasmus University Rotterdam.

[54] Desrosières, A. and Thévenot, L. (2002). Les catégories socio-professionnelles. 5th ed. La Découverte.

[55] Rose, D. and O'Reilly, K. (1998). The ESRC review of government social classifications. Office for National Statistics/ESRC.

[56] Cambois, E. (2004a). Occupational and educational differentials in mortality in French elderly people: Magnitude and trends over recent decades. *Demogr Res*, S2: 277–304.

[57] McCrink, L. M., Marshall, A. H., and Cairns, K. J. (2013). Advances in Joint Modelling: A Review of Recent Developments with Application to the Survival of End Stage Renal Disease Patients. *Int Stat Rev*, 81(2): 249–269.

[58] Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data. Springer, New York.
[59] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). Analysis of longitudinal data. Oxford University Press, New York.
[60] Agresti, A. (2013). Categorical Data Analysis. 3rd ed. Hoboken, NJ, USA: Wiley-Blackwell.
[61] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1): 13–22.
[62] McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. 2nd ed. Chapman & Hall/CRC.
[63] Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *Int Stat Rev*, 64(1): 89–118.
[64] McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J R Stat Soc Series B Methodol*, 42(2): 109–142.
[65] Stukel, T. (1993). Comparison of methods for the analysis of longitudinal interval count data. *Stat Med*, 12(14): 1339–1351.
[66] Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. Springer, New York.
[67] Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. *Stat Modelling*, 1(2): 81–102.
[68] Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Stat Med*, 22(9): 1433–1446.
[69] Fahrmeir, L. and Tutz, G. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models. Springer.
[70] Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50(4): 933–944.
[71] Pinheiro, J. C. and Bates, D. M. (2000). Mixed-Effects Models in S and S-PLUS. Springer.
[72] McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). Generalized, Linear, and Mixed Models. 2nd ed. Wiley.
[73] Hartley, H. O. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics*, 14(2): 174–194.
[74] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J R Stat Soc Series B Methodol*, 39(1): 1–38.
[75] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ, USA: John Wiley & Sons, Inc.
[76] Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *J R Stat Soc Series B Methodol*, 44(2): 226–233.
[77] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann Stat*, 7(1): 1–26.
[78] McLachlan, G. J. and Krishnan, T. (2008). The EM Algorithm and Extensions. 2nd ed. Wiley Series in Probability and Statistics.
[79] Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the Loglikelihood Function in the Nonlinear Mixed Effects Model. *J Comput Graph Stat*, 4(1): 12–35.

[80] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3): 581–592.

[81] Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. 2nd ed. New York, USA: John Wiley & Sons, Inc.

[82] O., S. and W., A. M. (1998). A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *J Biopharm Stat*, 8(4): 545–563.

[83] Little, R. J. A. (1995). Modeling the drop-out mechansim in repeated-measures studies. *J Am Stat Assoc*, 90(431): 1112–1121.

[84] — (1993). Pattern-Mixture Models for Multivariate Incomplete Data. *J Am Stat Assoc*, 88(421): 125–134.

[85] Wu, M. C. and Carroll, R. J. (1988). Estimation and Comparaison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics*, 44(1): 175–188.

[86] Hogan, J. W. and Laird, N. M. (1997a). Mixture models for the joint distribution of repeated measures and event times. *Stat Med*, 16(3): 239–257.

[87] Diggle, P. J. (1998). "Recent advances in the Statistical Analysis of Medical Data". Ed. by B. S. Everitt and G. Dunn. Arnold, London. Chap. Dealing with missing values in longitudinal studies: pp. 203–228.

[88] Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J R Stat Soc Series B Stat Methodol*, 70(2): 371–388.

[89] Hogan, J. W. and Laird, N. M. (1997b). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Stat Med*, 16(3): 259–272.

[90] Raghunathan, T., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values. Using a Sequence of Regression Models. *Surv Methodol*, 27: 85–95.

[91] Martikainen, P. (1990). Unemployment and mortality among Finnish men, 1981-5. *BMJ*, 301(6749): 407–411.

[92] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). Statistical Models Based on Counting Processes. New York, USA: Springer-Verlag.

[93] Klein, J. P. and Moeschberger, M. L. (2003). Survival Analysis: Techniques for Censored and Truncated Data. New York, USA: Springer-Verlag.

[94] Beyersmann, J., Schumacher, M., and Allignol, A. (2012). Competing risks and multistate models with R. New York, USA: Springer.

[95] Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4): 541–554.

[96] Aalen, O. O. (1987). Dynamic modelling and causality. *Scand Actuar J*, 1987(3-4): 177–190.

[97] Andersen, P. K. and Keiding, N. (2012a). Interpretability and importance of functionals in competing risks and multistate models. *Stat Med*, 31(11-12): 1074–1088.

[98] Ambrogi, F., Biganzoli, E., and Boracchi, P. (2008). Estimates of clinically useful measures in competing risks survival analysis. *Stat Med*, 27(30): 6407–6425.

[99] Wolbers, M., Koller, M. T., Witteman, J. C., and Steyerberg, E. W. (2009). Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*, 20(4): 555–561.

[100] Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012b). Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*, 41(3): 861–870.

[101] Latouche, A., Allignol, A., Beyersmannc, J., Labopin, M., and P., F. J. (2013). A Competing risk analysis should report results on all cause-specific hazards and cumulative incidence functions. *J Clin Epidemiol*, 66(6): 648–653.

[102] Cox, D. R. (1972). Regression models and life-tables. *J R Stat Soc Series B Methodol*, 34(2): 187–220.

[103] Schoenfeld, D. (1982). Partial Residuals for The Proportional Hazards Regression Model. *Biometrika*, 69(1): 239–241.

[104] Beyersmann, J. and Scheike, T. H. (2014). "Handbook of Survival Analysis". Ed. by J. P. Klein, H. C. Van Houwelingen, J. G. Ibrahim, and T. H. Scheike. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chap. 8 Classical regression models for competing risks.

[105] Whitehead, J. (1980). Fitting Cox's Regression Model to Survival Data using GLIM. *J R Stat Soc Ser C Appl Stat*, 29(3): 268–275.

[106] Gron, R., Gerds, T. A., and Andersen, P. K. (2016). Misspecified poisson regression models for large-scale registry data: inference for 'large n and small p'. *Stat Med*, 35(7): 1117–1129.

[107] Cortese, G. and Andersen, P. K. (2010). Competing Risks ans Time-dependent Covariates. *Biom J*, 51(6): 138–158.

[108] Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The Impact of Heterogeneity in Individual Frailty In the Dynamics of Mortality. *Demography*, 16(3): 439–454.

[109] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1): 141–151.

[110] Klein, J. P. (1992). Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. *Biometrics*, 48(3): 795–806.

[111] Guo, G. and Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *J Am Stat Assoc*, 87(420): 969–976.

[112] Sargent, D. J. (1998). A General Framework for Random Effects Survival Analysis in the Cox Proportional Hazards Setting. *Biometrics*, 54(4): 1486–1497.

[113] Ma, R., Krewski, D., and Burnett, R. T. (2003). Random Effects Cox Models: A Poisson Modelling Approach. *Biometrika*, 90(1): 157–169.

[114] Crowther, M. J., Riley, R. D., Staessen, J. A., Wang, J., Gueyffier, F., and Lambert, P. C. (2012). Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Med Res Methodol*, 12(34): 1–14.

[115] Ben-Shlomo, Y. and Kuh, D. (2002). A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *Int J Epidemiol*, 31(2): 285–293.

[116] Mishra, G. D., Cooper, R., and Kuh, D. (2010). A life course approach to reproductive health: Theory and methods. *Maturitas*, 65: 92–97.

[117] Niedzwiedz, C. L., Katikireddi, S. V., Pell, J. P., and Mitchell, R. (2012). Life course socio-economic position and quality of life in adulthood: a systematic review of life course models. *BMC Public Health*, 12: 628.

[118] Gabadinho, A., Mueller, N. S., Studer, M., and Ritschard, G. (2009). Package 'TraMineR'.

[119] Thiébaut, A. C. M. and Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med*, 23(24): 3803–3820.

[120] Therneau, T. M. (2013). *A Package for Survival Analysis in S*. URL: `http://CRAN.R-project.org/package=survival`.

[121] Raghunathan, T. E., Solenberger, P. W., and Hoewyk, J. V. (2007). *IVEware: Imputation and variance estimation software*. URL: `http://www.isr.umich.edu/src/smp/ive`.

[122] Beebe-Dimmer, J., Lynch, J. W., Turrell, G., Lustgarten, S., Raghunathan, T., and Kaplan, G. A. (2004). Childhood and Adult Socioeconomic Conditions and 31-Year Mortality Risk in Women. *Am J Epidemiol*, 159(5): 481–490.

[123] Claussen, B., Davey, S., and Thelle, D. (2003). Impact of childhood and adulthood socioeconomic position on cause specific mortality: the Oslo Mortality Study. *J Epidemiol Community Health*, 57(1): 40–45.

[124] Davey Smith, G., Hart, C., Blane, D., and Hole, D. (1998). Adverse socioeconomic conditions in childhood and cause specific adult mortality: prospective observational study. *BMJ*, 316(7145): 1631–1635.

[125] Mackenbach, J. P. et al. (1999). Socioeconomic inequalities in mortality among women and among men: an international study. *Am J Public Health*, 89(12): 1800–1806.

[126] Pensola, T. and Martikainen, P. (2003). Cumulative social class and mortality from various causes of adult men. *J Epidemiol Community Health*, 57(9): 745–751.

[127] Cambois, E. (2004b). Careers and mortality in France: evidence on how far occupational mobility predicts differentiated risks. *Soc Sci Med*, 58(12): 2545–2558.

[128] Brown, E. R. and Ibrahim, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics*, 59(3): 686–693.

[129] Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat Med*, 30(12): 1366–1380.

[130] Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J. M., and Lesaffre, E. (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Stat Med*, 33(18): 3167–3178.

[131] Elashoff, R. M., Li, G., and Li, N. (2008). A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, 64(3): 762–771.

[132] Chi, Y.-Y. and Ibrahim, J. G. (2006). Joint Models for Multivariate Longitudinal and Multivariate Survival Data. *Biometrics*, 62(2): 432–445.

[133] Liu, L. and Huang, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1): 65–81.

[134] Yu, M., Taylor, J. M. G., and Sandler, H. M. (2008). Individual Prediction in Prostate Cancer Studies Using a Joint Longitudinal Survival-Cure Model. *J Am Stat Assoc*, 103(481): 178–187.

[135] Song, H., Peng, Y., and Tu, D. (2012). A new approach for joint modelling of longitudinal measurements and survival times with a cure fraction. *Can J Stat*, 40(2): 207–224.

[136] Law, N. J., Taylor, J. M. G., and Sandler, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, 3(4): 547–563.

[137] Chen, M.-H., Ibrahim, J. G., and Sinha, D. (2004). A new joint model for longitudinal and survival data with a cure fraction. *J Multivar Anal*. Special Issue on Semiparametric and Nonparametric Mixed Models, 91(1): 18–34.

[138] Chen, Q., May, R. C., Ibrahim, J. G., Chu, H., and Cole, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Stat Med*, 33(26): 4560–4576.

