Inference and validation of prognostic marker for correlated survival data with application to cancer

Alessandra Meddis

université
PARIS-SACLAY

Inference and validation of prognostic marker for correlated survival data with application to cancer

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570, École doctorale de Santé Publique (EDSP)
Spécialité de doctorat: Santé publique - Biostatistiques
Unité de recherche: U900, INSERM, PSL Research University Institut Curie
Référent: université de Versailles -Saint-Quentin-en-Yvelines

Thèse présentée et soutenue à Paris, le 23/10/2020, par

Alessandra MEDDIS

Composition du jury:

Estelle Kuhn Directrice de Recherche, INRAE (UNITÉ MaIAGE)	Présidente
Helene Jacqmin-Gadda Directrice de Recherche, University of Bordeaux, ISPED, Bordeaux Population Health Research Center, UMR1219 (INSERM)	Rapportrice & Examinatrice
Hein Putter Professor, Leiden University Medical Center (Department of Biometrics)	Rapporteur & Examinateur
Yohann Foucher Docteur, Institut de Recherche en Santé 2 (Inserm UMR 1246 - SPHERE)	Examinateur
Aurélien Latouche Professeur, Institut Curie (PSL Research University, INSERM, U900)	Directeur de thése

Thèse de doctorat

NNT : 2020UPASR005

To my Family,
who has given to me
a lifetime of support and
unconditional love.

Acknowledgments

At the end of this "trip" I have to thank all the people who have supported me along the way. I would like to thank my supervisor, Aurélien Latouche who has the merit of having introduced me the world of research. I am very grateful for his guidance and encouragement in these years together. It was a pleasure being supervised by someone with a strongly positive attitude always ready with the right advice.

Thanks to all the members of the StaMPM team, in particular to Xavier Paoletti for his invaluable help and all the interesting discussions; thanks to my colleagues Bassirou and Jonas, we supported each others during these years, sharing bad and good news. Thank you to the HR department and Caroline for their essential help with the administrative part.

My gratitude are also due to the members of the jury for having accepted to be part of this last step before becoming a Doctor. Thank you for the interest showed on my work and for the inestimable feedback.

I had the pleasure to collaborate with great researchers during this three years. I would like to thank Paul Blanche for his ideas and his encouragement, his love for research inspired me; thanks to Stephen Cole for the opportunity to work together, for his precious time and his constant enthusiasm. I had also the opportunity to work with the biomarker team in Servier. Thanks for their worm welcome during my visits and, in particular, to Julia Geronimi, her pragmatism has taught me a lot.

I would like to thank my friends, the ones that I met along the way in Paris and those that even from a distance there have been. Thanks to Cucchi for his advises, he has always been there during my moments of uncertainty. Thanks to Valentina and Pasquale for their friendship, they have the ability to understand me in every situation. Thanks to Alida, Rosamaria, Giusy and Chiara for all the support and positivity.

My deepest thank goes to my family, my parents for their unconditional support in all my choices. They taught me the value of working with ambition and perseverance. Thanks to my first supporter, my sister Mariagrazia, for her love, she is my greatest example of determination. I would also like to thank my "new family": Serge, Isabelle, Victor and Justine. The time spent together made me feel home. Finally, I thank Alex for all the emotional support, for sharing with me ups and downs of these years and thank you for all the happiness you give me everyday.

Scientific production

Published articles

- A. Meddis, A.Latouche, B. Zhou, S. Michiel J. Fine (2020). Meta-analysis of clinical trials with competing time-to-event endpoints. Biometrical Journal, 62(3), 712-723.
- A. Meddis, P. Blanche, F.C. Bidard, A. Latouche (2020). A covariate-specific time-dependent receiver operating characteristic curve for correlated survival data. Statistics in Medicine. 2020 Aug 30;39(19):2477-2489.

Articles in preparation

- A. Meddis, A.Latouche (2020). Test for informative cluster size with survival data.

Related articles

- F.C. Bidard, S.Michiels, ..., A. Meddis, P. Blanche, K. D'Hollander, P. Cottu, J. W. Park, S. Loibl, A. Latouche, J. Pierga, K. Pantel (2018). Circulating tumor cells in breast cancer patients treated by neoadjuvant chemotherapy: a meta-analysis. JNCI: Journal of the National Cancer Institute, 110(6), 560-567.
- L. Thery , A. Meddis, L. Cabel, C. Proudhon, A. Latouche, J.Y. Pierga, F. C. Bidard (2019). Circulating tumor cells in early breast cancer.JNCI cancer spectrum, 3(2).

Oral communications and conferences

- A. Meddis, P. Blanche, A. Latouche. Semiparametric approach for covariate-specific time dependent ROC curve for correlated survival data Journée des Jeunes chercheurs de la SFB. Paris, France - May 2018

- A. Meddis, P. Blanche, A. Latouche. Semiparametric approach for covariate-specific time dependent ROC curve for correlated survival data. XXIXTH International Biometric Conference. Barcelona, Spain - July 2018.

- A. Meddis, A.Latouche, B. Zhou, S. Michiel J. Fine. Meta-analysis of clinical trials with competing time-to-event endpoints. Survival analysis for young researchers. Copenhagen, Denmark- April 2019

- A. Meddis, A. Latouche. Un test pour la taille informative de grappes en presence des données censurées. GDR Statistiques et Santé. Paris, France - October 2019

- A. Meddis, A. Latouche. On interpretation of biomarker performance with clustered survival data. Journée des statisticiens. Paris, France - January 2020

- A. Meddis, A. Latouche. On interpretation of biomarker performance with clustered survival data. Seminaire SPHERE. Tours, France - February 2020.

- A. Meddis, A. Latouche. Test of informative cluster size with survival data. 41st International Society for Clinical Biostatistics (ISCB). Krakov, Poland - August 2020

Resumé en Français

Les données en grappes apparaissent lorsque les observations sont collectées dans plusieurs groupes différents (grappes) et que les observations de la même grappe sont plus semblables que les observations d'une autre grappe en raison de plusieurs facteurs. Ce type de données sont souvent recueillies dans la recherche biomédicale. La méta-analyse est un exemple populaire qui combine observations provenant de plusieurs essais cliniques randomisés portant sur la même question médicale. Il s'agit d'une procédure typique qui permet une analyse plus précise des données mais soulève de nouveaux problèmes statistique. D'autres exemples sont les données longitudinales où de multiples mesures sont prises au fil du temps sur chaque individu; ou les données familiales où les facteurs génétiques sont partagés par les membres d'une même famille.

L'étude IMENEO est une méta-analyse sur données individuelles (IPD) sur des patientes avec un cancer du sein non métastatique et traitées par chimiothérapie néoadjuvante. L'objectif principal était d'étudier la capacité pronostique des cellules tumorales circulantes (CTC). Ce biomarqueur s'est avéré être pronostique dans les cas de cancer métastatique, et cette méta-analyse a voulu l'étudier également dans le cadre non métastatique. Les données ont été recueillies dans plusieurs centres et il est légitime de suspecter diverses sources d'hétérogénéité. Cela pourrait conduire à une corrélation entre les observations appartenant à un même groupe. Il n'était pas clair comment aborder le problème de la corrélation dans l'analyse discriminatoire d'un biomarqueur candidat. Nous avons proposé une méthode pour estimer la courbe ROC dépendante du temps en traitant les temps d'événement corrélés censurés à droite. De plus, un grand nombre de CTC sont détectés dans les cancers métastatiques et, dans un contexte non métastatique, ils sont plus fréquents dans les tumeurs inflammatoires. Les cliniciens étaient donc interessés d'évaluer la performance des CTC en discriminant les patients qui ont un profil de risque similaire, à savoir un stade de tumeur similaire. La courbe ROC cumulative/dynamique spécifique aux covariables et sa AUC ont été estimées en modélisant l'effet des covariables et des biomarqueurs sur le résultat par un modèle de fragilité partagée et un modèle de régression paramétrique pour la distribution des biomarqueurs conditionnée par les co-

variables. L'introduction de la fragilité permet de saisir la corrélation des observations à l'intérieur des groupes. Dans l'exemple motivant, un modèle de régression binomiale négative pour les CTC s'est montré approprié.
Une étude de simulation a été réalisée et a montré un biais négligeable pour l'estimateur proposé et pour un estimateur non paramétrique fondé sur la pondération par la probabilité inverse d'être censuré (IPCW), tandis qu'un estimateur semi-paramétrique, ignorant la structure en grappe est nettement biaisé. En outre, nous avons illustré la robustesse de la méthode en cas de mauvaise spécification de la distribution de la fragilité.

Dans l' application aux données sur le cancer du sein, l'estimation des AUC spécifiques au stade de tumeur a permis d'évaluer que les CTC discriminent mieux les patients atteints d'une tumeur inflammatoire que ceux atteints d'une tumeur non inflammatoire, en ce qui concerne leur risque de décès.

Par construction du modèle, on a supposé l'existence d'un biomarqueur homogène entre grappes. Cela dépend directement du fait que dans un modèle mixte, l'effet aléatoire doit être indépendant des covariables. L'hypothèse d'un biomarqueur homogène est raisonnable lorsque la technologie de mesure du biomarqueur est soit standardisée entre les groupes, soit centralisée. Une autre hypothèse est la taille non informative des grappes (NICS), qui est nécessaire pour éviter des résultats biaisés. La taille des grappes est dite informative lorsque la variable réponse dépend de la taille de grappe conditionnellement à un ensemble de variables explicatives. De plus, nous avons discuté de l'interprétation des résultats en présence de données en grappes et nous avons souligné quelles quantités peuvent être estimées sous quelles hypothèses. Pour les données en grappes, nous pouvons distinguer deux analyses marginales : la première a une interprétation pour la population de tous les membres observés (AOM) où des poids égaux sont donnés à chaque membre de la population observée. La seconde, a une interprétation pour la population d'un membre typique d'une grappe typique (TOM) et des poids égaux sont donnés à chaque grappe observée et les sujets au sein de la grappe sont pondérés de manière égale. Nous apportons la preuve que, dans le cadre du NICS, les deux estimations de la courbe ROC pour TOM et AOM coïncident, alors qu'elles diffèrent dans le contexte de ICS.

L'hypothèse d'une taille de grappe non informative n'est pas trop restrictive pour une méta-analyse, où, intuitivement, on ne s'attendrait pas à ce que le résultat varie en fonction de la taille des grappes. Toutefois, cette hypothèse est fausse dans certaines situations et ICS est souvent ignorée lors de l'analyse de données en grappes. Par exemple, prenons le cas où le temps de perte des dents chez un individu est consideré. Les sujets souffrant d'une maladie dentaire peuvent avoir déjà perdu des dents à cause de celle-ci. Ainsi, le temps de perte d'une dent chez un individu (grappe) est lié au nombre de dents

(taille de la grappe) de ce même individu. Dans le cas de la taille des grappes informatives, les méthodes statistiques standard pour les données en grappes produisent des résultats biaisés. Les estimations seront surpondérées en faveur de grappes plus grandes. L'idée de ICS a été initialement introduite par Hoffman qui a proposé la méthode de rééchantillonnage intra-groupe où plusieurs bases de données indépendants sont créés en échantillonnant aléatoirement une observation de chaque grappe. D'autres approches ont été successivement étudiées dans l'analyse de survie. Cependant, ces méthodes reposent sur l'hypothèse de ICS et celle-ci n'a jamais été formellement testée. À notre connaissance, il n'y a pas eu de méthode pour vérifier l'existence de ICS avec des données corrélées de survie. Pour cette raison, nous avons proposé un test qui pourrait vérifier cette hypothèse dans des données en grappes censurées à droite. Le test repose sur la propriété selon laquelle les résultats pour AOM et TOM coïncident dans le cadre du NICS et il tient compte de la différence de l'estimateur de Nelson-Aalen pour les deux analyses marginales. La statistique du test converge faiblement vers un processus gaussien avec une moyenne nulle et nous avons dérivé la matrice de covariance. Une étude de simulation a suggéré une bonne performance du test pour les données fortement groupées et pour le scenario avec quelques grands grappes. Cependant, une faible puissance a été détectée pour un petit nombre de grappes.

Quelques exemples ont été fournis dans plusieurs contextes. Nous avons appliqué le test à l'étude IMENEO et l'hypothèse nulle n'a pas été rejeté. En outre, une étude sur les maladies parodontales a été envisagée et, comme prévu, un fort ICS a été détecté. Un autre exemple illustratif était une étude multicentrique de patients avec une cirrhose billariale primaire due à une maladie du foie. Ici, de manière moins intuitive, nous avons découvert que les patients traités dans des centres plus petits avaient des temps d'événement plus longs, où le résultat d'intérêt était l'échec du traitement. Enfin, nous avons montré une limitation de la méthode pour petites tailles des grappes différentes dans une étude de patients atteints de cancer métastatique traités par immunothérapie. L'individu représente la grappe et le nombre de métastases les tailles de la grappe ; la progression de la tumeur a été suivie pour chaque métastase afin de détecter les réponses dissociées qui sont typiques avec l'immunothérapie. Les patients présentant un maximum de 5 métastases ont été inclus dans l'étude. Même si les fonctions de survie estimées pour les patients regroupés par nombre de métastases ont montré une nette différence sur la progression de la maladie, le test n'a pas rejeté l'hypothèse nulle.

Nous avons ensuite examiné une méta-analyse sur données individuels avec des risques concurrents. L'objectif était d'évaluer le bénéfice de l'adjonction de la chimiothérapie à la radiothérapie dans le carcinome du nasopharynx. Ce type de cancer est souvent diag-

nostiqué à un stade localement avancé, mais il est très sensible à la radiothérapie et à la chimiothérapie. Précédemment, l'effet du traitement sur la survie globale et sur la survie sans progression étaient analysés. Nous voulions évaluer le bénéfice de la chimiothérapie sur la rechute loco-régionale et la rechute à distance, d'où la nécessité d'un modèle de régression avec risques concurrents. L'analyse de la méta-analyse avec risques concurrents a été discutée en utilisant des données agrégées. Cependant, la disponibilité de données individuelles sur les patients entraîne divers avantages dans l'analyse. Étonnamment, aucune directive officielle n'a encore été proposée pour mener une méta-analyse IPD avec des risques concurrents. Pour combler cette lacune, nous avons détaillé : (i) comment gérer l'hétérogénéité entre les essais par un modèle de régression stratifié pour les risques concurrents et (ii) que les mesures habituelles d'hétérogénéité pour évaluer l'incohérence peuvent être facilement utilisées. Nous nous sommes principalement concentrés sur l'extension stratifiée des modèles de risques cause-spécifiques et de sous-distribution, qui sont les modèles de régression les plus populaires pour les risques concurrents. Une méthode de landmark a été introduite pour vérifier l'hypothèse de proportionnalité..

Une méta-analyse combine souvent les résultats d'études qui n'ont pas suivi un protocole commun, impliquant une population différente. La question de l'hétérogénéité est d'une grande importance pour évaluer s'il est raisonnable de résumer l'effet du traitement par une seule estimation globale qui s'applique à toutes les études. La statistique I^2 peut être utilisée pour quantifier l'hétérogénéité entre les essais. Lorsque elle est détectée, des analyses de sous-groupes peuvent être effectuées en utilisant des caractéristiques de niveau individuel. En outre, le temps de suivi (follow up, FUP) pourrait également avoir un impact sur les résultats. On peut se demander si des études avec différentes FUP donneront lieu à des estimations différentes de l'effet du traitement. Nous avons proposé une approche landmark de la fonction d'incidence cumulative. En outre, les caractéristiques individuelles pourraient être utilisées pour étudier les interactions possibles traitement - covariable.

Une légère hétérogénéité a été détectée pour les rechutes à distance, mais pas pour les rechutes locorégionales. L'ajout de la chimiothérapie à la radiothérapie améliore l'incidence cumulative pour les rechutes locorégionales et les rechutes à distance. Nous avons vérifié l'interaction statistique entre l'effet du traitement et l'âge et il n'y avait pas de preuve significative d'interaction pour tous les risques concurrents.

Dans cette thèse, nous avons abordé le problème des données en grappes dans différents contextes. Les exemples étaient les deux des meta-analyse sur données individuels, mais nous avons abordé la structure en grappes des données de différentes manières, en fonction de l'objectif principal de l'analyse. En évaluant la performance d'un biomarqueur sur la

survie globale, nous avons proposé une estimation de la courbe ROC dépendant du temps spécifique à la covariable en traitant la corrélation entre les observations. Cette méthode n'est pas spécifique à la méta-analyse, mais à un cadre plus général où les observations au sein d'un groupe sont corrélées en raison d'un facteur non mesuré. L'autre méta-analyse IPD visait à définir l'effet de l'ajout de la chimiothérapie à la radiothérapie sur la rechute locoréginale et la rechute à distance pour les patients atteints d'un carcinome du nasopharynx. Nous reconnaissons qu'il y a des problèmes d'interprétation pour les risques de sous-distribution. L'utilisation d'un modèle additif pour l'incidence cumulative pourrait fournir des informations qui ne sont pas saisies par le modèle des risques de sous-distribution.

Enfin, la question de la taille des grappes non informatifs a été discutée. Cette hypothèse est souvent ignorée ou prise en compte sans une évaluation formelle. Nous avons proposé un test pour la taille des grappes informatives avec des données de survie censurées. Aucune covariable n'est introduite pour le moment, mais l'utilisation d'un estimateur de Breslow est une extension possible de la méthode. Le test proposé est utile pour identifier ICS avec des données censurées à droite. Nous pensons qu'un indice pour la quantification de l'ICS pourrait également être défini. La différence entre les résultats obtenus pour TOM et AOM est une idée mais, intuitivement, elle ne résoudra pas le problème de la faible puissance pour quelques clusters. La détermination d'un indice défini par la variabilité entre les groupes et à l'intérieur des groupes, où les grappes sont regroupées en fonction de la taille de l'échantillon, pourrait être une autre solution.

Contents

List of Figures

List of Tables

List of Notations and Abbreviations

Notations

- $\alpha(\cdot)$: hazard function
- $h(\cdot)$: subdistribution hazard function
- β': β transpose
- $\mathbb{I}(\cdot)$: indicator function
- $L(\cdot)$: likelihood function
- $\mathcal{L}$: Laplace transform
- $LM\cdot$: landmark time

Abbrevations

- AD: Aggregate data
- AOM: all observed member
- ARR: Absolute Risk Regression
- AUC: Area Under the Curve
- CI: Confidence Interval
- CIF: cumulative incidence function
- cp: coverage probability
- CSH: Cause Specific Hazard
- CTCs: Circulating Tumor Cells
- EM: Expectation-Maximization
- FDA: Food and Drug Administration
- FPR: False Positive Rate
- FPR^D: Dynamic FPR
- FPR^S: Static FPR
- FUP: Follow up
- GEE: Generalized Estimating Equation

- HR: Hazard Ratio
- ICS: Informative Cluster Size
- IPCW: Inverse Probability Censoring Weighting
- IPD: Individual Patient Data
- IT: Inflammatory Tumor
- IWP: Independence Working Model
- NICS: Non Informative Cluster Size
- nonIT: non-Inflammatory Tumor
- PM: Proposed Method
- PSH: proportional Subdistribution Hazard
- PVF: Power Variance Function
- ROC: Receiver Operating Characteristic
- SH: Subdistribution Hazard
- SZ: Song and Zhou
- TOM: typical observed member
- TPR: True Positive Rate
- TPR^C: Cumulative TPR
- TPR^I: Incident TPR
- WCR: Within Cluster Resampling

Chapter 1

Introduction

Clustered data are frequently used in biomedical research. They arise when observations are collected into a number of different groups, referred to as clusters. Observations within a cluster are more alike than observations from different cluster because of genetic factors, persistent environmental characteristics or other determinants. Thus, observations within a cluster are correlated and clusters are considered independent. There are many examples of clustered data in biomedical research: longitudinal data where multiple measurements are taken over time on each individual; family data where observations from members of the same family are considered. Multicenter clinical trials are also common, where observations of the same center (hospital/city) belong to the same cluster. Moreover, meta-analyses combine observations from several randomized clinical trials focused on the same medical question which aims to generate a quantitative estimate of the studied phenomenon, for example, the treatment effect. This is an essential tool for gaining evidence on results but it can be challenging for statistical methodology. In fact meta-analysis may have potential misleading results, particularly if specific study designs, within-study biases, variation across studies, and reporting biases are not carefully considered [2]. However, it leads improvement in precision and it allows to answer question which cannot be addressed by individual studies.

Statistical methods for clustered data take into account the correlation between observations within clusters. There are two main classes of methods: marginal models and random effect models. The former estimate the population-average effect and no specification on the dependence structure is made. The latter estimate the cluster-specific effect and assumptions on the distribution of the dependence between observations are made. These two classes differ both in statistical definitions and interpretations of results. In the first part of this work we mostly discuss random effect models, in particular shared frailty models. The inclusion of frailties in survival models either models the dependence in clus-

tered data or explains the lack of fit of univariate survival models, like deviation from the proportional hazards assumption. The former case is considered, where the frailty represents the unobserved factors that are specific to the clusters and acts multiplicatively on the hazard function. Failure times within cluster are assumed to be independent given the frailty. One challenging point of frailty model is the choice of the frailty distribution. Most theoretical results have focused on the gamma distribution for the frailty. However, other distributions have been proposed [3, 4, 5].

Here we consider clustered survival data, and we mainly focus on individual patient data meta-analysis. Specifically, motivated by the IMENEO study, we explore the problem of biomarker validation by meta-analysis estimating covariate-specific ROC curve and AUC. It is an IPD meta-analysis conducted to validate the prognostic performance of circulating tumor cells in nonmetastatic breast cancer on overall survival. We point out some differences in results interpretation that arises with clustered data, and we discuss the problem of informative cluster size that is characterized by the dependence between the cluster size and the outcome. We underline the issues linked to this setting and we propose the first test for informative cluster size with survival data.

An other motivating data consists in an IPD meta-analysis conducted to assess the effect of adding chemotherapy to radiotherapy to patients with nasopharyngeal carcinoma on multiple endpoints. We consider the competing risks framework and we discuss the methods that can be employed in the analysis of individual patient data meta-analysis. We address the problem of heterogeneity and interpretation of results obtained by competing risks regression models extended to clustered data.

1.1 Meta-analysis in breast cancer

The International MEta-analysis of breast cancer NEOadjuvant CTC (IMENEO) study is an individual patient meta-analysis whose aim was to evaluate the prognostic detection of circulating tumor cells (CTCs) at different time points on overall survival in nonmetastatic breast cancer [6]. Data of 2156 women were collected in 16 different center, conducting 21 studies. Patients were treated with 4 to 12 cycles of neoadjuvant chemoterapy and CTCs were measured at different time points. IMENEO data provide information about center, patients, tumor subgroups, lymph nodes status, tumor stage, chemotherapy and surgery. After a preliminary analysis about CTCs and possible correlated variables, the main interest was restricted to tumor stage which affects both the outcome (survival) and the number of CTCs.

The number of circulating tumor cells appeared to be a promising biomarker for

metastatic breast cancer [7]. However, metastatic patients are a minority in breast cancer, thus study the performance of CTCs in non-metastatic context was of interest . Our objective was to assess the ability of CTCs in distinguishing patients who experienced a specific event (e.g. death) up to time t from patients who did not experience the event. In particular, clinicians agreed that it is more relevant to validate the biomarker in the subgroup of patients with same tumor stage. Therefore, we estimate the covariate-specific time dependent ROC curve and its AUC. We consider the CTCs count at baseline since no difference was detected when analysing the CTCs at several time points.

1.1.1 Statistical issues

The ROC curve is a mandatory tool to determine the performance of a biomarker. When a covariate that affects the outcome and the biomarker the covariate-specific ROC curve is of fundamental importance [8]. It considers covariate-specific threshold to discriminate individuals and thus it assesses the performance of the biomarker in sub-population of patients with similar risk based on the covariate values.

Numerous statistical methods have been developed in the literature to estimate the time dependent ROC curve. Nevertheless, these methods cannot assess the performance of a biomarker with clustered survival data. The semiparametric approach introduced by Song and Zoung jointly modeled the Cox regression model for the survival times and a location model for the biomarker distribution [9]. It estimated the covariate-specific time dependent ROC curve for continuous covariates, but it did not consider a possible correlations between observations. When data are clustered, the model assumptions do not hold and the estimated ROC curve is biased (Chapter 3). We extend the method to clustered survival data introducing a frailty term in the Cox model [10] to model the within cluster correlation.

There are several definition of the time-dependent ROC curve depending on the definition of cases and controls [11]. We refer to the cumulative-dynamic ROC curve because it is more appropriate for clinical decision making.

The proposed method assumes an homogeneous biomarker among clusters and non informative cluster size. These assumptions seem to be reasonable in the IMENEO data and, more in general, in the context of a meta-analysis where the outcome is likely to be independent to the cluster sample size.

1.2 Informative cluster size

A challenging issue that is often ignored in clustered data is informative cluster size, namely the conditional expected value of the outcome given the covariates depends on the cluster sample size. An example of clustered data with a survival outcome is found in a lymphatic filariasis which is often characterized by one or more nests of adult filarial worms in the scrotum [12]. The outcome of interest is the nest-specific time from treatment administration to clearance of the worms, knowing that a treatment is effective when it kills the worms in all of the nests. The cluster is the individual and the cluster size is the number of nests in each patient. Clearing a nest of worms in patients with multiple nests was longer than in patients with one nest, indicating the presence of informative cluster size.

Standard methodologies for clustered data do not take into account the informativeness of cluster sample size, thus appropriate methods have been introduced in these decades. Hoffman first mentioned about informative cluster size and proposed the within cluster resampling (WCR) approach[13]. Successively, this method was studied in the context of correlated survival data and in addition of its computational cost, it might be unstable under heavy censoring [14]. Moreover, for survival data, Williamson et al. described a weighted proportional hazard model [12] .

Williamson et al. suggested that there are two marginal analyses of interest in this setting: one for the population of all observed members and one for a typical member of a typical cluster [15]. Inference for the population of all observed member is obtained by GEE with independence working matrix, whereas for the typical member of the typical cluster, the WCR methods or cluster-specific models need to be employed. Under non informative cluster size results for the two populations coincide [16]. Under informative cluster size they differ in general and, for the population of all observed member it is challenging to generalize the results because they are specific to the collection design of the data. Moreover, improperly assuming informative cluster size results in loss of efficiency [17]. Thus, detection of informative cluster size plays an important role in the choice of the method. We propose a test for informative cluster size with right censored survival data. To our knowledge, this is the first test in survival for ICS. Nonetheless, Benhin et al. [17] and Nevailenen et al. [18] proposed two different tests in the context of logistic and linear regressions.

The proposed test is applied to some illustrative data set on periodontal disease, multicentric study of patients with liver disease and clustered data of individuals with metastatic cancer treated by immunotherapy.

1.3 Meta-analysis with competing events

Patients with nasopharyngeal carcinoma are often diagnosed with locally advanced stage because of the difficulty detection of this tumour. Moreover, surgery is limited to biopsy for histologic confirmation because of anatomical proximity to critical structures. Hopefully, this cancer is highly radiosensitive and chemosensitive. Radiotherapy is the standard treatment and chemotherapy has been proposed for further improvement for patients with advanced locoregional disease.

An individual patient data meta-analysis on nasopharyngeal carcinoma was considered to assess the benefit of addition of adjuvant chemotherapy to radiotherapy[19]. A total of 4940 patients were collected in 23 trials with median follow-up of 11.8 years (ranging from 5 to 22 years). Baseline characteristics of patients are provided in the data, such as the age of patients at diagnosis. A previous analysis on the chemotherapy benefit on overall survival and progression free survival was conducted. Here, we consider the effect of chemotherapy addition on locoregional relapse, distant relapse and death without relapse. Thus, competing risks models are needed for the analysis.

1.3.1 Statistical issue

Meta-analysis are increasingly popular in medical research where information on efficacy of a treatment is available from a number of clinical trials with similar treatment protocols. Standard meta-analysis consider aggregated data because it combines results from published works. However, individual patient data meta-analysis has several advantages for heterogeneity detection and investigating possible treatment interactions.

The analysis of a meta-analysis with competing risks has been already discussed using aggregated data [20]. One principal issue of meta-analysis is heterogeneity because of the inclusion of studies that could have been conducted under different conditions. The availability of individual patient data allows to perform subgroup analyses which are useful for investigating various sources of heterogeneity and check for treatment interactions. IPD meta-analysis was considered in [21] and [22] with survival outcomes. Here, we propose a guideline for the analysis of IPD meta-analysis with competing endpoints. We detail (i) how to handle the heterogeneity between trials via a stratified regression model for competing risks and (ii) that the usual metrics of inconsistency to assess heterogeneity can readily be employed.

Unlike aggregated data meta-analysis, the choice of the possible models lies in fixed or random effect and one- or two-stages approach. Burke et al. pointed out differences and similarities between these methods [23]. A previous analysis of the data employed

a two-stage fixed effect-model for the progression free survival and overall survival [19]. A stratified Cox model is usually used for individual patient meta-analysis with survival endpoints when applying a one-stage approach [23]. Both stratified Fine-Gray model [24] and cause-specific model are described more in details and methods to check proportionality assumptions across trials are provided. Finally, heterogeneity detection and treatment interactions investigations are discussed introducing a landmark method to analyse time-dependent treatment effect.

1.4 Structure of the manuscript

In biomedical research, the use of clustered data and, in particular of individual patient data meta-analysis is constantly increasing. We extended classic methodologies to clustered data. We discuss interpretation of results underling which quantities can be estimated and under which conditions.

The thesis is organised as follows. Chapter 2 introduces principal definitions and methods for analysing clustered survival data. The model families and the corresponding estimation methods are presented. The method of shared frailty models is considered in detail. In Chapter 3, we introduce the covariate-specific time dependent ROC curve and we propose a new method to estimate the cumulative/dynamic ROC curve and its AUC. We illustrate the application to non metastatic breast cancer to assess the discriminatory capability of circulating tumor cells. In Chapter 4 we consider the problem of informative cluster size and we propose a new test for ICS with right censored clustered survival data. A simulation study and illustrative data examples are described.

In Chapter 5 we provide a guideline for the analysis of individual patient data meta-analysis with competing risks. We describe the most common regression models for competing endpoints and we refer to methods for quantifying heterogeneity and determining treatment interactions with patient-level covariates.

Chapter 6 concludes with a discussion about the proposed methodologies and limitations and areas for further work.

All the developments are implemented in R, with ready-to-use functions that can be found on the link: https://github.com/AMeddis.

Chapter 2

Analysis for correlated survival data

2.1 Survival analysis

Survival analysis is a collection of statistical methods to analyse data where the outcome of interest is a time-to-event. Although more than one event may be considered in the same analysis, we assume that only one event is of interest. When more than one event is considered, the statistical problem can be characterized as a competing risk problem that is discussed in Chapter 5.

Let T represents the time-to-event, the survival function $S(t)$ is the probability of not experiencing the event prior to time t, $S(t) = 1 - F(t)$, where $F(t) = \mathbb{P}(T \leq t)$ is the cumulative distribution function of T. We introduce the hazard function $\alpha(t)$ which represents the instantaneous probability that the event occurs conditional on not having experienced the event by time t:

$$\alpha(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(T \in [t, t + \delta t) | T > t)}{\delta t} = \frac{f(t)}{S(t)}$$

with $f(t)$ density function of T. We also define the cumulative hazard function $A(t) = \int_0^t \alpha(s)ds = \int_0^t \frac{dF(s)}{S(s)} = -\ln(S(t))$.

In survival data, it is likely to have censored information, notably the outcome of interest, for some individuals, is not observed. Here, we always refer to right censoring. Let C be the censoring time (e.g. time of drop-out of the study, time of end of follow-up), a subject j is censored when $C_j < T_j$. In particular, we define $\tilde{T}_j = \min(T_j, C_j)$, the observed failure time, and the censoring indicator $\Delta_j = \mathbb{I}(T_j \leq C_j)$. We assume that T_j and C_j are independent.

For estimation of the cumulative hazard and the survival function, appropriate (semi)parametric or nonparametric model are available. For parametric inference, one needs to make some

assumptions on the time-to-event distribution; on the other hand, nonparametric methods need larger sample size to obtain reliable results and estimating the hazard function is challenging.

Given N individuals, we consider the counting process $N(t) = \sum_{j=1}^{N} \mathbb{I}(\tilde{T}_j \leq t, \Delta_j = 1)$ with intensity $\lambda(t)$, in particular:

$$\lambda(t)\delta t = \mathbb{E}[j : T_j \in [t, t+\delta t), \Delta_j = 1 | T_j > t] = Y(t)\alpha(t)\delta t$$

where $Y(t) = \sum_{j=1}^{N} \mathbb{I}(\tilde{T}_j > t)$ is the at-risk process which represents the number of subjects that are still at risk before t. The quantity $N(t)$ corresponds to the number of events observed before time t, and $\Delta N(t) = N(t) - N(t-)$ is the number of events that occurred at instant t. We introduce the counting process martingale $M(t) = N(t) - \Lambda(t)$ or, equivalently

$$dN(t) = d\Lambda(t) + dM(t) = \alpha(t)Y(t)dt + dM(t)$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ is the cumulative intensity process.

The nonparametric Nelson-Aalen estimator of the cumulative hazard function is defined as:

$$\hat{A}(t) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{j=1}^{N} \frac{\Delta N(T_j)}{Y(T_j)} \mathbb{I}(T_j \leq t)$$

Moreover, for right censored data, a nonparametric estimator of the survival function is given by the Kaplan-Meier estimator [25]:

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{\Delta N(T_j)}{Y(T_j)}\right)$$

This is a step-wise decreasing function that jumps at the event times.

Proportional hazard models are widely used (semi)parametric models in survival analysis. Let X be a vector of p covariates, the hazard is expressed as:

$$\alpha(t) = \alpha_0(t) exp(\beta' X)$$

where $\alpha_0(t)$ is the baseline hazard function and β is the vector of regression coefficients. The covariates act multiplicatively on the baseline hazard and the quantity $\exp(\beta)$ is referred to as hazard ratio. Common choices for the baseline hazard are Weibull, exponential or Gompertz distribution, but the model allows for any specification of the baseline. When the baseline is unspecified, the Cox proportional hazard is considered and

partial likelihood methods are used for the estimation of the β, and the Breslow estimator for the baseline hazard.

These quantities describe the aspects of survival data which build the basis for specific parts in the next chapters where we focus on modeling clustered right censored failure times data in different framework. In section 2.2 we give useful insights on the approaches exploitable on survival for clustered data and in section 2.3 we describe more in details the shared frailty models considering model definitions and estimation. Finally, in section 2.4 we point out some differences in the interpretation of results.

2.2 Inference for correlated survival data

Correlated survival data arise from many contexts due to recurrent events experienced by an individual or when observations are clustered into groups. For instance, studies on survival of a specific disease with familial data, or the assessment of a treatment strategy effect where data are collected from different centers (multi-centric data). An other example of clustered data is the meta-analysis which gains evidence for clinical interpretation combining results of multiple studies (clusters). Alternatively, the response may be repeatedly measured on each subject at several time occasions (repeated measurements). For example, in a clinical trial a measure of health outcome is recorded for each patient (cluster) at each visit, creating a vector of responses with natural time ordering among the measurements. In the latter scenario we specifically refer to longitudinal data. In this work we do not focus on longitudinal data. We assume that observations are grouped into clusters.

Observations which belong to the same cluster tend to be correlated because of some common shared features. Analysis of such data is challenging and ignoring the intracluster correlation leads to biased results. The estimation methods for analysing clustered data and the interpretation of regression estimates tend to be more complicated than in the independent setting. There are two broad classes of models that have been developed to handle clustered data: i) marginal or population-averaged models and ii) random effects or frailty models. These two approaches differ in both statistical approach and interpretation. Marginal models make inference on the population average effect addressing the correlation between failure times, but they do not model the correlation, thus no information on the relationship among failure times is provided [26, 27, 28]. On the contrary, frailty models include a random effect to account for the dependence between failure times and they make inference on cluster-specific effects [10].