[139] Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sinica*, 14: 809–834.

[140] Diggle, P. J., Sousa, I., and Chetwynd, A. G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture. *Stat Med*, 27(16): 2981–2998.

[141] Wu, L., Liu, W., Yi, G. Y., Huang, Y., Wu, L., Liu, W., Yi, G. Y., and Huang, Y. (2012). Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues, Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues. *Journal of Probability and Statistics*, 2012: 17 pages.

[142] Sousa, I. (2011). A Review on Joint Modelling of Longitudinal Measurements and Time-to-event. *Revstat Stat J*, 9(1): 57–81.

[143] Proust-Lima, C., Séne, M., Taylor, J. M. G., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: a review. *Stat Methods Med Res*, 23(1): 74–90.

[144] Tsiatis, A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2): 447–458.

[145] Wulfsohn, M. S. and Tsiatis, A. A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, 53(1): 330–339.

[146] Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited. *Biometrics*, 62(4): 1037–1043.

[147] Davison, A. C. (2008). Statistical Models. Cambridge University Press.

[148] Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Stat Med*, 28(6): 956–971.

[149] Dagum, L. and Menon, R. (1998). "OpenMP": an industry standard API for shared-memory programming. *IEEE Comput Sci Eng*, 5(1): 46–55.

[150] Spiessens, B., Verbeke, G., and Komárek, A. (2002). *A SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalised linear mixed models. Technical Report, Biostatistical Center, Catholic Univ. Leuven, Leuven.*

[151] Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *J Interdiscipl Hist*, 16(3): 471–494.

[152] Abbott, A. and Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musician's carrers. *Am J Soc*, 96(1): 144–185.

[153] Studer, M. and Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *J R Stat Soc Ser A Stat Soc*, 179(2): 481–511.

[154] Abbott, A. and Tsay, A. (2000). Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociol Methods Res*, 29(1): 3–33.

[155] Lesnard, L. (2010). Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. *Sociol Methods Res*, 38(3): 389–419.

[156] Hollister, M. (2009). Is Optimal Matching Suboptimal? *Sociol Methods Res*, 38(2): 235–264.

[157] Halpin, B. (2010). Optimal Matching Analysis and Life-Course Data: The Importance of Duration. *Sociol Methods Res*, 38(3): 365–388.

[158] Biemann, T. (2011). A transition-oriented approach to optimal matching. *Sociol methodol*, 41(1): 195–221.

[159] Gabadinho, A. and Ritschard, G. (2013). "Searching for typical life trajectories applied to childbirth histories". *Gendered life course - Between individualization and standardization*. Ed. by R. Levy and E. Widmer. Vienna: LIT: A European approach applied to Switzerland: pp. 287–312.

[160] Kaufman, L. and Rousseeuw, P. J. (2008). "Partitioning Around Medoids (Program PAM)". *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

[161] Houwelingen, H. C. van (2007). Dynamic Prediction by Landmarking in Event History Analysis. *Scand Journal Stat*, 34(1): 70–85.

[162] Willekens, F. (2014). Multistate Analysis of Life Histories with R. Springer.

[163] Karimi, M. (2012). "Trajectoire socio-professionnelle: Description et association avec la mortalité". MA thesis. Paris, France: University of Paris VI (UPMC).

[164] Eerola, M. and Helske, S. (2012). Statistical analysis of life history calendar data. *Stat Methods Med Res*: 1–27.

[165] Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5): 550–560.

[166] Moodie, E. E. M., Stephens, D. A., and Klein, M. B. (2014). A marginal structural model for multiple-outcome survival data: assessing the impact of

injection drug use on several causes of death in the Canadian Co-infection Cohort. *Stat Med*, 33(8): 1409–1425.

[167] Alpérovitch, A., Bertrand, M., Jougla, E., Vidal, J.-S., Ducimtrère, P., Helmer, C., Ritchie, K., Pavillon, G., and Tzourio, C. (2009). De we really know the cause of death of the very old? Comparison between official mortality statistics and cohort study classification. *Eur J Epidemiol*, 70(11): 669–675.

[168] Moreno-Betancur, M., Sadaoui, H., Piffaretti, C., and Rey, G. (2016). Survival analysis with multiple causes of death: Extending the competing risks model. *Epidemiology*. In press.

Part IV

Appendices

Appendix A

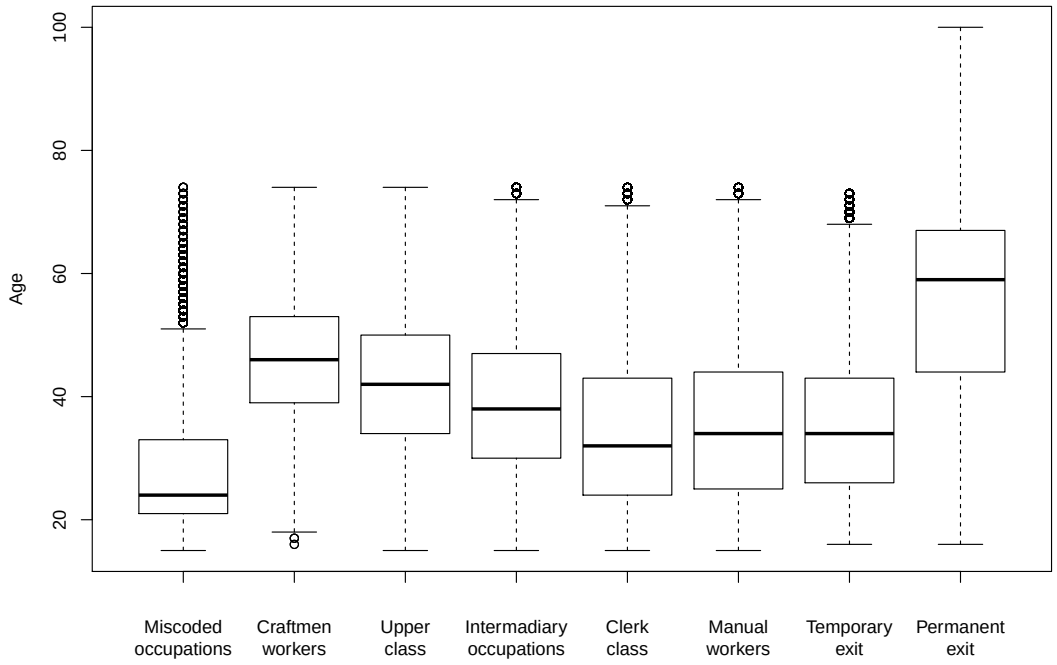
Descriptive Statistics of the Cosmop-DADS database

Table A.1 – Description of Cosmop-DADS

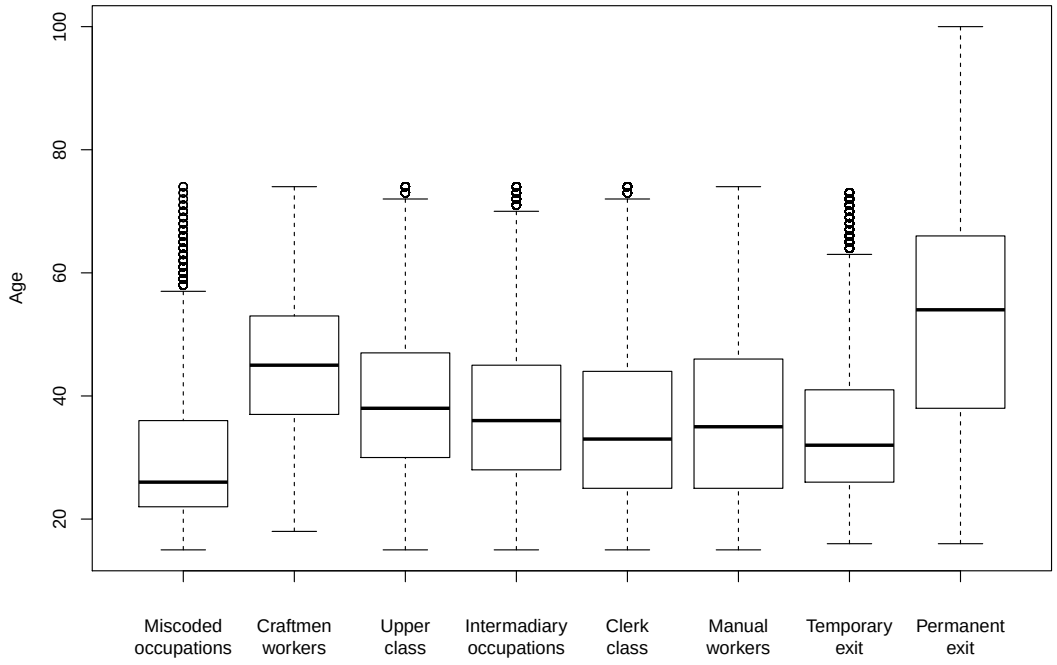
	Men (%) (n = **957 299**)	**Women (%)** (n = **798 291**)
Born in the French territory	806 513 (84.23)	704 943 (88.31)
Temporary exit	556 830 (69.04)	466 385 (66.16)
Definive exit	335 858 (41.64)	308 464 (43.76)
Death	89 639 (11.11)	29 218 (4.14)
Born outside the French territory	150 786 (15.75)	93 348 (11.69)
Temporary exit	92 497 (61.34)	53 635 (57.46)
Permanent exit	94 042 (62.37)	51 765 (55.45)
Death	9557 (6.34)	2401 (2.57)

Table A.2 – Description of missing professional episodes in Cosmop-DADS database

	Men (%) (n =16 892 644)	**Women (%)** (n =13 425 495)
Miscoded occupations	97 020 (0.57)	103 459 (0.77)
At home job	40 (0.04)	66 (0.06)
No Full-time job	699 (0.72)	483 (0.47)
Sporadic job	5585 (5.76)	6085 (5.88)
Part-time job	26 449 (27.26)	42 646 (41.22)
Full-time job	64 247 (66.22)	54 179 (52.37)
Craftsmen and trade-related workers	172 890 (1.02)	43 835 (0.33)
Temporary exit	2 773 523 (16.42)	2 521 171 (18.78)
Year 1981	396 681 (14.30)	268 374 (10.65)
Year 1983	401 304 (14.47)	283 772 (11.26)
Year 1990	448 501 (16.17)	358 769 (14.21)
Other	1 527 037 (55.06)	1 610 256 (63.88)
Permanent exit	4 304 526 (25.48)	3 608 861 (26.88)
Regional and local authorities	322 518 (1.91)	451 364 (3.36)
Miscoded occupation	20 449 (6.34)	22 391 (4.96)
Craftsmen and trade-related workers	72 (0.02)	23 (0.01)
Upper class	31 881 (9.89)	24 804 (5.50)
Intermediary occupations	88 858 (27.55)	107 449 (23.80)
Clerk class	140 205 (43.47)	285 657 (63.29)
Manual workers class	41 053 (12.73)	11 040 (2.44)



(a) Men



(b) Women

Figure A.1 – Distibution of age by observed professional situation

Appendix B

Supplementary results for the study of Chapter 3

Table B.1 – Cancer mortality hazard ratios among men according to socio-professional trajectories