Let $(G_1, G_2, ..., G_K)$ be a sample of K independent observations where each G_k repre-

sents the observed within each cluster k and consisting of

$$\big(N_k, (\tilde{T}_{k1}, \Delta_{k1}, X_{k1}), ..., (\tilde{T}_{kN_k}, \Delta_{kN_k}, X_{kN_k})\big)$$

with N_k the cluster sample size. Let X_{kj} denote the vector of covariates for the j-th subject in the k-th cluster. Let T_{kj} be the failure time and C_{kj} the censoring time, we observe $\tilde{T}_{kj} = \min(T_{kj}, C_{kj})$, the observed failure time, and the censoring indicator $\Delta_{kj} = \mathbb{I}(T_{kj} \leq C_{kj})$ for individual j in cluster k. We assume that T_{kj} and C_{kj} are independent for all k, j and that in each cluster k $(T_{k1}, T_{k2}, ..., T_{kN_k})$ can be correlated conditionally on $(X_{k1}, X_{k2}, ..., X_{kN_k})$.

In the following sections we describe more in details the two classes of methods, recalling that it is fundamental to choose the method to use depending on the question we want to address and not on statistical considerations.

2.2.1 Marginal models

Marginal modeling estimates the effect of explanatory covariates considering the marginal distribution of the outcome of interest. This type of models focus on the population average effect, and the correlation is often treated as a nuisance parameter to reduce the dependence of the marginal models on the specification of the unobserved correlation structure of the data. In fact, the dependence is not the interesting aspect and is not considered in detail. The regression coefficients estimates are found assuming independence between the observations. Afterwords, the uncertainty of the regression coefficient estimates is evaluated by means of an estimator that accounts for the dependence (a "sandwich estimator"). This approach is called "Independence Working Model (IWM) approach" and is closely related to the generalized estimating equations (GEEs). The correlation is properly modeled in order to assign weight to the data from each cluster specifying a working correlation matrix. The working correlation is assumed to be the same for all individuals, reflecting average dependence among the correlated observations within the cluster. Many working correlation structures can be specified: independent working correlation assumes no correlations between observations; an exchangeable working correlation assumes uniform correlations. An autoregressive working correlation assumes that observations are only related to their own past values through first or higher order autoregressive (AR) process.

This marginal method has been applied to the proportional hazards model where the baseline hazard function is common for all the failure times [26], or a stratified approach with different baseline hazard functions among cluster [27]. The regression parameters are

obtained by the partial likelihood function considering the observations to be independent (IWM assumption) and the corresponding variance-covariance estimators are properly corrected to account for the dependence structure. Moreover, Cai and Prentice proposed weighted procedures to estimate regression parameters under stratified and unstratified marginal proportional hazards model, respectively [28, 29].

Marginal modeling does not need any condition or assumption for the dependence distribution, but, on the other hand, this leads to uncertainty on the good specification of the model in practice. Therefore, this approach is advantageous to determine the covariates effect, but it is not useful for goodness-of-fit or for prediction. A complementary approach is the concept of copulas whose main aim is to study the dependence by assuming that the marginal distributions are known and a uniform distribution on the unit interval is considered. The dependence is then evaluated by specifying a family of distributions for the bivariate observations (correlated failure times). This approach is not important from a statistical point of view since the marginal distributions are rarely known in practice. Typically, there will be some parameters also in the marginal distributions and then we need a larger model for the analysis. This approach can be used for assessment of the dependence and for evaluating the goodness-of-fit of specific models [10].

2.2.2 Random effect models/frailty models

A frailty model is a random effects model for survival data. It is a conditional hazard model with a multiplicative factor, the so-called frailty, which models the correlation between observations. This model was introduced by Clayton in a study on chronic disease incidence in families [30], but the term frailty was introduced by Vaupel [31]. This approach aims to account for heterogeneity, caused by unmeasured covariates. It can be applied to describe the influence of unobserved factors in a proportional hazard model for univariate (independent) data. However, it is mostly used in case of multivariate (dependent) survival data to account for the dependence in clustered event times (e.g. multicentric clinical trials, recurrent events). For clustered data, the estimated variance of the frailty term summarizes unobserved heterogeneity between clusters; for recurrent events, the variance describes unobserved heterogeneity between individuals, as in the univariate case. The idea is to consider the variability in failure times as coming from two separate sources. One source is described by a hazard function (simple randomness), and the second one is described by a random effect which is either an individual variable (univariate), or a variable common to several individuals (multivariate).

One challenging point of frailty models is the definition of the frailty distribution. Several choices have been studied in detail in [4], underlying which distributions generate

specific type of dependence between observations within clusters. One other difficult point is the estimation of the regression coefficients when the baseline hazard is not specified. In case of parametric models, the hazard function is specified and the marginal function is obtained integrating out the frailty terms. Meanwhile, for semiparametric models, the EM algorithm and penalized likelihood techniques are needed to estimate the regression coefficients.

The basic probability results are shared between the multivariate and the univariate model, since the first includes the second. However, the statistical aspects, including interpretation, identifiability of parameters and estimation, are clearly different. It is than natural to consider the two cases separately. Clustered data are the main focus of this work, thus in the next section we give an insight on the shared frailty model: model and estimations methods, and we refer to [4] for a detailed description of univariate frailty models. We then highlight the properties and difference between the possible frailty distributions.

2.3 Shared Frailty models

Shared frailty models account for unobserved cluster characteristics introducing a frailty term shared among observations within the same cluster. The introduction of a random effect is a natural way to take into account of the dependence between observations. Conditional on the frailty, the failure times within a cluster are assumed to be independent. Note that, in case of univariate data the individuals are a random sample from a larger population, meanwhile in clustered data the clusters are a random sample of a population of clusters.

This method is more complex compared to the standard random effect model, since the basic variation is described by the hazard function instead of a random variable. In fact, two source of variation are distinguished: the groups variation which is described by the random effect variability, and the individual variation described by the hazard function. The model has the form:

$$\alpha(t|X, U_k) = U_k \alpha_0(t) \exp(\beta' X)$$

where $U_k > 0$ is the frailty term with density distribution $f_{U_k}(t)$. We refer to $\alpha(t|X, U_k)$ as the conditional hazard and when X is a categorical variable with $x \in \{0, 1\}$, the quantity $\exp(\beta^T)$ is the hazard ratio between individuals with the same frailty. The marginal effect of X can be constructed by the ratio of the two marginal hazards $\alpha(t|x = 0)$ and $\alpha(t|x = 1)$.

The marginal survival function is obtained integrating out the frailty from the conditional function $S(t|X, U_k) = \exp(-U_k A(t|X))$:

$$S(t|X) = \int_0^\infty S(T|X, U_k) f_{U_k}(u)du = \int_0^U \exp(-U_k A(t|X)) f_{U_k}(u)du = \mathbb{E}[\exp(-U_k A(t|X)].$$

This integration is the same as used in the Laplace transformation for the distribution of U_k. The Laplace transform is $\mathcal{L}(s) = \mathbb{E}[\exp(-sU_k)]$. Thus, we can rewrite the survival function as the Laplace transform of the frailty distribution:

$$S(t|X) = \mathcal{L}(A(t|X))$$

The gamma distribution is widely used in frailty model since the Laplace transform is computationally easier. However, in general, a family of distributions with tractable Laplace transform are considered for the choice of the frailty distribution [3].

The conditional likelihood for all the individuals is given by the product over the clusters of the likelihood contribution for each cluster k:

$$\begin{aligned} L_k(\beta, \alpha_0|U_k) &= \prod_j \alpha(t_{kj}, x_{kj}|U_k)^{\Delta_{kj}} \exp(-U_k A(t_{kj}, x_{kj})) \\ &= \prod_j \alpha(t_{kj}, x_{kj}|U_k)^{\Delta_{kj}} \times \exp(-U_k A_{k.}(x_{kj})) \end{aligned}$$

where $A_{k.} = \sum_j A(t_{kj}, x_{kj})$ is the sum of the conditional cumulative hazards of cluster k. Let $N_{k.}$ be the total number of events for cluster k, the marginal likelihood is obtained integrating out the frailty term with $f_{U_k}(u_k; \theta)$ and it is of the form:

$$\begin{aligned} L(\beta, \alpha_0, \theta) &= \prod_k \prod_j \int_0^\infty [u_k \alpha(t_{kj}, x_{kj})]^{\Delta_{kj}} \times \exp(-u_k A(t_{kj}, x_{kj})) \times f_U(u_k)du \\ &= \prod_k \prod_j \alpha(t_{kj}, x_{kj})^{\Delta_{kj}} \times \mathbb{E}[U_k^{N_{k.}} \exp(-U_k A_{k.})] \end{aligned}$$

Using the Laplace trasform and its derivatives, the term $\mathbb{E}[U_k^{N_{k.}} \exp(-U_k A_{k.})]$ is easily calculated for $U \sim Gamma(\theta)$, but it can be challenging for others frailty distributions. In the next section we provide a brief description of the different estimating methods that can be employed in case of frailty models.

2.3.1 Inference for shared frailty models

Several approaches have been proposed to estimate the parameters from a shared frailty model because some formula are complicated and iterations can be time consuming. The most obvious way would be to integrate out the frailty, but this is not the only method. Above all, we need to distinguish between parametric and semiparametric models, depending on the definition of the baseline hazards $\alpha_0(t)$. For parametric models, standard methods can be used maximizing the log-likelihood to estimate the regression coefficients. As described before, the likelihood can be obtained by the Laplace transform and we might need numerical differentiation methods to calculate the derivatives of the Laplace transform. For semiparametric models, the baseline hazards $\alpha_0(t)$ is estimated per time point with observed events, thus there is one parameter for each observed failure time.

The Expectation-Maximization (EM) algorithm has been proposed for semiparametric models with both gamma and power variance function (PVF) distributions [32, 33]. This method alternates between the Expectation step (E-step) and the Maximisation step (M-step) until convergence of estimates. During the E-step the expected log-likelihood $\sum_k \mathbb{E}[logL_k(\beta, \alpha_0|U_k)]$ is calculated, and in the (M-step) β, α_0, θ are obtained maximizing the log-likelihood. This last step is the same problem as for a Cox model, considering the frailty term as $\exp(\psi_k log(U_k))$ with $\psi_k = 1$. The EM-algorithm is simple but it can require a very large number of iterations. To obtain the standard errors of the estimates, the Louis' formula can be used [34].

A modified EM-algorithm, the "profile EM", is an alternative approach, where the EM algorithm is performed for fixed values of θ using a two-stage maximization procedures:

$$\max_{\theta,\beta,\alpha_0} L(\theta, \beta, \alpha_0) = \max_{\theta}\{\max_{\beta,\alpha_0} L(\beta, \alpha_0|\theta)\}$$

Therefore, in the E-step the likelihood for fixed θ is calculated where the estimates for the frailty $\hat{u_k}$ are obtained considering the expectation with respect to the posteriori distribution of the random effect ($\hat{u_k} = \mathbb{E}[U_k|data]$). Afterwords, in the M-step the likelihood is maximized as for a Cox model with $log(\hat{u}_k)$ as offset term in each cluster. A second M-step is needed to maximize the profile likelihood $\hat{L}(\theta) = \max_{\beta,\alpha_0} L(\beta, \alpha_0|\theta)$ over θ. The standard error for β and α_0 are calculates with the Louis formula, considering that θ is fixed. Than, these are adjusted considering the variability of θ [35].

Penalized likelihood methods as described in [36, 37] are also used for semiparametric gamma or log-normal frailty models. It is based on a modification of the Cox partial likelihood where the frailty terms U_k are treated as regular parameters for fixed θ. The likelihood is a product of the partial likelihood and a penalization term which is introduced

to avoid large differences between the frailties for the different groups. In other words, it is fitted by first setting the frailty values to 1 ($\theta = \infty$). Then, an iterative procedure is used with a first step of optimizing the partial likelihood, treating the frailties as fixed and known parameters. In the second step, the frailties are evaluated as the conditional means given their observations, like the EM-algorithm. This is repeated until convergence. These methods are fast, but it is hard to generalize them for others frailty distributions.

2.3.2 Frailty distribution

As mentioned above, several options for the frailty distribution are possible. There is no single family which have all the desirable properties, thus the choice of the distribution depends on the actual problem in consideration. In particular, besides the theoretical properties, it is important how dependence between time variables is translated.

The standard assumption is to use the gamma distribution justifying this choice based on its analytic simplicity and its variety of forms as the parameters vary. The gamma model was considered by Clayton [30] and in [38], and generalized to include covariates by Clayton and Cuzick [39]. From a computational point of view, it fits very well to survival models, because it is easy to derive the marginal quantities. The density of the gamma model is:

$$g(u) = \theta^{\delta} u^{\delta-1} \exp(-\theta u)/\Gamma(\delta) \quad \theta, \delta > 0$$

Because of identifiabiliy of the model, we fix $\mathbb{E}[U] = 1 \rightarrow \delta = \theta$ and thus $Var[U] = 1/\theta$. The conditional distribution of the frailty among survivors is still a gamma distribution with a different scale parameter. However, for frailty distributions belonging to the natural exponential family, the conditional distribution of the frailty is still within the same family. The inverse Gaussian distribution belongs to the natural exponential family, but it gives different results compared to gamma frailty. It was considered in [40].

The positive stable model was introduced by Hougaard [3] and further studied in Oakes [41]. This distribution is characterized by infinite mean and it is usually defined by the Laplace transform: $L(s) = \exp(-\delta s^{\alpha}/\alpha)$ where $\delta = \alpha$. When the conditional hazards are proportional, so are the marginal distributions. Moreover, if associated to a Weibull hazard, also the marginal distribution is in the Weibull family. Hougaard [3] introduced a group of distributions which include intermediate cases between the gamma and the Inverse Gaussian distributions: the power variance function (PVF). This was successively studied by Crowder [42]. It is a three-parameter distribution family which also includes the positive stable distribution, its Laplace transform is of the form: $L(s) = (-\delta((\theta+s)^{\alpha} - \theta^{\alpha})/\alpha]$. When $\alpha < 0$ some people do not experience the event, thus

a negative estimated parameter reflects that we have very little information on the risk at late time points. A further important frailty distribution is the lognormal distribution. It is a one-parameter model (the mean is fixed to 0), whose Laplace transform does not have a closed form but several packages proposed numerical methods to integrate out the frailty.

One important aspect in the definition of the frailty distribution is the type of dependence of the observed failure times. One consideration is whether the dependence is early or late. In Hougaard [4] the dependence type is described by an example of nine artificial pairs of twins. We consider late dependence when if one twin dies old, we are sure the other will also die old, whereas if one twin dies young, we do not know the actual age class of death of the other one. As an alternative, short dependence when one twin dies young leads that the other will also die young, whereas if one twin dies middle-aged or old, we do not know the actual age class of death of the other one. The positive stable model lead to early strong dependence, meanwhile, the gamma model to stronger late dependence because of the tails of the frailty distributions. The positive stable has a right tail, so we have a strong dependence initially, whereas the gamma distribution has a left tail corresponding to late dependence. However, intermediate case are usually more realistic and are represented by PVF and lognormal model.

Another consideration is the duration of dependence, which we split into three time frames - instantaneous, short-term, and long-term dependence. Instantaneous dependence occurs when two events happen at the same time. Short-term dependence is when the dependence is most pronounced immediately after other individuals in the group have experienced an event, and long-term dependence is when an event implies that the risk among group members is increased forever. Most standard models give long-term dependence and a few give instantaneous dependence.

Indeed, in practice mainly the gamma distribution and the lognormal distribution are used to model the frailty term and most of the software limits the choice of the frailty distribution to these cases. The choice of the frailty distribution is a challenging point. Shih and Louis proposed a graphical method for assessing the gamma distribution assumption when the basic functions are parametric and do not depend on covariates [43]. Glidden developed a test for the gamma frailty model without specifying the basic hazard functions when covariates are not involved [44]. Cui and Sun provided a graphical as well as a numerical method for checking the adequacy of the gamma under the marginal proportional hazards [45]. We refer to those papers for general discussions on this topic, more simulation results and advice about related model checking that we believe are also relevant in our context. Misspecification is then an important issue for the defini-

tion of the frailty distribution. In Chapter 3, in the specification of a new method for covariate-specific time dependent ROC curve, this problem is addressed by conducting a simulation study generating data with frailty distributions that do not belong to the natural exponential family.

2.3.3 Software available

Support for frailty model exists in several packages in `R`. The most popular fitting method for shared frailty models is via the penalized likelihood method [37]. This is implemented in the `survival` package [46]. Lognormal frailty models is estimated in `R` via Laplace approximation in `coxme` [47], h-likelihood in `frailtyHL` [48] or Monte Carlo Expectation-Maximization `phmm` [49]. Parametric and spline based shared frailty models are implemented for the gamma and log-normal distributions in the `frailtypack` package [50].

The `frailtySurv` package [51] implements the PVF distributions except the positive stable via a pseudo full likelihood approach. The `parfm` package [52] estimates fully parametric gamma, Inverse Gaussian, Positive stable and log-normal frailty models. In our work, we use the `frailtyEM` package [53] which provides maximum likelihood estimation of semiparametric shared frailty models using the Expectation-Maximization algorithm. In this package, a general full-likelihood estimation procedure is implemented for the gamma, positive stable and PVF frailty models, using a semi-parametric Breslow estimator for the baseline intensity.

2.4 Target population

Methods available for clustered data can be classified in two main classes: population-averaged and unit-specific approach. The marginal models, which relies on the GEE, provide results for population averaged effects, whereas the random effect models (frailty) estimate the individual-specific effect. Thus, we can define two marginal analyses that might be of interest in the context of clustered data, as mentioned in [13, 15]. One makes inference for the population of all observed members (AOM), where equal weights are given to each member of the observed population, and larger clusters are weighted more than smaller ones. The second one, makes inference for the population of typical observed members of a typical cluster (TOM), where equal weights are given to units within cluster and clusters equally contribute to inference since they have same weights. For the former, the parameters will have an interpretation for a unit randomly sampled from the overall observed populations. The latter will have a cluster-based interpretation, namely for a randomly selected unit sampled from a randomly selected cluster. Asymptotically the two

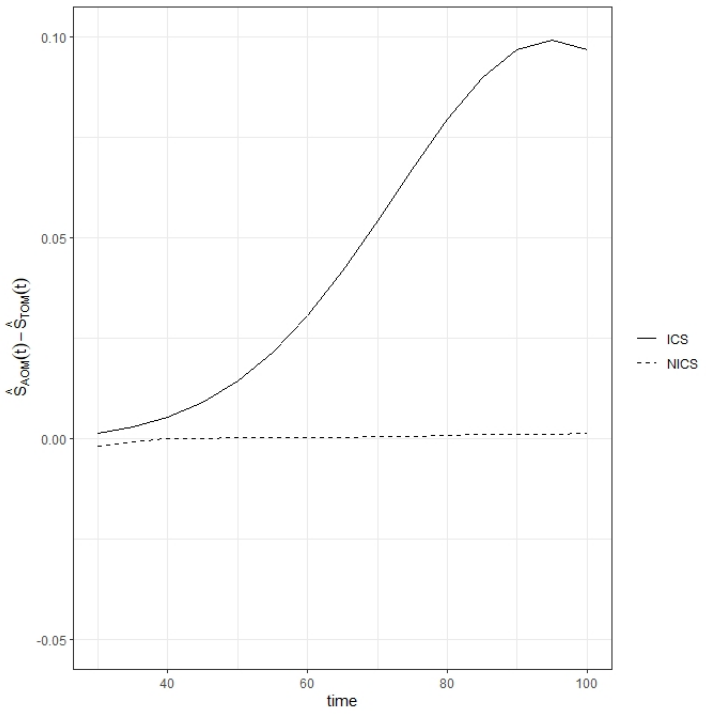

Figure 2.1: Difference of Survival function estimates for the two target population under ICS and NICS. Simulation results over 500 replications.

marginal analyses will reach the same conclusion if cluster size is unrelated to the outcome [16]. However, they differ in presence of informative cluster size, i.e. when the outcome measured among cluster members is related to the size of the cluster. In Figure 2.1 the difference of survival function estimates for the two target population is provided. Under NICS, the results obtained from the two marginal analyses coincide. On the other hand, under ICS, the two estimates are quite different. Hoffman first described the problem of informative cluster size proposing the within-cluster resampling (WCR) method [13]. Others approaches have been presented in these years.

Williamson et al. provided a guideline as to which population should be selected for inference according to the aim of the analysis [15]. A periodontal disease example is considered, where data on the disease status of the tooth (unit) from a sample if individuals (clusters) are analyzed. We expect that observations from the same individual are correlated and that subjects with fewer teeth are more likely to have worse dental health. Therefore, the cluster size is informative. If the goal is to assess how many teeth among the observed patients require a costly intervention, the population of all members analysis is privileged, since clustering by patient may not be of direct relevance. On the other hand, if we are interested in determining patient factors linked to the disease status

of teeth, the population of typical cluster members might be preferred.

A formal definition of informative cluster size is given in Chapter 4 where we detail the methods that can be employed in this setting and we introduce a test for informative cluster size with survival data.

Chapter 3

A covariate-specific time dependent ROC curve for correlated survival data

3.1 Introduction

Considerable research has focused on the development of new biomarkers to improve patient management in disease like cancer. An essential step in developing a clinically useful biomarker is to identify its ability in discriminating subjects at high or low risk of an event within the coming years. In survival studies, the time dependent receiver operating characteristic (ROC) curve is a popular tool to assess the performance of a candidate biomarker. It is the plot of time dependent True Positive Rate (TPR), or sensitivity (probability of the biomarker being above a given threshold in the diseased subjects) against time dependent False Positive Rate (FPR), or 1-specificity (probability of the biomarker being above the given threshold in the non diseased subjects) among all the possible thresholds used to classify individuals. Several definitions of the time-dependent ROC curve were proposed by Heagerty et al. [54] depending on the definition of cases and controls. In this work, we refer to the so called cumulative/dynamic ROC curve where, at time t, a patient is defined a case if he experiences the event in $[0; t]$, and a control if he experiences the event after time t. Furthermore, it has been advocated that adjusting for well established prognostic variables is important in the clinical interpretation of the results [8, 55] . Therefore, in addition to the time dependent ROC curve, the covariate-specific ROC curves are of prime interest in the study of the discrimination of a biomarker. For a discrete covariate, the most obvious option is to use a nonparametric approach. Uno et al. [56] proposed a nonparametric method based on the inverse probability

censoring weighting (IPCW) estimates, where the covariate-specific time dependent ROC curves can be obtained by stratifying on the covariate values. While, for a continuous covariate, Song and Zhou employed a semiparametric approach for the covariate-specific time dependent ROC curves [9]. They assumed a proportional hazards model for the hazard given the biomarker and covariates, and a semiparametic location model for the conditional distribution of the biomarker given the covariates.

To ensure reliable evidence of the biomarker prognostic capability, large multi-center trials or individual patient data meta-analysis are often conducted. An example is IME-NEO, a meta-analysis on individual patient data assessing the clinical usefulness of CTCs (Circulating Tumor Cells) count in a context of non metastatic breast cancer [6]. The CTCs, the candidate biomarker, were collected from multiple centers, thus it is legitimate to suspect various sources of heterogeneity. This might lead to correlation between observations coming from the same cluster. Thus, the natural question arises as to how to consider the cluster effect in the discriminatory analysis for a candidate biomarker. It is not clear yet how to address this problem and, which quantities can be estimated and under which assumptions, has not really been discussed so far. Common strategy to evaluate the discriminatory ability of the biomarker is to ignore heterogeneity, discarding any possible cluster effects, but it may lead to an incorrect evaluation. We propose an estimator of the covariate-specific time dependent ROC curve for correlated censored survival data which can simultaneously address all the challenges of our motivating data: adjusting for clinically useful covariates and allowing for clustered data. We compare the proposed method to the nonparametric one and we discuss the interpretation of the estimates and their consistency under different scenario.

In the next section the covariate-specific time dependent ROC curves and the respective area under the curve (AUC) are introduced. We recall the several definitions due to the time-varying framework and we detail existing method used to estimate covariate-specific time dependent ROC curves. In Section 3.4 a new estimator is described. We detail the simulation study conducted to evaluate the performance of the proposed method comparing it with the existing ones. The method is then applied in Section 3.5 to non-metastatic breast cancer data to assess the discriminatory ability of CTCs on overall survival. Some remarks are made in Section 3.6.

3.2 Motivating data

3.2.1 Biomarker of interest: CTCs

Breast cancer leads the incidence and mortality tables for cancers among women in 2018, it was responsible for an estimated 2.1 million cancers accounting for the fifth leading cause of cancer deaths worldwide.

The Circulating Tumor Cells (CTCs) detection proved to be a prognostic factor in metastatic breast cancer [7]. These are tumor cells deriving from primary and secondary sites that were discovered in 1869 by Thomas Ashworth, and started gaining interest in 1990s with the demonstration that CTCs exists prematurely in the course of cancer. A pros in the use of CTCs is that their detection is a non invasive technique because measured in the sample of patients blood (liquid biopsy). Different methods for isolating CTCs have been proposed but CellSearch was the only one approved by FDA (Food and Drug Administration). This method consists in counting the epithelial cells separated from the blood by magnetic technology using ferrofluids. However, interreader variability is often present in the counting step of CTC, which involves image recognition by a trained technician. An image analysis algorithm has been developed to fully automatize CTC counting and to improve interreader reproducibility [57].

CTCs detection is correlated with the tumor stage [58], thus large number of CTCs are detected in metastatic cancer, and in the non-metastatic setting, they are more frequent in inflammatory tumors. CTCs count facilitates the prognosis and improve cancer treatment . Originally CTCs studies focused on metastatic settings, but in breast cancer metastatic patients are a minority and so groups of research started to study CTCs prognostic significance in non-metastiatic context. As CTCs are rare in early breast cancer, technical and statistical concerns were initially raised about their validity as biomarkers. These initial theoretical concerns have been largely invalidated in several large studies that established CTC detection as a reliable and valuable biomarker of the metastatic process. Since 2004 several studies for the clinical validity of CTCs in metastatic breast cancer patients were conducted with the result that it is a prognostic biomarker, both at baseline (moment of diagnosis) and during treatment.

3.2.2 IMENEO data set

Neoadjuvant chemotherapy is a standard treatment for patients with non-metastatic breast cancer before surgery. It has two main actions: shrinkage of the primary tumor and eradication of blood-borne tumor cell dissemination. Here, we consider the International MEta-analysis of breast cancer Neoadjuvant CTC (IMENEO): a meta-analysis of

Center	N_k	events	mean(CTCs)	IT	nonIT
Brussels	44	3	1.00	1	43
Cremona	45	4	7.73	0	45
Gunma	112	14	1.92	0	112
Hamburg	602	72	2.25	41	561
Houston	161	39	1.53	55	106
Kyoto	72	5	2.40	0	72
London	16	0	2.06	0	16
Madrid	69	16	3.77	1	68
Oslo	120	13	1.82	0	120
Paris	286	54	4.10	155	131
Rotterdam	45	2	1.22	0	45
San Francisco	142	16	2.08	9	133
Santagio	9	0	0.00	0	9
Täbingen	61	8	0.75	3	58
U. Mich.	25	4	1.32	0	25
Valencia	102	17	0.68	1	101

Table 3.1: Desctiption of the IMENEO data.
N_k: number of observations in center k.
CTCs, circulating tumor cells; IT, inflammatory tumor; nonIT, noninfammatory tumor.

individuals with non-metastatic breast cancer treated by neoadjuvant chemotherapy [6]. The main aim of the study was to investigate the prognostic ability of CTCs on overall survival. Data of 2156 women from United States, Japan and Europe were included; inclusion criteria were studies with patients enrolled between 2005 and 2016 with CTCs collected at least once before surgery by the Cellsearch system. Patients were treated with 4 to 12 cycles of neoadjuvant chemotherapy and CTCs were measured at different time points: five of fewer weeks before neoadjuvant treatment, one to eight weeks after the treatment and five or fewer weeks prior surgery (Figure 3.1). The blood sample screened was of 7.5 mL in 19 studies, 15 mL in one study and 30 mL in one study. The CTCs count was normalized on the volume sampled. However, having a count of zero CTCs in the sampled blood does not necessarily imply that no CTCs are in the body.

Information about age, tumor stage, lymph nodes, tumor grade and tumor subgroup at diagnosis were provided. A previous analysis of the IMENEO study showed a statistically

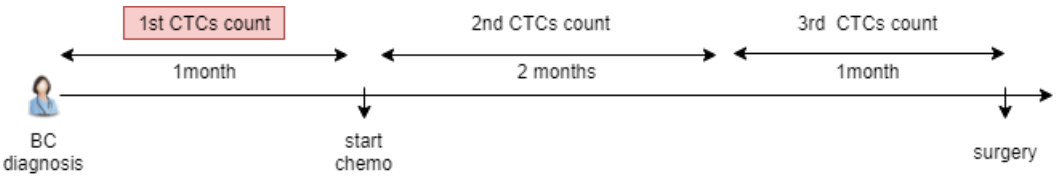

Figure 3.1: Timeline of CTCs collection in the IMENEO study. In the analysis, we consider CTCs before chemotherapy (at baseline).

significant association between between CTCs and tumor stage. In particular high CTC counts were observed in T4d tumors, thus in this work, we distinguish from inflammatory (IT) and noninflammatory tumor (nonIT). No other statistically significant association with any baseline clinical or pathological characteristics have been observed.

The aim of this work is to assess the ability of Circulating Tumor Cells in discriminating patients with non metastatic breast cancer on overall survival. The CTCs count at baseline (before treatment) was taken into consideration, since a prior investigation indicated that CTCs detection during neoadjuvant chemotherapy does not increase survival prognostication [6]. Moreover, because of its particular distribution, in a context of non-metastatic breast cancer, CTCs changes between measurement may not be due to treatment effect [58].

The number of CTCs is characterized by 80% of null values (Figure 3.3). Thus, we consider zero inflated model, in particular, a negative binomial regression model. Moreover we provide the estimated survival function stratifying on biomarker values to show that CTCs count is prognostic for overall survival (Figure 3.2).

3.3 Time dependent ROC curve

ROC curve analysis is extensively used in biomedical studies for evaluating the discriminant capability (power) of a continuous diagnostic test or marker. Given a marker Y, we are interested in its ability in discriminating individuals that experienced a specific event (cases) from the event-free individuals (controls). Let D be a binary outcome which is 1 if the event occurred, and 0 otherwise, the ROC curve is the plot of the true positive rate, or sensitivity, $TPR(c) = \mathbb{P}(Y > c|D = 1)$ and false positive rate, or 1-specificity, $FPR(c) = \mathbb{P}(Y > c|D = 0)$ among all the possible threshold values c. We assume that higher values of Y leads to greater risk of event. The higher the ROC curve is in the quadrant [0,1] x [0,1], the better the marker discriminates subjects. Moreover, the Area Under this Curve (AUC) is a summary measure for the performance of the biomarker.

However, with survival data, where the performance of prognostic marker is of interest,

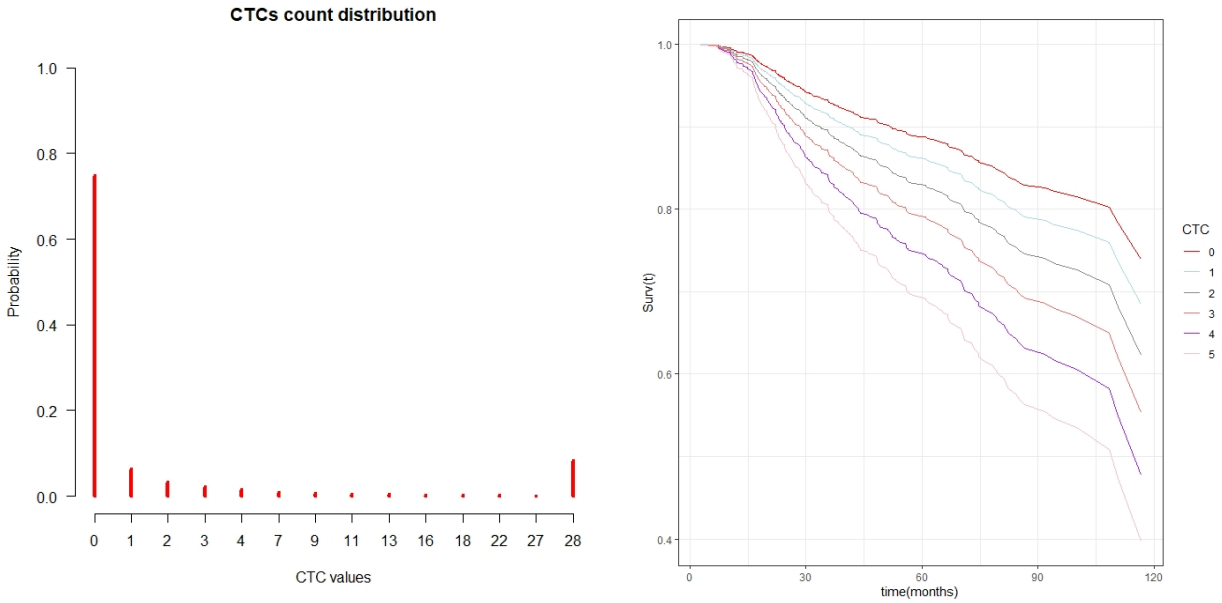

Figure 3.2: Estimated survival function for different value of CTCs.

Figure 3.3: CTCs distribution observed in the IMENEO data set.

the outcome status changes over time. Therefore, sensitivity, specificity and ROC curves are functions of time as well and, several definitions have been introduced [11]. We denote by $D_j(t)$ the time-dependent outcome status for a subject j at time t, $D_j(t) = 1$ if subject j is considered as a case and $D_j(t) = 0$ if subject j is considered as a control at time t. At a given time t, we define the time-dependent ROC curve and its AUC:

$$ROC(t) = \{\big(FPR(c,t), TPR(c,t)\big), c \in R\}$$
$$AUC(t) = \int_{-\infty}^{\infty} TPR(t,y)|\frac{\delta FPR(t,y)}{\delta y}|dy$$

Thus, the definitions of ROC(t) and AUC(t) rely on those of time-dependent TPR and FPR and thus on how the cases and controls are defined.

Following Heagerty and Zheng, cases are said to be incident when $T_j = t$ is used to define cases at time t, and cumulative if $T_j \leq t$ is used instead. Similarly, controls are said to be static or dynamic depending on whether $T_j > \tau$ for a time $\tau > t$ or $T_j > t$ is used for defining controls at time t. Thus, we can define the cumulative-dynamic ROC

curve as the plot of the cumulative TPR and dynamic FPR

$$TPR^C(c,t) = \mathbb{P}(Y > c|T \leq t)$$
$$FPR^D(c,t) = \mathbb{P}(Y > c|T > t)$$

and the incident-static ROC(t) which is the plot of incident TPR and static FPR

$$TPR^I(c,t) = \mathbb{P}(Y > c|T = t)$$
$$FPR^S(c,t) = \mathbb{P}(Y > c|T > \tau)$$

The cumulative-dynamic definition may be more appropriate for clinical decisions making while the incident-static definition may be more appropriate when trends over time of AUCs are of interest [59]. In this work we focus on the cumulative-dynamic definition since we believe it is more clinically relevant in our settings.

Several approaches have been proposed to estimate the cumulative-dynamic ROC(t) and its AUC(t) dealing with right censoring. Heagerty et al. [60] proposed estimators based on Bayes' theorem and the Kaplan–Meier estimator. They also developed an other method based on the nearest neighbor estimator of the bivariate distribution of the marker and the time-to-event to handle dependent censoring. Chambless and Diao [61] detailed a Kaplan–Meier-like estimator method conditioning on observed event times. These approaches do not include covariates in the definition of the ROC curve, which may be important in assisting classification. On this purpose, a semiparametric approach was introduced by Song and Zhou [9] which models the conditional survival probability of failure times given the marker Y. Finally, Uno et al. [56] proposed a nonparametric estimators employing inverse probability of censoring weighting method. We explore more in details the last two methods, and we refer to [62] for a more accurate illustration of the different proposed methodologies.

Furthermore, other measures can be employed to assess the ability of a prognostic marker in discrimination. The concordance index, also known as c-index or c-statistic [63, 64] is very popular. It is the probability that, between two randomly chosen patients, the one who first occurs the event has an higher predicted risk (higher marker value). The time-dependent AUC is interpreted as the probability that a case has a higher biomarker value than a control subject. To explain the main difference between the two discrimination measures for survival analysis, let consider two individuals i, j with observed failure times T_i and T_j and biomarker values Y_i, Y_j. The concordance index is defined by $CI = \mathbb{P}(Y_i(t) > Y_j(t)|T_i < T_j)$, and the $AUC(t) = \mathbb{P}(Y_i(t) > Y_j(t)|D_i(t) = 1, D_j(t) = 0)$. Unlike the time-dependent AUC, the c-index does not depend on the horizon time t, thus

it cannot be used when a specific time t is of interest, but it provides a summary measure over all the times. In [65] it has been shown that the c-index is not a proper scoring rule to evaluate t-year predicted risks.

3.4 Methods for covariate-specific time dependent ROC curve

The time-dependent ROC curve is useful in assessing the discriminatory ability of a biomarker with survival data. When marker observations depend on a set of covariate X, it is needed to take into account for these [8]. Therefore, the covariate-specific time dependent ROC curve is of interest in this setting. It calibrates the marker respect to the covariates. However, covariate adjustment is also necessary when the covariate is associated with both the biomarker and the time-to-event. Following Janes and Pepe, we define the pooled ROC curve which considers all the individuals regardless their covariates values, and it quantifies the discrimination including the portion of discriminatory ability due to covariates. The covariate-specific ROC curve is evaluated in a sub-population with fixed values of covariates. Moreover, the pooled ROC classifies individuals using a common threshold which is independent of their covariate value. Whereas, the covarate-specific ROC considers covariate-specific thresholds for classification. When a covariate affects either the marker observations or the outcome, the pooled ROC curve is biased respect to the covariate-specific ones [8]. The difference between the two curves reveals the increased accuracy that can be achieved when covariate-specific thresholds are used. The covariate-adjusted ROC curve (AROC(t)) is a covariate-adjusted summary of classification accuracy. It is the overall true positive rate at a specific FPR value, where thresholds are covariate-specific. It describes the performance of the marker in a population with a fixed covariate value. It can also be interpreted as a weighted average of covariate-specific ROC curves with weights corresponding to the proportion of cases in each covariate group:

$$AROC(t) = \int ROC(t|x) f_X(x) dx = \mathbb{E}[ROC(t|X)]$$

with the expectation taken with respect to X.

The main objective is to determine how well a marker can discriminate individuals. When the marker depends on the covariates but the discriminatory ability is not impacted by X, than the adjusted ROC curve is more of interest; meanwhile, when the covariates affects both outcome and marker, the covariate-sepcific ROC curve is preferred since it quantifies the discriminatory ability for subjects that are considered as having similar

risk profiles based on X. However, if we want to compare several markers, the covariate-adjusted ROC curve can be employed to have a summary measure of covariate-adjusted accuracy.

We focus on the covariate-specific time dependent ROC curve because we think that it is more relevant for the motivating example. The tumor stage represents the covariate taken into account, and it affects both the biomarker (individuals with higher tumor stage have larger number of CTCs) and the outcome (individuals with higher tumor stage are likely to die before). In the next sections we introduce two methods available to estimate covariate-specific ROC curve in presence of time-to-event data: i) a non parametric method [56] based on the Inverse Censoring Probability Weighting method; ii) a semiparametric method [9] based on the assumption of proportional hazards. The former allows to adjust for discrete covariates, meanwhile the latter can also handle continuous covariates. Other methods have been proposed [66, 67] but we refer to the references for more details.

Note that the methods mentioned above do not refer to the clustered data settings. In Section 3.5 we propose a new method, motivated by the IMENEO data set, to assess the discriminatory accuracy of the biomarker estimating the covariate-specific ROC curve and its AUC taking into account for the correlation between observations. We compare the results to the ones obtained by the Song and Zhou method and the nonparametric method by IPCW.

3.4.1 Inverse Probability Censoring Weighting

Uno et al. [56] introduced a nonparametric estimator for time-dependent ROC curve and AUC by using Inverse Probability Censoring Weighting (IPCW). The cumulative/dynamic definition was presented:

$$\widehat{TPR}^{C}(t,y) = \frac{\sum_{j=1}^{N} \Delta_j \mathbb{I}(Y_j > y, \tilde{T}_j \leq t)/\hat{S}_C(\tilde{T}_j)}{\sum_{j=1}^{N} \Delta_j \mathbb{I}(\tilde{T}_j \leq t)/\hat{S}_C(\tilde{T}_j)}$$

$$\widehat{FPR}^{D}(t,y) = \frac{\sum_{j=1}^{N} \mathbb{I}(Y_j > y, \tilde{T}_j > t)}{\sum_{j=1}^{N} \mathbb{I}(\tilde{T}_j > t)}$$

$\hat{S}_C(\tilde{T}_j)$ is the Kaplan-Meier estimator for the survival function of the censoring time.

To estimate the covariate-specific time dependent ROC curve, when the covariate X is categorical, it is possible to stratify over the covariate values. However, Le Borgne et al [68] extended the IPCW method to account for covariates correcting the weights by standardizing the marker according to the covariates among the controls.

An estimator of the $AUC(t)$ is given by

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{I}(\tilde{T}_i \leq t)\mathbb{I}(\tilde{T}_j > t)\mathbb{I}(Y_i > Y_j)\frac{\Delta_i}{\hat{S}_C(\tilde{T}_i)\hat{S}_C(t)}}{N^2\hat{S}(t)\ (1-\hat{S}(t))}$$

The method is implemented in `R` in the package `timeROC` [69].

3.4.2 Semiparametric method

A semiparametric approach to estimate time-dependent ROC curves adjusting for covariates [9]. The Cox proportional hazards model is considered to model the effect of covariates on survival times and the biomarker depends on covariates through a semiparametric location model. Estimates for cumulative/dynamic and incident/dynamic covariate-specific ROC curves are proposed. The former is useful in distinguishing individuals experiencing the event by a given time t and those experiencing it after t, while the latter discriminates subjects experiencing the event at a given time t and those experiencing it after t.

Using some algebra we can rewrite the TPR and FPR as follows:

$$TPR^C(t,y,x) = \frac{\int_y^\infty (1-S(t|z,x))\mathbb{P}(Y=z|X=x)dz}{\int_{-\infty}^\infty (1-S(t|z,x))\mathbb{P}(Y=z|X=x)dz}$$
$$TPR^I(t,y,x) = \frac{\int_y^\infty f(t|z,x)\mathbb{P}(Y=z|X=x)dz}{\int_{-\infty}^\infty f(t|z,x)\mathbb{P}(Y=z|X=x)dz}$$
$$FPR^D(t,y,x) = \frac{\int_y^\infty S(t|z,x)\mathbb{P}(Y=z|X=x)dz}{\int_{-\infty}^\infty S(t|z,x)\mathbb{P}(Y=z|X=x)dz}$$

where $S(t|z,x) = \mathbb{P}(T > t|Y=z, X=x)$ is the conditional survival distribution function and $f(t|z,x) = dS(t|z,x)/dt$ is the corresponding conditional survival density.
To estimate these quantities a proportional hazards model is considered:

$$\alpha(t) = \alpha_0(t)\exp(\beta Y + \gamma^{'} X)$$

where $\alpha_0(\cdot)$ is an unspecified baseline hazard function, and a semiparametric location

model:

$$P(Y \leq y|X) = H(y - \psi' X)$$

$H(\cdot)$ is the unspecified biomarker distribution function which can be estimated by:

$$\hat{H}(y;\hat{\psi}) = \frac{1}{N}\sum_{j=1}^{N}(Y_j - \hat{\psi}^T X_j \leq y)$$

with $\hat{\psi}$ obtained by solving $\sum_{j=1}^{N}(Y_j - \hat{\psi}^T X_j)X_j = 0$. The survival function is estimated by maximizing the partial likelihood of the proportional hazards model. Thus the estimators for TPR^I, TPR^C and FPR^D are:

$$\widehat{TPR}^C(t,y,x) = \frac{\sum_{j=1}^{N}\big(1 - \hat{S}(t|Y_j - \hat{\psi}^T(X_j - x), x)\big)\mathbb{I}(Y_j - \hat{\psi}^T(X_j - x) \geq y)}{\sum_{j=1}^{N}\big(1 - \hat{S}(t|Y_j - \hat{\psi}^T(X_j - x), x)\big)}$$

$$\widehat{TPR}^I(t,y,x) = \frac{\sum_{j=1}^{N}\exp(\hat{\beta}Y_j - \hat{\psi}^T(X_j - x))\ \hat{S}(t|Y_j - \hat{\psi}^T(X_j - x), x)\mathbb{I}(Y_j - \hat{\psi}^T(X_j - x) \geq y)}{\sum_{j=1}^{N}\exp(\hat{\beta}Y_j - \hat{\psi}^T(X_j - x))\ \hat{S}(t|Y_j - \hat{\psi}^T(X_j - x), x)}$$

$$\widehat{FPR}^D(t,y,x) = \frac{\sum_{j=1}^{N}\hat{S}(t|Y_j - \hat{\psi}^T(X_j - x), x)\mathbb{I}(Y_j - \hat{\psi}^T(X_j - x) \geq y)}{\sum_{j=1}^{N}\hat{S}(t|Y_j - \hat{\psi}^T(X_j - x), x)}$$

The estimate for the covariate-specific time dependent ROC curve is straightforward. The consistency of the estimators depends on the correct specification of the models but, this method has the the advantage of simple computation. It is implemented in `R` in the packages `condtimeROC` [66] and `survAUC` [70].

3.5 A covariate-specific ROC(t) curve for correlated survival data

So far, characterization for different time-dependent ROC curve have been provided, together with the definition in the covariate-specific setting. However no precision has been made on the estimation of the covariate-specific time dependent ROC curve and its AUC

in presence of correlated survival data. In this chapter we propose a new method to address this problem and we argue which quantities can be estimated and under which assumptions. We compare the new method to the nonparametric one (IPCW) and we discuss the interpretation of the estimates and their consistency under different scenarios.

3.5.1 Two marginal parameters

For each cluster k, let r_k be the index of a randomly selected member of the observed cluster. As discussed in the previous Chapter, in presence of clustered data two kind of marginal parameters might be of interest to estimate [16] [15]. The first one has an interpretation for the population of all observed members (aom), e.g. equal weights are given to each member of the observed population. The second one has an interpretation for the population of typical observed members of a typical cluster (tom), e.g. equal weights are given to each observed cluster and subjects within each observed cluster have equal weights. Considering the time-dependent true positive rate, we can define:

$$TPR_{aom}(t,y) = \frac{\mathbb{E}[N_k \mathbf{I}(Y_{r_k} \geq y) | D_{r_k}(t) = 1]}{\mathbb{E}[N_k | D_{r_k}(t) = 1]}$$
$$TPR_{tom}(t,y) = \mathbb{E}[\mathbf{I}(Y_{r_k} \geq y) | D_{r_k}(t) = 1],$$

where $TPR_{aom}(t,y)$ corresponds to the probability for a random subject in the population of all observed subjects. It provides information about the predictive accuracy of the biomarker in the population of all the observed members of all the clusters. On the other hand, the $TPR_{tom}(t,y)$ corresponds to the probability for a random subject belonging to a random cluster. Similarly, we can distinguish the False Positive Rate (FPR), which is the probability of the biomarker's value being above the threshold y in the control population ($D_{r_k}(t) = 0$):

$$FPR_{aom}(t,y) = \frac{\mathbb{E}[N_k \mathbf{I}(Y_{r_k} \geq y) | D_{r_k}(t) = 0]}{\mathbb{E}[N_k | D_{r_k}(t) = 0]}$$
$$FPR_{tom}(t,y) = \mathbb{E}[\mathbf{I}(Y_{r_k} \geq y) | D_{r_k}(t) = 0]$$

In presence of clustered data, we can distinguish two settings depending on the association between the cluster sample sizes and the outcome. Hoffman et al [13], Williamson et al [15] and Benhin at al [17] define non informative cluster size (NICS) when $\mathbb{P}(D_{r_k}(t) = 1 | Y_{r_k} = y, N_k) = \mathbb{P}(D_{r_k}(t) = 1 | Y_{r_k} = y)$, otherwise the cluster size is said to be informa-

tive (ICS). Interestingly, under NICS, when the biomarker does not depend on the cluster size ($Y_{r_k} \perp\!\!\!\perp N_k$), the two parameters are equal ($TPR_{tom}(t,y) = TPR_{aom}(t,y)\ \forall y$), while under ICS they differ in general (proof in the Appendix). Thus, for the interpretation of results, it might be important to underline the quantity of interest.

Note that, under ICS, the TPR_{aom} and FPR_{aom} depend on the study design to collect the data through the sample sizes N_k, $k = 1, \ldots, K$. Therefore a challenge for the interpretation arises when these sample sizes might not be representative of any underling, well-defined, population of interest. In that case any conclusion based on TPR_{aom} and FPR_{aom} and their estimates would be difficult to generalize to such population of interest, from which random sampling would lead to different cluster sizes. For the IMENEO data, as in most of meta-analysis, the sample size in each cluster is arbitrary and not representative of any underlying well-defined population of interest. For instance, the cluster sample sizes are not necessarily representative of the size of the population treated in the hospitals. Hence, the estimates for the population of all members could be difficult to generalize in case of ICS. However, in the motivating IMENEO data example there is no reason to suspect ICS and an explanatory analysis did not suggest ICS (see Appendix). In this chapter we assume NICS and we discuss more in details ICS in Chapter 4.

As assuming NICS implies $TPR_{aom} = TPR_{tom}$ and $FPR_{aom} = FPR_{tom}$, we now discard the subscripts aom/tom and we write:

$$TPR(t,y) = \mathbb{E}[\mathbf{I}(Y_{kj} \geq y)|D_{kj}(t) = 1] = \mathbb{P}(Y_{kj} \geq y|D_{kj}(t) = 1) \tag{3.1}$$

$$FPR(t,y) = \mathbb{E}[\mathbf{I}(Y_{kj} \geq y)|D_{kj}(t) = 0] = \mathbb{P}(Y_{kj} \geq y|D_{kj}(t) = 0) \tag{3.2}$$

We define the pooled time dependent ROC curve as the plot of $TPR(t,y)$ and $FPR(t,y)$ for different thresholds y used to classify individuals at time t. The term pooled refers to the fact that we marginalize over X, by contrast to the covariate-specific ROC curve defined in the next section.

3.5.2 Definition of the method

In the assessment of the performance of a biomarker, the presence of a covariate associated with both the outcome and the biomarker may lead to results that are challenging to interpret and that can be misleading [8]. Recalling that Y_{kj} represents the biomarker value for individual j in cluster k and X_{kj} is a vector of covariates. Let T_{kj}, C_{kj} be the failure time and censoring time respectively. We assume that T_{kj} and C_{kj} are independent for all k, j and that in each cluster k $(T_{k1}, T_{k2}, ..., T_{kN_k})$ can be correlated conditionally on $(X_{k1}, X_{k2}, ..., X_{kN_k})$. For this scenario we propose an estimator for the covariate-

specific cumulative dynamic ROC curve, $\text{ROC}(t,x) = \{(FPR(t,y,x), TPR(t,y,x)), y \in V\}$ where $TPR(t,y,x) = \mathbb{P}(Y_{kj} \geq y | T_{kj} \leq t, X_{kj} = x)$ and $FPR(t,y,x) = \mathbb{P}(Y_{kj} \geq y | T_{kj} > t, X_{kj} = x)$ are the covariate-specific cumulative true positive rate and dynamic false positive rate respectively.

In short, the pooled ROC(t) describes the accuracy of Y in classifying individuals using a common threshold independent of the subject's covariate values. By contrast, the covariate-specific ROC curve describes the accuracy of Y in classifying subjects with specific covariate values using covariate-specific thresholds. In other words, it naturally quantifies how well the new biomarker discriminates between subjects that are considered as having similar risk profiles based on the covariate X.

To estimate the covariate-specific time dependent ROC curve with clustered failure times we extend the semiparametric method proposed by Song and Zhou[9]. Their approach assumes a proportional hazard model for the conditional distribution of T given (Y, X) and a semiparametric location model for the conditional distribution of Y given X. To accommodate clustered failure times the shared frailty model is used instead of the proportional hazards model. The introduction of a random effect (frailty term) $U_k \geq 0$ captures the correlation of within-clusters observations. The frailty can be thought as a proxy for the unmeasured covariates which are common to all members of the same cluster and associated to the time-to-event. For each cluster k, $U_{kj} = U_{kj'}\ \forall j, j' = 1, .., N_k$, i.e. subjects belonging to the same cluster k have same random effect U_k.

We propose the following model :

$$\begin{cases} \alpha(t|y_{kj}, x_{kj}, u_k) = u_k \alpha_0(t) \exp(\beta y_{kj} + \gamma' x_{kj}) \\ \mathbb{P}(Y_{kj} \leq y | x_{kj}) = H(y|x_{kj}; \psi) \end{cases} \tag{3.3}$$

with the additional assumption $U_k \perp\!\!\!\perp (Y_{kj}, X_{kj})\ \forall k, j$.

Let $H(\cdot)$ be the conditional cumulative distribution function of the biomarker with ψ vector of parameters. In the following sections we define $H(\cdot)$ as the cumulative distribution function of a negative binomial, as this distribution has been shown to be appropriate for our motivating IMENEO data (see supplementary material of Bidard et al.[6]). Details on the negative binomial distribution are provided in Appendix. Of note, an interaction term can be further added to the survival model.

Seaman et al.[16] pointed out that in a general context of mixed models the assumption $U_k \perp\!\!\!\perp (Y_{kj}, X_{kj})$ implies $U_k \perp\!\!\!\perp N_k$ and, under this assumption, the random effect model provides a consistent estimator for the typical observed members interpretation. Since in (3.3) Y_{kj} is a covariate in the shared frailty model (mixed effect model), the assumption

$U_k \perp\!\!\!\perp Y_{kj}$ is needed. This is consistent with the usual case-mix setting [71], where the cluster affects the outcome but it does not affect the biomarker's value (as illustrated in Figure 3.4). Considering an heterogeneous biomarker would require another approach and it is beyond the scope of this work.

Using some algebra from (3.1) and (3.2), for a continuous biomarker, we obtain:

$$TPR(t,y,x) = \frac{\int_y^{\infty}(1-S(t|z,x))\mathbb{P}(Y_{r_k}=z|X_{r_k}=x)dz}{\int_{-\infty}^{\infty}(1-S(t|z,x))\mathbb{P}(Y_{r_k}=z|X_{r_k}=x)dz} \tag{3.4}$$

$$FPR(t,y,x) = \frac{\int_y^{\infty} S(t|z,x)\mathbb{P}(Y_{r_k}=z|X_{r_k}=x)dz}{\int_{\infty}^{\infty} S(t|z,x)\mathbb{P}(Y_{r_k}=z|X_{r_k}=x)dz} \tag{3.5}$$

The conditional survival function $S(t|z,x,u) = \mathbb{P}(T_{r_k} > t|X_{r_k}=x, Y_{r_k}=z, U_k = u)$ is linked to the cumulative hazard function $A(t|z,x,u) = \int_0^t \alpha(s|z,x,u)ds$ through $S(t|z,x,u) = \exp\{-uA(t|z,x)\}$. The marginal survival function (relative to the random effect) $S(t|z,x)$ is obtained by integrating over the frailty:

$$S(t|z,x) = \int_0^{\infty} \exp\Big\{-u_k \int_0^t \alpha(s|z,x)ds\Big\} f_U(u_k)du_k = \int_0^{\infty} \exp\{-u_k A(t|z,x)\} f_U(u_k)du_k$$

where f_U denotes the density of U_k. We can rewrite it as $S(t|z,x) = \mathbb{E}[\exp\{-U_k A(t|z,x)\}] = \mathcal{L}(A(t|z,x))$. For analytic reason, it is often preferred to consider a random effect with distribution belonging to the natural exponential family (gamma, power variance, positive stable and lognormal function) as the Laplace transform is easier to compute in this case. The possible choices are discussed in details by Hougaard [4]. In our working example we consider $U_k \sim Gamma(\theta,\delta)$ which is computationally easier since $\mathcal{L}(s) = (\theta^{\delta})/(\theta+s)^{\delta}$, with $\theta,\delta > 0$. In particular, for a shared frailty model we address the problem of identifiability by imposing the restriction $\mathbb{E}[U_k] = 1$. Therefore, for the gamma distribution, $\theta = \delta$ and $Var(U_k) = 1/\theta$.

The likelihood of the data given X is $L = L_{T|Y,X} \times L_{Y|X}$, where $L_{T|Y,X}$ is the likelihood of the failure times given Y and X and $L_{Y|X}$ is the likelihood of the biomarker given X. Here $L_{T|Y,X}$ is the following product over the clusters:

$$L_{T|Y,X}(\theta,\beta,\gamma,\alpha_0) = \prod_{k=1}^{K}\prod_{j=1}^{N_k}\int_0^{\infty} [u_k\alpha(t_{kj},y_{kj},x_{kj})]^{\Delta_{kj}} \times \exp(-u_k A(t_{kj},y_{kj},x_{kj})) \times f_U(u_k)du$$

Maximizing the marginal likelihood L is equivalent to maximizing the two likelihood $L_{T|Y,X}$ and $L_{Y|X}$ separately. The plug-in estimates of the time dependent covariate-specific TPR and FPR are derived from the maximum likelihood estimation of the conditional

hazard $A(t|y,x)$ and the distribution function of the biomarker Y given X, by substituting the corresponding estimates into equations (3.4) and (3.5). Methods based on maximum likelihood and the EM-algorithm, as implemented in the `R` package `frailtyEM` [72], are used to estimate the marginal hazard function and the frailty parameter.

Finally, the estimated time dependent covariate-specific ROC curve for population with a fixed value of the covariate $X = x$ is:

$$\widehat{ROC}(t,x) = \{(\widehat{FPR}(t,y,x), \widehat{TPR}(t,y,x)), y \in V\}$$

.

For a discrete biomarker the integral becomes a finite sum; e.g. the numerator of $\widehat{TPR}(t,y,x)$ becomes $\sum_{z \geq y} \{1 - \hat{S}(t|z,x)\}\hat{\mathbb{P}}(Y_{r_k} = z | X_{r_k} = x)$

Moreover, from the $\widehat{ROC}(t,x)$, we can estimate an overall measure of discrimination up to time t with the covariate-specific time dependent Area Under the ROC Curve ($\widehat{AUC}(t,x)$). It can be interpreted as the probability that a typical case of a typical cluster has a higher biomarker value that of a typical control of a typical cluster. Typical subject of a typical cluster means that a cluster is firstly randomly sampled and then a subject of that cluster is randomly sampled.

Under NICS, we assume:

1. $T_{kj} \perp\!\!\!\perp T_{kj'} | U_k$ for all $j \neq j'$, that is, in each cluster k, the times $T_{kj}, j = 1, \ldots, N_k$, are correlated but they are independent conditionally on the frailty U_k,

2. $U_k \perp\!\!\!\perp (Y_{kj}, X_{kj})$ for all j, which implies $Y_{kj} \perp\!\!\!\perp N_k$

it can be shown that the AUC estimated by the proposed method is an unbiased estimator of the discriminatory accuracy for the interpretation in terms of a typical observed member of a typical cluster. At the same time, for the nonparametric estimator of AUC based on IPCW [56], the weights are given at the member level, and it provides the discriminatory ability for the interpretation with respect to the all observed members population. For similar reason, as explained in [16], the nonparametric approach provides a consistent estimator for AUC_{aom}.

3.5.3 Bootstrap method

A parametric bootstrap method is implemented for the confidence interval of the covariate-specific time dependent AUC. Given the original data set, we generate the bootstrap data from the estimated parameters $\hat{\beta}, \hat{\gamma}, \hat{\psi}$ and $\hat{\theta}$.

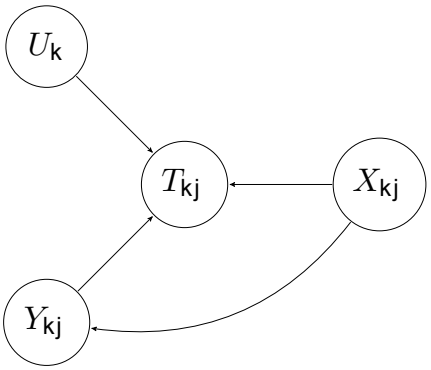

Figure 3.4: DAG for the case-mix assumption: the random effect U_k at the cluster level is independent on the biomarker Y_{kj} and covariate X_{kj} , but U_k affects the failure time T_{kj}.

For each bootstrap replication $b = 1, ..., B$:

1. we randomly sample, with replacement, K clusters sample sizes $N_k^{(b)}$ among the K clusters sample sizes of the original data. This step defines the sample size of the bootstrap data set, which is $N^{(b)} = \sum_{k=1}^{K} N_k^{(b)}$

2. we randomly generate K frailty terms $U_k^{(b)}, k = 1, ..., K$ from the distribution Gamma$(\hat{\theta})$

3. we randomly sample, with replacement, $N^{(b)}$ covariate values $X_{kj}^{(b)}, j = 1, ..., N_k^{(b)}$, $k = 1, ..., K$, from the values observed in the original data

4. we generate $N^{(b)}$ biomarker values $Y_{kj}^{(b)}, j = 1, ..., N_k^{(b)}$, $k = 1, ..., K$ given $X_{kj}^{(b)}$ from the estimated conditional distribution $H(\cdot|X_{kj}^{(b)}; \hat{\psi})$

5. for each cluster $k = 1, ..., K$, we generate the time-to-event $T_{kj}^{(b)}, j = 1, ..., N_k^{(b)}$, given $Y_{kj}^{(b)}, X_{kj}^{(b)}, U_k^{(b)}$ from the estimated cumulative distribution of $T_{kj}|Y_{kj}, X_{kj}, U_k$:

$$\widehat{F}(t|Y_{kj}^{(b)}, X_{kj}^{(b)}, U_k^{(b)}) = 1 - \exp(-U_k^{(b)} \widehat{\Lambda}_0(t) \exp(\hat{\beta} Y_{kj}^{(b)} + \hat{\gamma} X_{kj}^{(b)}))$$

   We invert the cumulative distribution function and we obtain the failure times as $T_{kj}^{(b)} = \hat{F}^{-1}(Z_{kj}|Y_{kj}^{(b)}, X_{kj}^{(b)}, U_k^{(b)})$ with $Z_{kj} \sim Unif(0, 1)$

6. let $\hat{S}_C(\cdot)$ be the Kaplan-Meier estimator of the censoring distribution, estimated from $\{(T_{kj}, 1 - \Delta_{kj}), k = 1, ..., K \;\; j = 1, ..., N_k\}$, we generate the censoring times $C_{kj}^{(b)}$ $j = 1, ..., N_k^{(b)}$, $k = 1, ..., K$ by sampling, with replacement, among the censoring times of the original data with probability equal to the jump of the Kaplan-Meier estimator

7. we compute $\widehat{AUC}^{(b)}(t,x)$ by applying the proposed method to the bootstrap data set.

We compute the 95% percentile confidence interval of AUC(t,x) where the upper and lower values are given respectively by the 2.5% and 97.5% quantiles of $\{\widehat{AUC}^{(b)}(t,x), b = 1, ..., B\}$.

The regular bootstrap method cannot be employed for clustered data because it assumes exchangeability between patients being resampled and than underestimates the errors [73]. Xiao [74] proposed two alternative procedure to bootstrap clustered data: the cluster bootstrap where clusters are resampled with replacement and then all patients of the selected clusters are included; the two-step bootstrap which first samples clusters with replacement and than resmaple with replacement observations in the cluster. The two-step procedure considers between and within cluster variability and best represent real life scenario, but in introduces too much variability and produces overestimated results [73].

We consider a parametric cluster bootstrap approach which randomly samples with replacement K clusters (with their sample sizes) having an overall sample size $N^{(b)}$. For each resampled cluster the frailty term is generated from a Gamma distribution with the estimated parameter $\hat{\theta}$. We assume an homogeneous biomarker and covariates across clusters, thus we than resample, with replacement, $N^{(b)}$ covariates, independently on the cluster and we generate the biomarker values from the estimated biomarker model. For each resampled cluster, we generate the time-to-events and censoring time by their estimated distributions.

3.5.4 Simulation study

We conducted a simulation study to assess the performance of the proposed method. We mimicked the settings of the IMENEO data for the biomarker and covariate distribution. We generated a biomarker Y following a negative binomial distribution with set of parameter $\psi = (d, \xi)$ where $d = 0.5$ is the dispersion parameter and $\mu_X = 0.2 + \xi X$. Here, X is a categorical covariate with 2 levels: $P(X = 1) = \frac{2}{3}$ and $P(X = 2) = \frac{1}{3}$. The failure times were generated from a frailty model, i.e from the conditional cumulative distribution function $P(T \leq t|Y, X, U) = 1 - \exp(-UA_0(t)\exp(\beta Y + \gamma X))$ with the frailty term $U \sim Gamma(\theta = 1)$ and the cumulative baseline hazard function $A_0(t) = st^{\omega}(s = 6.31e^{-6}, \omega = 4.6)$. The censoring times were generated from an exponential distribution in order to reach 80% of censoring as in the motivating example. We fixed $\beta = 0.8,\ \gamma = 0.5$ which define respectively the dependence on the failure times

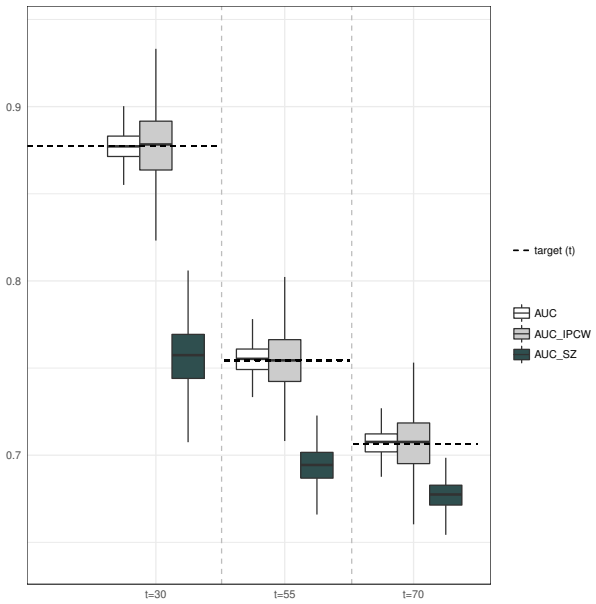

Figure 3.5: Simulation results for 1000 replications with 100 clusters and 80% of censoring: boxplot at different time points(t=30,55,70) for the estimated covariate-specific AUC($t|X = 2$) with the proposed method (AUC), the semiparametric method of Song and Zhou (AUC_{SZ}) and the nonparametric method (AUC_{IPCW}). The dotted horizontal lines represent the true values at each time.

of the biomarker Y and the covariate X. We set $\xi = 0.4$ to control the correlation between Y and X. Since we do not allow the biomarker's distribution to vary across center, we did not introduce any dependence between the biomarker and the cluster. A total of 1000 data sets were drawn with 100 clusters of cluster's sample sizes randomly chosen with uniform distribution in the interval $[10, 210]$ assuming NICS.

We compared the proposed method with the Song and Zhou's approach and the nonparametric method of Inverse Probability of Censoring Weighting (IPCW) [56]. Note that, to estimate the covariate-specific ROC(t), a stratified analysis is necessary for the IPCW method, while the full data set is used for the others approaches.

The Figure 3.5 summarizes the results of the simulation for the three methods at time $t = 30, 55, 70$ with marginal probability of event $\mathbb{P}(\tilde{T}_{kj} \leq t) = 0.03, 0.21, 0.40$ and proportion of censored observations during the follow-up of $49.5\%, 69.3\%, 75.2\%$ respectively. Results of the three methods for all time are presented in the Appendix. The proposed approach and the nonparametric one (IPCW) present negligible bias, but the IPCW estimator shows larger variability at each time. The Song and Zhou's method, on the other hand, is biased in this setting of correlated failure times because of the violation of the hypothesis of proportionality of the marginal hazards. We further investigate

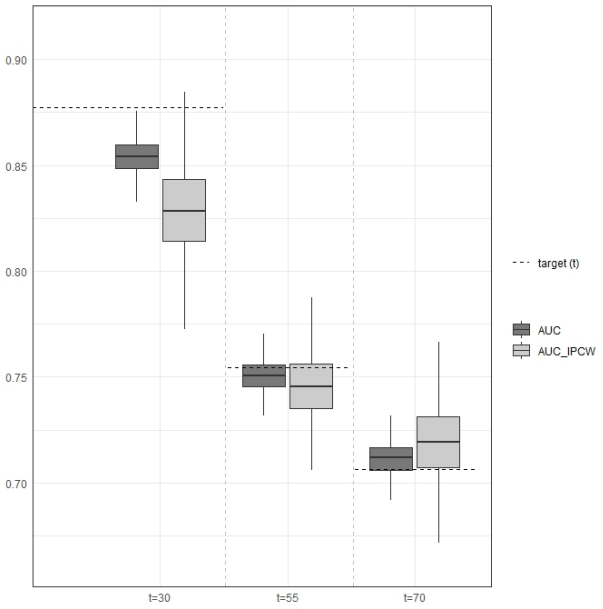

Figure 3.6: Simulations results for 1000 replications with 100 clusters and 80% censoring under ICS. Boxplot at different time points(t=30,55,70) for the estimated covariate-specific AUC($t|X = 2$) with the proposed method (AUC) and the nonparametric method ($\widehat{AUC}_{\text{IPCW}}$). The dotted horizontal lines represent the true values at each time.

the variability of the two eligible methods in terms of bias. The clustered parametric bootstrap method described above was implemented for the confidence intervals of the $AUCs$' estimates, and the nonparametric cluster bootstrap [75] for the $\widehat{AUC}_{\text{IPCW}}$ with B=2000 resamplings. Table 3.2 presents the results at time $t = 30, 55, 70$ for the estimated covariate-specific time dependent $AUCs$ for $X = 2$. Bias and coverage probability are illustrated. We observe small bias and rather good coverage probability for both methods, but the $\widehat{AUC}_{\text{IPCW}}$ presents wider confidence interval (L_{ci}) as compared to the proposed estimator ($\widehat{AUC}_{\text{PM}}$). The $\widehat{AUC}_{IPCW}$ is a consistent estimator of the discriminatory accuracy of the biomarker in sense of the all observed members population (AUC_{aom}). The $\widehat{AUC}_{\text{PM}}$ is consistent in sense of the typical members population (AUC_{tom}). As we previously underlined, under NICS $AUC_{tom} = AUC_{aom}$. This is in line with our simulation results where both methods produce nearly unbiased estimators. Further results on the estimated parameters are provided in the Appendix.