	Cancer (n=3116)	**Lung (n=848)**	**UADT (n=472)**	**Other (n=1796)**
	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up				
Upper class	1	1	1	1
Intermediary occupations	0.98 [0.77, 1.24]	0.88 [0.58, 1.36]	3.13 [1.06, 9.23]*	0.89 [0.66, 1.19]
Clerk class	1.02 [0.81, 1.29]	0.98 [0.64, 1.50]	3.34 [1.14, 9.77]*	0.91 [0.68, 1.22]
Manual workers class	1.10 [0.88, 1.37]	0.93 [0.62, 1.39]	3.84 [1.32,11.15]*	1.01 [0.76, 1.33]
Current occupational class[a]				
Upper class	1	1	1	1
Intermediary occupations	1.10 [0.88, 1.37]	1.27 [0.84, 1.93]	2.79 [0.93, 8.33]	0.97 [0.74, 1.28]
Clerk class	1.50 [1.16, 1.93]**	1.71 [1.02, 2.85]*	5.38 [1.82,15.91]**	1.28 [0.93, 1.75]
Manual workers class	1.26 [1.02, 1.56]*	1.69 [1.11, 2.57]*	3.92 [1.40,11.00]**	1.06 [0.81, 1.39]
Outside the scope	2.21 [1.81, 2.71]***	2.06 [1.39, 3.06]***	1.26 [4.60,34.49]***	1.82 [1.42, 2.34]
Cumulative time spent in occupational class				
Upper class	1	1	1	1
Intermediary occupations	1.20 [0.98, 1.46]	1.20 [0.84, 1.72]	0.52 [0.27, 0.98]*	1.39 [1.07, 1.80]
Clerk class	1.53 [1.23, 1.89]***	1.12 [0.74, 1.71]	1.64 [0.98, 2.73]	1.73 [1.31, 2.30]
Manual workers class	1.75 [1.48, 2.06]***	1.56 [1.13, 2.15]**	1.93 [1.27, 2.93]**	1.76 [1.42, 2.19]
Outside the scope	1.33 [1.12, 1.57]***	1.26 [0.91, 1.74]	1.19 [0.78, 1.83]	1.41 [1.13, 1.76]
Social mobility indicator[b]				
Low (= 0)	1	1	1	1
Medium	0.96 [0.87, 1.06]	0.95 [0.79, 1.14]	1.08 [0.85, 1.38]	0.95 [0.83, 1.08]
High (> 1.11)	1.07 [0.97, 1.18]	1.00 [0.83, 1.21]	1.18 [0.90, 1.54]	1.10 [0.97, 1.25]

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Table B.2 – Cancer mortality hazard ratios among women according to socio-professional trajectories

	Cancer (n=1388)	**Lung (n=133)**	**UADT (n=39)**	**Breast (n=447)**	**Other (n=769)**
	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.02 [0.71, 1.47]	1.71 [0.58, 5.42]	-	0.63 [0.35, 1.13]	1.17 [0.70, 1.97]
Clerk class	1.07 [0.76, 1.52]	1.27 [0.40, 4.00]	-	0.83 [0.48, 1.43]	1.18 [0.72, 1.96]
Manual workers class	1.35 [0.94, 1.94]	2.38 [0.76, 7.46]	-	1.00 [0.57, 1.77]	1.43 [0.85, 2.41]
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	0.77 [0.55, 1.09]	1.07 [0.35, 3.21]	-	0.77 [0.44, 1.36]	0.73 [0.45, 1.18]
Clerk class	0.75 [0.54, 1.04]	1.07 [0.37, 3.12]	-	0.72 [0.42, 1.25]	0.69 [0.44, 1.10]
Manual workers class	0.71 [0.49, 1.04]	1.28 [0.40, 4.12]	-	0.61 [0.31, 1.18]	0.74 [0.44, 1.24]
Outside the scope	1.20 [0.86, 1.66]	1.51 [0.54, 4.23]	-	1.06 [0.62, 1.82]	1.21 [0.76, 1.91]
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	1.12 [0.82, 1.53]	0.86 [0.35, 2.16]	-	1.21 [0.71, 2.06]	1.15 [0.74, 1.80]
Clerk class	1.16 [0.88, 1.54]	0.80 [0.36, 1.78]	-	1.34 [0.83, 2.15]	1.18 [0.79, 1.76]
Manual workers class	1.10 [0.81, 1.49]	0.54 [0.24, 1.24]	-	1.18 [0.69, 2.03]	1.17 [0.76, 1.79]
Outside the scope	1.08 [0.82, 1.43]	0.73 [0.34, 1.60]	-	1.25 [0.78, 2.00]	1.05 [0.70, 1.57]
Social mobility indicator[b]					
Low (= 0)	1	1	1	1	1
Medium	0.85 [0.73, 0.99]*	0.68 [0.42, 1.11]	-	0.95 [0.73, 1.24]	0.81 [0.65, 1.01]
High (> 0.91)	1.04 [0.91, 1.18]	1.07 [0.71, 1.63]	-	1.14 [0.92, 1.43]	0.96 [0.80, 1.13]

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Table B.3 – All-cause and cause-specific mortality hazard ratios among men according to socio-professional trajectories considering an order between occupational categories

	All-cause (n=12 162)	**Cardiovascular (n=1452)**	**Cancer (n=3116)**	**External causes (n=4026)**	**Other causes (n=3568)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.01 [0.81, 1.26]	0.96 [0.40, 2.32]	1.04 [0.72, 1.49]	0.91 [0.60, 1.38]	1.46 [1.15, 1.85]**
Clerk class	1.16 [0.94, 1.44]	0.95 [0.40, 2.22]	1.17 [0.82, 1.66]	0.95 [0.63, 1.44]	1.83 [1.45, 2.31]***
Manual workers class	1.38 [1.09, 1.74]**	1.36 [0.55, 3.33]	1.58 [1.05, 2.23]*	0.98 [0.63, 1.53]	1.88 [1.49, 2.38]***
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.02 [0.80, 1.30]	2.11 [0.74, 6.07]	0.77 [0.54, 1.08]	1.37 [0.83, 2.25]	1.05 [0.84, 1.30]
Clerk class	0.91 [0.72, 1.15]	1.28 [0.44, 3.77]	0.71 [0.51, 1.00]**	1.48 [0.90, 2.43]	1.48 [1.18, 1.87]***
Manual workers class	0.96 [0.73, 1.25]	2.04 [0.65, 6.39]	0.65 [0.44, 0.97]*	1.50 [0.86, 2.59]	0.95 [0.76, 1.17]
Outside the scope	1.64 [1.30, 2.07]***	2.12 [0.74, 6.06]	1.14 [0.82, 1.61]	2.04 [1.25, 3.32]**	2.94 [2.42, 3.59]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.97 [0.77, 1.21]	1.57 [0.61, 4.04]	1.11 [0.81, 1.52]	0.86 [0.54, 1.35]	0.81 [0.64, 1.02]
Clerk class	1.10 [0.90, 1.33]	2.58 [1.11, 6.00]*	1.13 [0.86, 1.50]	0.79 [0.52, 1.19]	1.54 [1.24, 1.91]***
Manual workers class	1.07 [0.87, 1.31]	1.94 [0.82, 4.62]	1.07 [0.78, 1.45]	1.06 [0.70, 1.62]	1.44 [1.21, 1.73]***
Outside the scope	1.19 [0.98, 1.43]	3.10 [1.38, 6.93]**	1.05 [0.79, 1.39]	1.02 [0.69, 1.51]	1.34 [1.12, 1.61]**
Social mobility indicator (positive)[b]					
Low (= 0)	1	1	1	1	1
Medium	0.86 [0.77, 0.95]**	0.82 [0.58, 1.15]	0.83 [0.71, 0.97]*	1.08 [0.86, 1.36]	0.91 [0.82, 1.01]
High (> 0.59)	0.86 [0.76, 0.98]*	0.78 [0.50, 1.21]	0.89 [0.73, 1.09]	0.90 [0.70, 1.17]	0.89 [0.79, 1.01]
Social mobility indicator (negative)[b]					
Low (= 0)	1	1	1	1	1
High ($\neq$ 0)	1.32 [1.20, 1.45]***	1.47 [1.06, 2.04]*	1.21 [1.04, 1.40]*	1.09 [0.89, 1.35]	1.41 [1.28, 1.56]***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)
a: observed with two-year time lag
b: transition rates between occupational classes (10 years of follow-up)
c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, positive and negative social mobility indicator and observation periods
†: age as the time-scale in Cox proportional hazards model

Table B.4 – All-cause and cause-specific mortality hazard ratios among women according to socio-professional trajectories considering an order between occupational categories

	All-cause (n=3551)	**Cardiovascular (n=304)**	**Cancer (n=1388)**	**External causes (n=894)**	**Other causes (n=965)**
	$HR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]	$CSHR^{c}_{\dagger}$ [95% CI]
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.01 [0.81, 1.26]	0.96 [0.40, 2.32]	1.04 [0.72, 1.49]	0.91 [0.60, 1.38]	1.04 [0.71, 1.60]
Clerk class	1.16 [0.94, 1.44]	0.95 [0.40, 2.22]	1.17 [0.82, 1.66]	0.95 [0.63, 1.44]	1.34 [0.90, 2.00]
Manual workers class	1.38 [1.09, 1.74]**	1.36 [0.55, 3.33]	1.58 [1.05, 2.23]*	0.98 [0.63, 1.53]	1.44 [0.94, 2.21]
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.02 [0.80, 1.30]	2.11 [0.74, 6.07]	0.77 [0.54, 1.08]	1.37 [0.83, 2.25]	1.24 [0.71, 2.15]
Clerk class	0.91 [0.72, 1.15]	1.28 [0.44, 3.77]	0.71 [0.51, 1.00]**	1.48 [0.90, 2.43]	0.97 [0.57, 1.65]
Manual workers class	0.96 [0.73, 1.25]	2.04 [0.65, 6.39]	0.65 [0.44, 0.97]*	1.50 [0.86, 2.59]	1.15 [0.65, 2.04]
Outside the scope	1.64 [1.30, 2.07]***	2.12 [0.74, 6.06]	1.14 [0.82, 1.61]	2.04 [1.25, 3.32]**	2.72 1.61, 4.58]***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.97 [0.77, 1.21]	1.57 [0.61, 4.04]	1.11 [0.81, 1.52]	0.86 [0.54, 1.35]	0.65 [0.38, 1.11]
Clerk class	1.10 [0.90, 1.33]	2.58 [1.11, 6.00]*	1.13 [0.86, 1.50]	0.79 [0.52, 1.19]	1.09 [0.73, 1.62]
Manual workers class	1.07 [0.87, 1.31]	1.94 [0.82, 4.62]	1.07 [0.78, 1.45]	1.06 [0.70, 1.62]	0.97 [0.63, 1.49]
Outside the scope	1.19 [0.98, 1.43]	3.10 [1.38, 6.93]**	1.05 [0.79, 1.39]	1.02 [0.69, 1.51]	1.31 [0.89, 1.93]
Social mobility indicator (positive)[b]					
Low (= 0)	1	1	1	1	1
Medium	0.86 [0.77, 0.95]**	0.82 [0.58, 1.15]	0.83 [0.71, 0.97]*	1.08 [0.86, 1.36]	0.82 [0.67, 1.00]
High (> 0.59)	0.86 [0.76, 0.98]*	0.78 [0.50, 1.21]	0.89 [0.73, 1.09]	0.90 [0.70, 1.17]	0.87 [0.67, 1.13]
Social mobility indicator (negative)[b]					
Low (= 0)	1	1	1	1	1
High ($\neq$ 0)	1.32 [1.20, 1.45]***	1.47 [1.06, 2.04]*	1.21 [1.04, 1.40]*	1.09 [0.89, 1.35]	1.67 [1.39, 2.01]***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)
a: observed with two-year time lag
b: transition rates between occupational classes (10 years of follow-up)
c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, positive and negative social mobility indicator and observation periods
†: age as the time-scale in Cox proportional hazards model