Moreover, we conducted a simulation study assuming informative cluster size. The data were generated as above, with the difference that the sample size of cluster k is defined considering the frailty term associated to k. In particular, we order the U_k and we split them in 5 levels, clusters with U_k belonging to the same level have similar sample size, and we assign smaller sample sizes to smaller values of frailty. Figure 3.6 confirms that,

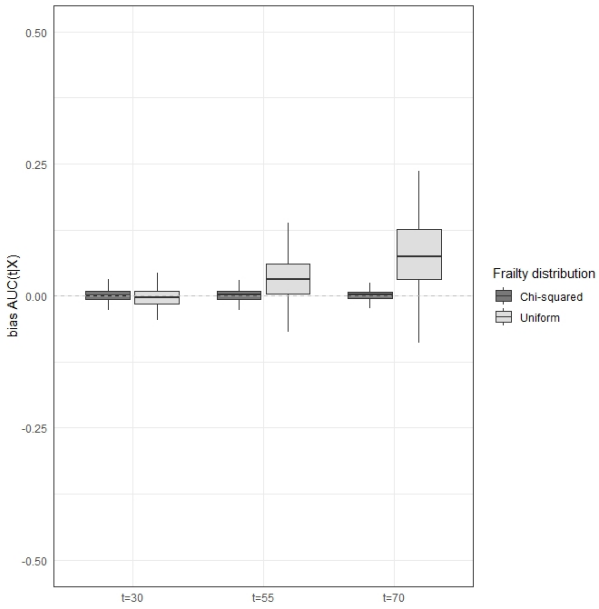

Figure 3.7: Simulations for a misspecified frailty distribution: data were generated with $U_k \sim [0, 10]$ and $U_k \sim \chi^2(2)$, and the covariate-specific AUC(t) was estimated by a shared gamma frailty model. Results of bias at $t = 30, 55, 70$ are provided.

under ICS, the estimates for typical observed members and for all observed members differ. Additionally, both approaches produce biased estimators, thus appropriate methods that take into account ICS should be employed.

Furthermore, we were interested in illustrating the robustness of the method in case of misspecification of the frailty distribution. The frailty is a latent variable, and the definition of its distribution is a challenging point. We simulated data with $U_k \sim \chi^2(2)$ and $U_k \sim U[0, 10]$ and estimated the survival function with a shared gamma frailty model. The results show a limited impact on the consistency of the AUC estimator when $U_k \sim \chi^2(2)$. Bias is detected in case of strong misspecification, when $U_k \sim U[0, 10]$ (Figure 3.7).

3.6 Application to breast cancer

We applied the proposed method to the IMENEO data. The goal was to evaluate the capability of CTCs count measured at the time of diagnosis (baseline) to discriminate patients who die prior a time t from those who survive up to time t. Data on CTCs count at baseline of 1911 women were collected from 2005 to 2016 in 16 different centers, conducting 21 trials. The detection of CTCs was performed by the CELLSEARCH System (the only one FDA approved in 2014) in all the trials. We observed the death of 14% of the patients, with failure times ranging from 0.2 to 9.7 years (median 30 months).

t	AUC	Method	Bias	cp	L_{ci}
30	0.8773	PM	1e-04	0.941	0.0330
		IPCW	6e-04	0.939	0.0816
55	0.7544	PM	7e-04	0.928	0.0328
		IPCW	3e-04	0.934	0.0662
70	0.7064	PM	8e-04	0.938	0.0295
		IPCW	5e-04	0.948	0.0663

Table 3.2: Simulation results of the proposed method (PM) and the nonparametric method (IPCW) for 1000 replications with 100 cluster and β=0.8, d=0.5, $\xi = 0.5$. The estimators and the respective bias are provided at $t = 30, 55, 70$; the coverage probability (CP) and the average length of the bootstrap confidence intervals (L_{ci}) are obtained with 2000 bootstrap samples.

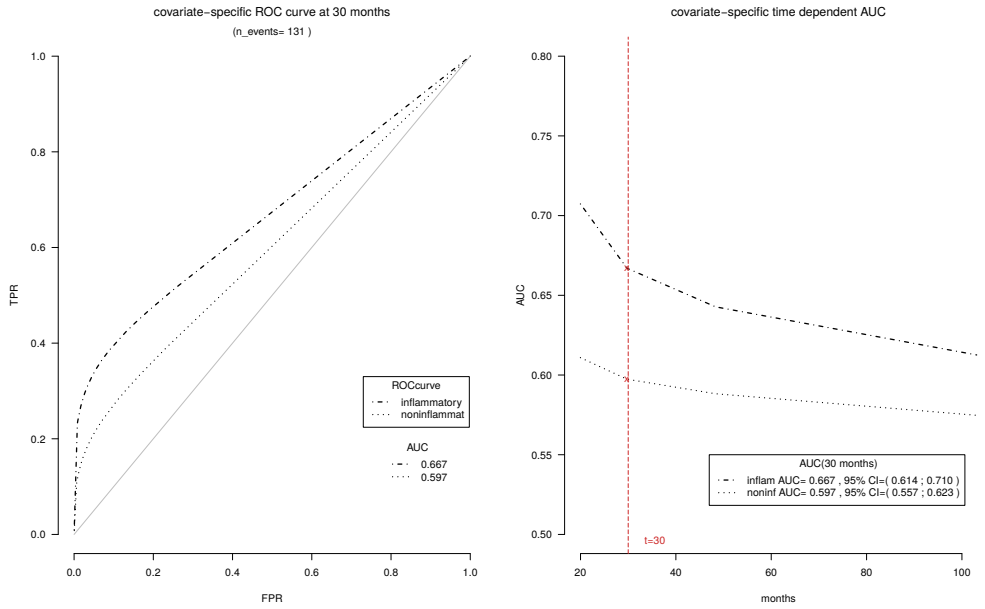

Figure 3.8: Covariate-specific time dependent ROC curves at t=30 months and time dependent *AUC* of CTCs count at baseline adjusted for tumor stage. We provide the 95% confidence interval at t=30 months.

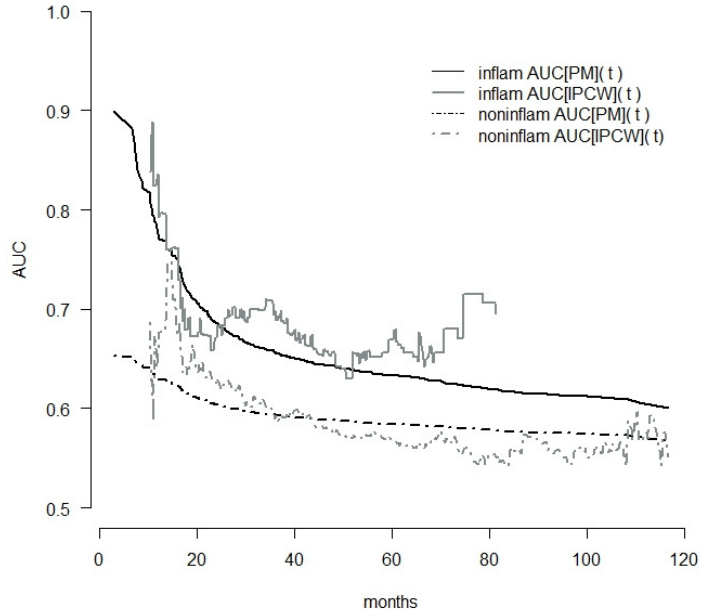

Figure 3.9: Covariate-specific time dependent AUC of CTCs count per tumor stage. We provide the estimates obtained with the proposed method (black) and the nonparametric one (IPCW in gray).

The clinicians were particularly interested in assessing the performance of CTCs counts within each tumor stage group since patients with inflammatory tumor (T4d) have more aggressive tumor and larger number of CTCs. In fact, the probability of death within 30 months is 0.06 (95% CI (0.05,0.07)) for patients with noninflammatory tumor and 0.19 (95% CI (0.15, 0.22)) for the group of patients with inflammatory tumor. As such, the discriminatory capability irrespective of the tumor stage is not of clinical interest.

We estimated the covariate-specific time dependent ROC curves and $AUCs$ for the two groups using the proposed method. To assess the validity of these estimators we checked the assumption of the model (3.3). We first tested for the homogeneity of the failure times distribution implementing the Commenges-Andersen score test [76] with $U_k \sim \text{Gamma}(\theta)$. The null hypothesis $H_0 : Var(U_k) = 0$ was rejected with a p-value of 0.01, supporting the shared frailty model. Of note, this test can be employed with all kind of distribution of the frailty. To check for the adequacy of the gamma distribution, we compare the estimated marginal survival function by a shared frailty model with its non parametric counter part (Kaplan-Meier estimator). This analysis did not indicate a

suspicion of misspecification (see Appendix).

Next, we test for the homogeneity of the biomarker at the center level ($U \perp\!\!\!\perp Y$). The CTCs count did not appear heterogeneous in a subset of the IMENEO data of 15 centers which will be our working data set (see supplementary material of Bidard et al.[6] for an in depth study of the homogeneity of the CTCs).

We estimated the covariate-specific $ROC(t, x)$ for subgroups of patients with same tumor stage (inflammatory tumor/noninflammatory tumor). The estimated regression coefficients for the shared gamma frailty model were $\exp(\hat{\beta}) = 1.067$ and $\exp(\hat{\gamma}) = 2.552$; for the biomarker model, using a negative bionomial regression, we estimated the dispersion parameter $\hat{d} = 0.091$ and the regression coefficient $\hat{\xi} = 0.966$. In Figure 3.8, we provide the estimation of $ROC(t^*, x)$ at t^*=30 months and the time dependent $AUC(t, x)$. The biomarker showed an estimated AUC of 0.667 with 95% confidence interval $(0.614, 0.710)$ at 30 months after the diagnosis for patients with inflammatory tumor, and an estimated AUC of 0.597 with confidence interval $(0.557, 0.623)$ among noninflammatory tumor. The confidence intervals of $\widehat{AUC}(t, x)$ were computed via parametric bootstrap. Similar results were obtained when estimating the AUC by IPCW (Figure 3.9). At $t^* = 30$ months $\widehat{AUC}_{\text{IPCW}} = 0.603$ with confidence interval $(0.513, 0.693)$ among noninflammatory tumor patients and an estimated AUC of 0.691 $(0.567, 0.815)$ for patients with inflammatory tumor.

3.7 Discussion

In this section we have proposed a method to determine how well a biomarker discriminate patients in a context of clustered failure times. We have discussed which quantities can be estimated and their interpretations. We have introduced an estimator for the covariate-specific time dependent ROC curve which takes into account the dependence of failure times between members of the same cluster. More specifically, we have extended the Song and Zhou[9] approach assuming a shared frailty model instead of a proportional hazards model for the hazard of failure. Our contribution was based on the cumulative dynamic time dependent ROC curve and AUC. In many contexts, including that of our motivating example, we believe that this definition is more clinically relevant than alternatives [11]. This is because at each time t, it illustrates the ability of a biomarker to discriminate patients who experience a specific event in the time interval [0,t] from ones who do not experience the event after the time t. Moreover, we assume a case-mix context where the biomarker's distribution does not change across clusters. This assumption is often reasonable. In fact, it is coherent with the usual setting in meta-analysis for individual

patients data when inclusion criteria are similar in all the trials and the measurement of the biomarker is robust between trials.

When clustered data arises, two quantities for the performance of a biomarker can be estimated: (i) for all observed members population which assesses the discriminatory ability of the biomarker in the overall observed population, (ii) for typical observed members population which assesses the discriminatory ability for a typical member in a typical cluster of the observed population. The proposed method provides an estimator of the covariate-specific ROC(t) for a typical member setting ($ROC_{tom}(t, x)$). While, the nonparametric method of IPCW provides an estimator of the covariate-specific ROC(t) in the sense of all observed members population ($ROC_{aom}(t, x)$).

Moreover, in this context of clustered data, two scenarios can be distinguished based on the dependence of the outcome (T_{kj}) on the clusters sample sizes (N_k). Under informative cluster sizes (ICS), the outcome depends on the cluster sample sizes and caution is required in the interpretation of the estimated quantities since the observed population is generally not representative of a well-defined population of interest. Several approaches have been proposed to handle ICS in survival analysis [77, 12], we explore this issue in details in the next chapter. However, The assumption of NICS was not straightforward. Initially, for the simulation study we decided to generate the cluster sample sizes conditioned to the frailty term, namely assuming informative cluster size. As result, the proposed method provided unexpected biased results. We than wondered on the generation algorithm and we found out about the problem of informative cluster size.

Under NICS the outcome is independent to the sample sizes and we proved that the two quantities coincide ($ROC_{aom}(t, x) = ROC_{tom}(t, x)$). Therefore, our contribution was also to point out that the nonparametric method is an other eligible method to assess the performance of a biomarker in case of clustered survival data under NICS. In fact, as highlighted within our simulation study, both the proposed estimator and the IPCW estimator have negligible bias, in contrast to the semiparametric estimator obtained by Song and Zhou, where the assumption of proportional hazards was violated. Note that for the nonparametric method, the covariate-specific ROC(t) is obtained by stratification on the covariate values. Thus, for the estimation of $ROC(t, x)$, it is required to have enough data in each strata $X = x$. With a discrete covariate, we recommend to employ the nonparametric method since no assumption is needed either on the biomarker nor on the frailty distribution. However, the proposed method displays narrower confidence interval and can address the problem of correlated censored survival data adjusting on both continuous and discrete covariates.

We have considered a parametric model for the biomarker distribution but the ap-

proach can accommodate other models. Motivated by the CTCs count and its particular distribution, we have employed a parametric model to estimate the conditional distribution of Y given X, instead of the semi parametric location model used by Song and Zhou. We agree that a more flexible model could make the method more versatile, but we also think that a reasonable model check will generally prevent significant bias. For completeness, we further provide the code which implements the direct extension of the Song and Zhou method with a location model for the biomarker in the supplementary material.

The method is first presented for an arbitrary frailty distribution, but the gamma distribution is then used in both the application and the simulation study. In the application, we performed one ad hoc analysis to check for the adequacy of the gamma distribution. It is a graphical approach which compares the estimated marginal survival function with the Kaplan-Meier curve and it did not raise suspicion of misspecification. We illustrated the impact of the frailty's misspecification on the estimation of the AUC in the simulation section. The results suggest that the choice of the frailty might have a limited effect on the AUC but, in case of strong misspecification, the method provide biased estimators. The choice of the frailty distribution is a challenging topic. The various possibilities are discussed in [10]. Several approaches have been proposed to check the gamma distribution assumption [44, 43, 45]. We refer to these papers for details on related model checking aspects.

The usefulness of the proposed method was illustrated in our application on non metastatic breast cancer where data are characterized by heterogeneity of failure times and homogeneity of biomarker's distribution among centers (case-mix setting). The discriminatory ability of CTCs count was assessed for patients with inflammatory tumor and noninflammatory tumor estimating the covariate-specific time dependent AUC. The implementation in `R` of the proposed method is provided at `https://github.com/AMeddis/AUCtime`.

Acknowledgment

We would like to thank François Clement Bidard (Institut Curie, Saint Cloud,France) and Stefan Michiels (Gustave Roussy, Villejuif, France) for providing the IMENEO data from which our analyses were derived. This project was in collaboration with Paul Blanche (Univeristy of Copenhagen, Denmark).

Chapter 4

Informative cluster size

4.1 Introduction

Several methods have been proposed to handle clustered data, such as frailty model or marginal models, but they assume that the outcome is unrelated to the clusters sample sizes. This assumption is not always satisfied, and in this case, the cluster size is said to be informative. For instance, the time to tooth loss in one individual is of interest. Subjects with a dental disease may already have lost some teeth due to the disease. Thus, time to loss in one individual (cluster) is linked to the number of teeth (cluster sizes) of the same. An other example can be found in studies of men with lymphatic filariasis, which is characterized by one or more nests of adult filarial worms in the scrotum [12]. Ideally, effective treatment would kill the worms in all of the nests. The nest-specific time to clearance the worm is longer in men with multiple nests than in men with one nest.

Under informative cluster size (ICS), the standard statistical methods produce biased results, since the estimates will be over-weighted in favor of bigger clusters. Various approaches have been introduced to take into account for ICS: Hoffman [77] proposed the within-cluster resampling method where multiple independent data sets are created randomly sampling one observation from each cluster, with replacement; Williamson [15] considered a GEE method inversely weighted by clusters sample sizes. Cong et al [14] investigated the WCR method for clustered survival data with ICS analyzing the resampled data sets using a Cox model. They also generalize the marginal models by incorporating the inverse of cluster sizes as weights into the score function. Williamson et al[12] explored the estimation of the marginal distribution for multivariate survival data with informative cluster size using cluster-weighted Weibull and Cox models. For all these methods, they rely on the assumption of ICS, without testing it in the application study. It is possible to check for ICS comparing the marginal distributions between strata defined by cluster

size. This is an ad-hoc approach, but our scope is to provide a more general method to test for ICS for right censored survival data. Benhin [78] employed a Wald-type test for ignorability of cluster size in the estimating equations framework for linear and logistic regression models. Nevalainen [18] proposed a test for ICS using a balanced bootstrap to estimate the null distribution. To our knowledge, no other test for clustered survival data is available. We propose a method to test for ICS considering the Nelson-Aalen estimator for the cumulative hazard function for the two target population. The asymptotic distribution of the test statistic is obtained using standard martingale results.

The chapter is organized as follows. In Section 4.2, we illustrate the problem of ICS and we provide some more in deep description for possible target populations in clustered data. We briefly mention the method that can be employed to handle informative cluster size and we describe the tests previously proposed for iCS in linear regression model. In Section 4.3 we describe a new method to test for ICS in right censored survival data and we provide the asymptotic distribution. Simulation studies were conducted to assess the power of the test. In Section 4.4, we illustrate the usefulness of the method by several applications. We provide some discussion in Section 4.5.

4.2 Informative cluster size

A challenging problem of clustered data which is often ignored is the possibility of informative cluster size (ICS). In this setting, the outcome of interest is related to the clusters sample sizes. Examples of informative cluster size can be found in volume-outcome studies, where surgeons treating a larger number of patients may have better outcomes. The typical example of ICS is in periodontal studies [15] where the number of teeth affects the prognosis of patients. More examples on toxicity with longitudinal data on litters born can be provided [12]. The reasons of ICS are usually unknown because some latent variable could affect the baseline hazard for each cluster. The variability of sample size, which is now a random variable, can also be due to missing data, thus observed clusters are incomplete [79]. In this case, if the association between outcomes and covariates in complete cluster is of interest, than assumption about missing data need to be made. In this work, we do not focus on this particular context. We assume that either there are no missing data or that we want to do inference on the observed members.

Hoffman et al [13] defined informative cluster size any violation of the condition $\mathbb{P}(T_{kj} \leq t | N_k = n) = \mathbb{P}(T_{kj} \leq t)\ \forall n$, therefore the distribution of failure times is independent on the cluster sample sizes. Chen et al [80] defined ICS when the mechanism that generate cluster size is not independent on the mechanism that generates the out-

come. Standard approaches, marginal models and random effect models, provide biased results in presence of ICS because they fail to account for the information carried by the cluster sizes. For marginal models the individuals equally contribute to the likelihood and larger cluster are overweighted, meanwhile in random effect models, the random effect is linked to the the mechanism of cluster sizes which is ignored. For correlated survival data, this issue is even more complicated because of censoring and the unknown hazard function.

Furthermore, adjusting for the cluster sample size in the model including N_k as a covariate is not appropriate. Firstly, introducing N_k in the model we will have information on the outcome conditional on the cluster sample sizes. We are not interested in the effect of N_k, but we want to take into account its information within the inference model to have unbiased results. Nonetheless, the sample size may be a mediator in the causal pathway from the covariates X and the outcome T. Additionally, if there is a latent variable that affects both sample size and outcome, introducing N_k as a covariate may produce collider-stratification bias.

An important point is that when ICS is absent, methods that unnecessarily allow for ICS lead to substantial loss of efficiency [81]. Thus, the analyst should first test the assumption of ICS and than decide the method to choose. Furthermore, as introduced in Chapter 2, two target population can be defined. One makes inference for the population of all observed members (AOM), and the second one, makes inference for the population of typical observed members of a typical cluster (TOM). For each cluster k, let r_k be the index of a randomly selected member of the observed cluster. As in Seaman [16] we define

$$e_{AOM} = \frac{\mathbb{E}[N_k T_{r_k} | N_{r_k} \geq 1]}{\mathbb{E}[N_{r_k} | N_{r_k} \geq 1]}$$
$$e_{TOM} = \mathbb{E}[T_{r_k} | N_{r_k} \geq 1].$$

Under non informative cluster size (NICS) the two marginal analyses coincide $e_{AOM} = e_{TOM}$, while they differ under ICS. Thus, it is important to detect if cluster sample size is informative for inference because the parameters of interest are not the same for alternative target populations under ICS and their estimation needs correctly specified estimators. We will rely on this property to construct the test described in the next section.

4.2.1 Methods for clustered data with Informative Cluster Size

Many methods, marginal models and random effect models, have been extended to address the problem of ICS. Hoffman [13] proposed the within-cluster resampling (WCR) method where a series of data sets are constructed by randomly sampling one observation from each cluster. The resulting resampled data sets can be analysed by any marginal analysis since the observations are independent. The parameters obtained by the resampled data are then averaged. A variation of this method was described by Chiang and Lee [82] where m members are sampled in each cluster, with m the minimum sample size ($m > 1$), and the GEE are then employed with realistic working correlation to each resampled data set. Williamson [15] introduced a weighted GEE method with an independence working correlation matrix, where the weights are the inverse of the cluster sample sizes. This is asymptotically equivalent to the WCR method. Benhin [17] discussed mean estimating equation approach to handle ICS. Huang and Leroux [83] proposed double weighted GEE with categorical covariates, where the member is inversely weighted by the total number of member in the same cluster with $X = x$. Alternatively, Chen et al. [80] proposed a joint model approach with random effect model for the outcome and the link between outcome and cluster size models is established by a shared random effect. Neuhaus and McCulloch [84] address the analysis of informative cluster size data from a cluster-specific approach through the use of generalized linear mixed models. They demonstrate that maximum likelihood method that ignores informative cluster sizes exhibits little bias in estimating covariate effects that are uncorrelated with the random effects associated with cluster sizes. Alternatively, estimates of covariate effects may be biased if the covariate effects are associated with the random effects.

Cong et al. [14] investigated WCR for correlated survival data and they generalize the marginal models introducing the inverse of cluster sample sizes as weights into the score function. The WCR method is computationally intensive and also, since it considers one member in each cluster, the estimated parameters from a simple resampling might be unstable under heavy censoring. Furthermore, Williamson [12] introduced a cluster weighted Weibull and Cox proportional hazard model to estimate marginal distribution incorporating cluster size weighting to the independence working models. More in general, for correlated survival data, Datta et al [85, 86] generalized a runk-sum and signed-rank test to account for ICS . Fen and Datta [87] used inverse cluster size weighting in accelerated failure times models.

Marginal models are more attractive because no assumptions are made on the correlation between outcome and cluster sample size, but they can be less efficient than random effect model. The last should be used if the correlation between Y and N_k is of interest

but random effect models are subjected to missspecifications.

4.2.2 Existing test for ICS

So far, we discussed the importance of taking into account informative cluster size and we briefly described some methods that can be used in this setting. However, the assumption of ICS(NICS) has to be verified since methods that consider ICS when it is not needed may lead to a loss of efficiency [81, 17]. An empirical method, ad-hoc approach, consists in the plot of the marginal distribution between strata defined by cluster size (Figure A.2). In this example, we provide the Kaplan-Meier estimator of the survival function at $t = 30$ for each cluster sample size. In the IMENEO study, the analysis suggests NICS because no trend of survival can be identified at varying of the sample size.

Benhin et al. (2005) [17] employed a test for ICS in the estimating equations framework for linear regression model. A Wald-test on the difference of the estimated coefficient from independence and mean estimating equation is considered. Nevalainen et al (2017) [18] introduced a test for the presence of ICS using a novel bootstrap method to estimate the test statistic distribution. The null hypothesis of the test is that the marginal distribution does not depend on the cluster size. Let Y_{kj} be the outcome value for individual j in cluster k, the test statistic is defined as $T_F = \sup_y |\hat{F}(y) - \tilde{F}(y)|$ with

$$\hat{F}(y) = \frac{1}{N} \sum_{k=1}^{K} \sum_{j=1}^{N_k} (Y_{kj} \leq y)$$

$$\tilde{F}(y) = \frac{1}{k} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{j=1}^{N_k} (Y_{kj} \leq y)$$

Under ICS the two estimators are consistent of the same population cumulative distribution function. An approximation for the test statistic distribution is obtained by a bootstrap procedure where K clusters are sampled from the data set obtained by permuting members within each cluster. At each bootstrap, a matching method based on the definition of distance between clusters is considered to preserve the cluster sample sizes. Given B bootstrap data sets, the null distribution is approximated by the obtained test statistic $T_F^{(1)}, ..., T_F^{(B)}$ and the p value is computed as $(1/B) \sum_b (T_F^{(b)} \geq T_F)$. Other versions of the statistics grouping the cluster by sample sizes are presented and extensions to the regression setting (GLM) using the model residuals are considered.

The Nevaleinen test is computationally intensive because of the bootstrap procedure, especially since, for the matching, the distance has to be computed at each resampling. Both the methods can be extended to the survival data, but no formal test has been

presented in the context of correlated survival data. In the next section, we propose a novel test for ICS with right censoring clustered survival data where the asymptotic distribution is theoretically obtained by using martingale theory.

4.3 Test for ICS with survival data

In this section we propose a new method to test for ICS with right censored survival data. The definition of the test and the asymptotic distribution for the test statistics are provided. In Section 3 we describe the simulation study conducted to determine the power of the test for different scenarios. In Section 4 some illustrative examples are described in different contexts and some conclusions are made in Section 5.

4.3.1 Definition of the test

Let $\tilde{T}_{kj}$ be the observed failure time for individual j in cluster k with sample size N_k and a maximum number of cluster K such that $N = \sum_{k=1}^{K} N_k$. The quantity $N_{kj}(t) = \sum_{k} \mathbf{I}(\tilde{T}_{kj} \leq t, \Delta_{kj} = 1)$ is the counting process at time t, with intensity $\lambda_{kj}(t) = \alpha_{kj}(t)Y_{kj}(t)$, where $Y_{kj}(t) = I(\tilde{T}_{ik} \geq t)$ represents the at-risk process. The quantity $M_{kj}(t) = N_{kj}(t) - \Lambda_{kj}(t)$ is not a martingale with respect to the joint filtration generated by all the times, because of the correlation within clusters. It is a martingale with respect to filtration $\mathcal{F}_{kj}(t) = \sigma\{N_{kj}(u), Y_{kj}(u) : 0 \leq u \leq t)$. We extend the definition of the Nelson-Aalen estimator for the two marginal analyses:

$$\hat{A}_{tom}(t) = \int_0^t \frac{dN_{tom}(s)}{Y_{tom}(s)}$$
$$\hat{A}_{aom}(t) = \int_0^t \frac{dN_{aom}(s)}{Y_{aom}(s)}$$

$\hat{A}_{tom}(t)$ estimates the number of events for a typical observed member and $\hat{A}_{aom}(t)$ estimates the number of events in the sense of all observed member populations. In fact, the weighted counting process and at risk process are defined as:

$$N_{tom}(t) = \frac{1}{K} \sum_{k} \frac{1}{N_k} \sum_{j} N_{kj}(t)$$
$$Y_{tom}(t) = \frac{1}{K} \sum_{k} \frac{1}{N_k} \sum_{j} Y_{kj}(t)$$

where units within cluster are equally weighted by the inverse of the cluster sample size, and

$$N_{aom}(t) = \frac{1}{N}\sum_k \sum_j N_{kj}(t)$$
$$Y_{aom}(t) = \frac{1}{N}\sum_k \sum_j Y_{kj}(t)$$

where equal weights are given to each unit, regardless the cluster they belong to. Ying and Wei [88] stated that even though data are clustered and observations are dependent in each cluster, the above estimators are consistent estimators for the cumulative hazard functions.

To define the null hypothesis of the test, we rely on the fact that under NICS the two marginal analyses coincide:

$$H_0 : \alpha_{tom}(t) = \alpha_{aom}(t) \ \forall t \in [0, \tau]$$
$$H_1 : \alpha_{tom}(t) \neq \alpha_{aom}(t) \text{ in } t^* \in [0, \tau]$$

The test statistic

$$Z(\tau) = \int_0^T L(t)(d\hat{A}_{tom} - d\hat{A}_{aom})$$

with $L(t) = \frac{Y_{aom}(t)Y_{tom}(t)}{K}$. Under the null hypothesis $Z(\tau)/\sqrt{K}$ is asymptotically equivalent to a Gaussian with mean 0 and covariance matrix V.

4.3.2 Asymptotic distribution

We obtain the asymptotic distribution of the test statistic under the null hypothesis of NICS.

By definition of $d\hat{A}_{\cdot}$:

$$Z(\tau) = \int_0^T L(t) \ (\frac{dN_{tom}(t)}{Y_{tom}(t)} - \frac{dN_{aom}(t)}{Y_{aom}(t)})$$

where $dN_h(t) = dM_h(t) + \alpha_h(t)Y_h(t)$

therefore

$$Z(\tau) = \int_0^\tau L(t)\big(\frac{dM_{tom}(t) + \alpha_{tom}(t)Y_{tom}(t)}{Y_{tom}(t)}\big) - \big(\frac{dM_{aom}(t) + \alpha_{aom}(t)Y_{aom}(t)}{Y_{aom}(t)}\big)$$
$$= \int_0^\tau L(T)\big(\frac{dM_{tom}(t)}{Y_{tom}(t)} - \frac{dM_{aom}(t)}{Y_{aom}(t)}\big) + \int_0^\tau L(t)\big(\alpha_{tom}(t) - \alpha_{aom}(t)\big)dt$$

Under the null hypothesis $\alpha_{tom}(t) = \alpha_{aom}(t)\ \forall t \in [0, \tau]$ and by definition of $N_h(t)$, $dM_{tom}(t) = \sum_k \frac{1}{N_k} \sum_j dM_{kj}(t)$ and $dM_{aom}(t) = \sum_k \sum_j dM_{kj}(t)$.

We specify $L(t) = \frac{Y_{aom}(t)Y_{tom}(t)}{K}$, and we obtain:

$$Z(\tau) = \int_0^\tau \frac{L(t)}{Y_{tom}(t)} \sum_k \frac{1}{N_k} \sum_j dM_{kj}(t) - \int_0^\tau \frac{L(t)}{Y_{aom}(t)} \sum_k \sum_j dM_{kj}(t)$$
$$= \int_0^\tau \frac{Y_{aom}(t)}{K} \sum_k \frac{1}{N_k} \sum_j dM_{kj}(t) - \int_0^\tau \frac{Y_{tom}(t)}{K} \sum_k \sum_j dM_{kj}(t)$$

Because N_k's are bounded we can interchange sums and integral:

$$Z(\tau) = \sum_k \frac{1}{N_k} \sum_j \int_0^\tau \frac{Y_{aom}(t)}{K} dM_{kj}(t) - \sum_k \sum_j \int_0^\tau \frac{Y_{tom}(t)}{K} dM_{kj}(t)$$
$$= \sum_k \frac{1}{N_k} \int_0^\tau \frac{Y_{aom}(t)}{K} dM_k(t) - \sum_k \int_0^\tau \frac{Y_{tom}(t)}{K} dM_k(t)$$

where $M_k(t) = \sum_j^{N_k} M_{kj}(t)$. Thus, the statistic can be rewritten as

$$Z(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}} \sum_k \int_0^\tau \big(\frac{Y_{aom}(t)}{N_k K} - \frac{Y_{tom}(t)}{K}\big) dM_k(t)$$

The quantity $\frac{1}{\sqrt{K}} \sum_k M_{ik}(t)$ converges weakly to a Gaussian process $U^Z(t)$ [88]. Define $y_{aom}(t), y_{tom}(t)$ such that for $N \to \infty$ $Y_{aom}/N_k K \to y_{aom}(t)$ and $Y_{tom}/K \to y_{tom}(t)$, the quantity $\int_0^\tau |\frac{Y_{aom}(t)}{N_k K} - \frac{Y_{tom}(t)}{K}|$ is bounded away from infinity in N, and as in [89]

$$Z(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}} \sum_k \int_0^\tau \big(\frac{Y_{aom}}{N_k K} - \frac{Y_{tom}}{K}\big) dM_k(t)$$

and

$$Z^*(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}} \sum_k \int_0^\tau \big(y_{aom}(t) - y_{tom}(t)\big) dM_k(t)$$

converge almost surely to the same limit $\int_0^\tau (y_{aom}(t) - y_{tom}(t)) dU^Z(t)$ and the statistic is

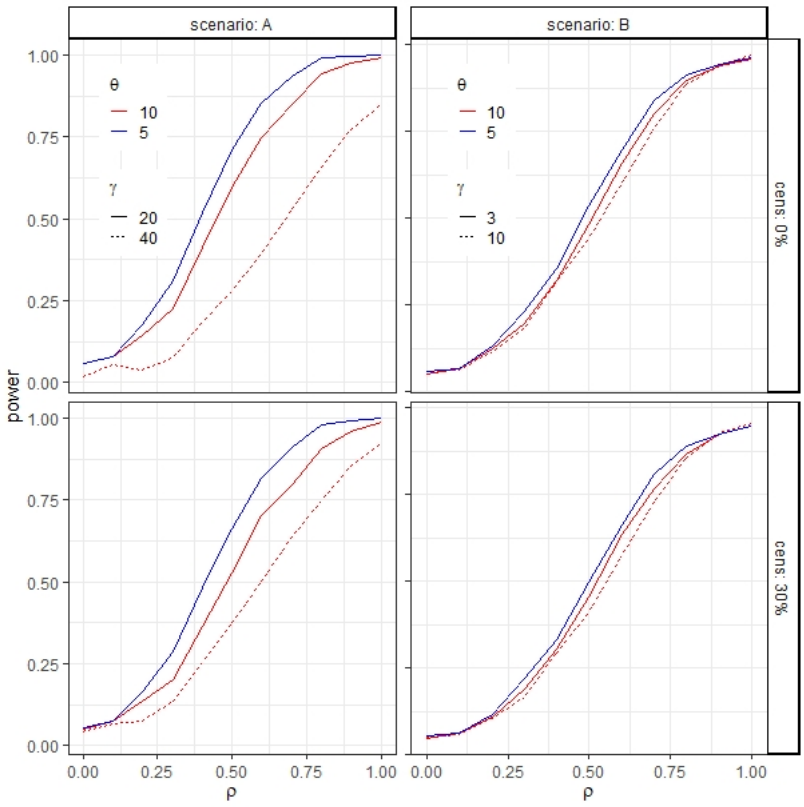

Figure 4.1: Power of the test at varying of the correlation ρ for both scenarios considering different values of θ, γ and censoring. Each framework is based on 1000 replications, fixing $\alpha = 0.05$. Scenario A: highly clustered data ($K = 100$, $\lambda = 5$), scenario B: few big clusters ($K = 25, \lambda = 20$).

asymptotically equivalent to a Gaussian with mean 0 and covariance matrix V which is asymptotically equivalent to $V^* = \frac{1}{K}\sum_k \sum_j \sum_{j'} \epsilon_{kj}\epsilon_{kj'}$

with $\epsilon_{kj} = \int_0^\tau \omega_k(t) dM_{kj}(t)$ where $\omega_k(t) = (y_{aom}(t) - y_{tom}(t))$. We can estimate the covariance with

$$\hat{\epsilon}_{kj} = \Delta_{kj}\hat{\omega}_k(T_{kj}) - \sum_i \sum_l \frac{\Delta_{li}\hat{\omega}_k(T_{li})Y_{kj}(T_{li})}{\sum_m \sum_f Y_{mf}(T_{li})}, \quad \hat{\omega}_k = \left(\frac{Y_{aom}(t)}{KN_k} - \frac{Y_{tom}(t)}{K}\right)$$

4.3.3 Simulation Study

We conducted a simulation study to assess the performance of the test for a fixed I type error of 5% also evaluating the power of the test under different scenarios. The correlated failure times were generated from a frailty model, i.e from the conditional

Sample Size	0% censoring					30% censoring				
N	K	λ	γ	θ	$\rho = 0$	K	λ	γ	θ	$\rho = 0$
1500	100	5	40	10	0.020	100	5	40	10	0.043
	100	5	20	5	0.057	100	5	20	5	0.056
	100	5	20	10	0.060	100	5	20	10	0.049
	50	5	20	10	0.049	50	5	20	10	0.047
700	100	2	20	10	0.056	100	2	20	10	0.052
	50	5	20	5	0.045	50	5	20	5	0.041
	100	2	20	5	0.055	100	2	20	5	0.048

Table 4.1: Scenario A: highly clustered data. Nominal power of the test for 1000 replications (power under NICS, $\rho = 0$).

cumulative distribution function $P(T \leq t|U_k) = 1 - \exp(-U_k A_0(t))$ with the frailty term $U_k \sim Gamma(\theta)$ and the cumulative baseline hazard function $A_0(t) = st^{\omega}(s = 6.31e^{-6}, \omega = 4.6)$. To obtain informative cluster size, we generate K clusters with sample size $N_k \sim Pois(\lambda \exp(V_k))$ where λ, common between clusters, represents the expected number of observations in each cluster and V_k defines the cluster-specific sample size. To create the dependence between the sample size N_k and the failure times T_{kj}, we generate (U_k, V_k) from a multivariate Gamma with unit mean and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_U^2 & \rho\sigma_V\sigma_U \\ \rho\sigma_V\sigma_U & \sigma_V^2 \end{pmatrix}$$

The variance $\sigma_U^2 = 1/\theta$ defines the variability of time-to-event among clusters. The variance $\sigma_V^2 = 1/\gamma$ represents the variability between clusters sample sizes. The parameter $\rho \in [0, 1]$ is the correlation between the two random effects, and thus it defines the dependence between T_{ik} and N_k, when $\rho = 0$ there is NICS. The strength of ICS depends on θ, ρ, γ: it decreases with larger values of θ, since the difference in time-to-event across clusters decreases. With larger values of γ, the range of cluster sample sizes is more narrow and, for fixed θ, it translates in higher ICS.

Let us consider two parameters γ_a and γ_b, and for each $\gamma_\cdot$ value, we examine two clusters with sample sizes $n_{\dot{1}}, n_{\dot{2}}$ and the respective mean time of event $\overline{T}_{n_{\dot{1}}}, \overline{T}_{n_{\dot{2}}}$. For $\gamma_b < \gamma_a$, to $\overline{T}_{n_1}, \overline{T}_{n_2}$ will correspond the sample sizes n_1^b, n_2^b, with $n_1^b > n_1^a$ and $n_2^b < n_2^a$. Therefore, with increasing γ, for a fixed difference in sample sizes $|n_{\dot{1}} - n_{\dot{2}}|$, the difference in failure times $|\overline{T}_{n_{\dot{1}}} - \overline{T}_{n_{\dot{2}}}|$ is larger (see Appendix). When γ increases, the variability decreases and so does ICS. Therefore, there is a trade-off between variability of clusters sample sizes and the magnitude of difference in time-to-event, which also depends on θ.

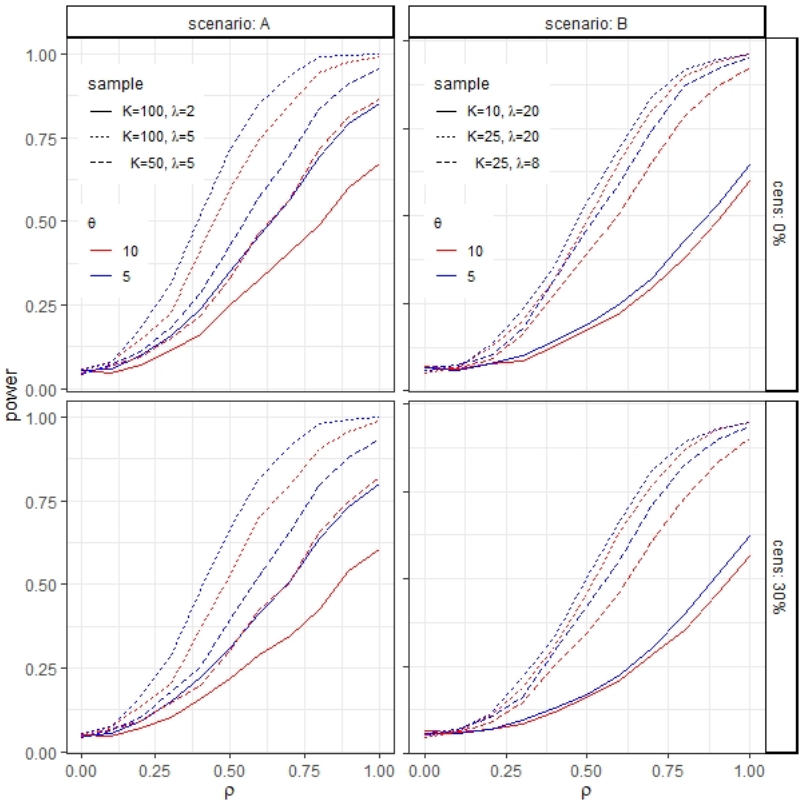

Figure 4.2: Power of the test at varying of the correlation ρ for both scenarios considering different values of K, λ and censoring. Each framework is based on 1000 replications, fixing $\alpha = 0.05$.

Sample Size	0% censoring					30% censoring				
N	K	λ	γ	θ	$\rho = 0$	K	λ	γ	θ	$\rho = 0$
1500	25	20	10	10	0.057	25	20	10	10	0.052
	25	20	3	5	0.059	25	20	3	5	0.053
	25	20	3	10	0.052	25	20	3	10	0.045
	10	20	3	10	0.066	10	20	3	10	0.064
700	25	8	3	10	0.069	25	8	3	10	0.056
	10	20	3	5	0.067	10	20	3	5	0.058
	25	8	3	5	0.065	25	8	3	5	0.059

Table 4.2: Scenario B: few big clusters. Nominal power of the test for 1000 replications (power under NICS, $\rho = 0$).

We simulate two settings: a) highly clustered data with $K = 100, \lambda = 5, \gamma = 3$ and b) few big clusters with $K = 25, \lambda = 20, \gamma = 20$ (e.g., in meta-analysis). For both scenarios, the overall sample size $N = 1500$. We let θ, γ vary to determine the behaviour of the test in different frameworks. Moreover, we decrease the overall sample size N either varying the number of cluster (K=50, K=10) or the clusters sample sizes ($\lambda = 2, \lambda = 8$). We generate 1000 replications for each combination of parameters and we consider both uncensored and 30% right censoring (uniform distribution). In Figure 4.1 we provide the empirical power of the test at varying of the correlation ρ. The simulations suggested a good performance of the test reaching a power of 80% in most scenarios. The results confirmed that θ is inversely proportional to ICS, showing higher power for $\theta = 5$. A decrease in the sample size (N=700) does not seem to result in a worse performance overall (Figure 4.2). In case of $\lambda = 2$ a lower θ is needed to detect ICS since the clusters sample sizes are smaller and the the between-clusters variability is not enough strong. However, for $K = 10$, even decreasing θ, low power is detected, thus a sufficient number of clusters is necessary for the validity of the test. This result was expected, since the asymptotic distribution is valid for $K \to \infty$. Simulations results also suggested that censoring is not degrading the performance of the test. The nominal level of the test (power under NICS) for scenario A and B are provided in Table 4.1 and 4.2 respectively.

4.4 Application

In this section we apply the test for ICS in different settings. Note that we are not interested in the analysis of the data, but this is in support to the theoretical findings and simulations.

4.4.1 IMENEO data set

Individual patient data meta-analysis in breast cancer to validate the performance of the circulating tumor cell counts as biomarker for overall survival. A total of 1911 patients with non metastatic breast cancer are collected in 16 centers and 21 studies. These are treated by neoadjuvant chemotherapy and the number of CTCs measured at baseline is considered. The 14% of the patients die with a median failure time of 30 months ranging from 0.2 to 9.7 years. In Chapter 3 we applied the proposed method to estimate the covariate-specific time dependent ROC curve for the CTCs. The method assumed NICS and the Kaplan-Meier estimator at different cluster sample sizes was plotted to check for this assumption A.2. The test confirmed NICS with a test statistic of -0.48 (pvalue=0.67).

4.4.2 Dental data

We consider data of patients treated at the Creighton University School of Dentistry from August 2007 to March 2013. A total of 5336 patients with periodontal disease were collected with a total of 65228 teeth. We excluded from the analysis individuals with only one tooth resulting in a sample size of 65034. The average age was 58 years, with 51% women, 9% had Diabetes Mellitus, and 23% were smokers. The number of tooth that fall is 4334 with a median tooth loss time of 0.556 $[0.003, 5.594]$ years.

The data are available in the `MST` package in `R` as Teeth [90]. The principal aim of the analysis was to construct multivariate survival trees to predict tooth loss. Several teeth and individual characteristics are also provided in the data set but we do not take them into consideration. We are interested in performing the proposed test to check for informative cluster size on the time-to-loss of each tooth. We would suspect ICS because the number of teeth (cluster size) in each individual (cluster) is linked to the disease and thus, a tooth is more likely to fall in one individual with smaller cluster size. The test confirms a strong ICS with a test statistic of 8.932 (pvalue=0). Figure 4.3 indicates as well ICS. The estimator of the survival function at the median time for each cluster sample size are illustrated: the tooth loss time is longer in individuals with more teeth (bigger cluster size).

4.4.3 Multicentric data

We consider a multicentric study of patients with liver disease primary biliary cirrhosis (PBC). It is a randomized clinical trial conducted in six European hospitals between 1983 and 1987. A total of 349 patients were randomized to either treatment with Cyclosporin A (176 patients) or placebo (173 patients). The effect of treatment on the survival time was

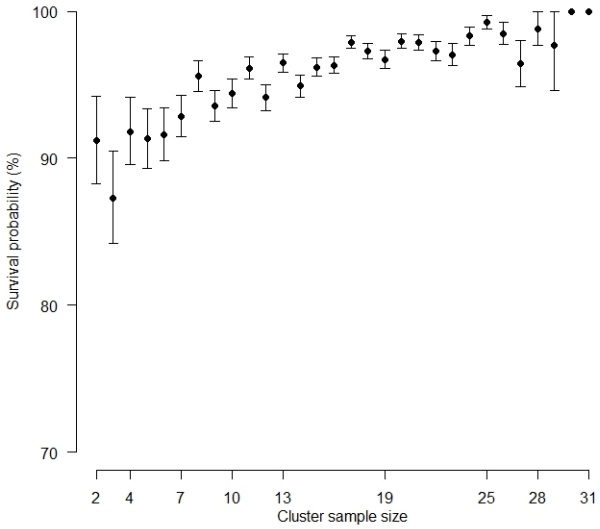

Figure 4.3: Estimated survival function at time $t = 0.556$ at changes of cluster sample size.

the primary outcome of interest. Successively, because of an increment success of liver transplantation for patients with this disease made, the composite outcome "failure of medical treatment" was considered. It was defined as either death or liver transplantation. Data are characterised by 75% of censoring where 61 patients died with a median time of 21 $[0.8, 62]$ months and 29 had liver transplantation with a median time of 23 $[3.27, 48]$ months.

The data are provided in the `pec` package in `R` as Pbc3 [91]. We employed the proposed test that detected slight informative cluster size with a test statistic equal to -1.98 (pvalue=0.04). We observed longer time-to-event in smaller clusters (Figure 4.4). There is not a strong difference between the estimated survival for successive sample sizes, but a clear difference is illustrated between small and big clusters.

4.4.4 Cancer data: Immunotherapy

Immunotherapy is a type of cancer treatment that helps the immune system fight cancer. This type of treatment has become widely used in the last few decades. However, it is more effective for some types of cancer than for others. It is used by itself for some of these cancers, but for others it seems to work better when used with other types of treatment. We consider a data set of 100 patients with metastatic cancer treated by immunotherapy at the Institute Curie in Paris. The metastasis are evaluated singularly in each subject

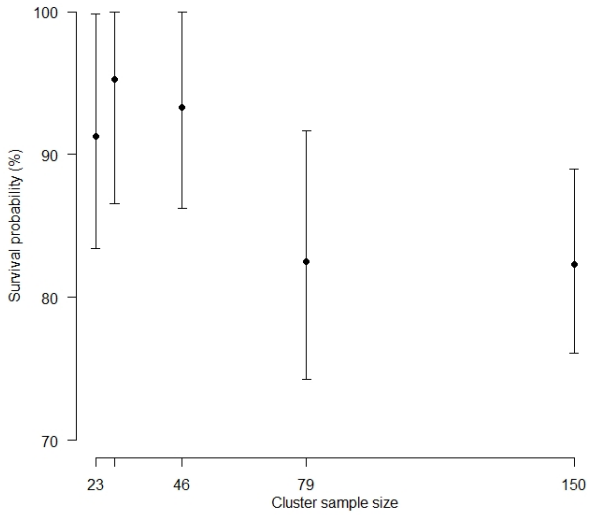

Figure 4.4: Estimated survival function at time $t = 21$ months for different cluster sample sizes.

since the treatment may have different effect on each of it. A total of 272 metastases are examined and each individual has from 2 to 4 metastases (subjects with more than 5 metastases were not included in the study). The primary cancer was of different nature: breast cancer, head neck cancer, lung cancer, urological cancer and others. The principal objective of the study was to have some insight on dissociate response that are typical of immunotherapy, notably in the same individual, the response to treatment might be of different nature among metastases.

The individual represents the cluster and the number of metastases is the cluster sample size. The outcome of interest is the time to progression which depends on the tumor growth. Intuitively, the number of metastasis should affect the outcome. However, this idea was not confirmed by the test that did not reject the null hypothesis of NICS with a test statistic of -0.85 (pvalue=0.39). This illustrative example shows a limitation of the proposed test when the cluster sample sizes are small and similar and the number of clusters is not sufficient to detect ICS. We agree that if we could follow an higher number of metastases in each individual or more individuals, we would have had different results. In fact, as shown in Figure 4.5, there is an impact of the cluster size on the survival function for metastasis disease progression.

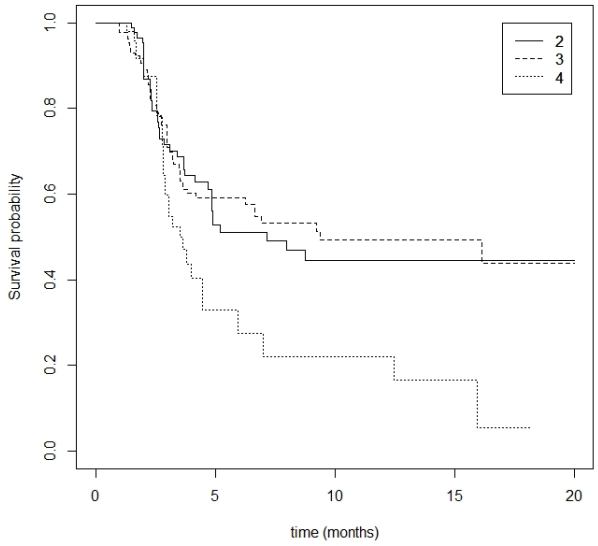

Figure 4.5: Estimated survival function for different number of metastases.

4.5 Discussion

In presence of clustered data, standard statistical methods implicitly assume that the size of the clusters is unrelated to the outcome of interest. This assumption is not always verified and we define the cluster size to be informative. Several approaches have been proposed to handle the issue of ICS with survival clustered data, but, to our knowledge, no test has been introduced to explore ICS. Standard methods for clustered data analysis may result biased estimates with ICS, and ICS methods can lead to a loss of efficiency with NICS [78]. In this work, we propose a test for the assumption of NICS with right censored survival data. The test statistic relies on the fact that under NICS the two marginal analyses for typical observed member and all observed member coincide. The asymptotic distribution of the statistic under the null hypothesis is provided. A simulation study for different settings of clustered data shows a good performance of the test with an estimated power being greater than 80% and a nominal power around 5%. Censoring does not seem to affect the performance of the test, but simulations results suggest that a sufficient number of cluster K is needed.

In section 2 we mention that the variability in sample sizes can be a result of missing data. Hoffman [77] and Williamson [15] vaguely stated that missing completely at random (MCAR) mechanism is equivalent to non-informative cluster size. Pavlou [79] associated NICS to missing data mechanism, of which MCAR is special case, and they proved the

equality of results for the target populations in three cases (tom,aom,missing data). In our work, we assume that the observed clusters are complete, and thus uninformative and independent censoring, discarding the problem of missing data.

No covariate X is introduced in the test. In this case, the definition of NICS can be extended to $\mathbf{P}(T_{ik} \leq t|X_{ik}, N_k = n) = \mathbf{P}(T_{ik} \leq t|X_{ik})\ \forall n$ and the Breslow estimator can be employed instead of the Neslon-Aalen estimator. However, we believe that if the outcome is related to cluster sample sizes, conditionally on the covariate, and not marginally, than the covariates might be size-unbalanced (their distribution is dependent of the cluster sizes) and informative covariate structure may arise. This could be an interesting point for future work and possible extensions of the proposed method.

A test for ICS has been already proposed for clustered data by a balanced bootstrap method, since the distribution of the statistic under the null is analytically intractable [18]. An adaptation of this method to survival data could be an other proposition to check for ICS, but it is characterized by high computational cost.

Acknowledgment

We would like to thank Christophe Le Tourneau, Vaflard Pauline and Xavier Paoletti (Institut Curie, Paris, France) for providing the data of patients with metastatic cancer treated by immunotherapy .

Chapter 5

IPD meta-analysis with competing endpoints

5.1 Introduction

Competing risks data are inherent to medical researches where subjects may experience different type of events. To gather strength of evidence, the usual idea is to combine results across studies. A traditional meta-analysis focuses on the combination of aggregated data obtained from study publications. An alternative approach is the IPD meta-analysis where the raw data from each study are considered. It is a powerful tool that allows clinicians to reach conclusions based on independently performed studies with the possibility to explore heterogeneity across trials and whether particular individual patient characteristics (such as the age of the patient) or trial characteristics (such as particular treatment modalities) may explain some of the observed treatment heterogeneity.

This work was motivated by an IPD meta-analysis of chemotherapy in nasopharyngeal carcinoma [19], a cancer which is very frequent in South-East Asia. In this meta-analysis, the addition of chemotherapy to a standard radiotherapy regimen was associated with a significant improvement in: i) a composite endpoint, progression-free survival defined as the time from randomization to first progression (either locoregional or distant) or death from any cause (Hazard Ratio (HR) = $0.75, 95\%CI[0.69; 0.81]$), ii) overall survival (HR= 0.79 $[0.73; 0.86]$), defined as the time from randomization until death of any cause. Using cause-specific hazard regression, the authors also identified a treatment effect on time-to-locoregional relapse (CSHR = $0.73[0.64; 0.83]$) and time-to-distant relapse (CSHR =$0.67[0.59; 0.75]$). To assess the effect on cancer-related mortality, a logrank subtraction method originally proposed by Peto [92] was used which provides an estimate of CSHR =$0.76[0.69; 0.84]$. This logrank subtraction method imputes a cancer death whenever the

cause is unknown or when death occurs subsequent to recurrence, whatever the recorded cause. It calculates cause-specific mortality as the difference between overall mortality and that attributable to other causes.

In order to quantify the effect of the combination of chemotherapy and radiotherapy on time to locoregional relapse, time to distant relapse or time to death without relapse, it is appropriate to employ a competing risks approach to provide unbiased estimation of the corresponding cumulative probability of event over time (e.g. the cumulative incidence function up to a given time horizon). Such cumulative probabilities are particularly useful for quantifying the absolute benefits of treatment on particular event types, which may be more meaningful in meta-analyses than the cause-specific hazard functions, which provide relative measures of treatment benefit.

Meta-analysis of a survival endpoint is typically performed using a one-stage model using a Cox model stratified by trial or with random effects [23], or by a two-stage approach with a logrank test by trial. When individuals are exposed to competing events the analysis is much more complicated. Standard methods for data without competing risks are not applicable if there is interest in understanding the different effects of treatment on the different event types. In the competing risks framework, considerable care is needed to understand the treatment effect, with the use of competing risk endpoints like the cause-specific hazard and cumulative incidence function are necessary. The application of competing risks models has already been proposed in the context of meta-analysis using aggregated or summary data [20] where the cumulative incidence is reconstructed from available published data. However, individual-level data offers distinct advantages over aggregated data. When patient-level covariates are available, it is possible to analyze how these covariates are associated to the treatment effect. For multivariate IPD meta-analysis with survival outcomes, [21] and [22] considered estimating the correlation between event specific treatment effects to assess a potential surrogacy. Yet, no formal recommendations have been proposed for an IPD meta-analysis with competing endpoints. In this chapter, we propose and illustrate a framework for IPD meta-analysis with competing endpoints. The first step is to establish the consistency of the included studies, otherwise it is not possible to determine a general result of the included studies. We employ a one-step approach where all the individual participant data are modelled simultaneously taking into account the potential inconsistency of the included studies [93]. Both the cumulative incidence function and cause-specific hazards are considered to quantify the benefit of the treatment on different endpoints. A possible time-varying treatment effect is investigated via a landmark approach on the subdistribution hazard.

In the next Section we recall basic definitions in competing risks and the most popular

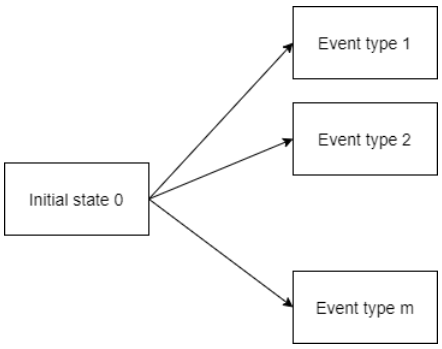

Figure 5.1: Representation of the allowed transition in the competing risks setting with m possible events.

regression model. In Section 5.3 we discuss about individual patient data meta-analysis characteristics and we describe the statistical tools that can be used in this framework. We propose a Landmark approach to understand if a the FUPs impact the estimated treatment effect. In section 5.5 we illustrate the presented methodology in our motivating example. Some discussions are made in Section 5.6.

5.2 Competing risks

Competing risks is an extension of survival analysis from a single endpoint to multiple types of events. So far, we have considered the time T as the observed time of a specific event (e.g. death); in competing risks, T represents the time until some first event. For instance, if we consider cancer data, we might be interested in multiple events, such as local relapse, or death without relapse. Usually, the survival analysis is seen as a two-state models, where patients are in the initial state 0 at time origin and at time T move to the absorbing state 1, meaning that once in state 1 the individual cannot move. In competing risks, several competing absorbing states (possible events) are introduced (Figure 5.1) and the possible transition is from the initial state to one of the competing absorbing state. In case of censoring, it may prevent the observation of the event.

Let T_j and C_j be the failure and censoring times for individual j and $\epsilon_j \in \{1, 2, ...m\}$ the failure type with m possible different types of failure. We observe the failure time $\tilde{T}_j = \min\{T_j, C_j\}$ and the status indicator $\Delta_j = \mathbb{I}(T_j \leq C_j) \times \epsilon_j$. We define the cause-specific hazard for each competing event i :

$$\alpha_i(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \delta t, \epsilon = i | T \geq t)}{\delta t}$$

it represents the transition intensity from the initial state 0 to state i, namely the instanta-

neous risk that event i occurs at time t, knowing that no other event has happened before t. The cause specific hazards completely define the stochastic behavior of the process, and their sum defines the all-cause hazard $\alpha_{\cdot}(t) = \sum_i \alpha_i(t)$. The cumulative cause-specific hazard $A_i(t) = \int_0^t \alpha_i(s)ds$ and the survival function of time T is

$$S(t) = \mathbb{P}(T > t) = \exp(-\int_0^t \alpha_{\cdot}(s)ds)$$

thus, it is a function of all the cause-specific hazards $\alpha_i(\cdot)$. The cumulative incidence function of cause i is

$$F_i(t) = \mathbb{P}(T \leq t, \epsilon = i) = \int_0^t \mathbb{P}(T > s-)\alpha_i(s)ds$$

where the probability of being in the initial state just before time s is $S(s-) = \mathbb{P}(T > s-)$. Summing over all the cause-specific cumulative incidence functions, we obtain the all-cause distribution function $F(t) = \sum_i F_i(t)$ and the survival function can also be expressed as $S(t) = 1 - \sum_i F_i(t)$.

We define the counting process $N_i(t) = \sum_{j=1}^{N} \mathbb{I}(\tilde{T}_j \leq t, \epsilon_j = i), i = 1, 2, .., m$ and the at-risk process $Y(t) = \sum_{j=1}^{N} \mathbb{I}(\tilde{T}_j > t)$, where one individual is considered at risk as long as he is in the initial state just before time t. The number of transitions out of the initial state in $[0, t]$ is $N(t) = \sum_i N_i(t)$ and the number of transitions at time t is $\Delta N(t) = \sum_i N_i(t) - N_i(t-)$. The nonparametric estimator for the cause-specific cumulative hazard is the Nelson-Aalen estimator:

$$\hat{A}_i(t) = \sum_{j=1}^{N} \frac{\Delta N_i(\tilde{T}_j)}{Y(\tilde{T}_j)} \mathbb{I}(\tilde{T}_j \leq t)$$

which can be generalized to the all-cause cumulative hazard $\hat{A}(t) = \sum_i \hat{A}_i(t)$. Moreover, the Kaplan-Meier estimator of the survival function is obtained by

$$\widehat{S}(t) = \prod_{\tilde{T}_j \leq t} \left(1 - \frac{\Delta N(\tilde{T}_j)}{Y(\tilde{T}_j)}\right)$$

In the absence of competing risks, one minus the Kaplan-Meier provides an estimate of the cumulative incidence of events over time. However, using the Kaplan-Meier estimate

to estimate the incidence function in the presence of competing risks generally results in biased results [94].

5.2.1 Regression model

In this section, we recall the popular regression models for competing endpoints discussing both cause-specific hazard and subdistribution hazard. Let X_j be a vector of baseline covariates for individual $j = 1, 2, .., N$, proportional cause-specific hazards (CSH) models assume a Cox model for each cause:

$$\alpha_i(t|X_j) = \alpha_{i0}(t) \exp(\beta_i^{CS} X_j), i = 1, 2, ..., m$$

where $\alpha_{i0}(t)$ is an unspecified baseline hazard function and β_i^{CS} the regression coefficient specific to cause i. The estimation of $\beta_i^{CS}, i = 1, 2, ..., m$ is based on maximizing the partial likelihood:

$$L(\beta^{CS}) = \prod_t \prod_{j=1}^{N} \prod_{i=1}^{m} \left(\frac{\exp(\beta_i^{CS} X_j)}{\sum_{j=1}^{N} \exp(\beta_i^{CS} X_j) Y_j(t)}\right)^{\Delta N_{ij}(t)}$$

Given $\hat{\beta}_i^{CS}$, the cumulative cause-specific hazards are estimated by the Breslow estimator.

The cumulative incidence for event i is

$$F_i(t; X_j) = \int_0^t \exp(-\int_0^v \alpha_\cdot(u|X_j)du)\alpha_i(v|X_j)dv$$

the cumulative incidence $F_i(t; X_j)$ depends on the all-cause hazard in addition to the cause-specific hazard $\alpha_i(t; X_j)$. Thus, interpretations of β_i^{CS} on the cumulative incidence scale are challenging because the one-to-one relation between all-cause and survival function is lost, and the effect of a covariate on the CSHs for a specific event can be different from its effect on the cumulative incidence function (CIF) for the same event. To tackle this problem, the subdistribution hazard has been proposed. It reestablishes the one-to-one relation with the cumulative incidence function, in fact the subdistribution hazard for event i is obtained by $h_i(t) = \frac{dF_i(t;X)}{(1-F_i(t,X))}$. It denotes the instantaneous risk of failure from the i-th event in subjects who have not yet experienced an event of type i. Thus, in this case, the risk set includes individuals who are event free at time t as well as those who have previously experienced a competing event. This differs from the risk set for the cause-specific hazard function, which only includes individuals who are currently event free.

The Fine and Gray model [95] is a popular approach to model competing risks, where a semiparametric proportional subdistribution hazards specification is considered:

$$h_i(t; X_j) = h_{i0}(t) \exp(\beta_i^{SH} X_j)$$

with $h_{i0}(t)$ an unspecified baseline subdistribution hazard function and β_i^{SH} the regression coefficients vector with $\beta_i^{SH} \neq \beta_i^{CS}, i = 1, 2, ..m$,in general. Thus, the results obtained by the two models, do not necessarily coincide and when the proportional CSH model holds, the SH model is misspecified and vice versa. However, Hjort discussed that a misspecified model still provides a consistent estimate in terms of the least false parameter, a time-averaged hazard ratio [96].

It is common to report the two analyses side-by-side without clearly distinguishing between the interpretations. The subdistribution hazard analysis only allows for a direct probability interpretation, but the absolute value of the regression coefficients are difficult to interpret. In fact, the positive regression coefficient has a qualitative meaning but interpretation on the quantitative meaning of the regression coefficient is not simple [97]. In addition, all the cause-specific hazards completely describe the data, whereas the aim of the subdistribution hazard is the analysis for one single event. The idea is to define the subdistribution time $T^\star$ until occurrence of the event of interest, with $T^\star = T$ if the event of interest has happened, otherwise, if one of the competing events occurred, $T^* = \infty$. Estimation for the Fine-Gray model is analogous to the one presented for the proportional CSH, where the subdistirbution time T^* is considered instead of T. Nowadays, the Fine-Gray model is fitted for several competing events, but it is needed to recall that it cannot generally hold simultaneously for all causes. In fact, SH model specifies the cumulative incidence function for the event of interests but not for the competing events.

Proportional hazards regression model are the most popular regression methods for competing risks. However, other models exist for both cause-specific hazards and for the cumulative incidence function. Aalen [98] introduced an additive regression model for cause-specific hazards:

$$\alpha_i(t; X_j) = \beta_{i0}(t) + \beta_i^{CS}(t) X_j$$

the $\beta_i^{CS}(t)$ is the vector of regression coefficients, thus the model allows for time-varying covariates effects. It is a nonparametric model, since regression functions are unspecified. The additive model can also be employed for the subdistribution hazards. We refer to [99] and [100] for more details.

A regression model for the cumulative incidence function $F_j(\cdot)$ have been described

in [97], where quantitative interpretation can be made on the regression coefficients. Let $F_{i0}(\cdot)$ be an unspecified function which represents the cumulative incidence for individuals with $X = 0$, the Absolute Risk Regression (ARR) model is of the form:

$$F_i(t|X_j) = F_{i0}(t)\exp(\gamma_i X_j)$$

and the probability of event during the next t years is $\exp(\gamma_i)$ times as high for a patient with $X_j = 1$ than for a patient with $X_j = 0$.

In terms of individual predictions and prediction error, no substantial differences between the two standard competing risks regression models and ARR model were identified. However, when it comes to parameter interpretation, some of the fitted models are more attractive than others. For the cause-specific hazards models, $\exp(\beta_i^{CS})$ parameters have standard rate ratio interpretations, but in competing risks, they do not directly translate to relationships between risks (cumulative incidences). For the Fine-Gray regression model $\exp(\beta_i^{SH})$ parameters are subdistribution hazard ratios and they have a quite indirect interpretation, but this model establishes a useful direct link between covariates and cumulative incidence. For the ARR model, $\exp(\gamma_i)$ are the ratios between cumulative incidences.

However, a problem possessed by all direct regression models for cumulative incidences is that the sum of all predicted cumulative incidences may exceed 1. If focus is on a single cause then one might argue that this is a minor problem but for a thorough competing risks analysis, all causes should be studied and the problem does become relevant. This problem does not occur for cause-specific hazard models[97].

5.2.2 Goodness of fit

Both the regression model previously discussed rely on the assumption of proportionality of hazards. Goodness-of-fit methods can be employed to check for this assumption. Popular methods to check, formally and graphically, for the proportionality of the CSH have been adapted to subdistribution hazard model [101]. For binary covariate, the cumulative hazards plot is a popular method for this scope. The estimated cumulative hazards within one level of covariate are plotted against the respective estimate within the other covariate level. Moreover, plotting the Schoenfeld's residuals, we would expect a null mean across time and, a non constant average would be related to a misspecification of the proportional hazards.

A formal test of the PSH assumption was proposed by [102]. It is based on the assumption that non proportionality is a result of time-varying covariates effect. A score

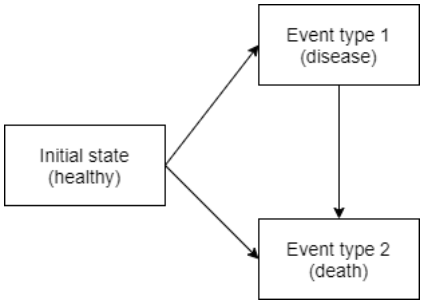

Figure 5.2: Representation of the allowed transitions for multistate setting (illness-death model).

test which employs Schoenfeld's residuals adapted to competing risks data is considered. A parametric time-varying covariate effect is considered as $\boldsymbol{\beta}^{SH}(t) = \boldsymbol{\beta}^{SH} + \mathbf{D}(t)\boldsymbol{\theta}$ with $\boldsymbol{\beta}^{SH} = (\beta_1, \beta_2, ..., \beta_p)'$, $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_b)'$ and $D(t)$ $p \times b$ matrix of pre-specified time functions with non null values for the components of X_i being tested for time-varying effects. The null hypothesis of the test is $H_0 : \boldsymbol{\theta} = \mathbf{0}$. When it is rejected the assumption of proportionality of subdistribution hazards is not satisfied. It is important to correctly specify the time functions in $\mathbf{D}(t)$ in order to detect the non-proportionality. This test can be performed via the function `PSH.test` in the `crrSC` package in `R` [103].

Furthermore, the landmark method can also be applied for evidence of non proportionality since a time-varying covariate effect is a clear indication of violation of the proportionality assumption. We consider a landmark sequence of times and, at each landmark time LM^s, the Fine-Gray model is fitted considering the individuals who are still at risk $(\tilde{T}_i > LM^S)$. If the estimated treatment effects vary with the landmark times, then this provide evidence for a time-varying effect (non proportionality).

5.2.3 Multistate model

Competing risks model time until some first event and the type of event, but do not take into analysis subsequent events, in fact only absorbing states are considered. In multistate model, events can be modelled as transitions between different states. In Figure 5.2 the illness-death model without recovery is presented where the disease process of a patient will also consist of intermediate event that can neither be classified as initial state nor as absorbing state. It is a multistate model since an individual that starts in the healthy state will have either one or two event times, in fact it could make either a healthy-dead transition or a healthy-diseased transition and successively a diseased-death transition. This model is time-inhomogeneous Markov and it satisfies the Markov property where the future course of an individual depends only on its current time and state. We can define the transition probabilities $\mathbb{P}_{li}(s,t) = \mathbb{P}(\epsilon(t) = i|\epsilon(s) = l)$, where $\epsilon(t)$ indicates the state

occupied by the individual at time t and the transition hazard for $l \to i$

$$\alpha_{li}(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(\epsilon(t+\delta t) = i | \epsilon(t) = l)}{\delta t}$$

and the cumulative hazard for transition $l \to i$: $A_{li}(t) = \int_0^t \alpha_{li}(s)ds$.

In this Chapter we do not consider multistate model, but we refer to some extensions of cause-specific and subdistribution hazards regression model that can be employed when analysing IPD meta-analyses.