Appendix C

Supplementary details for the M-step of Chapter 4

$$\alpha_k^{(m+1)} = \alpha_k^{(m)} - \frac{S_{\alpha_k}^{(m)}}{I_{\alpha_k}^{(m)}} \tag{C.1}$$

for $k = 1, \cdots, K-1$, with $S_{\alpha_k}^{(m)}$ and $I_{\alpha_k}^{(m)}$ being

$$\begin{aligned} S_{\alpha_k} = \frac{\partial l(\Psi|Y, \widetilde{T}, a)}{\partial \alpha_k} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{\partial}{\partial \alpha_k}\Big[\sum_{k=1}^{K} I(Y_{ij}=k) log(\pi_{ijk})\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{h=1}^{K} I(Y_{ij}=h) E\{I(h=k) - \pi_{ijk}\} \end{aligned} \tag{C.2}$$

and

$$I_{\alpha_k}^{(m)} = \frac{\partial S_{\alpha_k}}{\partial \alpha_k} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{h=1}^{K} I(Y_{ij}=h) E\{\pi_{ijk}^2 - \pi_{ijk}\} \tag{C.3}$$

$$\beta_k^{(m+1)} = \beta_k^{(m)} - \frac{S_{\beta_k}^{(m)}}{I_{\beta_k}^{(m)}} \tag{C.4}$$

for $k = 1, \cdots, K-1$, with $S_{\beta_k}^{(m)}$ and $I_{\beta_k}^{(m)}$ being

$$\begin{aligned} S_{\beta_k} = \frac{\partial l(\Psi|Y, \widetilde{T}, a)}{\partial \beta_k} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{\partial}{\partial \beta_k}\Big[\sum_{k=1}^{K} I(Y_{ij}=k) log(\pi_{ijk})\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{h=1}^{K} \Big[I(Y_{ij}=h) E\{I(h=k) - \pi_{ijk}\} X_{ij}\Big] \end{aligned} \tag{C.5}$$

and

$$I_{\beta_k}^{(m)} = \frac{\partial S_{\beta_k}}{\partial \beta_k} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{h=1}^{K}\Big[I(Y_{ij}=h)E\{\pi_{ijk}^2 - \pi_{ijk}\}X_{ij}X_{ij}^{'}\Big] \tag{C.6}$$

$$\gamma_d^{(m+1)} = \gamma_d^{(m)} + \frac{S_{\gamma_d}^{(m)}}{I_{\gamma_d}^{(m)}} \tag{C.7}$$

for $d = 1, \cdots, g$, with $S_{\gamma_d}^{(m)}$ and $I_{\gamma_d}^{(m)}$ being

$$S_{\gamma_d} = \frac{\partial l(\Psi|Y,\widetilde{T},a)}{\partial \gamma_d} = \sum_{i=1}^{n}\Big\{I(\epsilon_i = d)Z_i - \sum_{t_{dj}\le T_i}\lambda_{0d}^{(m+1)}\big(t_{dj}\big)\exp\big(Z_i'\gamma_d^{(m)}\big)\times E\Big(\exp\big(\nu_d^{(m)}u_i\big)\Big)Z_i\Big\} \tag{C.8}$$

and

$$I_{\gamma_d} = \frac{\partial S_{\gamma_d}}{\partial \gamma_d} = \sum_{i=1}^{n}\sum_{t_{dj}\le T_i}\lambda_{0d}^{(m+1)}\big(t_{dj}\big)\exp\Big(Z_i'\gamma_d^{(m)}\Big)\times E\Big(\exp\big(\nu_d^{(m)}u_i\big)\Big)Z_iZ_i' \tag{C.9}$$

$$\nu_d^{(m+1)} = \nu_d^{(m)} + \frac{S_{\nu_d}^{(m)}}{I_{\nu_d}^{(m)}} \tag{C.10}$$

for $d = 1, \cdots, g$, with $S_{\nu_d}^{(m)}$ and $I_{\nu_d}^{(m)}$ being

$$S_{\nu_d} = \frac{\partial l(\Psi|Y,\widetilde{T},a)}{\partial \nu_d} = \sum_{i=1}^{n}\Big\{I(\epsilon_i = d)E(u_i) - \sum_{t_{dj}\le T_i}\lambda_{0d}^{(m+1)}\big(t_{dj}\big)\exp\big(Z_i'\gamma_d^{(m)}\big)\times E\Big(u_i\exp\big(\nu_d^{(m)}u_i\big)\Big)\Big\} \tag{C.11}$$

and

$$I_{\nu_d} = \frac{\partial S_{\nu_d}}{\partial \nu_d} = \sum_{i=1}^{n}\sum_{t_{dj}\le T_i}\lambda_{0d}^{(m+1)}\big(t_{dj}\big)\exp\Big(Z_i'\gamma_d^{(m)}\Big)\times E\Big(u_i^2\exp\big(\nu_d^{(m)}u_i\big)\Big) \tag{C.12}$$

Appendix D

Socioprofessional trajectories and mortality in France, 1976–2002: a longitudinal follow-up of administrative data

Title Page

Title: Socio-professional trajectories and mortality in France, 1976-2002: a longitudinal follow-up of administrative data

Corresponding author:

Maryam Karimi

Inserm-CépiDc
80, Rue du Général Leclerc
94270 Le Kremlin-Bicêtre Cedex
France

Tel.: (+33) 1.49.59.53.34

Fax: (+33) 1.49.59.19.30

E-mail: maryam.karimi@inserm.fr

Co-authors:

1. Béatrice Geoffroy-Perez, Institut de Veille Sanitaire, Département Santé-Travail, Saint-Maurice, France
2. Aurélie Fouquet, Institut de Veille Sanitaire, Département Santé-Travail, Saint-Maurice, France
3. Aurélien Latouche, Conservatoire national des arts et métiers, Paris, France
4. Grégoire Rey, Inserm, CépiDc, Le Kremlin-Bicêtre, France

Keywords: Mortality, Social inequalities, Trajectory, Social class, Life-course epidemiology

Word count: 3467

ABSTRACT

Background

Occupying a low socioeconomic position is associated with increased mortality risk. To disentangle this association, previous studies considered various dimensions of socioeconomic trajectories across the life-course. However, they used a limited number of stages. Here, we simultaneously examined various dimensions of the whole professional trajectory and its association with mortality.

Methods

We used a large sample (337 706 men, 275 378 women) of the data obtained by linking individuals' annual occupation (collected in 1976-2002 from a representative panel of the French salaried population in the semi-public and private sectors) with causes of death obtained from registries. All-cause and cause-specific hazard ratios were estimated using Cox's regression models adjusted for the occupational class at the beginning of the follow-up, the current occupational class, the transition rates between occupational categories, and the duration of time spent in occupational categories.

Results

An increase in the time spent in the clerk class increased men and women's cardiovascular mortality risk compared to that in the upper class (HRs: 1.59(1.14-2.20) and 2.65(1.14-6.13) for 10 years increase, respectively for men and women). Men with a high rate of transitions had about a 1.2-fold increased risk of all-cause and external-cause mortality compared to those without transitions during their professional life. This association was also observed for women's all-cause mortality.

Conclusion

Strong associations between professional trajectories and mortality from different causes of death were found. Long exposure to lower socioeconomic conditions was associated with increased

mortality risk from various causes of death. The results also suggest gradual associations between transition rates and mortality.

What is already known on this subject?

- Previous studies reported strong associations between socio-economic trajectories and mortality.
- Most of these studies have used two or three stages of life to show these associations across life-course models.

What this study adds?

- We consider all stages of professional trajectory to investigate these relationships in a representative sample of the French salaried population of semi-public and private sectors.
- Long-time exposure to poor socioeconomic position was strongly associated with adult mortality, especially for cardiovascular diseases.
- Having more transitions during professional life was adversely associated with mortality.

INTRODUCTION

Socioeconomic inequalities in mortality, as quantified by mortality differentials between social groups, have been studied in many industrialized countries.[1-3] Despite the low level in mortality and its continuous decrease, studies conducted in the UK, US and Europe have shown that these inequalities are still large in some countries[4-6] and have increased over time in both men and women.[1,7-11]

A large body of research has shown that mortality rates are higher among those in lower socioeconomic positions;[12,13] regardless of the socioeconomic indicator (occupational status, educational level or income).[14] Most of these studies have measured socioeconomic positions only at one stage of life. This approach does not consider the impact of transitions between different socioeconomic groups. Thus, to obtain better understanding of the relationship between health and socioeconomic position, various dimensions of socioeconomic trajectories, such as the evolution of socioeconomic position and the modality of transitions between social groups, need to be taken into account.[15,16]

Although an observed individual's social level at a given time reflects his/her social position at different stages of his/her past life,[17] several life-course models have been proposed in the literature to explain the possible association between socioeconomic status and health: critical period, accumulation, and social mobility models. The critical period model considers some stages or specific moments in life as key periods affecting health. The cumulative model hypothesises that mortality differentials are explained by the accumulation of all present and past working conditions, lifestyles and behaviours. Analyses using this model are based on the life-cumulated length of stay in the most disadvantaged social group. They suggest that the accumulation of poor socioeconomic exposure in life increases the risk of mortality.[15,18-20] The social mobility model was developed to take into account the modality of transitions between social groups once or several times in life. These models help to explain the potential impact of socioeconomic status on health. However,

some studies point to a bias in the results due to the impact of health on socioeconomic position. Therefore, this reverse causation is another issue that should be taken into account.[21,22]

The aim of this study was to examine the associations between life-course professional trajectory and adult mortality. Previous studies have considered two or three stages in professional life and used a simple classification for socioeconomic positions (low-medium-high). This paper goes further by considering the whole professional trajectories and all-cause and cause-specific mortality. For this purpose, we use life-course models on a representative sample of the French salaried population in the semi-public and private sectors from 1976 to 2002 to investigate the possible ways in which professional trajectories may be associated with adult mortality.

METHODS

Cosmop-DADS database

The Cosmop-DADS database was obtained by linking the occupational life-course provided from the panel of the Annual Declarations of Social Data (DADS)[23] that has been regularly updated by the French National Institute for Statistics and Economic Studies (INSEE) since 1976, with the causes of death recorded by the French National Death Registry (INSERM-CépiDc). The DADS Panel contains the employment records of approximately $1/24^{th}$ of all employees in the private and semi-public sectors, i.e. 80% of all paid occupations in France. Episodes of careers declared as self-employed, employees of the state, employees in agriculture, domestic services, extra-territorial activities, interns and apprentices are excluded from its scope. A deterministic record linkage using the following identifiers linked these two data sets: sex, date of birth, date of death and the commune of residence at the time of death. The matching rate was 98%. In total, the Cosmop-DADS population is a sample of the French population (for whom vital status and date of death are available), employed at least once as a salaried worker in the semi-public and private sectors between 1976 and 2002.