5.3 IPD Meta-analysis

Meta-analysis is a formal evaluation of the quantitative evidence from two or more trials addressing the same question. The rapid increase in the number of meta-analyses is mainly due to a greater emphasis on evidence-based medicine and the need for reliable summaries of the expanding volume of clinical research. Meta-analysis is often used to assess the effect of a new treatment compared to the standard one and it is useful to obtain a more precise estimate of the overall treatment effect. A traditional meta-analysis involves synthesis of aggregate data (AD) obtained from study publications. The aggregated data usually include the mean treatment difference and its variance; a weighted average across studies is than calculated to give an overall measure of treatment effect. A more powerful approach is individual patient data (IPD) meta-analysis, where the raw data from each study are obtained. Having information at the individual level has many advantages over the aggregated data, such as the possibility to assess interaction between covariates and treatment. An other benefit of IPD is the possibility to perform subgroup analyses and to investigate different sources of heterogeneity, which is one of the principal issue in meta-analyses. For AD meta-analysis, meta-regression methods have been proposed to identify significant relationships between the treatment effect and covariates of interest [23]. However, meta-regression has low power to detect treatment–covariate interactions as it assesses across-trials relationships between study-level summaries (e.g. mean age) and treatment effect, rather than within-trial relationships between patient-level values and treatment effect.

Meta-analyses often combines results from studies which have not followed a common protocol, involving different patient populations. Thus, the main concern in the analysis is the determination of heterogeneity in the treatment difference across trials. Deciding whether or not the amount of heterogeneity is of matter and, how to deal with it is not straightforward. Test statistics and other measures for heterogeneity are discussed in

section 5.5. These would also help in the choice between fixed or random effects model. However, heterogeneity is not the only criteria to base on the model decision. For instance, the number of trials and the distribution of the study estimates of treatment effects should also be considered. Many have argued that the model decision should be based on whether the intervention effects are expected to be truly identical [104]; others have argued that a fixed-effect analysis can be interpreted in the presence of heterogeneity, and that it makes fewer assumptions than a random-effects meta-analysis [105]. Moreover, in presence of IPD, two main approaches have been proposed: one-stage and two-stage. Meta-analysis of a survival endpoint is typically performed using a one-stage model using a Cox model stratified by trial or with random effects [23], or by a two-stage approach with a logrank test by trial. For multivariate IPD meta-analysis with survival outcomes, [21] and [22] considered estimating the correlation between event specific treatment effects to assess a potential surrogacy. When individuals are exposed to competing events the analysis is much more complicated since considerable care is needed to understand the treatment effect. The application of competing risks models has already been proposed in the context of meta-analysis using aggregated or summary data [20] where the cumulative incidence is reconstructed from available published data. However, no formal recommendations have been proposed for an IPD meta-analysis with competing endpoints. In the next sections we provide some insights for IPD meta-analysis with competing risks underling: i) how to evaluate heterogeneity, ii) how to assess covariate-treatment interactions, iii) interpretation of results for competing risks.

5.3.1 One vs two stage approach

Individual patient data meta-analysis allows to use raw data to synthesise results on the quantity of interest. It is becoming increasingly popular and two main approaches can be employed to analyse this kind of data: one-stage approach and two-stage approach. The former consider all individuals from all trials together using appropriate models that can take into account the clustering. The latter, initially analyses separately individual in each trials, and than, combines the obtained results using standard meta-analysis methods. Fixed-effect and random-effect model can be chosen for both one and two-stage approaches.

Let $k = 1, ...K$ be the trial indicator and the treatment effect is of interest. For the two-stage approach, in the setting of competing risks, a possible choice is the proportional hazards model (cause-specific and Fine-Gray) to estimate the treatment effect β_k and its variance σ_k^2 in each trial k. These quantities are than combined in the second stage to obtain the overall treatment effect $\beta_\cdot$. A fixed-effect or a random-effect model is employed

depending on the assumption made on the treatment effect among trials. If the treatment effect is assumed to be the same across trials, than the fixed-effect model is used and $\hat{\beta}_k \sim \mathcal{N}(\beta^{\cdot}, \hat{\sigma}_k^2)$. The pooled treatment effect β and its variance are than estimated. The most common method is the inverse variance method [2]. If the treatment effect is assumed to vary across trials, a random-effect model is used and the between-studies variance τ^2 need also to be estimated [106, 107]. In this case, the summary estimate $\hat{\beta}^{\cdot}$ is interpreted as the average estimate of the true treatment effects.

One-stage and two-stage approaches often provide similar results. Theoretical equivalence have been discussed in [108]. Nonetheless, attention is needed, because differences may arise. When small trials are included in the study or, more precisely, study with low number of events, the assumption of normality in the second stage of the two-stage approach may be inappropriate. Instead, the one-stage method does not need any assumption on the treatment effect distribution among trials. Moreover, introducing adjusting covariates may produce different results between the two analyses, especially when the covariates are heterogeneous across trials. More in general, most differences between one-stage and two-stage approaches are due to different modelling assumptions. However, when the same assumptions are made, the results obtained by the two approaches are very similar [23].

5.3.2 Competing risks regression for clustered data

In the context of meta-analysis, for the one-stage approach, standard methods for competing risks cannot be employed since the dependence between observations in the same trials need to be taken into consideration. More in general, in many other applications of competing risks data can be correlated within clusters, such as multicentric data, or family studies and models that can handle clustered competing risks are necessary. In the analysis of cause-specific hazards functions, a proportional hazards model with common covariate effects but different baseline hazard for each cluster as proposed by Wei et al [26] can be employed. Marginal proportional hazards approach [27, 109] with both the regression coefficients and the baseline hazards having population average interpretations can also be used for cause-specific hazards. These methods are not generally appropriate for the cumulative incidence function. For subdistribution hazards, a random-effect Fine-gray model was proposed by Katsahian [110] fitting Gamma and Gaussian frailty models, and [111] introduced modifications of the nonparametric Gray's test , which is applicable with categorical covariates. Zhou et al. [112] extended the Fine–Gray model to the clustered data setting constructing a marginal proportional subdistribution hazards model under an independence working assumption. Moreover, Zhou et al [24] considered strati-

fied Fine-Gray model with common covariate effects but different baseline hazard for each cluster. We describe the latter more in details, because we think it is more relevant for our motivating example. In fact, in an IPD meta-analysis, postulating a common baseline hazard across trials is not tenable due to varying patient populations and treatment.

Let T_{kj} and C_{kj} be the failure and right censoring time for the j-th subject in the k-th study with $j = 1, 2, ..., N_k$ and $k = 1, 2, ..., K$; $\epsilon_{kj} \in \{1, 2, .., m\}$ the failure type with m possible different type of failure and $\mathbf{X}_{kj}$ a $p \times 1$ vector of time-fixed covariates measured at baseline (e.g. randomization). The observed data consists of the i.i.d. observations $\{\tilde{T}_{kj} = \min\{T_{kj}, C_{kj}\}, \Delta_{kj} = I(T_{kj} \leq C_{kj}), \Delta_{kj}\epsilon_{kj}, \mathbf{X}_{kj}\}$. We assume that (T_{kj}, ϵ_{kj}) and C_{kj} are independent given X_{kj}. For the cause-specific hazard function, we consider the following stratified version of the proportional hazards model:

$$\alpha_{ik}(t; \mathbf{X}_{kj}) = \alpha_{ik0}(t) \exp\{\boldsymbol{\beta}_i^{CS} \mathbf{X}_{kj}\} \;\; i = 1, 2, ..., m \; k = 1, 2, ...K$$

where $\alpha_{ik0}(t)$ is the baseline cause specific hazard function for cause i in the stratum (e.g. trial) k and $\boldsymbol{\beta}_i^{CS}$ is the vector of regression coefficients specific for the cause i, assumed constant across study. The assumption of constant covariate effects can be explored by fitting models separately to either individual studies or group of studies. This model can be implemented in `R` via the `coxph` function in the `survival` package considering the `strata` option in order to stratify by study [46].

We define the cumulative incidence for event i in the stratum k as

$$F_{ik}(t; \mathbf{X}) = P(T \leq t, \epsilon = i | \mathbf{X}, k) = \int_0^t \exp\big(- \int_0^v \alpha_k(u|\mathbf{X}) du \big) \alpha_{ik}(v|\mathbf{X}) dv$$

where $\alpha_k(t|\mathbf{X}) = \sum_{i=1}^m \alpha_{ik}(t|\mathbf{X})$ is the all-cause hazard in the stratum k. The cumulative incidence for event i depends on the all-cause hazard in addition to the cause-specific hazard for event i. Therefore, the effect of a covariate on the CSHs for a specific event can be different from its effect on the CIF for the corresponding event. The key idea is to consider the subdistribution hazard to recover this one-to one relation. The subdistribution hazard for event i in the stratum k is defined as $h_{ik}(t; \mathbf{X}) = dF_{ik}(t; \mathbf{X})/(1 - F_{ik}(t; \mathbf{X}))$.

Similarly to the CSHs, the baseline subdistribution hazard may vary across studies. A stratified extension of the Fine and Gray model was proposed in [24]. For event i, the model may be expressed as

$$h_{ik}(t; \mathbf{X}_{kj}) = h_{ik0}(t) \exp\{\boldsymbol{\beta}_i^{SH} \mathbf{X}_{kj}\}$$

where $h_{ik0}(\cdot)$ is the baseline subdistribution specific for the stratum k and $\boldsymbol{\beta}_i^{SH}$ are the

regression coefficients vector specific to the event i. As with the cause-specific hazard models, the coefficient is fixed across studies, thus we derive one overall covariate effect for all the studies but different baseline function may be estimated for either individual studies or groups of studies. The estimate for β_i^{SH} is obtained maximizing the partial likelihood of the subdistribution hazards

$$L(\beta_i^{SH}) = \prod_{k=1}^{K}\prod_{j=1}^{N_k}\Big(\frac{\exp(\beta_i^{SH}X_{kj})}{\sum_{l=1}^{N_k} Y_{kl}^*(\tilde{T}_{kj}\exp(\beta_i^{SH}X_{kl})}\Big)^{\Delta_{kj}\mathbb{I}(\epsilon_{kj}=i)}$$

When right censoring is present, the likelihood is weighted by the inverse censoring weighting technique, where the survival function for the censoring is estimated by Kaplan Meier estimator. The stratigied Fine-Gray method is implemented in `R` in the package `crrSC` via the `crrs` function ([103]).

In each study k, we distinguish the treatment group, where an experimental treatment is tested (e.g. chemo plus radio), and a control group where subjects are treated with the standard treatment (e.g. radiotherapy alone). Let z_{kj} be the binary treatment covariate with $z_{kj} = 1$ in the treatment group and $z_{kj} = 0$ in the control group and β_i its correspondent regression coefficient. The quantity $\exp(\beta_i)$ refers to cause-specific hazard ratio (CSHR) or subdistribution hazard ratio (SHR) depending on the stratified regression model employed. However, only $\exp(\beta_i^{SH})$ quantifies the effect of the treatment on the cumulative incidence scale. Both cause specific hazards and cumulative incidence are useful for a complete understanding of the results. As suggested by [101] the two estimates are best presented side by side.

5.3.3 Treatment interaction

Clinicians are interested in assessing how patients characteristics may be associated with variation in the magnitude of the treatment effect which is the estimation of the interaction between a covariate and treatment. For aggregated data meta-analysis, the usual unavailability of individual data has led to the application of meta-regression for predicting summary treatment effects [113, 114]. Meta-regression is a subgroup analysis that allows to investigate the effect of multiple factors to be investigated simultaneously [115]. The trial is the unit of analysis and the outcome variable is the treatment effect. The explanatory variables are factors of trials that might influence the size of treatment effect. A significant correlation with these factors suggests treatment interaction. The covariates are either true trial-level covariates which are equals to all patients in a trial, or individual-level covariates that have different values for each patient, but are then aggre-

gated into a summary trial statistic. Thus, meta-regression has some limitations in case of individual-level covariates. In fact, it is not always able to capture within-trial treatment variation across a covariate. These issues are not existent when IPD are accessible. Moreover, meta-regression should generally not be considered when there are fewer than ten studies in a meta-analysis [2].

When an IPD meta-analysis is performed, patient- and trial-level characteristics are available to explore potential heterogeneity and assess treatment effects in subgroups. In our illustrative example, we want to explore a possible interaction between treatment and age. Given the stratified Fine-Gray model, we include the information on age with x_{kj} and an interaction term with the treatment:

$$h_{ik}(t; \mathbf{Z}_{kj}) = h_{ik0}(t) \exp\{\beta_i z_{kj} + \psi_i x_{kj} + \gamma_i z_{kj} \times x_{kj}\}$$

where $\exp(\gamma_i)$ is the SHR over the trials for the change in treatment effect for a one-unit increase in x_{kj}. The interaction term is given by the sum of the within-study interactions (difference in treatment effect among the levels of x_{kj} in the study k) and of the across-studies interaction (difference in mean treatment effect between studies for different values of x_{kj}).

Furthermore, the individual data allow us to perform some subgroup analyses, where the subgroup is defined considering the individuals with same level of covariate (e.g. group of age, sex). We may be also interested on how a trial-level covariate (e.g. chemotherapy modality) affects the treatment effect or interacts with others individual-level covariates. Considering a stratified model (trial is a stratum), as described in the previous section, does not allow to simply add the trial-level covariate in the model.

Let g_{kj} be a trial-level covariate such that $g_{kj} = g_k$ for all subjects j in the trial k. When a stratified approach (at the trial level) is employed, g_{kj} cannot be introduced in the model since in each stratum k all the individuals are associated to the same value g_k. For this reason, we first stratify on the different levels $s = 1, 2, ..., S$ of g_k and we apply the stratified Fine-Gray model for a second stratification on the trial. We assume $S < K$, namely several clusters share the same covariate value. In the following section we consider the chemotherapy type as the trial-level covariate ($S = 4$, $K = 23$).

For each level s of g_k we define:

$$h_{iks}(t; z_{kjs}) = \lambda_{iks0}(t) \exp\{\beta_{i,s} z_{kjs}\}$$

where $\exp(\beta_{i,s})$ is the SHR for $g_k = s$ versus the reference stratum for cause i (i.e. the treatment effect for the specific chemotherapy modality). Differences of SHRs across

levels of g_k are related to an interaction between the treatment effect and the trial-level covariate. In order to check for the interaction (differences of SHRs) a χ^2 test with $S-1$ degree of freedom can be implemented to test the null hypothesis of homogeneity among groups. We can also extend the model introducing others individual-level covariates x_{kj} in order to test for an interaction between the trial-level covariate and others patient-level covariates. In section 5.5 we present the results of this model where the interaction treatment-age is tested for the four chemotherapy modalities.

5.4 Heterogeneity

A meta-analysis attempts to gain objectivity and generalization by combining studies that may be conducted under different conditions, involving different patient populations, different disease definitions, and different treatment regimens. We can distinguish different types of heterogeneity [116]: clinical heterogeneity due to variability in the participants, treatments and outcomes studied; methodological heterogeneity due to variability in study design and outcome measurement tools. We are interested into statistical heterogeneity, which can be a consequence of clinical and/or methodological inconsistency, and leads to variability in the treatment effects evaluated in the different trials. We will refer to statistical heterogeneity simply as heterogeneity.

A careful assessment of heterogeneity is critically important in meta-analysis in assessing whether it is reasonable to summarize the treatment effect with a single overall estimate which applies to all studies. Of course, the conclusions will be less clear when there is substantial heterogeneity and it is thus important to quantify the extent of heterogeneity when reporting the results.

A standard test for the heterogeneity is the Cochran's Q statistic test. It is a weighted sum of the squared deviation of each study's estimate (e.g. treatment effect) from the overall estimate. This test is known to have low statistical power to detect heterogeneity when the number of studies is small and, on the contrary, excessive power when the meta-analysis includes many studies [117]. Higgins pointed out that heterogeneity is likely to be expected in a meta-analysis thus, rather than testing for heterogeneity, it is more relevant to quantify the inconsistency and its impact on the analysis [118]. Several statistics have been developed and the most commonly used is the I^2 [115]. This statistic is easily interpretable and, unlike the Cochran's test does not depend on the number of trials in the meta-analysis.

Let $\hat{\beta}_k$ estimate the overall cause-specific treatment effect in the study k with precision $w_k = 1/\sigma_k^2$, obtained by the regression model for the individual study k. We define

the between-studies heterogeneity $\tau^2 = \text{Var}(\beta_k)$ and the within-study variance σ^2 as a summary statistic of the σ_k^2:

$$\sigma^2 = \frac{\sum_k w_k(df - 1)}{(\sum_k w_k)^2 - \sum_k w_k^2}$$

$df = K-1$ is the degrees of freedom of the Cochran's Q statistic. We recall the I^2 statistic which describes the percentage of total variation between the results of the studies that is due to heterogeneity :

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2} \times 100 = \frac{Q - df}{Q} \times 100$$

The I^2 measures the impact of heterogeneity rather than its quantification, as with σ^2 and τ^2. It ranges from 0 to 100 and is typically labelled as low $(0 - 40\%)$, moderate $(30 - 60\%)$, substantial $(50 - 90\%)$ and considerable $(75 - 100\%)$ [2]. However, it is necessary to consider the magnitude and the direction of the treatment effect since the interpretation of heterogeneity will depend on these as well as clinical diversity between the studies. When heterogeneity is found, it is of interest to explore its causes. A subgroup analysis may be one strategy. The forest plot can also be helpful in identifying the source of the heterogeneity. When considerable inconsistency is detected, we might question if it is appropriate to proceed with the analysis and consider one overall estimate for the treatment effect.

To quantify the heterogeneity with the I^2 statistic we consider a two-stage approach where a stratification on the trials is needed. The two-stage approach allows for the estimation of the within-trials and between-trials variability. This is the first step in a meta-analysis, but when the heterogeneity is quantified we can proceed with a one-stage approach to estimate the cause-specific treatment effect. To our knowledge, tools are not available to quantify the heterogeneity considering a one-stage approach in this context.

5.4.1 Effect of follow-up

In a meta-analysis where many trials are included, there may be concerns that studies with different FUP will yield different treatment effect estimates [20]. We might inspect whether having shorter (longer) FUP would impact on the estimation of the SHRs. Following [119], we suggest to landmark the CIF analyses to investigate the impact of FUP on the treatment effect. These landmark analyses directly reflect the effect of treatment on prediction of the cumulative incidence for the event of interest. Let LM^k be the FUP for the k-th study, we consider as landmark times the ordered sequence of the study-

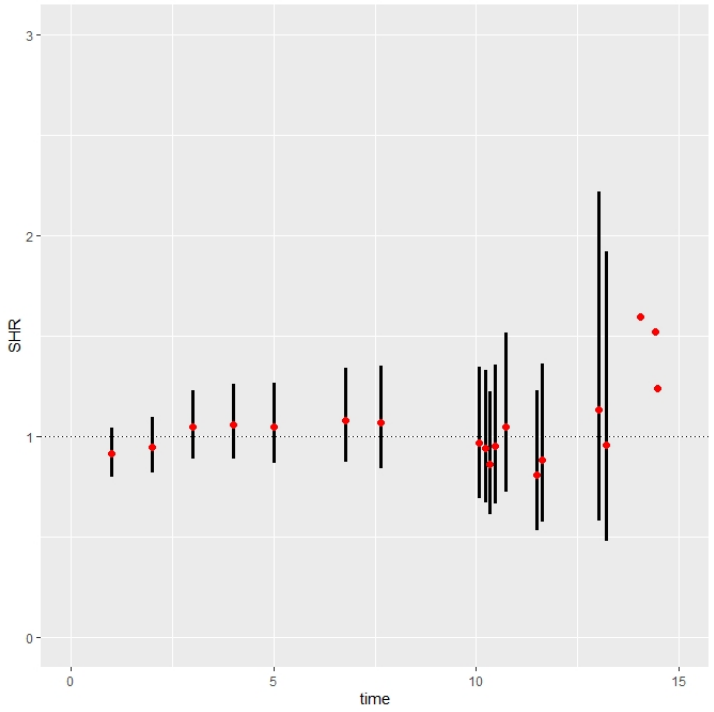

Figure 5.3: Landmark approach of the SHR defining the FUP of each study as landmark times.

specific FUP $LM_* = \{0 < ... < LM^{(1)} < ... < LM^{(K)}\}$. At each landmark time $LM^{(j)}$ we fit the stratified Fine-Gray model based on subjects who have not yet experienced any of competing risks and are still under observation ($\tilde{T}_{ik} > LM^{(j)}$). If the estimated treatment effects vary with the landmark times, then this provide evidence for a time-varying effect. In such cases, care is needed when reporting the results because different conclusions on the treatment efficacy can be made for different follow up times. An example of this Landmark approach is presented in Figure 5.3 where the FUP seems to not impacting the SHR.

5.5 Application

5.5.1 Data description

Nasopharyngeal carcinoma is a cancer that occurs in the nasopharynx which is much more frequent in Southeast Asia. It is difficult to detect early, probably because the nasopharynx is not easy to examine and symptoms of nasopharyngeal carcinoma mimic those of other, more-common conditions. Thus, most of the patients with this type of cancer present locally advanced stage. Treatment is difficult because of anatomical

proximity to critical structures; and the role of surgery is limited to biopsy for histologic confirmation and salvage of persistent or recurrent disease. Fortunately, this cancer is highly radiosensitive and chemosensitive; radiotherapy is the standard treatment and excellent control can be achieved for patients with early disease, chemotherapy has been proposed for further improvement for the majority of patients presenting with advanced locoregional disease. Chemotherapy has been used in three ways: as induction treatment, concomitant with radiotherapy and adjuvant therapy after radiotherapy.

An individual patient data meta-analysis of clinical trials that included patients with nasopharyngeal carcinoma was conducted to investigate the benefit of the addition of chemotherapy to a standard radiotherapy regimen (Ref Blanchar). A total of 4940 patients collected in 23 trials are analysed. Four trials investigated the addition of adjuvant chemotherapy to radiotherapy, seven trials the addition of concomitant chemotherapy, six trials the addition of concomitant plus adjuvant chemotherapy and six trials the addition of induction chemotherapy. The median follow-up (FUP) in the meta-analysis is estimated to be 11.8 years (ranging from 5.01 to 21.68). In this study, we are interested in the competing endpoints locoregional relapse, distant relapse and death without relapse (neither locoregional nor distant). Of the 4940 subjects 2529 (51%) were censored, 843 (17%) had a locoregional relapse, 1112 (23%) had a distant relapse and 456 (9%) died without relapse. In addition to the chemotherapy modality, which is a trial-level covariate, the age of the patient will be used as patient-level covariate in the proposed meta-regression approach.

The trials are characterized by different FUP times ranging from about 5 years to almost 22 years. No events are observed after 15 years, except deaths without relapse. We note that while the FUP are different among studies, the median of failure times are similar among studies for both locoregional relapse (1-3 years) and distant relapse (1-3 years). Thus, one wouldn't expect that the treatment effect estimates for these endpoints would be sensitive to FUP.

The original analysis combining the trials results employed a two-stage fixed effect-model for the PFS and the OS [19]. As mentioned in the introduction, the competing endpoints were addressed with proportional cause-specific hazard models for locoregional and distant relapse.

5.5.2 Statistical analysis and results

The reported analyses are stratified by trial (23 in total) and CSHRs and CIFs are provided for a complete understanding of the results. The I^2 statistic is used to investigate the impact of heterogeneity that we may observe in the meta-analysis with the inclusion

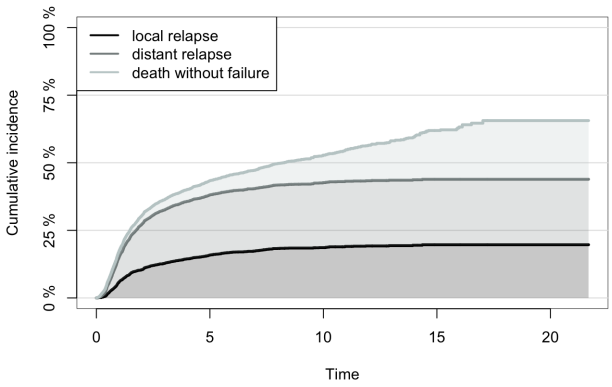

Figure 5.4: Stacked plot of the cumulative incidence functions for all individuals for all the competing time-to-event: time to local relapse (black), time to distant relapse (grey), time to death without relapse (light grey)

of different studies. We test for the assumption of proportionality in each trial (PSH test) and we look for time-varying treatment effect using a landmark approach on the subdistribution hazards. Finally, we check for an interaction between the treatment effect and age employing the stratified Fine-Gray model.

In Figure 5.4 we provide a stacked plot with the cumulative incidence functions for all the competing endpoints. We observe most of the locoregional and distant relapses at earlier times, even in studies with long FUP. The stacked plot for each arm is shown in Figure 5.5. A similar plot of the cumulative incidence for each treatment subgroup can be found in the appendix. Forest plot of the estimated treatment effects are provided in Figures 5.8 and 5.9, respectively for locoregional relapse and distant relapse.

The addition of chemotherapy to radiotherapy improves the cumulative incidence for both locoregional relapse (SHR of 0.77 $[0.67, 0.88]$) and distant relapse (SHR of 0.70 $[0.63, 0.79]$). We observe coherent results in the cause specific analysis, where CSHR for locoregional and distant relapse are 0.71 $[0.62, 0.82]$ and 0.67 $[0.60, 0.70]$ respectively. For locoregional relapse, the addition of adjuvant chemotherapy leads to stronger treatment effect with SHRs of 0.60 $[0.39, 0.92]$ for adjuvant alone and 0.63 $[0.47, 0.83]$ for concomitant plus adjuvant. The remaining two treatment subgroups present SHRs of 0.85 $[0.69, 1.05]$ for the concomitant and 0.86 $[0.65, 1.13]$ for induction, showing a non significant effect on the cumulative incidence. On the other hand, the treatment effect is significant for the cause specific hazards. The addition of adjuvant chemotherapy alone, does not show a significant improvement on distant relapse (SHR and CSHR of 0.8 [0.58,1.10]). When the adjuvant therapy is used together with a concomitant treatment we observe a strong

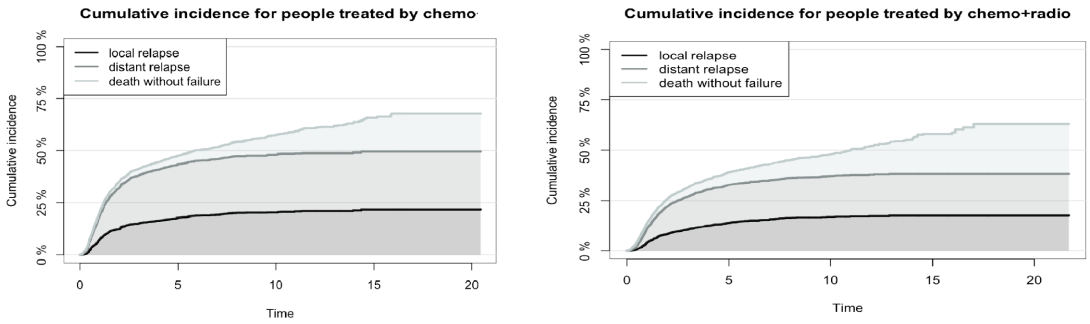

Figure 5.5: Stacked plot of the cumulative incidence functions in the two different arms (chemotherapy and chemotherapy plus radiotherapy) for all the competing time-to-event: time to local relapse (black), time to distant relapse (grey), time to death without relapse (light grey)

improvement for distant relapse (SHR of 0.62 $[0.49, 0.77]$ and CSHR of 0.57 $[0.45, 0.71]$). However, we implemented the test for interaction between the treatment effects according to the four treatment modalities. The results show no significant interaction for both local and distant relapse with a p-value of 0.17 and 0.32 respectively.

Regarding treatment effects across trials, there is evidence of mild heterogeneity across trial-specific SHRs for distant relapse ($I^2 = 34\%$), but not for locoregional relapse ($I^2 = 0\%$). The forest plot in Figure 5.9 shows that mainly the adjuvant and induction treatment subgroups are characterized by strong heterogeneity. In the former, heterogeneity is likely due to trials with small sample sizes (trial 16-17) where the adjuvant chemotherapy does not strongly improve the outcome (time to distant relapse). In the latter, it may be driven by trial 14 (77 observations) which is characterized by a less efficacious treatment effect as compared to others.

For the death without relapse endpoint, the additional chemotherapy reveals a beneficial treatment effect on the CSHR (CSHR=0.72 $[0.62, 0.82]$). However, the addition of chemotherapy has a deleterious effect on the cumulative incidence of death without relapse with a SHR=1.30 $[1.08, 1.58]$. These results may be explained by an indirect effect of the treatment on this endpoint e.g. the increased of the CIF of the experimental group is a consequence of a simultaneous decrease of the two others CIFs of locoregional and distant failure. The complete results (forest plots, goodness of fit analysis) for these endpoints are given in the appendix. Finally a mild heterogeneity is detected with an $I^2 = 29\%$.

In each trial, the PSH test is implemented considering the time functions $\log(t), t, t^2$ for evaluating the potentially non-proportionality treatment effect. In two trials, for

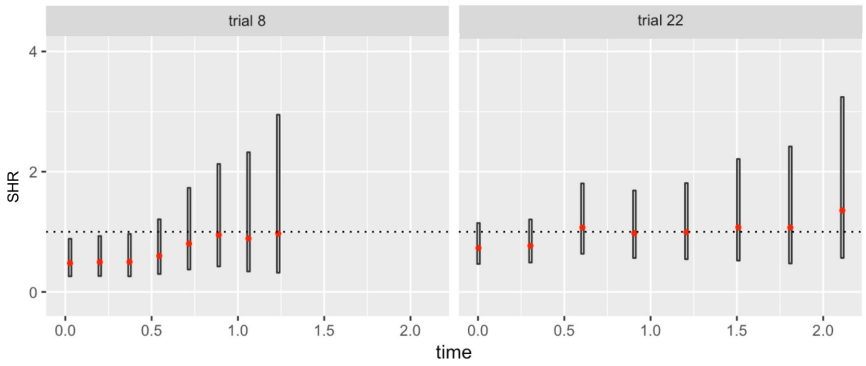

Figure 5.6: Landmark of the Fine-Gray model for local relapse in the studies where non-proportionality was detected by PSH.test. The landmark times are chosen in the interval between the minimum time of relapse and the third quartile of the failure times in the study. A time-varying SHR is linked to non proportionality of hazards. The SHRs (red dots) and the confidence interval are provided for each landmark time.

locoregional relapse, we reject the null hypothesis of proportionality: trial 22 (76 events) and 8 (42 events). In trial 22 the two cumulative incidence curves are crossing, there is a plateau in the cumulative incidence in the radiotherapy alone group and an increase in the chemotherapy plus radiotherapy group. Therefore, the treatment effect, detected during the first years of follow-up, becomes less strong for longer follow-up. In trial 8 the non proportionality could be due to the high number of events in the control arm during the first year. For distant relapse, the null hypothesis of proportionality is rejected in trial 4 (112 events) where we observe a delayed treatment effect.

These indications of non-proportionality are confirmed by the landmark analysis of the subdistribution hazards. We consider as study-specific landmark times a sequence in between the minimum and the third quartile of the observed failure times in the study. Figure 5.6 shows increasing SHR for both trial 8 and 22. At earlier landmark times, the treatment effect is detected ($\exp(\hat{\beta}_{SH}) < 1$) , but for later times, it becomes non-significant ($\exp(\hat{\beta}_{SH}) \simeq 1$) in trial 8 and $\exp(\hat{\beta}_{SH}) > 1$ (radiotherapy alone is more efficacious) in trial 22. The landmark for trial 4 (Figure 5.7) shows decreasing SHR, e.g. a significant benefit of chemotherapy is observed at longer landmark times. The Schoenfeld's residuals plot and the cumulative subdistribution plot for these trials are provided in the appendix for evidence of non-proportionality.

We test for a statistical interaction between the treatment effect and age in the stratified model, i.e. we employ the model in Section 3.4 where x_{kj} is the age of subject j in study k. We first consider age as a continuous variable and, in a second model, we split it

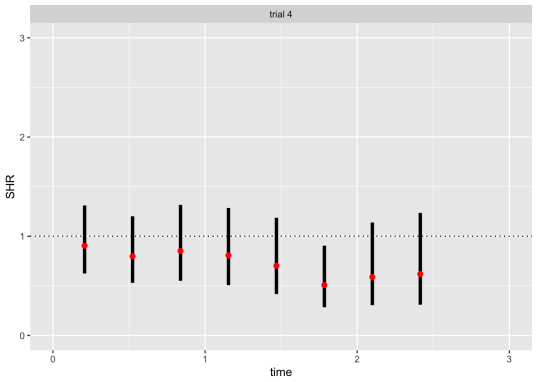

Figure 5.7: Landmark of Fine-Gray model for distant relapse in the studies where non-proportionality was detected by PSH.test. The landmark times are chosen in the interval between the minimum time of relapse and the third quartile of the failure times in the study. A time-varying SHR is linked to non proportionality of hazards. The SHRs (red dots) and the confidence interval are provided for each landmark time.

in 3 categories: < 50 years old, $[50, 59]$ years old and ≥ 60 years old. In both cases there is no significant evidence of interaction for all the competing endpoints. We further explore the interaction treatment-age according to the four chemotherapy modality groups to understand whether being younger has an impact on the treatment effect. Referring to Section 3, we consider the chemotherapy modality as the trial-level covariate (g_k with S=4). Therefore we stratify on the possible levels (adjuvant, concomitant etc.) and in each subgroup we fit the stratified Fine-Gray model where an interaction term treatment-age is introduced. A beneficial effect of the addition of the induction chemotherapy is observed for subjects in the middle category ($50 - 59$ years old) on the time-to-distant relapse with an estimated SHR for the interaction term of 2.11 $[1.13, 3.94]$.

5.5.3 Software

Several package have been developed in `R` useful for IPD meta-analysis with competing risks. The `comprsk` package [120] implements methods for regression modeling of subdistribution functions as described in Gray 1988 [95], the `crr` function models the subdistribution hazard and provide estimation for the SHR. The package `survival` [46] can also be adapted to the proportional cause-specific hazards, using the `coxph` function specifying the event as the one of interest. Employing these two packages the CSHR and the SHR are obtained and through the package `meta` [1] the forest plot can be constructed.

Concerning the proportionality assumption, the cumulative hazards plots can be obtained by using the package `etm` [121]. The latter is usually used for multistate model, and to obtain the cumulative subdistribution hazards, we consider the competing risks

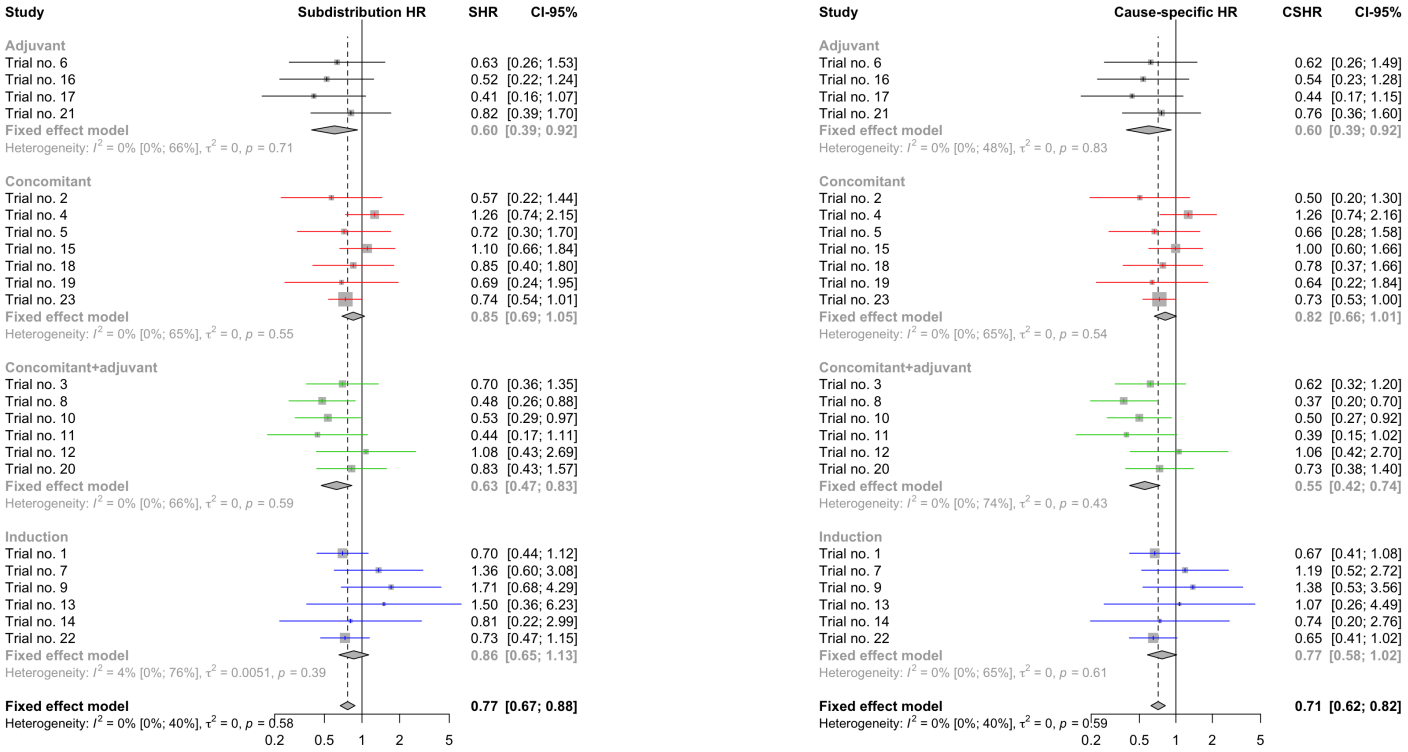

Figure 5.8: Forest plot for local relapse. Subdistribution HRs and CSHRs are provided for each trial which are grouped according to the chemotherapy modalities. The I^2 represents the heterogeneity and τ^2 the between-studies variation. Figures were created with the `R` package `meta` ([1]).

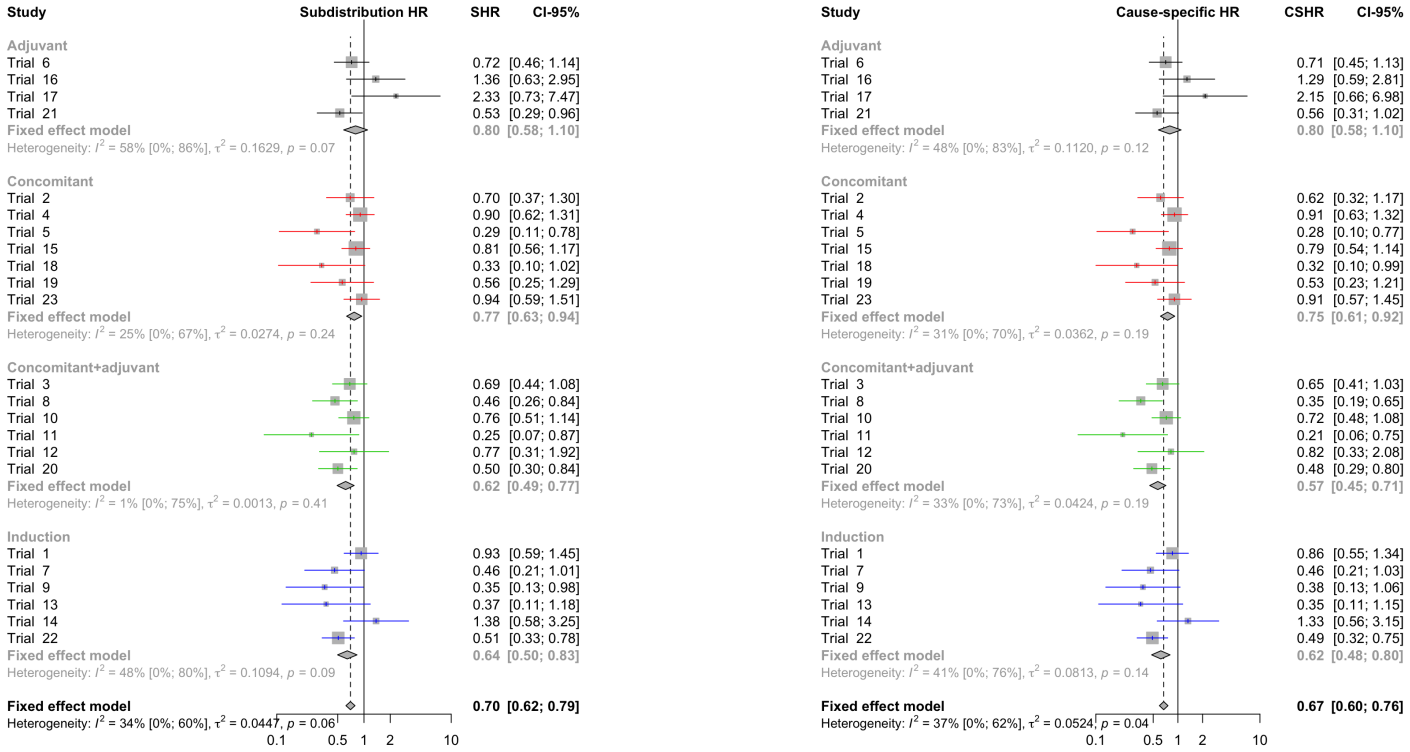

Figure 5.9: Forest plot for distant relapse. Subdistribution HRs and CSHRs are provided for each trial which are grouped according to the chemotherapy modalities. The I^2 represents the heterogeneity and τ^2 the between-studies variation. Figures were created with the `R` package `meta` ([1]).

as a special example of multistate model, where the possible transitions are from initial state to the absorbing competing events states. The Aalen-Johansen estimators for the cumulative incidence functions are obtained and than the cumulative subdistribution hazards are derived. Finally, the package crrSC [103] fits the stratified Fine-Gray method for subdistirbution hazards with clustered data, and the PSH.test is also implemented to check the proportionality assumption.

An example code for an IPD meta-analysis as described for the nasopharyngeal carcinoma can be found at https://github.com/AMeddis/Meta-analysis-for-competing-risk.

5.6 Discussion

In this chapter, we have proposed a guideline to analyse IPD meta-analysis in presence of competing risks. An IPD meta-analysis is the gold standard for synthesizing evidence for clinical interpretation based on multiple studies. Patient-level and trial-level covariates are provided and this improves power for subgroup analysis and to detect possible treatment-by-covariate interactions. A major issue is the consistency of studies that are included, and quantifying the impact of heterogeneity in the estimation of treatment effect should be the starting point of the analysis. On the basis of this quantification either a fixed effect or a random effects statistical model is usually employed. The former assumes that all studies share the same treatment effect, whereas a random effects model allows that the observed estimates of treatment effect vary across studies. Moreover, one or two-stage approaches can be employed in IPD meta-analysis. The one-stage approach consider all individuals simultaneously considering the clustering in the data; the two-stage approach firstly analyses individuals in each trials and than estimate an overall measure combining the trial-specific results. The two methods provide similar results when same assumptions are made in [23].

Note that we proceed with a one-stage approach to estimate the overall treatment effect across studies but a two-stage approach is necessary to quantify the heterogeneity, i.e. to calculate the I^2 statistic. In the presence of heterogeneity, the availability of IPD offers opportunities to explore and adjust for heterogeneity in the model. Heterogeneity in treatment effect may also be caused by follow-up length. This issue can be examined in sensitivity analyses, where one either artificially restricts follow-up or landmarks the analysis, both of which were done for analyses of the subdistribution hazards. The approach allowed us to detect possible time-varying treatment effect and thus whether different FUP impacts the estimated treatment effects. How to interpret heterogeneity across different competing risk endpoints is not straightforward, and no tools has been

proposed yet to understand the partition of the heterogeneity across such competing endpoints. We do not discuss about this aspect in our work but we think that it may be a topic of interest for further works.

Another advantage of IPD data is the ability to check for interaction between treatment effect and individual-level variables (age in our example). These analyses may be useful in identifying patient subgroups for which treatment is beneficial that can play a key role in the planning of new studies.

The primary focus of this chapter has been issues faced when individuals may experience competing events. In this framework, the interpretation of results needs considerable care; in fact, both the CSHRs and SHRs are necessary for a overall understanding of the treatment effect on the competing endpoints. The CSHRs alone are not sufficient since they do not provide information on the cumulative incidence scale, i.e. about the probability of a specific event. Andersen et al. discuss issues related to interpretation of CSHR and SHR and review alternative regression models for SHR [122]. Moreover, Klein states in favor of the additive model for the cumulative incidence functions where the covariates effect is partitioned into its comoeting parts [123]. For IPD meta-analysis, we have emphasized use of the proportional subdistribution hazard regression model, which is widely used and the default in practice.

The methods were illustrated by the re-analysis of an IPD meta-analysis in nasopharyngeal carcinoma. As in [19], a benefit of the addition of chemotherapy to radiotherapy was detected for both local and distant failure. No evidence of severe heterogeneity was identified, and we did not identify a strong treatment by age interaction. We do not allow observations within studies to be clustered because of unobserved factors shared between individuals. Such factors are assumed to be captured by stratifying the baseline hazard function by study, inducing conditional independence among individuals within studies. One might consider approaches that accommodate clustered competing risks data [112], where robust variance estimates for treatment effect estimates adjust for within study correlations.

In the application, we consider one-stage fixed-effect meta-analysis using stratified regression models with a trial-specific baseline hazard. For random effect meta-analysis, with IPD, one may separately estimate treatment effects for each trial and then summarize the distribution of the random treatment effects using the estimated fixed effect parameters, as in a standard meta-regression. Alternatively, one might consider fitting a random effects model directly using data from all trials, using, for example, the method of [110]. In such IPD analyses, one may not have a simple study-specific estimate of treatment effect, which may complicate the interpretation and the resulting inferences

are complex relative to the fixed effects models.

Acknowledgment

We would like to thank Stefan Michiels (Gustave Roussy, Villejuif, France) for providing the individual patient data meta-analysis, and Jason Fine (University of North Caroline UNC) for the opportunity to collaborate at this project.

Chapter 6

General discussion

In this dissertation we have been developed methods for clustered data in different contexts. The motivating examples were both individual patient data meta-analyses which are the gold standard to gain evidence on results in biomedical research. However, we considered the clustered structure of the data in different way depending on the main objective of the analysis. In Chapter 3 the performance of a biomarker on overall survival was of interest; meanwhile, in Chapter 5 we estimated the treatment effect on several competing events. In the former, the proposed method was not specific to meta-analysis, but to a more general setting where observations within cluster are assumed to be correlated because of unmeasured factors. The use of shared frailty model instead of a Cox model resulted in unbiased results for the estimation of the covariate-specific time dependent ROC curve with clustered survival data. Simulation results showed that the method introduced by Song and Zhou [9] was inappropriate in presence of clustered data for both continuous and discrete biomarker. On the contrary, the nonparametric estimator by IPCW [56] was unbiased with clustered survival and coincide with the proposed one under non informative cluster size. Moreover, the simulations showed biased results under strong misspecification of the frailty distribution. The definition of the frailty distribution is a challenging point that has to be addressed when employing frailty models. We refer to [10] for a more in dept discussion. The misspecification problem has not been explored for the biomarker model because we agree that since the biomarker is an observable variable, we should be confident in choosing a parametric model. In fact, a more general model could be employed for the biomarker but various sources of bias may be avoided with a more detailed study of its distribution.

By construction of the model, the assumption of a homogeneous biomarker across clusters has been considered. This directly follows from the fact that in a mixed effect model the random effect must be independent on the covariates [71]. Assuming the biomarker to

be homogeneous is reasonable when the technology for biomarker's measurement is either standardized among clusters or centralized.

Furthermore, we discussed about interpretations of results in case of clustered data underlying that the target population has to be specified in advance and a careful choice of the model is needed. In the assessment of biomarker performance in clustered data, when the biomarker is homogeneous, inference for all observed member and for typical observed member population produces same results under non informative cluster size. In the Appendix, the proof for equivalence of the two definitions of time dependent TPR is provided. This was confirmed by simulations showing same results for the proposed method and IPCW method. Still, the assumption of non informative cluster size was not initially discussed. A previous simulation generating the clusters sample sizes by depending on the frailty was conducted providing incorrect results for both methods; successively, the issue of informative cluster size was analysed.

Both the assumptions of NICS and homogeneous biomarker were met by the IMENEO study. This is an IPD meta-analysis of patients with non metastatic breast cancer and the biomarker in analysis was the circulating tumor cells (CTCs) count. The tumor-stage specific time dependent ROC curve and AUC were of interest since the tumor stage was both related to CTCs counts and overall survival.

In Chapter 5 an other individual patient data meta-analysis was considered. Patients with nasopharyngeal carcinoma were followed to assess the benefit of addition of chemotherapy to radiotherapy. In particular, we were interested in the treatment effect on locoregional relapse and distant relapse. The availability of individual characteristics leads to big advantages in the analysis. It allows to explore the possible source of heterogeneity and to detect treatment interactions. Moreover, the one-stage approach where all the observations are considered at once can be employed instead of combining results obtained by each group to have an overall treatment effect.

In meta-analyses it is essential to consider the extent to which the results of studies are consistent with each other. In fact, the presence of heterogeneity defines how much conclusions can be generalizable. It is clearly of interest to investigate the causes of heterogeneity among results of trials. This is problematic since there are often many characteristics that vary across trials from which one may choose. When considerable variation in results is detected, and particularly if there is inconsistency in the direction of effect, it may be misleading to use an average value for the intervention effect. Heterogeneity may be explored by conducting subgroup analyses. Moreover, the follow up (FUP) time usually differs among trials and a landmark on the cumulative incidence function can be used to determine whether FUP impacts the treatment effect.

In competing risks, interpretation of results is not trivial. Extension of the cause specific hazards model and Fine-Gray model for clustered data were discussed. The stratified extensions of the two methods were used for the analysis, where a different baseline for each trial is considered. We acknowledge that some issues for the interpretation of the subdistribution hazards were mentioned in [122] with a focus on alternatives to the proportional subdistribution hazard regression models. The use of additive risks models or transformation models could provide insights not captured by the subdistribution hazard model. Our focus is on the most commonly used models in order to provide a guideline on how to handle IPD meta-analysis with competing risks.

The assumption of (non) informative cluster size is often considered without further formal evaluation. However, informative cluster sizes may arise with clustered data and appropriate methods are needed to obtain unbiased results. Moreover, methods that handle ICS when it is not needed lead to loss of efficiency [81]. An other contribution is the definition of a test for informative cluster size with clustered survival data. The test performs well for both highly clustered data and few big clusters (meta-analysis). Low power was detected for a not sufficient enough number of clusters. By contrast to the test introduced in [18] for linear regression, the asymptotic distribution is provided and the test is fast to compute.

No covariates are introduced at the moment, but employing the Breslow-estimator is a possible extension of the method. In presence of covariate, the effect of the covariate might differ between clusters of different sizes. In this case, Pavlou [79] defined informative covariate structure when the conditional expectation of the outcome for a member given covariates for that member and the cluster size depends on the covariate values of other members in the cluster where the member in question belongs. This is an other issue that can occur simultaneously with informative cluster size and standard methods are considered inappropriate.

The proposed test is useful to identify ICS with right censored survival data. Yet, we think that it can also be of interest the formulation of an index to quantify the information carried by sample size. The difference between the results obtained by the two marginal analyses could be an idea. Still, intuitively, the problem of low power for a small number of clusters will not be solved. The determination of an index considering the between-groups and the within-group variability regrouping clusters by the sample size could be an other solution.

Further projects

During the last two years of the PhD I have been working in Servier, Paris as a consultant(" Doctorat conseil") in the biomarker team. In the first year together, we have worked on several short projects mostly based on modelisation of clustered data (not necessarily time-to-event). The common question to address was: Which model does fit best this specific data set? My role consisted in examining more in details the structure of the data and conducting some simulations to determine the best model to employ in that particular situation.

The last year we have been working on a bigger project on the assessment of sample size for future experiments in the context of omics data. These, are characterized by the problem of multiple testing and thus some methods to correct the pvalues have been proposed. The scope of the project was not to develop a new methodology, but to do a review on all the existing ones and create some standard functions in `R` that can be used from the team for the analysis of pilot data to determine the sample sizes by defining the desired power and controlling the false discovery rate.

Furthermore, during this year, I had the possibility to start a project in collaboration with Stephen R. Cole, Professor of Epidemiology in the University of North Carolina (UNC). The aim of this work is to estimate the per treatment effect by parametric g-formula in randomized clinical trials with time varying outcome. We consider a randomized clinical trial conducted to compare epirubicin, cisplatin, and capecitabine (ECX) with fluorouracil, leucovorin, and irinotecan (FOLFIRI) as treatments in patients with advanced gastric cancer, where more than 30% of patients discontinued protocol.

Appendix A

ROC(t,x) curve with clustered survival data

Equivalence between TPR_{aom} and TPR_{tom}

Under non informative cluster size (NICS), if $Y_{r_k} \perp\!\!\!\perp N_k$

$$TPR_{tom}(t, y) = TPR_{aom}(t, y) \ \forall y$$

while under ICS they differ in general. As defined in section 2.2

$$\begin{aligned} TPR_{aom}(t, y) &= \frac{\mathbb{E}[N_k \mathbf{I}(Y_{r_k} \geq y) | D_{r_k}(t) = 1]}{\mathbb{E}[N_k | D_{r_k}(t) = 1]} \\ &= \frac{\mathbb{E}\big[\mathbb{E}[N_k \mathbf{I}(Y_{r_k} \geq y) | D_{r_k}(t) = 1, N_k] | D_{r_k}(t) = 1\big]}{\mathbb{E}[N_k | D_{r_k}(t) = 1]} \\ &= \frac{\mathbb{E}\big[N_k \mathbb{E}[\mathbf{I}(Y_{r_k} \geq y) | D_{r_k}(t) = 1, N_k] | D_{r_k}(t) = 1\big]}{\mathbb{E}[N_k | D_{r_k}(t) = 1]} \\ &= \frac{\mathbb{E}\big[N_k \mathbb{P}(Y_{r_k} \geq y | D_{r_k}(t) = 1, N_k) | D_{r_k}(t) = 1\big]}{\mathbb{E}\big[N_k | D_{r_k}(t) = 1\big]} \end{aligned}$$

Considering that $\mathbb{P}(Y_{r_k} \geq y | D_{r_k}(t) = 1, N_k) = \int_y^\infty \mathbb{P}(Y_{r_k} = z | D_{r_k}(t) = 1, N_k) dz$ and using the Bayes theorem we can rewrite

$$\mathbb{P}(Y_{r_k} \geq y | D_{r_k}(t) = 1, N_k) = \int_y^\infty \frac{\mathbb{P}(D_{r_k}(t) = 1 | Y_{r_k} = z, N_k) \mathbb{P}(Y_{r_k} = z | N_k)}{\mathbb{P}(D_{r_k}(t) = 1 | N_k)} dz \qquad (A.1)$$

For the denominator

$$\mathbb{P}(D_{r_k}(t)=1|N_k)=\int_{-\infty}^{\infty}\mathbb{P}(D_{r_k}(t)=1|Y_{r_k}=z,N_k)\mathbb{P}(Y_{r_k}=z|N_k)dz$$

Assuming NICS and $Y_{kj} \perp\!\!\!\perp N_k$

$$\mathbb{P}(D_{r_k}(t)=1|N_k)=\int_{-\infty}^{\infty}\mathbb{P}(D_{r_k}(t)=1|Y_{r_k}=z)\mathbb{P}(Y_{r_k}=z)dz=\mathbb{P}(D_{r_k}(t)=1)$$

Substituting in (A.1)

$$\mathbb{P}(Y_{r_k}\geq y|D_{r_k}(t)=1,N_k)=\int_{y}^{\infty}\frac{\mathbb{P}(D_{r_k}(t)=1|Y_{r_k}=z,N_k)\mathbb{P}(Y_{r_k}=z|N_k)}{\mathbb{P}(D_{r_k}(t)=1)}dz$$

Assuming NICS and $Y_{kj} \perp\!\!\!\perp N_k$

$$\begin{aligned}
&=\int_{y}^{\infty}\frac{\mathbb{P}(D_{r_k}(t)=1|Y_{r_k}=z)\mathbb{P}(Y_{r_k}=z)}{\mathbb{P}(D_{r_k}(t)=1)}dz\\
&=\int_{y}^{\infty}\mathbb{P}(Y_{r_k}=z|D_{r_k}(t)=1)dz\\
&=\mathbb{P}(Y_{r_k}\geq y|D_{r_k}(t)=1)
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
TPR_{aom}(t,y)&=\frac{\mathbb{E}\big[N_k\mathbb{P}(Y_{r_k}\geq y|D_{r_k}(t)=1)|D_{r_k}(t)=1\big]}{\mathbb{E}\big[N_k|D_{r_k}(t)=1\big]}\\
&=\frac{\mathbb{P}(Y_{r_k}\geq y|D_{r_k}(t)=1)\mathbb{E}\big[N_k|D_{r_k}(t)=1\big]}{\mathbb{E}\big[N_k|D_{r_k}(t)=1\big]}\\
&=\mathbb{P}(Y_{r_k}\geq y|D_{r_k}(t)=1)\\
&=\mathbb{E}\big[\mathbf{I}(Y_{r_k}\geq y)|D_{r_k}(t)=1\big]=TPR_{tom}(t,y)
\end{aligned}$$

Negative binomial distribution

Let $Y|X$ be a random variable following a negative binomial distribution, we define

$$\mathbb{P}(Y=y|X)=(\Gamma(y+d))/(d+\mu_X)\times(d/(d+\mu_x))^d\times(\mu_X/(\mu_x+d))^y$$

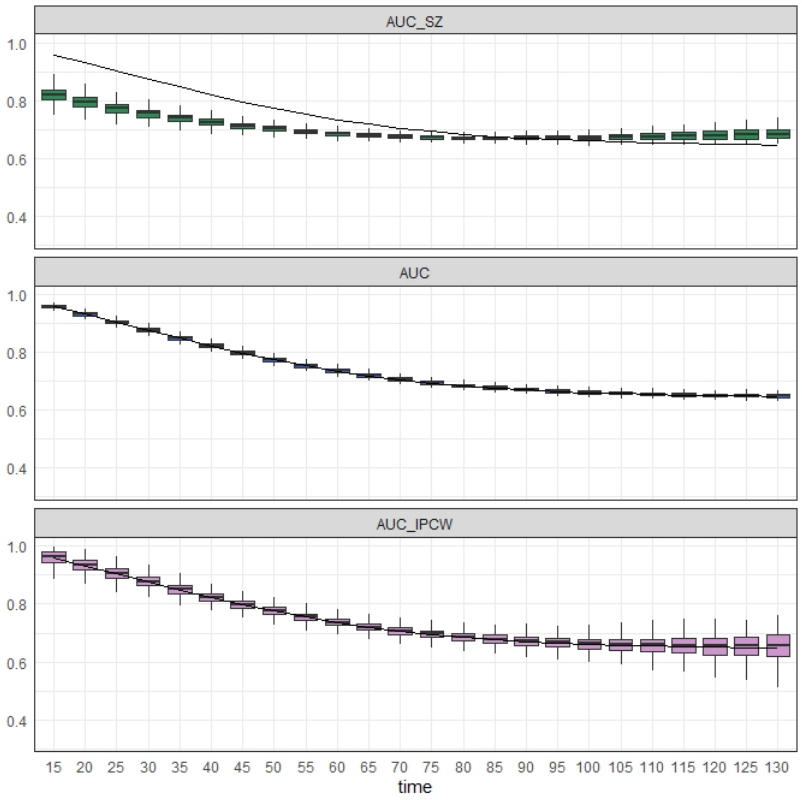

Figure A.1: AUc for all the methods in all the time, instead of just the three times, it shoes that SZ is biased and the others are not (under NICS)

with $\mu_X = \mathbb{E}(Y|X)$ and $Var(Y|X) = \mu_X(1 + \mu_X/d)$. $\Gamma(s) = \int_0^\infty z^{s-1} \exp^{-z} dz$ is the gamma function and d is the dispersion parameter.

Simulation results

The estimated AUC by the proposed method, by the nonparametric one using IPCW and by the Song and Zhou method for each time. Data were generated assuming a negative binomial biomarker and a shared frailty model, as described in the simulation section.

Results on the estimated parameters

We provide the coefficients estimated in the simulation study in the Table A.1. As in the manuscript, β and γ are the coefficients of the shared frailty model with a Gamma frailty distribution with parameter θ; the biomarker $Y|X$ follows a negative binomial distribution with set of parameter $\psi = (d, \xi)$ where d is the dispersion parameter and ξ the regression coefficient for the covariate X.

	tvalue	estimate (sd)
β	0.8	0.799 (0.017)
γ	0.5	0.501 (0.051)
θ	1	1.028 (0.145)
d	0.5	0.501 (0.016)
ξ	0.4	0.399 (0.012)

Table A.1: Results of simulation: parameters.

Assumptions for IMENEO data

We propose some visualization to check the assumptions of the proposed method: non informative cluster size, an homogeneous biomarker among clusters and the gamma frailty distribution for the shared frailty model.

Non informative cluster size

We provide the Kaplan-Meier estimator of the survival function at time t^*=30 months for each cluster in order to study the relationship between the cluster sample sizes N_k and the outcome. The figure A.2 does not suggest informative cluster size as no trend can be defined for increasing (decreasing) sample sizes.

Homogeneous biomarker

To understand whether the marker varies among center in Figure is proposed the boxplot of the CTCs count in different trials. Cremona showed a different distributions where the CTCs count has higher values, 20% of women presents a number of CTCs greater than 5. This is a geriatric hospital with 4 number of events over 45 observations. We agreed to discard this center for the analysis.

To strengthen the homogeneity of CTCs a loglikelihood ratio test (LRT) for the null variance of the random effect was employed. A random intercept model was considered and the null hypothesis is that the variance of the random effect is null. The statistic is a 50:50 mixture of $\chi^2(1)$ and $\chi^2(0)$ and the corresponding p-value is half of the p-value obtained if considering incorrectly the asymptotic distribution $\chi^2(1)$ [124]. The test confirmed the assumption of an homogeneous distribution of CTCs count rejecting the null hypothesis with a pvalue=0.33.

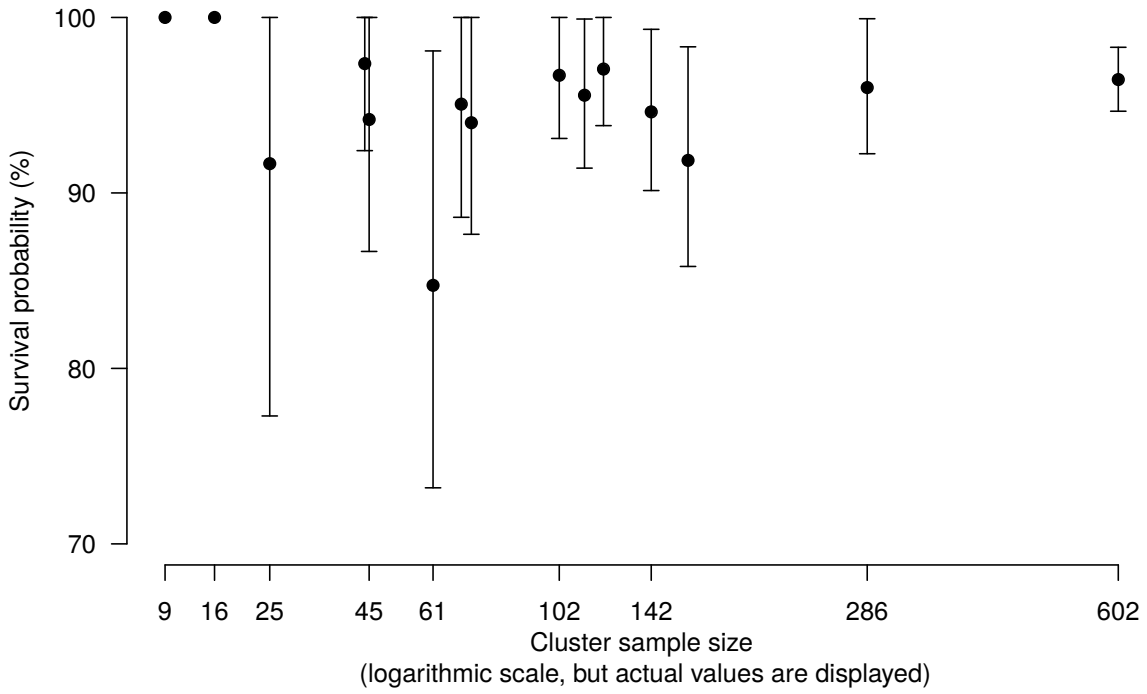

Figure A.2: Kaplan-Meier estimator of the survival function at time t=30 months in each cluster for subjects with noninflammatory breast cancer.

Gamma frailty distribution in the IMENEO

In the motivating example for non metastatic breast cancer, we assume a gamma frailty distribution. To check for the adequacy of this assumption, we compare the estimated marginal survival function by a shared frailty model with the Kaplan-Meier estimator (Figure A.4).

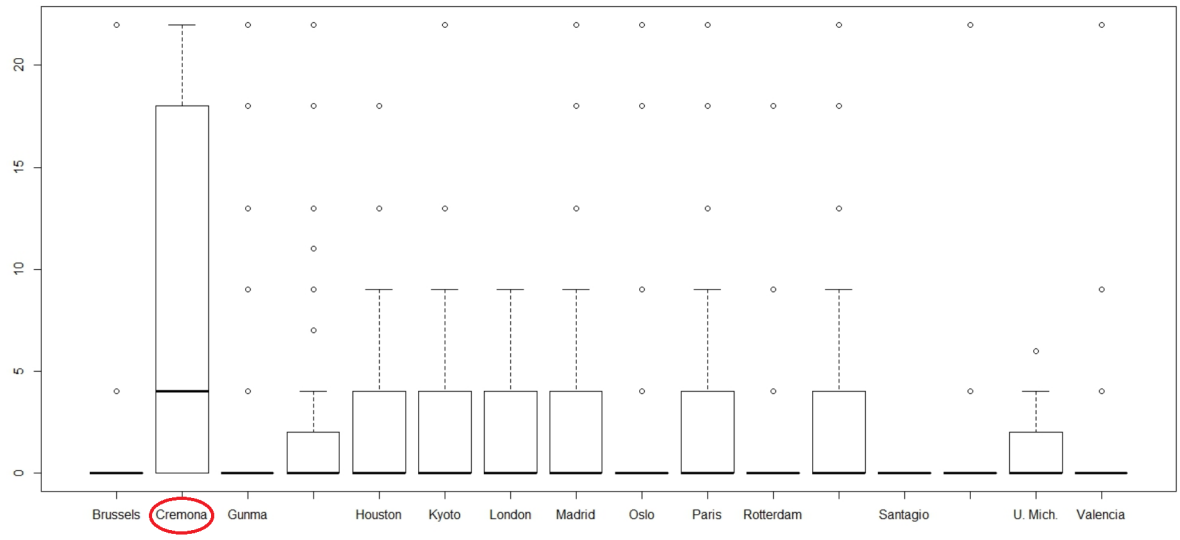

Figure A.3: Boxplot of the observed CTCs in different center.

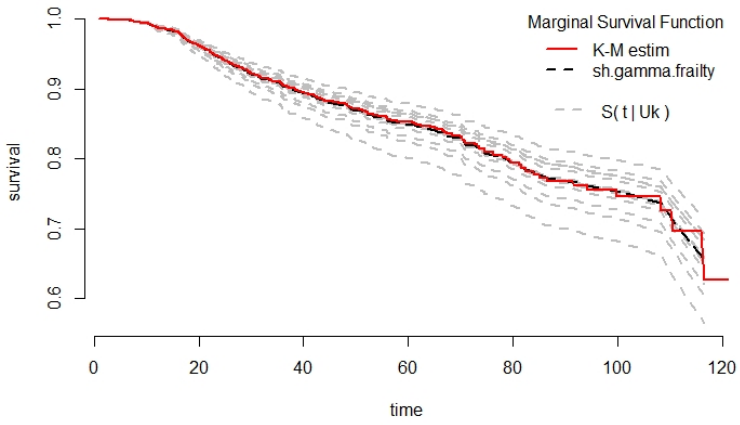

Figure A.4: Marginal survival function estimated by the shared gamma frailty model (in black) and by the Kaplan-Meier estimator (in red). We also provide the estimated conditional survival functions for each cluster $S(t|U_k)$ (in gray).

Appendix B

Informative cluster size

Impact of γ on ICS

To obtain informative cluster size, we generate K clusters with sample size $N_k \sim Pois(\lambda \exp(V_k))$ where λ, common between clusters, represents the expected number of observations in each cluster and V_k defines the cluster-specific sample size. Let U_k be the frailty term for the shared frailty model employed to generate the failure times. To create the dependence between the sample size N_k and the failure times $T_{kj}s$, we generate (U_k, V_k) from a multivariate Gamma with unit mean and covariance matrix Σ. The variance $\sigma_U^2 = 1/\theta$ defines the variability of failure times among clusters. The variance $\sigma_V^2 = 1/\gamma$ represents the variability between clusters sample sizes. The parameter ρ is the correlation between the two random effects. The strength of ICS depends on θ, ρ, γ.

In this section we explore how ICS changes with γ. We generate (U_k, V_k) for 100 clusters with $\gamma \in \{3, 10, 40\}$. Figure B.1 shows that U_k increases faster for higher γ, because the range of V_k (sample size) becomes narrower but θ is fixed, and thus the range of U_k (failure times) does not change. This translates in higher informative cluster size. In Figure B.2 we provide the mean failure times $\overline{T}_k$ for each cluster sample sizes: for small values of V_k (sample sizes) the U_k will be lower for increasing γ and so failure times will be larger when $\gamma = 40$; bigger values of V_k are associated to bigger U_k and thus to shorter failure times. Therefore, for two fixed sample sizes, the difference of the associated failure times will be larger with an increasing value of γ and informative cluster size is stronger. However, this difference is not visible anymore when the mean clusters sample size decreases ($\lambda = 5$) because, the cluster sample sizes are similar when $\gamma = 40$.

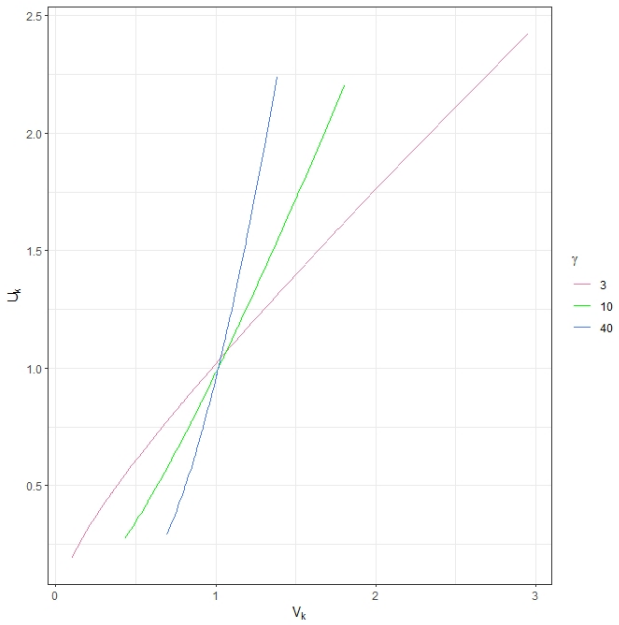

Figure B.1: Representation of the two random effects U_k and V_k generated for 100 clusters with different values of γ.