The study was approved by the French data protection committee and institutional ethical review board: Commission Nationale de l'Informatiqueet des Libertés (CNIL) (authorisation n° 904210v1).

Occupational classes

Occupations were coded according to the French classification created by INSEE regarding various social characteristics, without any specific hierarchical order between the defined classes.[24] Originally, the DADS covers five classes: **Craftsmen and trade-related workers**; **upper class**; **intermediary occupations**; **clerk class** and **manual workers class**.

Since DADS declarations are mandatory for employers, there were theoretically no missing occupational episodes for employees working in companies within the DADS scope. However, professional trajectories were not fully observed for several individuals. The first set of missing episodes concerned the years 1981, 1983 and 1990, which were not collected owing to administrative reasons. These episodes were complemented with information from the previous years. For the other years, some occupations could not be classified in the five occupational classes. These occupations were imputed using a multivariable multinomial logistic regression[25] incorporating sex, age and type of employment in the imputation model.

Since regional and local authorities were not fully covered by DADS declarations before 1987, any occupations of this type were excluded from our professional scope. The same decision was taken for occupations declared in the craftsmen and trade-related workers class, as those in DADS are not representative of this class in the general population.

In summary, our professional scope contained the DADS scope mentioned previously excluding regional and local authorities, and craftsmen and trade related workers class. Those outside this scope could either be working, inactive or retired. It was not possible to distinguish these different situations. As it is well established that inactivity is associated with an increased mortality risk,[26,27] any episodes outside this scope should not be ignored so the category "outside the scope" was

added to the four other categories. This strategy induced a bias but given the structure of the data, building a sound imputation model would require additional assumptions for which no auxiliary data, such as data on employees of the public sector, were available.

Study population

All individuals born in the French territories for whom a salaried period was declared in Cosmop-DADS between ages 25 and 30, excluding those working outside the study scope in their first year (337 706 men and 275 378 women) were included in the study (Due to the uncertainty of the vital status of people born outside France, they were excluded from our study population). Less than 1% of occupations were imputed, and in total, 22% and 30% of follow-up years were outside the study scope for men and women, respectively. 52% of men and 61% of women were outside the study scope for at least one year of their follow-up. Owing to the non-negligible number of episodes outside the study scope and the lack of available information for making more hypotheses about these episodes, a replicated analysis was carried out on a subsample of the analysed population for whom the first five years of their follow-up was covered by the study scope in order to ensure that an observed trajectory was complete (in the first five years) for the analysis (198 381 males and 134 784 females, with fewer than 14% of follow-up years outside the study scope in total).

Professional trajectory

A professional trajectory may be defined as the sequence of consecutive professional positions occupied by an individual (Figure 1). To characterise it, three time-dependent variables were used:

- Occupational class at each year;
- Cumulative social class indicator, defined as individual's length of stay in each occupational class. This indicator was calculated for all classes except the upper class, so the latter served as reference;

- 10-year social mobility indicator, defined by the transition rates between classes, excluding the "outside the scope" category and calculated as follows:

$$\frac{\text{number of transitions between occupational classes}}{\text{duration of follow-up}} \times 10$$

This indicator was categorised into three groups using tertiles, separately for men and women.

To limit the impact of reverse causation,[21,22] occupational classes were considered with a two-year time lag, i.e. instead of using the current occupational class, that of two years before, was taken into account.

(Figure here)

Mortality

The Cosmop-DADS database is a sample of the French population for whom the vital status and date of death are available. All individuals of this sample were followed up to 2002 and the administrative censoring date was set at 31st December 2002. The underlying causes of death, recorded by INSERM-CépiDc, were coded according to the International Classification of Diseases, 8th, 9th and 10th revisions (ICD-8, ICD-9 and ICD-10). Three broad categories of causes were specifically considered: cardiovascular diseases, cancer and external causes (See Appendix I).

Statistical analysis

Cox proportional hazards models were used to estimate all-cause hazard ratios (HRs), cause-specific hazard ratios [CSHs] and their 95% confidence intervals [CIs] while accounting for left truncation induced by the delayed entries. Age was used as the time-scale.[28] The model for each cause was fitted using a Cox model by censoring the participants who failed from competing cause.[29]

Adjustment for the variables, occupational class at the beginning of the follow-up as a baseline covariate and the three indicators of professional trajectory as time-dependent covariates, was done

by performing univariable analysis in the first step and then using all these covariates in a multivariable analysis. Considering the decrease in mortality rates over time in France, the models were adjusted for observation periods.

The occupational class and the social mobility indicator were introduced into the models as categorical variables, and the upper class and those without any mobility between classes were considered as the reference categories. For the cumulative social class indicator, HRs were interpreted as the hazard corresponding to an increase in the time spent in an occupational class versus that in the upper class. These HRs were calculated for a 10-year increase. No violation of the proportional hazards assumptions was found according to Schoenfeld residuals.

Proportional hazards models were conducted separately for men and women using the Survival package of the R software,[30] and the imputation was carried out by the IVEware software.[31]

RESULTS

The average number of transitions between occupational classes differed between the age categories. Transitions were more numerous between the ages of 25 and 44 in women and between the ages of 25 and 34 in men. At the beginning of the follow-up, the largest class was the clerk class (about 54%) in women and manual workers (about 60%) in men. For young men (25-34 years), 49.3% of the cumulated time spent was in the manual workers class and much less in the upper class (6.5%). The same magnitude was observed in young women for the clerck and the upper class (25-34 years) (Table 1).

During the follow-up, 12 162 (3.6%) men and 3551 (1.3%) women died. Most deaths occurred between the ages of 35 and 44. 48.7% of deaths among women and 39.8% of deaths among men occurred while individuals were outside the study scope two years before death. Most other deaths in men and women occurred while they were in the manual workers class and the clerk class, respectively (Table2).

Table 1 Characteristics of study population according to occupational trajectories

		Average number of transitions/10 years follow-up	Proportion of time spent in occupational classes					
			Upper class	Intermediary occupations	Clerk class	Manual workers class	Outside the scope	Total
Men	At the beginning	0	5.5	17.3	17.7	59.5	0	100
	25-34	1.0	6.5	17.0	12.4	49.3	14.8	100
	35-44	0.9	9.6	17.8	7.4	38.4	26.8	100
	45-54	0.6	12.9	17.9	5.8	31.8	31.6	100
	55-56	0.6	15.5	18.5	5.2	28.4	32.4	100
	All ages	0.9	8.8	17.4	9.4	42.1	22.3	100
Women	At the beginning	0	4.2	19.4	53.5	22.9	0	100
	25-34	0.8	4.3	17.0	41.0	16.0	21.7	100
	35-44	0.8	4.5	15.8	30.9	12.3	36.5	100
	45-54	0.6	5.5	16.6	28.1	11.4	38.4	100
	55-56	0.6	7.0	17.8	25.4	9.7	40.1	100
	All ages	0.8	4.6	16.5	35.0	13.8	30.1	100

Table 2 Distribution of study population according to occupational trajectories

		Number of deaths (%)		Person-year (%)	
		Men	**Women**	**Men**	**Women**
Observation period	**Beginning of follow-up**				
	Upper class	344 (2.8)	104 (3.0)		
	Intermediary occupations	1371 (11.3)	568 (16.0)		
	Clerk class	2042 (16.8)	1699 (47.8)		
	Manual workers class	8405 (69.1)	1180 (33.2)		
	Outside the scope	0 (0)	0 (0)		
	End of follow-up				
	Upper class	525 (4.3)	118 (3.3)		
	Intermediary occupations	1299 (10.7)	417 (11.7)		
	Clerk class	941 (7.7)	868 (24.5)		
	Manual workers class	4558 (37.5)	419 (11.8)		
	Outside the scope	4839 (39.8)	1729 (48.7)		
	1976-1980	306 (2.5)	66 (1.9)	4.52	3.87
	1981-1985	970 (8.0)	268 (7.5)	12.41	11.65
	1986-1990	1739 (14.3)	464 (13.1)	17.49	17.05
	1991-1996	2999 (24.6)	856 (24.1)	23.38	23.50
	1996-2002	6148 (50.6)	1897 (53.4)	42.20	43.93
Age category	**25-34**	2930 (24.1)	831 (23.4)	44.18	45.53
	35-44	4637 (38.1)	1396 (39.3)	38.77	38.49
	45-54	4329 (35.6)	1251 (35.2)	16.40	15.48
	55-56	266 (2.2)	73 (2.1)	0.65	0.50
Total		12 162	3551	337 706	275 378

Overall, the same magnitude was found for the results of the univariable and multivariable analysis, except for the estimated hazard ratios for the social mobility indicator, although, adjusting for all indicators led to some attenuation in the increased risk of death in association to professional trajectory indicators. Here, the results of the multivariable analysis are subsequently presented (those of the univariable analysis could be found in Appendix II).

Occupation at beginning of follow-up

Men in the manual workers class at the beginning had a higher mortality risk compared to those who were in the upper class (except for cancer mortality) but to a different degree depending on the causes of death (Table 3). In women, this association was not statistically significant (Table 4).

Current occupational class

Among men, being in the clerk class increased the mortality risk compared to being in the upper class (HRs: 1.49(1.31-1.69), 1.58(1.09-2.30), 1.50(1.16-1.93), 1.43(1.14-1.79) and 1.58(1.26-1.98) respectively for mortality from all causes, cardiovascular diseases, cancer, external causes and other causes. Among men, those in the manual workers class had an increased mortality risk compared to those in the upper class (HRs: 1.39(1.25-1.56), 1.43(1.03-1.99), 1.26(1.02-1.56) and 1.73(1.42-2.12) respectively for all-cause, cardiovascular, cancer and external-cause mortality). Those outside the study scope had the highest mortality risk except for cardiovascular and cancer mortality among women, i.e. about two to three-fold higher than the mortality risk in the upper class (Table 3 and Table 4).

Cumulative time spent in occupational classes

The cumulative time spent in occupational classes was strongly associated with men's all-cause and cause-specific mortality and women's all-cause and cardiovascular mortality, with less pronounced associations for men's external-cause mortality. Among men, more time spent in an occupational class increased the mortality risk compared to that in the upper class. This increase in manual

workers was associated with a 1.8-fold higher cancer mortality risk (HR: 1.75(1.48-2.06)) and that outside the study scope was associated with a 1.5-fold higher external-cause mortality risk (HR: 1.46(1.19-1.77)) compared to that in the upper class. Among women, more time spent in the clerk class was associated with a 2.7-fold higher cardiovascular mortality risk compared to that in the upper class (HR: 2.65(1.14-6.13)) (Table 3 and Table 4).

Social mobility indicator

In the univariable analysis, an inverse association between the social mobility indicator and mortality was systematically found among men, and only for cancer mortality among women. Adjusting for other indicators changed the direction of the results, except for women's cancer mortality.

In multivariable analysis, the same magnitude was observed for this indicator among men and women except for women's external-cause mortality, with significant results for men and women's all-cause, external-cause and other causes mortality, and women's cancer mortality. Having a high social mobility indicator increased the all-cause mortality risk (HRs: 1.15(1.09-1.21) and 1.13(1.04-1.22) respectively for men and women), the other causes mortality risk (HRs: 1.23(1.12-1.34) and 1.40(1.19-1.64) respectively for men and women) and the external-cause mortality risk (HR: 1.17(1.08-1.28) for men) compared to not experiencing any mobility during professional life (Table 3 and Table 4).

Table 3 All-cause and cause-specific mortality hazard ratios among men according to socio-professional trajectories

	All-cause (n=12 162)	Cardiovascular (n=1452)	Cancer (n=3116)	External causes (n=4026)	Other causes (n=3568)
	$HR^{c}_{\dagger}$(95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.17(1.04-1.33)*	1.41(0.94-2.14)	0.98(0.77-1.24)	1.10(0.89-1.37)	1.39(1.10-1.77)**
Clerk class	1.34(1.18-1.51)***	1.57(1.04-2.37)*	1.02(0.81-1.29)	1.26(1.02-1.56)*	1.68(1.34-2.12)***
Manual workers class	1.43(1.27-1.61)***	1.90(1.27-2.83)**	1.10(0.88-1.37)	1.41(1.15-1.73)**	1.60(1.28-2.00)***
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.16(1.03-1.30)*	1.26(0.90-1.76)	1.10(0.88-1.37)	1.23(1.01-1.51)*	1.07(0.86-1.33)
Clerk class	1.49(1.31-1.69)***	1.58(1.09-2.30)*	1.50(1.16-1.93)**	1.43(1.14-1.79)**	1.58(1.26-1.98)***
Manual workers class	1.39(1.25-1.56)***	1.43(1.03-1.99)*	1.26(1.02-1.56)*	1.73(1.42-2.12)***	1.09(0.89-1.33)
Outside the scope	2.57(2.31-2.85)***	2.45(1.80-2.34)***	2.21(1.81-2.71)***	2.20(1.81-2.68)***	3.25(2.69-3.94)***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	1.04(0.92-1.17)	1.13(0.83-1.54)	1.20(0.98-1.46)	1.03(0.81-1.31)	0.84(0.66-1.06)
Clerk class	1.50(1.33-1.69)***	1.59(1.14-2.20)**	1.53(1.23-1.89)***	1.23(0.95-1.60)	1.62(1.31-2.00)***
Manual workers class	1.52(1.38-1.66)***	1.54(1.18-2.00)**	1.75(1.48-2.06)***	1.33(1.10-1.60)**	1.53(1.28-1.83)***
Outside the scope	1.35(1.22-1.48)***	1.29(0.99-1.69)	1.33(1.12-1.57)***	1.46(1.19-1.77)***	1.39(1.16-1.67)***
Social mobility indicator[b]					
Low (=0)	1	1	1	1	1
Medium	1.03(0.97-1.08)	1.03(0.88-1.20)	0.96(0.87-1.06)	1.11(0.99-1.24)	1.03(0.93-1.13)
High (>1.11)	1.15(1.09-1.21)***	1.12(0.97-1.29)	1.07(0.97-1.18)	1.17(1.08-1.28)***	1.23(1.12-1.34)***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Table 4 All-cause and cause-specific mortality hazard ratios among women according to socio-professional trajectories

	All-cause (n=3551)	Cardiovascular (n=304)	Cancer (n=1388)	External causes (n=894)	Other causes (n=965)
	$HR^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	0.99(0.79-1.23)	0.93(0.39-2.24)	1.02(0.71-1.47)	0.90(0.60-1.35)	1.03(0.69-1.55)
Clerk class	1.05(0.85-1.29)	0.81(0.35-1.86)	1.07(0.76-1.52)	0.92(0.62-1.36)	1.15(0.78-1.70)
Manual workers class	1.15(0.93-1.43)	1.04(0.45-2.43)	1.35(0.94-1.94)	0.91(0.60-1.37)	1.09(0.73-1.63)
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.04(0.82-1.32)	2.18(0.76-6.23)	0.77(0.55-1.09)	1.40(0.86-2.27)	1.27(0.74-2.19)
Clerk class	1.00(0.80-1.26)	1.49(0.52-4.26)	0.75(0.54-1.04)	1.58(0.98-2.53)	1.12(0.66-1.89)
Manual workers class	1.13(0.88-1.45)	2.63(0.88-7.85)	0.71(0.49-1.04)	1.65(0.99-2.74)	1.47(0.84-2.58)
Outside the scope	1.81(1.45-2.27)***	2.48(0.89-6.87)	1.20(0.86-1.66)	2.18(1.38-3.47)***	3.16(1.90-5.26)***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.98(0.78-1.23)	1.61(0.63-4.10)	1.12(0.82-1.53)	0.86(0.55-1.36)	0.66(0.39-1.12)
Clerk class	1.12(0.92-1.36)	2.65(1.14-6.13)*	1.16(0.88-1.54)	0.79(0.52-1.19)	1.14(0.77-1.69)
Manual workers class	1.12(0.91-1.38)	2.05(0.86-4.89)	1.10(0.81-1.49)	1.08(0.71-1.64)	1.07(0.70-1.64)
Outside the scope	1.21(1.01-1.47)*	3.16(1.41-7.05)**	1.08(0.82-1.43)	1.01(0.69-1.49)	1.37(0.93-2.01)
Social mobility indicator[b]					
Low (=0)	1	1	1	1	1
Medium	1.00(0.90-1.11)	1.05(0.76-1.47)	0.85(0.73-0.99)*	1.26(1.01-1.58)*	1.09(0.89-1.33)
High (>0.91)	1.13(1.04-1.22)**	1.10(0.84-1.46)	1.04(0.91-1.18)	1.05(0.88-1.24)	1.40(1.19-1.64)***

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Ad-hoc sensitivity analysis

When replicated analyses were performed on the subsample, including individuals working in the study scope during their first five years of follow-up, the estimated all-cause and cause-specific hazard ratios did not change for any of the indicators except for men's cardiovascular mortality (See Appendix III).

DISCUSSION

Previous studies on this topic have generally considered individuals' socioeconomic position at two or three stages of life including childhood (father's socioeconomic position), entry into the labour market and mid-life position. To our knowledge, the present study is the first to investigate the association between the whole professional trajectory and all-cause mortality and within that, three major causes of death: cardiovascular disease, cancer and external causes. Overall, our results add to the existing evidence of the strong relationship between professional trajectory and all-cause mortality among men, with less pronounced associations among women.[13,15,16,32-36]

Compared to previous studies, a new aspect of our study is the use of the duration of time spent in occupational classes as a measure of socioeconomic exposure and the transition rates between occupational classes as a measure for capturing the social mobility dimension.

The three most commonly used life-course models, namely the critical period, cumulative and social mobility models were taken into account. Our results suggest that all three dimensions are associated to men's all-cause mortality. For women, only the cumulative and the social mobility models were confirmed by this analysis.

Interpretations and comparisons with other studies

As shown in previous studies, strong associations between professional trajectories and men's and women's mortality was found.[13,15,16,32-36] However, a direct comparison with other studies

cannot be easily made given the different occupational classifications in each country, and the fact that we used whole professional trajectories.

The present study only focused on professional trajectories with no information on childhood circumstances. However, the individual's first occupation is likely to be the most representative dimension of the end of childhood. We found that the association between the first occupation and mortality was strong for men's cardiovascular and external-cause mortality. Previously, strong associations have also been reported between socioeconomic circumstances in childhood and mortality from some causes of death, such as cardiovascular diseases.[15, 32, 33, 35]

On the other hand, for some other causes of death such as external causes and lung cancer,[35] stronger associations were found between socioeconomic circumstances in adulthood and adult mortality than those in childhood. Our results are in accordance with the literature, since in other studies, for some causes of death such as external causes and cancer, occupational classes found to be strongly linked with men's mortality. Supplementary analysis on different cancers also reported the same associations or even stronger ones (for deaths by UADT cancers) (data not shown). For women, the results were not statistically significant.

Another hypothesis in the literature is the putative association between the accumulation of exposure to different socioeconomic conditions and mortality. However, the use of only three stages of life limited the number of possible trajectories, so the different trajectories could be compared. By investigating the duration of time spent in each occupational class instead of comparing different trajectories, we found a strong relationship between the duration of exposure to low professional position and mortality. This association was stronger for cardiovascular and cancer mortality in men but was significant only for all-cause and cardiovascular mortality in women. This is consistent with the results of previous studies.[15,16,33,37] The large mortality risk of those who stay longer in the low occupational categories can be explained by exposure to poor working conditions and by the fact that the least skilled are less likely to move upward. Furthermore, staying a long time in the same

professional conditions could reflect a greater adherence to a professional class and its specific lifestyle.

The changes between occupational categories and their dynamics were also pointed out in previous studies. Some studies have shown that within classes, male movers have a mortality risk situated between that of non-movers in their class of origin and that of their destination.[38,39] Here, we investigated the association between the frequency of changes between occupational classes and mortality. Instability in professional life may be interpreted in two ways. If instability is chosen, it could be the reflection of high dynamism with the ability to change and adapt to several professional environments. Conversely, if instability is forced, it could be due to difficulties in finding one's place, to a high dependence on the work market or to personal events. We found an inverse association for this indicator in the univariable analysis, as it does not take into account the occupational classes before and after the transitions. Our results of the multivariable analysis show that subjects with high transition rates have an increased risk of all-cause and external-cause mortality. These results suggest that the instability measured is more forced than chosen, with a deleterious association on mortality. In a very explorative approach to disentangle the chosen and forced instability, we considered the following naive order of occupations from high to low level: "upper class", "intermediary occupations", "clerk class", and "manual workers". Although this order is not strictly hierarchical, upward and downward changes were studied as separate variables. The risk of mortality was positively associated with downward changes (for example, going from the "upper class" to the "clerk class"), and negatively with upward changes (for example, going from the "manual workers class" to the "intermediary occupations class") (data not shown).

Limitations

The main limitation in this investigation is the high percentage of follow-up years outside the scope of the study. The decision to consider all these data in the "outside the scope" category could induce a bias. However, we examined a wide range of occupational sectors and the occupational stages are

sufficiently reliable as they were collected within the context of administrative procedures. Furthermore, the replicated analysis on the subsample with sufficient follow-up provided almost the same results, which strengthens the findings.

All participants had worked at least once between the ages of 25 and 30 and were likely to be healthier than the general population, so the sample should not be interpreted as representative of the French population.

Finally, taking into account the individual's occupation with a two-year time lag could reduce the reverse causation bias. However, for some causes of death such as transport accidents, the problem of reverse causation is less likely to be a source of bias.

Despite these drawbacks, the large size of the sample, the annual nature of the information collected and the causes of death coded with high precision are the major strength of this study. Using repeated measures of occupational category over the follow-up could provide insight into changes that may have occurred during a person's professional life. To gain a better understanding of the complex social inequalities in mortality, future analysis should focus on models that take into account simultaneously all aspects of professional trajectories and mortality. Joint modelling of nominal occupational data and cause-specific mortality following the approach of Li et al.[40] is the object of an on-going project.

Acknowledgements:

The authors would like to thank Walid Ghosn for his help and comments during this study.

Financial support:

This work was supported by joint funding from the Ministry of Health, the General Directorate of Health and the Mission research of the Directorate of research, studies, evaluation and statistics, the National Health Insurance Fund of the Salaried Workers, the Social Security Scheme for Self-employed Workers, the National Solidarity Fund for Autonomy and the National Institute of Prevention and Education for Health, within the framework of the 2011 call for research launched by IReSP (Institute of Research in Public Health) [grant number A11226LS].

Competing interest: None declared.

References

1. Marang-van de Mheen PJ, Davey Smith G, Hart CL, et al. Socioeconomic differentials in mortality among men within Great Britain: time trends and contributory causes. *J Epidemiol Community Health.* 1998; 52(4): 214–8.
2. Martikainen P, Valkonen T, Martelin T. Change in male and female life expectancy by social class: decomposition by age and cause of death in Finland 1971-95. *J Epidemiol Community Health.* 2001; 55(7): 494–9.
3. Davey Smith G, Bartley M, Blane D. The Black report on socioeconomic inequalities in health 10 years on. *BMJ.* 1990; 301(6748): 373–7.
4. Kunst AE, Mackenbach JP. The size of mortality differences associated with educational level in nine industrialized countries. *Am J Public Health.* 1994; 84(6): 932–7.
5. Leclerc A, Lert F, Fabien C. Differential mortality: some comparisons between England and Wales, Finland and France, based on inequality measures. *Int J Epidemiol.* 1990; 19(4): 1001–10.
6. Mackenbach JP, Kunst AE, Cavelaars AE, et al. Socioeconomic inequalities in morbidity and mortality in western Europe. *Lancet.* 1997; 349(9066): 1655–9.
7. Mackenbach JP, Bos V, Andersen O, et al. Widening socioeconomic inequalities in mortality in six Western European countries. *Int J Epidemiol.* 2003; 32(5): 830–7.
8. Pappas G, Queen S, Hadden W, et al. The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986. *N Engl J Med.* 1993; 329(2): 103–9.
9. Menvielle G, Chastang J-F, Luce D, et al. Changing social disparities and mortality in France (1968-1996): cause of death analysis by educational level. *Rev EpidemiolSantePublique.* 2007; 55(2): 97–105.
10. Monteil C, Robert-Bobée I. Les différences sociales de mortalité: en augmentation chez les hommes, stables chez les femmes. *Insee Première.* 2005; 1025.

11. Leclerc A, Chastang J-F, Menvielle G, et al. Socioeconomic inequalities in premature mortality in France: have they widened in recent decades? *SocSci Med.* 2006; 62(8): 2035–45.
12. Steenland K, Henley J, Thun M. All-Cause and Cause-specific Death Rates by Educational Status for Two Million People in Two American Cancer Society Cohorts, 1959–1996. *Am J Epidemiol.* 2002; 156(1) :11–21.
13. Kunst AE, Groenhof F, Mackenbach JP, et al. Occupational class and cause specific mortality in middle aged men in 11 European countries: comparison of population based studies. EU Working group on Socioeconomic Inequalities in Health. *BMJ.* 1998; 316(7145): 1636–42.
14. Galobardes B, Shaw M, Lawlor DA, at al. Indicators of socioeconomic position (part 1). *J Epidemiol Community Health.* 2006; 60(1): 7–12.
15. Davey Smith G, Hart C, Blane D, et al. Lifetime socioeconomic position and mortality: prospective observational study. *BMJ.* 1997; 314(7080): 547–52.
16. Melchior M, Berkman LF, Kawachi I, et al. Lifelong socioeconomic trajectory and premature mortality (35-65 years) in France: findings from the GAZEL Cohort Study. *J Epidemiol Community Health.* 2006; 60(11): 937–44.
17. Stringhini S, Dugravot A, Kivimaki M, et al. Do different measures of early life socioeconomic circumstances predict adult mortality? Evidence from the British Whitehall II and French GAZEL studies. *J Epidemiol Community Health.* 2011; 65(12): 1097–103.
18. Heslop P, Davey Smith G, Macleod J, et al. The socioeconomic position of employed women, risk factors and mortality. *SocSci Med.* 2001; 53(4): 477–85.
19. Singh-Manoux A, Ferrie JE, Chandola T, et al. Socioeconomic trajectories across the life course and health outcomes in midlife: evidence for the accumulation hypothesis? *Int J Epidemiol.* 2004; 33(5): 1072–9.
20. Kuh D, Ben-Shlomo Y, Lynch J, et al. Life course epidemiology. *J Epidemiol Community Health.* 2003; 57(10): 778–83.

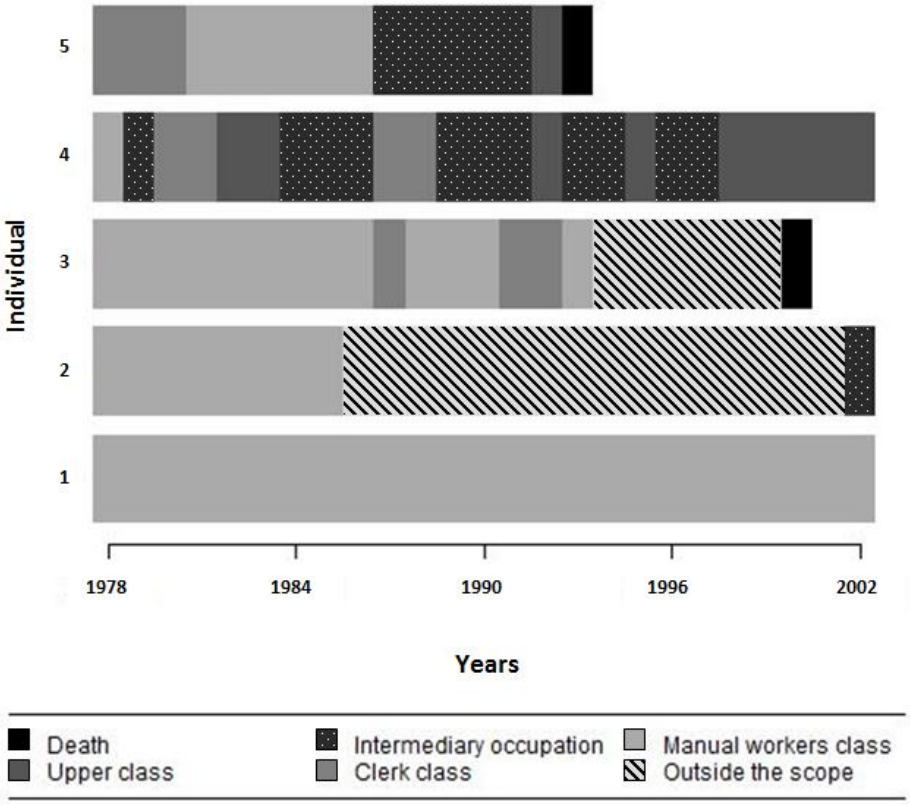
21. Fox AJ, Goldblatt PO, Jones DR. Social class mortality differentials: artefact, selection or life circumstances? *J Epidemiol Community Health.* 1985; 39(1): 1–8.
22. Smith JP. Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status. *J Econ Perspect.* 1999; 13(2): 144–66.
23. Annual Declaration of Social Data / DADS [Internet]. Available from: http://www.insee.fr/en/methodes/default.asp?page=definitions/dec-ann-don-soc.htm
24. Desrosières A, Thévenot L. Les catégories socio-professionnelles. *La Découverte.* 2002.
25. Raghunathan T, Lepkowski J, Van Hoewyk J, et al. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survy Methodology.* 2001; 27: 85–95.
26. Martikainen PT. Unemployment and mortality among Finnish men, 1981-5. *BMJ.* 1990; 301(6749): 407–11.
27. Menvielle G, Luce D, Geoffroy-Perez B, et al. Social inequalities and cancer mortality in France, 1975–1990. *Cancer Causes Control.* 2005; 16(5): 501–13.
28. Thiébaut AC, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med*. 2004; 23(24): 3803–20.
29. Andersen PK, Geskus RB, de Witte T, et al. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol.* 2012; 41(3): 861–70.
30. Therneau TM. A Package for Survival Analysis in S [Internet]. 2014. Available from: http://CRAN.R-project.org/package=survival
31. Raghunathan T, Solenberger P, Hoewyk J. IVEware: Imputation and variance estimation software [Internet]. 2007. Available from: http://www.isr.umich.edu/src/smp/ive
32. Beebe-Dimmer J, Lynch JW, Turrell G, et al. Childhood and Adult Socioeconomic Conditions and 31-Year Mortality Risk in Women. *Am J Epidemiol.* 2004; 159(5): 481–90.

33. Claussen B, Davey S, Thelle D. Impact of childhood and adulthood socioeconomic position on cause specific mortality: the Oslo Mortality Study. *J Epidemiol Community Health.* 2003; 57(1): 40–5.
34. Pollitt RA, Rose KM, Kaufman JS. Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. *BMC Public Health.* 2005;5: 7.
35. Davey Smith G, Hart C, Blane D, Hole D. Adverse socioeconomic conditions in childhood and cause specific adult mortality: prospective observational study. *BMJ.* 1998; 316(7145): 1631–5.
36. Mackenbach JP, Kunst AE, Groenhof F, et al. Socioeconomic inequalities in mortality among women and among men: an international study. *Am J Public Health.* 1999; 89(12): 1800–6.
37. Pensola TH, Martikainen P. Cumulative social class and mortality from various causes of adult men. *J Epidemiol Community Health.* 2003; 57(9): 745–51.
38. Cambois E. Careers and mortality in France: evidence on how far occupational mobility predicts differentiated risks. *SocSci Med.* 2004; 58(12): 2545–58.
39. Blane D, Harding S, Rosato M. Does social mobility affect the size of the socioeconomic mortality differential?: evidence from the Office for National Statistics Longitudinal Study. *J R Stat SocSer A Stat Soc.* 1999; 162: 59–70.
40. Li N, Elashoff RM, Li G, et al. Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. *Stat Med.* 2010; 29(5): 546–57.

Figure Legend:

Figure: Examples of Fictional Trajectories

Example: Individual 2 was working in the manual workers class from 1978 till 1985. He was outside the scope of the study between 1986 and 2001 and finally worked in an intermediary occupation in 2002.


Individual
5
4
3
2
1
1978
1984
1990
1996
2002
Years
Death
Upper class
Intermediary occupation
Clerk class
Manual workers class
Outside the scope

APPENDIX I

Table I Causes of death according to International Classification of Diseases (ICD)

Causes of death	ICD-8	ICD-9	ICD-10
Cardiovascular diseases	390–444.1, 444.3–458, 782.4	390–459	I00–I99
Cancer	140–239	140–239	C00–D48
External causes	E800–E999	E800–E999	V01–Y89

ICD-8: before 1979, ICD-9: from 1979 to 1999, ICD-10: since 2000

APPENDIX II

Table II.A All-cause and cause-specific mortality hazard ratios among men according to socio-professional trajectories (univariable analysis)

	All-cause (n=12 162)	Cardiovascular (n=1452)	Cancer (n=3116)	External causes (n=4026)	Other causes (n=3568)
	$HR^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.23(1.09-1.38)***	1.56(1.06-2.30)*	1.10(0.88-1.37)	1.20(0.97-1.47)	1.29(1.04-1.62)*
Clerk class	1.79(1.60-2.01)***	2.21(1.52-3.22)***	1.43(1.16-1.77)***	1.62(1.33-1.98)***	2.23(1.81-2.76)***
Manual workers class	2.05(1.84-2.28)***	2.84(1.98-4.05)***	1.85(1.52-2.26)***	2.03(1.69-2.45)***	2.01(1.64-2.46)***
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.44(1.30-1.60)***	1.76(1.31-2.35)***	1.49(1.24-1.80)***	1.44(1.20-1.72)***	1.26(1.04-1.54)*
Clerk class	2.33(2.09-2.60)***	2.73(1.99-3.74)***	2.41(1.96-2.97)***	1.88(1.56-2.28)***	2.73(2.24-3.33)***
Manual workers class	2.34(2.14-2.56)***	2.74(2.10-3.58)***	2.52(2.13-2.98)***	2.46(2.10-2.90)***	1.83(1.54-2.18)***
Outside the scope	3.67(3.35-4.01)***	3.68(2.83-4.79)***	3.20(2.71-3.78)***	3.05(2.59-3.60)***	4.80(4.06-5.68)***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.94(0.86-1.03)	1.00(0.81-1.25)	1.04(0.90-1.21)	0.91(0.75-1.10)	0.80(0.67-0.95)*
Clerk class	1.68(1.54-1.83)***	1.66(1.33-2.06)***	1.57(1.36-1.83)***	1.35(1.12-1.63)**	2.09(1.80-2.42)***
Manual workers class	1.66(1.56-1.76)***	1.69(1.45-1.96)***	1.74(1.57-1.93)***	1.68(1.48-1.91)***	1.51(1.34-1.70)***
Outside the scope	2.09(1.95-2.23)***	1.84(1.55-2.18)***	1.79(1.60-2.00)***	2.10(1.82-2.42)***	2.61(2.30-2.96)***
Social mobility indicator[b]					
Low (=0)	1	1	1	1	1
Medium	0.84(0.79-0.88)***	0.83(0.72-0.95)**	0.77(0.70-0.84)***	0.87(0.78-0.97)*	0.89(0.80-0.98)*
High (>1.11)	0.83(0.79-0.86)***	0.80(0.71-0.90)***	0.75(0.69-0.81)***	0.85(0.79-0.92)***	0.88(0.81-0.95)**

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted separately for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Table II.B All-cause and cause-specific mortality hazard ratios among women according to socio-professional trajectories (univariable analysis)

	All-cause (n=3551)	Cardiovascular (n=304)	Cancer (n=1388)	External causes (n=894)	Other causes (n=965)
	$HR^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	0.95(0.77-1.17)	1.20(0.54-2.63)	0.98(0.70-1.37)	0.93(0.63-1.38)	0.87(0.58-1.31)
Clerk class	1.06(0.87-1.30)	1.23(0.58-2.64)	1.05(0.76-1.45)	0.97(0.67-1.40)	1.15(0.79-1.69)
Manual workers class	1.26(1.03-1.54)*	1.73(0.81-3.69)	1.28(0.92-1.77)	1.18(0.81-1.71)	1.21(0.82-1.79)
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.05(0.85-1.29)	2.92(1.04-8.18)*	0.84(0.63-1.11)	1.27(0.83-1.97)	1.09(0.68-1.75)
Clerk class	1.12(0.92-1.35)	2.66(0.97-7.34)	0.86(0.66-1.12)	1.40(0.93-2.11)	1.29(0.83-2.01)
Manual workers class	1.35(1.10-1.66)**	4.50(1.61-12.56)**	0.91(0.69-1.22)	1.77(1.15-2.72)**	1.60(1.01-2.55)*
Outside the scope	2.11(1.75-2.54)***	5.12(1.90-13.75)**	1.34(1.04-1.72)*	2.22(1.48-3.32)***	3.82(2.50-5.85)***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.83(0.70-0.97)*	1.38(0.74-2.60)	0.88(0.70-1.11)	0.79(0.56-1.10)	0.59(0.40-0.86)**
Clerk class	0.95(0.83-1.09)	1.71(0.98-2.99)	0.93(0.76-1.13)	0.80(0.60-1.08)	0.96(0.73-1.27)
Manual workers class	1.11(0.96-1.28)	2.15(1.24-3.73)**	1.03(0.84-1.26)	1.14(0.85-1.53)	1.02(0.76-1.36)
Outside the scope	1.47(1.29-1.68)***	2.88(1.67-4.96)***	1.17(0.96-1.41)	1.31(0.99-1.73)	1.95(1.50-2.54)***
Social mobility indicator[b]					
Low (=0)	1	1	1	1	1
Medium	0.93(0.84-1.03)	0.99(0.72-1.38)	0.83(0.71-0.97)*	1.15(0.93-1.44)	0.93(0.76-1.14)
High (>0.91)	0.93(0.86-1.00)	0.84(0.65-1.10)	0.95(0.85-1.07)	0.90(0.77-1.05)	0.94(0.81-1.09)

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted separately for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

APPENDIX III

Table III. A All-cause and cause-specific mortality hazard ratios according to socio-professional trajectories among men working in scope of study on their first five years of follow-up

	All-cause (n=6884)	Cardiovascular (n=949)	Cancer (n=2067)	External causes (n=1979)	Other causes (n=1889)
	$HR^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.25(1.04-1.50)*	2.04(1.10-3.79)*	0.88(0.64-1.21)	1.04(0.75-1.46)	1.84(1.27-2.66)**
Clerk class	1.38(1.15-1.66)***	2.18(1.16-4.08)*	0.96(0.71-1.32)	1.30(0.93-1.82)	1.86(1.29-2.69)***
Manual workers class	1.31(1.10-1.57)**	2.57(1.39-4.76)**	0.90(0.66-1.22)	1.22(0.88-1.69)	1.61(1.13-2.31)**
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.09(0.94-1.25)	1.09(0.75-1.59)	1.01(0.77-1.31)	1.20(0.91-1.56)	1.04(0.79-1.37)
Clerk class	1.40(1.19-1.66)***	1.29(0.83-2.01)	1.29(0.94-1.76)	1.33(0.97-1.82)	1.70(1.26-2.31)***
Manual workers class	1.28(1.12-1.48)***	1.09(0.74-1.58)	1.12(0.87-1.45)	1.78(1.37-2.32)***	1.08(0.83-1.41)
Outside the scope	2.50(2.18-2.86)***	2.05(1.42-2.96)***	2.01(1.58-2.56)***	2.22(1.70-2.89)***	3.54(2.77-4.54)***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	1.07(0.93-1.23)	1.05(0.73-1.49)	1.26(0.99-1.60)	1.10(0.82-1.47)	0.87(0.65-1.15)
Clerk class	1.48(1.27-1.71)***	1.58(1.08-2.33)*	1.57(1.21-2.04)***	1.22(0.88-1.68)	1.55(1.18-2.03)**
Manual workers class	1.60(1.43-1.80)***	1.52(1.11-2.07)**	1.91(1.56-2.33)***	1.33(1.04-1.69)*	1.65(1.32-2.06)***
Outside the scope	1.57(1.38-1.78)***	1.31(0.92-1.86)	1.67(1.34-2.08)***	1.73(1.31-2.29)***	1.61(1.26-2.05)***
Social mobility indicator[b]					
Low (=0)	1	1	1	1	1
Medium	0.98(0.92-1.06)	0.97(0.81-1.17)	0.94(0.83-1.06)	1.06(0.92-1.22)	0.99(0.87-1.13)
High (>1.18)	1.08(1.02-1.16)*	1.07(0.89-1.27)	1.01(0.89-1.14)	1.09(0.97-1.22)	1.18(1.05-1.33)**

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model

Table III.B All-cause and cause-specific mortality hazard ratios according to socio-professional trajectories among women working in scope of study on their first five years of follow-up

	All-cause (n=1544)	Cardiovascular (n=136)	Cancer (n=723)	External causes (n=316)	Other causes (n=369)
	$HR^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)	$CSH^{c}_{\dagger}$ (95% CI)
Occupation at beginning of follow-up					
Upper class	1	1	1	1	1
Intermediary occupations	1.04(0.72-1.51)	2.10(0.25-17.80)	1.12(0.64-1.95)	1.15(0.52-2.52)	0.78(0.42-1.47)
Clerk class	0.98(0.68-1.39)	2.02(0.24-16.62)	1.09(0.63-1.87)	0.86(0.40-1.86)	0.80(0.43-1.49)
Manual workers class	0.99(0.68-1.44)	2.35(0.28-19.74)	1.24(0.71-2.19)	0.90(0.41-2.01)	0.58(0.30-1.14)
Current occupational class[a]					
Upper class	1	1	1	1	1
Intermediary occupations	1.16(0.82-1.65)	2.21(0.56-8.71)	0.97(0.61-1.56)	0.85(0.41-1.77)	2.60(0.98-6.72)
Clerk class	1.12(0.80-1.56)	1.65(0.41-6.67)	0.94(0.60-1.48)	1.24(0.60-2.55)	1.92(0.77-4.81)
Manual workers class	1.16(0.80-1.68)	4.06(0.94-17.59)	0.84(0.50-1.40)	0.98(0.44-2.16)	2.10(0.77-5.69)
Outside the scope	2.08(1.49-2.91)***	3.03(0.76-12.11)	1.45(0.92-2.29)	1.52(0.73-3.15)	6.73(2.73-16.6)***
Cumulative time spent in occupational class					
Upper class	1	1	1	1	1
Intermediary occupations	0.94(0.69-1.28)	1.55(0.45-5.28)	0.91(0.60-1.39)	1.14(0.58-2.24)	0.69(0.33-1.45)
Clerk class	1.05(0.80-1.39)	2.38(0.78-7.29)	0.95(0.65-1.39)	0.86(0.45-1.65)	1.15(0.64-2.09)
Manual workers class	1.12(0.83-1.50)	1.44(0.43-4.78)	0.98(0.65-1.48)	1.38(0.72-2.64)	1.26(0.65-2.43)
Outside the scope	1.21(0.90-1.61)	2.56(0.82-8.01)	1.10(0.74-1.64)	1.12(0.55-2.26)	1.27(0.69-2.32)
Social mobility indicator[b]					
Low (=0)	1	1	1	1	1
Medium	0.94(0.81-1.08)	1.18(0.73-1.90)	0.84(0.68-1.03)	1.03(0.74-1.43)	1.01(0.75-1.36)
High (>1)	1.03(0.91-1.15)	1.17(0.79-1.72)	0.97(0.81-1.15)	0.86(0.66-1.12)	1.27(0.99-1.61)

*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

a: observed with two-year time lag

b: transition rates between occupational classes (10 years of follow-up)

c: adjusted for occupation at the beginning, current occupational class, cumulative time spent in occupational class, social mobility indicator and observation periods

†: age as the time-scale in Cox proportional hazards model