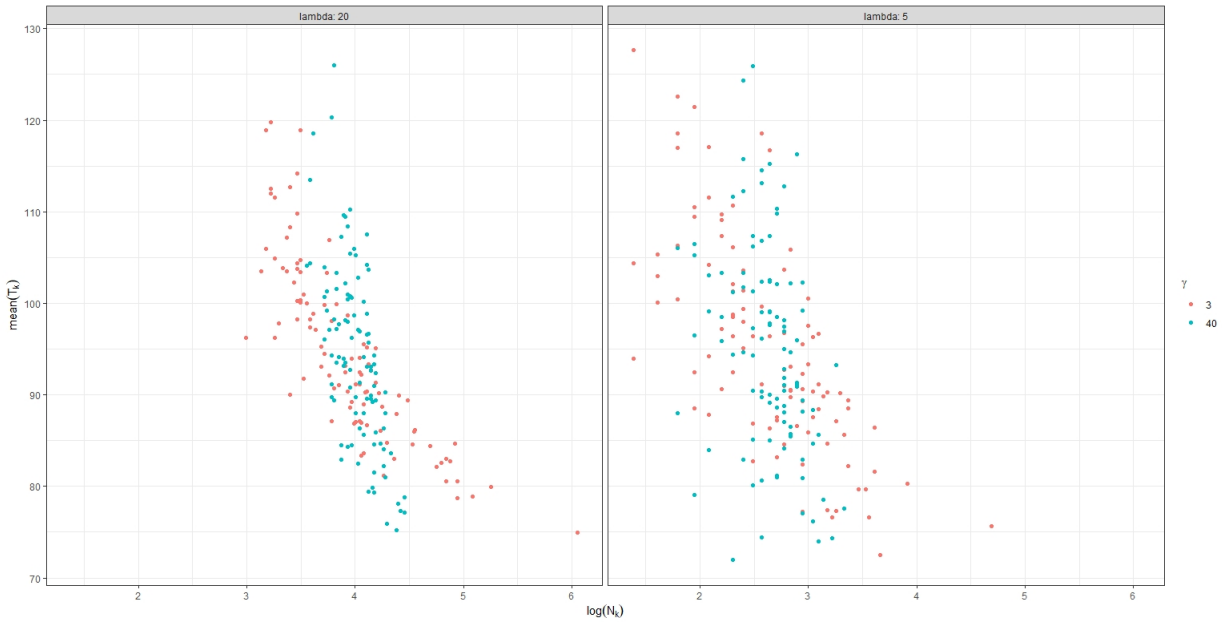

Figure B.2: Plot of median failure times T_k and the cluster sample sizes N_k (logarithm scale) associated to the random effects (U_k, V_k). Data for 100 clustered are generated by a shared frailty model and a Poisson distribution as described in the simulation section in Chapter 4. The parameter λ of the Poisson distribution represents the mean sample size of clusters if no variability is present in the sample sizes distribution ($\gamma = \infty$).

Appendix C

IPD meta-analysis with competing risks

Cumulative incidence function

In Figure C.1 we provide the cumulative incidence functions for each of the chemotherapy modalities. The proportion of event in the population treated by concomitant plus adjuvant chemotherapy is likely to be higher compared to the other groups.

Forest plot for distant without relapse

Results for the competing event of death without failure are shown in Figure C.2. The subdistribution hazard ratios and cause specific hazard ratios are provided to examine the treatment effect (addition of chemotherapy to radiotherapy) and the I^2 statistics quantifies the impact of heterogeneity on the results.

Additional results for non proportionality

The Schoenfeld's residuals plot and the cumulative subdistribution plot are graphical methods to study the proportionality of the subdistribution hazards. The plots for studies where non-proportionality was detected by PSH.test are illustrated in Figures C.3 and C.4.

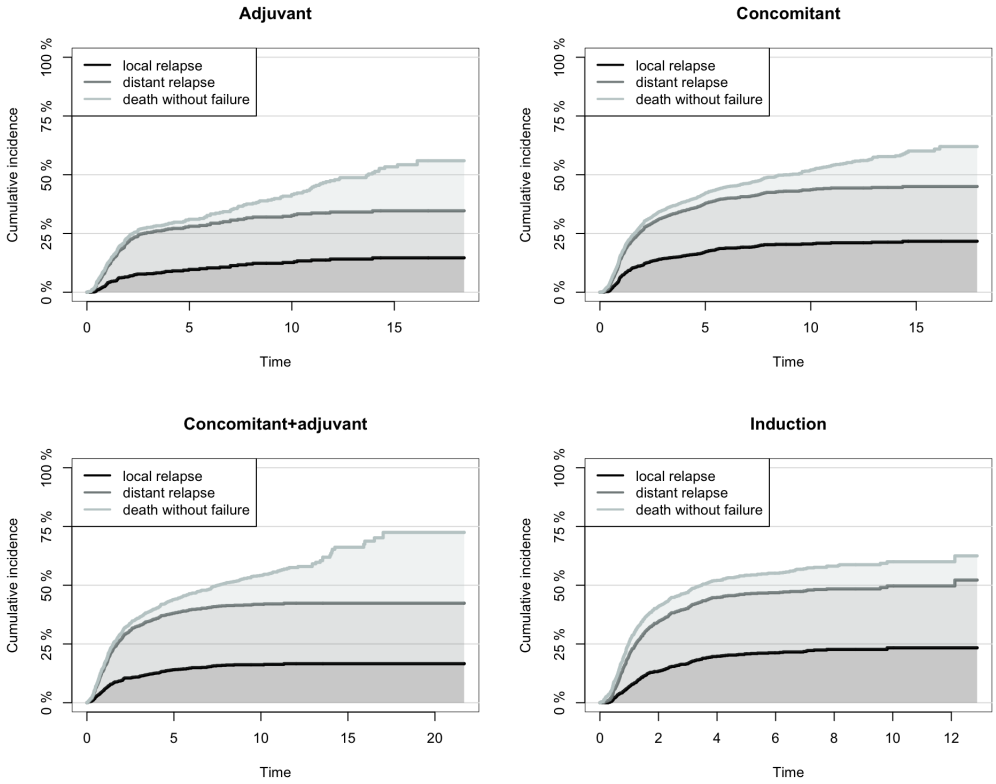

Figure C.1: Stacked plot of the cumulative incidence functions in each treatment subgroup for all the competing time-to-event: time to local relapse (black), time to distant relapse (grey),time to death without relapse (light grey)

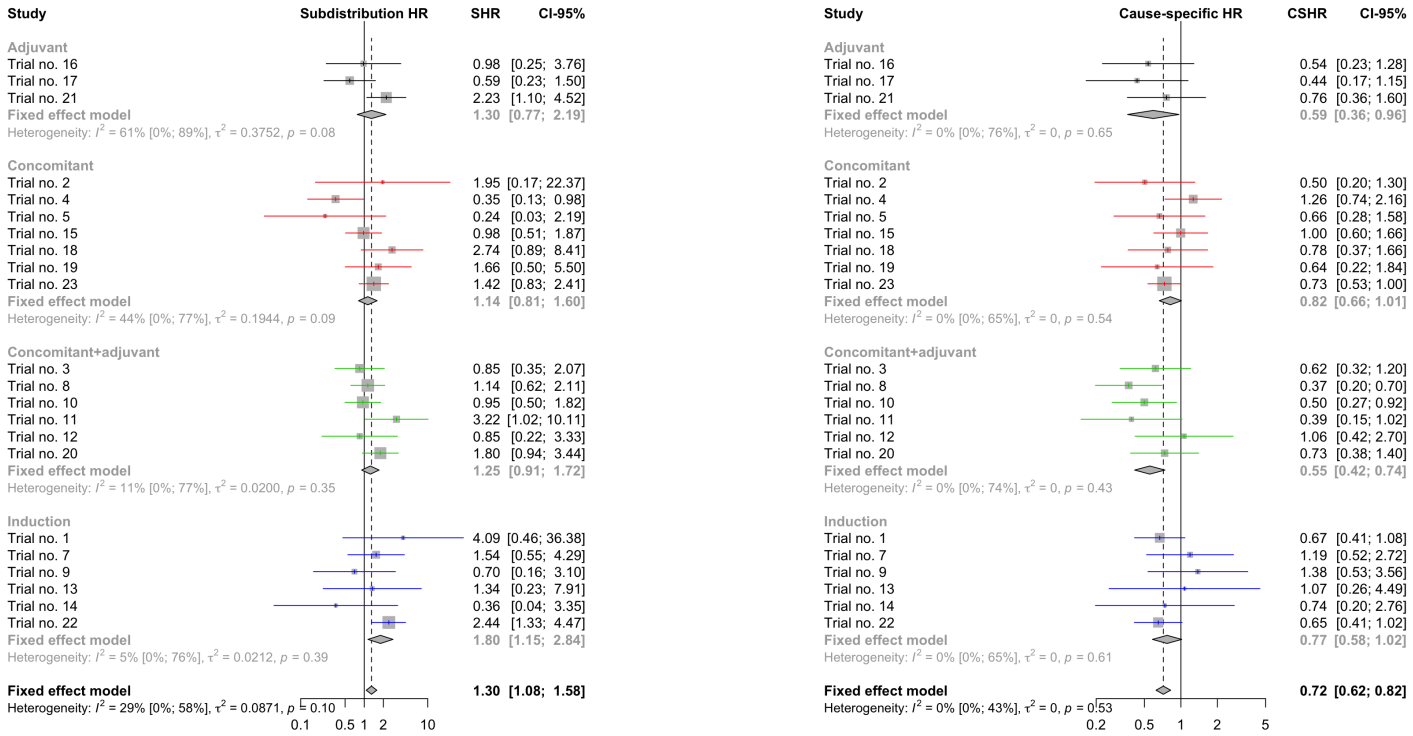

Figure C.2: Forest plot for death without failure. Subdistribution HRs and CSHRs are provided for each trial which are grouped according to the chemotherapy modalities. The I^2 represents the heterogeneity and τ^2 the between-studies variation.

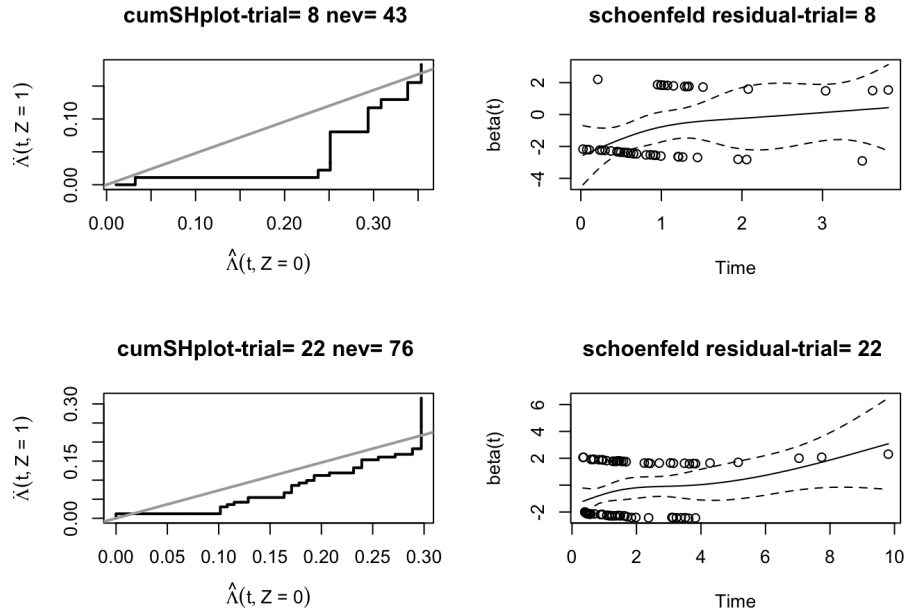

Figure C.3: Schoenfeld's residuals plot and cumulative subdistribution hazards plot for the studies where non-proportionality was detected by PSH.test (local relapse).

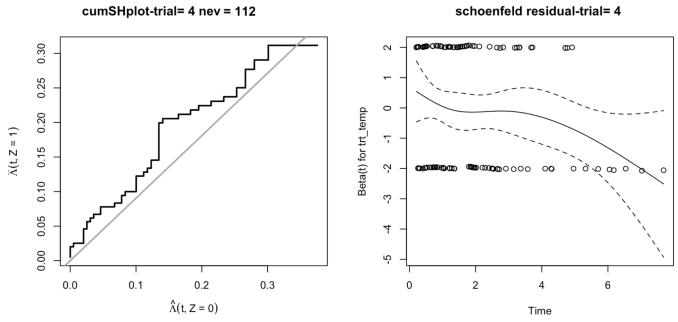

Figure C.4: Schoenfeld's residuals plot and cumulative subdistribution hazards plot for the studiy where non-proportionality was detected by PSH.test (distant relapse).

Bibliography

[1] G. Schwarzer, "meta: An R package for meta-analysis," R News, vol. 7, no. 3, pp. 40–45, 2007.

[2] J. P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, Cochrane handbook for systematic reviews of interventions. John Wiley & Sons, 2019.

[3] P. Hougaard, "A class of multivanate failure time distributions," Biometrika, vol. 73, no. 3, pp. 671–678, 1986.

[4] P. Hougaard, Analysis of multivariate survival data. Springer Science & Business Media, 2012.

[5] C. Paddy Farrington, S. Unkel, and K. Anaya-Izquierdo, "The relative frailty variance and shared frailty models," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 74, no. 4, pp. 673–696, 2012.

[6] F. Bidard, S. Michiels, S. Riethdorf, V. Mueller, L. J. Esserman, A. Lucci, B. Naume, J. Horiguchi, R. Gisbert-Criado, S. Sleijfer, et al., "Circulating tumor cells in breast cancer patients treated by neoadjuvant chemotherapy: a meta-analysis," JNCI: Journal of the National Cancer Institute, vol. 110, no. 6, pp. 560–567, 2018.

[7] M. Cristofanilli, G. T. Budd, M. J. Ellis, A. Stopeck, J. Matera, M. C. Miller, J. M. Reuben, G. V. Doyle, W. J. Allard, L. W. Terstappen, et al., "Circulating tumor cells, disease progression, and survival in metastatic breast cancer," N Engl J Med, vol. 2004, no. 351, pp. 781–791, 2004.

[8] H. Janes and M. S. Pepe, "Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting," American Journal of Epidemiology, vol. 168, no. 1, pp. 89–97, 2008.

[9] X. Song and X. Zhou, "A semiparametric approach for the covariate-specific roc curve with survival outcome," Statistica Sinica, vol. 18, pp. 947–965, 2008.

[10] P. Hougaard, "Frailty models for survival data," Lifetime data analysis, vol. 1, no. 3, pp. 255–273, 1995.

[11] P. J. Heagerty and Y. Zheng, "Survival model predictive accuracy and roc curves," Biometrics, vol. 61, no. 1, pp. 92–105, 2005.

[12] J. M. Williamson, H.-Y. Kim, A. Manatunga, and D. G. Addiss, "Modeling survival data with informative cluster size," Statistics in medicine, vol. 27, no. 4, pp. 543–555, 2008.

[13] E. B. Hoffman, P. K. Sen, and C. R. Weinberg, "Within-cluster resampling," Biometrika, vol. 88, no. 4, pp. 1121–1134, 2001.

[14] X. J. Cong, G. Yin, and Y. Shen, "Marginal analysis of correlated failure time data with informative cluster sizes," Biometrics, vol. 63, no. 3, pp. 663–672, 2007.

[15] J. M. Williamson, S. Datta, and G. A. Satten, "Marginal analyses of clustered data when cluster size is informative," Biometrics, vol. 59, no. 1, pp. 36–42, 2003.

[16] S. R. Seaman, M. Pavlou, and A. J. Copas, "Methods for observed-cluster inference when cluster size is informative: A review and clarifications," Biometrics, vol. 70, no. 2, pp. 449–456, 2014.

[17] E. Benhin, J. N. K. Rao, and A. J. Scott, "Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes," Biometrika, vol. 92, no. 2, pp. 435–450, 2005.

[18] J. Nevalainen, H. Oja, and S. Datta, "Tests for informative cluster size using a novel balanced bootstrap scheme," Statistics in medicine, vol. 36, no. 16, pp. 2630–2640, 2017.

[19] P. Blanchard, A. Lee, S. Marguet, et al., "Chemotherapy and radiotherapy in nasopharyngeal carcinoma: an update of the MAC-NPC meta-analysis," Lancet Oncol., vol. 16, pp. 645–655, Jun 2015.

[20] F. Bonofiglio, J. Beyersmann, M. Schumacher, M. Koller, and G. Schwarzer, "Meta-analysis for aggregated survival data with competing risks: a parametric approach using cumulative incidence functions," Res Synth Methods, vol. 7, pp. 282–293, Sep 2016.

[21] R. D. Riley, M. J. Price, D. Jackson, M. Wardle, F. Gueyffier, J. Wang, J. A. Staessen, and I. R. White, "Multivariate meta-analysis using individual participant data," Res Synth Methods, vol. 6, pp. 157–174, Jun 2015.

[22] F. Rotolo, X. Paoletti, T. Burzykowski, B. Marc, and S. Michiels, "A poisson approach to the validation of failure time surrogate endpoints in individual patient data meta-analyses," Stat Methods Med Res, p. 962280217718582, 2017.

[23] D. L. Burke, J. Ensor, and R. D. Riley, "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ," Statistics in medicine, vol. 36, no. 5, pp. 855–875, 2017.

[24] B. Zhou, A. Latouche, V. Rocha, and J. Fine, "Competing risks regression for stratified data," Biometrics, vol. 67, no. 2, pp. 661–670, 2011.

[25] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," Journal of the American statistical association, vol. 53, no. 282, pp. 457–481, 1958.

[26] L.-J. Wei, D. Y. Lin, and L. Weissfeld, "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions," Journal of the American statistical association, vol. 84, no. 408, pp. 1065–1073, 1989.

[27] E. W. Lee, L. Wei, D. A. Amato, and S. Leurgans, "Cox-type regression analysis for large numbers of small groups of correlated failure time observations," in Survival analysis: state of the art, pp. 237–247, Springer, 1992.

[28] J. Cai and R. L. Prentice, "Estimating equations for hazard ratio parameters based on correlated failure time data," Biometrika, vol. 82, no. 1, pp. 151–164, 1995.

[29] J. Cai and R. L. Prentice, "Regression estimation using multivariate failure time data and a common baseline hazard function model," Lifetime Data Analysis, vol. 3, no. 3, pp. 197–213, 1997.

[30] D. G. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," Biometrika, vol. 65, no. 1, pp. 141–151, 1978.

[31] J. W. Vaupel, K. G. Manton, and E. Stallard, "The impact of heterogeneity in individual frailty on the dynamics of mortality," Demography, vol. 16, no. 3, pp. 439–454, 1979.

[32] J. P. Klein, "Semiparametric estimation of random effects using the cox model based on the em algorithm," Biometrics, pp. 795–806, 1992.

[33] G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. Sørensen, "A counting process approach to maximum likelihood estimation in frailty models," Scandinavian journal of Statistics, pp. 25–43, 1992.

[34] T. A. Louis, "Finding the observed information matrix when using the em algorithm," Journal of the Royal Statistical Society: Series B (Methodological), vol. 44, no. 2, pp. 226–233, 1982.

[35] H. Putter and H. C. Van Houwelingen, "Dynamic frailty models based on compound birth–death processes," Biostatistics, vol. 16, no. 3, pp. 550–564, 2015.

[36] S. Ripatti and J. Palmgren, "Estimation of multivariate frailty models using penalized partial likelihood," Biometrics, vol. 56, no. 4, pp. 1016–1022, 2000.

[37] T. M. Therneau, P. M. Grambsch, and V. S. Pankratz, "Penalized survival models and frailty," Journal of computational and graphical statistics, vol. 12, no. 1, pp. 156–175, 2003.

[38] D. Oakes, "A model for association in bivariate survival data," Journal of the Royal Statistical Society: Series B (Methodological), vol. 44, no. 3, pp. 414–422, 1982.

[39] D. Clayton and J. Cuzick, "Multivariate generalizations of the proportional hazards model," Journal of the Royal Statistical Society: Series A (General), vol. 148, no. 2, pp. 82–108, 1985.

[40] G. A. Whitmore and M.-L. T. Lee, "A multivariate survival distribution generated by an inverse gaussian mixture of exponentials," Technometrics, vol. 33, no. 1, pp. 39–50, 1991.

[41] D. Oakes, "Multivariate survival distributions," Journaltitle of Nonparametric Statistics, vol. 3, no. 3-4, pp. 343–354, 1994.

[42] M. Crowder, "A multivariate distribution with weibull connections," Journal of the Royal Statistical Society: Series B (Methodological), vol. 51, no. 1, pp. 93–107, 1989.

[43] J. H. Shih and T. A. Louis, "Assessing gamma frailty models for clustered failure time data," Lifetime Data Analysis, vol. 1, no. 2, pp. 205–220, 1995.

[44] D. V. Glidden, "Checking the adequacy of the gamma frailty model for multivariate failure times," Biometrika, vol. 86, no. 2, pp. 381–393, 1999.

[45] T. O. L. Insurance, "Checking for the gamma frailty distribution under the marginal proportional hazards frailty model," Statistica Sinica, vol. 14, pp. 249–267, 2004.

[46] T. M. Therneau and T. Lumley, "Package 'survival'," Survival analysis Published on CRAN, 2014.

[47] T. M. Therneau and M. T. M. Therneau, "Package 'coxme'," Mixed effects cox models. R package version, vol. 2, 2015.

[48] I. Do Ha, M. Noh, and Y. Lee, "frailtyhl: a package for fitting frailty models with hlikelihood," R Journal, vol. 4, no. 2, pp. 28–36, 2012.

[49] M. Donohue, R. Xu, and M. M. Donohue, "Package 'phmm'."

[50] V. Rondeau, Y. Mazroui, and J. R. Gonzalez, "frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation," J Stat Softw, vol. 47, no. 4, pp. 1–28, 2012.

[51] J. V. Monaco, M. Gorfine, and L. Hsu, "General semiparametric shared frailty model: estimation and simulation with frailtysurv," Journal of statistical software, vol. 86, 2018.

[52] F. Rotolo, M. Munda, and C. Legrand, "parfm: Parametric frailty models," R package version, vol. 2, no. 2, 2012.

[53] T. A. Balan and H. Putter, "frailtyem: An r package for estimating semiparametric shared frailty models," Journal of Statistical Software, vol. 90, no. 1, pp. 1–29, 2019.

[54] P. J. Heagerty and Y. Zheng, "Survival model predictive accuracy and roc curves," Biometrics, vol. 61, no. 1, pp. 92–105, 2005.

[55] K. Kerr and M. Pepe, "Joint modeling, covariate adjustment, and interaction: contrasting notions in risk prediction models and risk prediction performance," Epidemiology, vol. 22, no. 6, pp. 805–812, 2011.

[56] H. Uno, T. Cai, L. Tian, and L. Wei, "Evaluating prediction rules for t-year survivors with censored regression models," Journal of the American Statistical Association, vol. 102, no. 478, pp. 527–537, 2007.

[57] R. Maltoni, G. Gallerani, P. Fici, A. Rocca, and F. Fabbri, "Ctcs in early breast cancer: A path worth taking," Cancer letters, vol. 376, no. 2, pp. 205–210, 2016.

[58] F.-C. Bidard, C. Proudhon, and J.-Y. Pierga, "Circulating tumor cells in breast cancer," Molecular oncology, vol. 10, no. 3, pp. 418–430, 2016.

[59] P. Blanche, A. Latouche, and V. Viallon, "Time-dependent auc with right-censored data: a survey study," in Lee, M-L and Gail, G. and Cai, T. and Pfeiffer, R. and Gandy, A. (Risk Assessment and Evaluation of Predictions, ed.), Springer, 2013.

[60] P. Heagerty, T. Lumley, and M. Pepe, "Time-dependent ROC curves for censored survival data and a diagnostic marker," Biometrics, vol. 56, no. 2, pp. 337–344, 2000.

[61] L. Chambless and G. Diao, "Estimation of time-dependent area under the ROC curve for long-term risk prediction," Statistics in Medicine, vol. 25, no. 20, pp. 3474–3486, 2006.

[62] P. Blanche, J.-F. Dartigues, and H. Jacqmin-Gadda, "Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring," Biometrical Journal, vol. 55, no. 5, pp. 687–704, 2013.

[63] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," Jama, vol. 247, no. 18, pp. 2543–2546, 1982.

[64] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," Statistics in medicine, vol. 15, no. 4, pp. 361–387, 1996.

[65] P. Blanche, M. W. Kattan, and T. A. Gerds, "The c-index is not proper for the evaluation of t-year predicted risks," Biostatistics, p. kxy006, 2018.

[66] M. X. Rodríguez-Álvarez, L. Meira-Machado, E. Abu-Assi, and S. Raposeiras-Roubín, "Nonparametric estimation of time-dependent roc curves conditional on a continuous covariate," Statistics in Medicine, vol. 35, no. 7, pp. 1090–1102, 2016.

[67] Y. Zheng, T. Cai, and Z. Feng, "Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers," Biometrics, vol. 62, no. 1, pp. 279–287, 2006.

[68] F. Le Borgne, C. Combescure, F. Gillaizeau, M. Giral, M. Chapal, B. Giraudeau, and Y. Foucher, "Standardized and weighted time-dependent receiver operating

characteristic curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors," Statistical Methods in Medical Research, vol. 27, no. 11, pp. 3397–3410, 2018.

[69] P. Blanche and M. P. Blanche, "Package 'timeroc'," 2019.

[70] S. Potapov, W. Adler, M. Schmid, and M. S. Potapov, "Package 'survauc'," Statistics in Medicine, vol. 25, pp. 3474–3486, 2012.

[71] A. Meisner, C. Parikh, and K. Kerr, "Biomarker combinations for diagnosis and prognosis in multicenter studies: Principles and methods," Statistical methods in medical research, p. 0962280217740392, 2017.

[72] T. Balan and H. Putter, frailtyEM: Fitting Frailty Models with the EM Algorithm, 2017. R package version 0.7.0-1.

[73] W. Bouwmeester, K. Moons, T. Kappen, W. Van Klei, J. Twisk, M. Eijkemans, and Y. Vergouwe, "Internal validation of risk models in clustered data: a comparison of bootstrap schemes," American journal of epidemiology, vol. 177, no. 11, pp. 1209–1217, 2013.

[74] Y. Xiao and M. Abrahamowicz, "Bootstrap-based methods for estimating standard errors in cox's regression analyses of clustered event times," Statistics in medicine, vol. 29, no. 7-8, pp. 915–923, 2010.

[75] W. Bouwmeester, K. Moons, T. Kappen, W. Van Klei, J. Twisk, M. Eijkemans, and Y. Vergouwe, "Internal validation of risk models in clustered data: a comparison of bootstrap schemes," American journal of epidemiology, vol. 177, no. 11, pp. 1209–1217, 2013.

[76] D. Commenges and P. Andersen, "Score test of homogeneity for survival data," Lifetime Data Analysis, vol. 1, no. 2, pp. 145–156, 1995.

[77] E. B. Hoffman, P. K. Sen, and C. R. Weinberg, "Within-cluster resampling," Biometrika, vol. 88, no. 4, pp. 1121–1134, 2001.

[78] E. Benhin, J. Rao, and A. Scott, "Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes," Biometrika, vol. 92, no. 2, pp. 435–450, 2005.

[79] M. Pavlou, Analysis of clustered data when the cluster size is informative. PhD thesis, UCL (University College London), 2012.

[80] Z. Chen, B. Zhang, and P. S. Albert, "A joint modeling approach to data with informative cluster size: robustness to the cluster size model," Statistics in medicine, vol. 30, no. 15, pp. 1825–1836, 2011.

[81] S. Seaman, M. Pavlou, and A. Copas, "Review of methods for handling confounding by cluster and informative cluster size in clustered data," Statistics in medicine, vol. 33, no. 30, pp. 5371–5387, 2014.

[82] C.-T. Chiang and K.-Y. Lee, "Efficient estimation methods for informative cluster size data," Statistica Sinica, pp. 121–133, 2008.

[83] Y. Huang and B. Leroux, "Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations," Biometrics, vol. 67, no. 3, pp. 843–851, 2011.

[84] J. M. Neuhaus and C. E. McCulloch, "Estimation of covariate effects in generalized linear mixed models with informative cluster sizes," Biometrika, vol. 98, no. 1, pp. 147–162, 2011.

[85] S. Datta, J. Nevalainen, and H. Oja, "A general class of signed-rank tests for clustered data when the cluster size is potentially informative," Journal of nonparametric statistics, vol. 24, no. 3, pp. 797–808, 2012.

[86] S. Dutta and S. Datta, "A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative," Biometrics, vol. 72, no. 2, pp. 432–440, 2016.

[87] J. Fan and S. Datta, "Fitting marginal accelerated failure time models to clustered survival data with potentially informative cluster size," Computational statistics & data analysis, vol. 55, no. 12, pp. 3295–3303, 2011.

[88] Z. Ying and L. Wei, "The kaplan-meier estimate for dependent failure time observations," Journal of Multivariate Analysis, vol. 50, no. 1, pp. 17–29, 1994.

[89] E. W. Lee, L. Wei, and Z. Ying, "Linear regression analysis for highly stratified failure time data," Journal of the American Statistical Association, vol. 88, no. 422, pp. 557–565, 1993.

[90] X. Su, P. Calhoun, and J. Fan, "Mst: Multivariate survival trees," 2018.

[91] T. A. Gerds, "Prediction error curves for survival models; r package pec; version 1.1. 5," R Foundation for Statistical Computing, 2009.

[92] F. Rotolo and S. Michiels, "Testing the treatment effect on competing causes of death in oncology clinical trials," BMC Med Res Methodol, vol. 14, no. 72, 2014.

[93] M. C. Simmonds, J. P. Higgins, L. A. Stewart, J. F. Tierney, M. J. Clarke, and S. G. Thompson, "Meta-analysis of individual patient data from randomized trials: a review of methods used in practice," Clinical Trials, vol. 2, no. 3, pp. 209–217, 2005.

[94] H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in biostatistics: competing risks and multi-state models," Statistics in medicine, vol. 26, no. 11, pp. 2389–2430, 2007.

[95] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," Journal of the American statistical association, vol. 94, no. 446, pp. 496–509, 1999.

[96] N. L. Hjort, "On inference in parametric survival data models," International Statistical Review/Revue Internationale de Statistique, pp. 355–387, 1992.

[97] T. A. Gerds, T. H. Scheike, and P. K. Andersen, "Absolute risk regression for competing risks: interpretation, link functions, and prediction," Statistics in medicine, vol. 31, no. 29, pp. 3921–3930, 2012.

[98] O. Aalen, A model for nonparametric regression analysis of counting processes. Springer, 1980.

[99] T. Martinussen and T. H. Scheike, Dynamic regression models for survival data. Springer Science & Business Media, 2007.

[100] T. H. Scheike and M.-J. Zhang, "Flexible competing risks regression modeling and goodness-of-fit," Lifetime data analysis, vol. 14, no. 4, p. 464, 2008.

[101] A. Latouche, A. Allignol, J. Beyersmann, M. Labopin, and J. P. Fine, "A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions," Journal of clinical epidemiology, vol. 66, no. 6, pp. 648–653, 2013.

[102] B. Zhou, J. Fine, and G. Laird, "Goodness-of-fit test for proportional subdistribution hazards model," Statistics in medicine, vol. 32, no. 22, pp. 3804–3811, 2013.

[103] B. Zhou and A. Latouche, crrSC: Competing risks regression for Stratified and Clustered data, 2013. R package version 1.1.

[104] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein, "A basic introduction to fixed-effect and random-effects models for meta-analysis," Research synthesis methods, vol. 1, no. 2, pp. 97–111, 2010.

[105] K. Rice, J. P. Higgins, T. Lumley, et al., "A re-evaluation of fixed effect (s) meta-analysis," JR Stat Soc Ser A Stat Soc, vol. 181, no. 1, pp. 205–27, 2018.

[106] R. DerSimonian and R. Kacker, "Random-effects model for meta-analysis of clinical trials: an update," Contemporary clinical trials, vol. 28, no. 2, pp. 105–114, 2007.

[107] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," Controlled clinical trials, vol. 7, no. 3, pp. 177–188, 1986.

[108] S. Senn, "The many modes of meta," Drug Information Journal, vol. 34, no. 2, pp. 535–549, 2000.

[109] K.-Y. Liang, S. G. Self, and Y.-C. Chang, "Modelling marginal hazards in multivariate failure time data," Journal of the Royal Statistical Society: Series B (Methodological), vol. 55, no. 2, pp. 441–453, 1993.

[110] S. Katsahian and C. Boudreau, "Estimating and testing for center effects in competing risks," Statistics in medicine, vol. 30, no. 13, pp. 1608–1617, 2011.

[111] B. E. Chen, J. L. Kramer, M. H. Greene, and P. S. Rosenberg, "Competing risks analysis of correlated failure time data," Biometrics, vol. 64, no. 1, pp. 172–179, 2008.

[112] B. Zhou, J. Fine, A. Latouche, and M. Labopin, "Competing risks regression for clustered data," Biostatistics, vol. 13, no. 3, pp. 371–383, 2012.

[113] N. Freemantle, J. Cleland, P. Young, J. Mason, and J. Harrison, "β blockade after myocardial infarction: systematic review and meta regression analysis," Bmj, vol. 318, no. 7200, pp. 1730–1737, 1999.

[114] C. H. Schmid, P. C. Stark, J. A. Berlin, P. Landais, and J. Lau, "Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors," Journal of clinical epidemiology, vol. 57, no. 7, pp. 683–697, 2004.

[115] J. P. Higgins and S. G. Thompson, "Quantifying heterogeneity in a meta-analysis," Statistics in medicine, vol. 21, no. 11, pp. 1539–1558, 2002.

[116] J. P. Higgins and S. Green, Cochrane handbook for systematic reviews of interventions, vol. 4. John Wiley & Sons, 2011.

[117] R. J. Hardy and S. G. Thompson, "Detecting and describing heterogeneity in meta-analysis," Statistics in medicine, vol. 17, no. 8, pp. 841–856, 1998.

[118] J. P. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," BMJ: British Medical Journal, vol. 327, no. 7414, p. 557, 2003.

[119] G. Cortese, T. A. Gerds, and P. K. Andersen, "Comparing predictions among competing risks models with time-dependent covariates," Stat Med, vol. 32, pp. 3089–3101, Aug 2013.

[120] B. Gray, M. B. Gray, and R. Gray, "The cmprsk package," The Comprehensive R Archive network, 2004.

[121] A. Allignol and M. A. Allignol, "Package 'etm'," Reproductive Toxicology, vol. 26, pp. 31–35, 2015.

[122] P. K. Andersen and N. Keiding, "Interpretability and importance of functionals in competing risks and multistate models," Statistics in medicine, vol. 31, no. 11-12, pp. 1074–1088, 2012.

[123] J. P. Klein, "Modelling competing risks in cancer studies," Statistics in medicine, vol. 25, no. 6, pp. 1015–1034, 2006.

[124] D. Zhang and X. Lin, "Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics," in Random effect and latent variable model selection, pp. 19–36, Springer, 2008.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre: Inférence et validation d'un marqueur pronostique pour des données de survie corrélées avec l'application au cancer

Mots clés: Données en grappes, curbe ROC, meta-analyse, survie, effet centre, taille informative de grappes

Résumé: Les données de survie en grappes sont caractérisées par des corrélations entre des observations appartenant à un même groupe. Ici, nous discutons des extensions a des données en grappes dans différents contextes. Nous avons envisagé une méta-analyse pour le cancer du sein afin d'évaluer dans quelle mesure les cellules tumorales circulantes discriminent les patients ayant le même stade de la tumeur. Bien que la courbe ROC dépendante du temps ait été largement utilisée pour la discrimination des biomarqueurs, il n'existe pas de méthodologie permettant de traiter des données en grappes censurées. Nous avons proposè un estimateur pour les courbes ROC dépendantes du temps et pour l'AUC lorsque les temps d'évènements sont correlés. Nous avons employé un modèle de fragilité partagée afin de tenir compte de l'effet de la grappe. Une étude de simulation a été réalisée et a montré un biais négligeable pour l'estimateur proposé et pour un estimateur non paramétrique fondé sur la pondération par la probabilité inverse d'être censuré (IPCW), tandis qu'un estimateur semi-paramétrique, ignorant la structure en grappe est nettement biaisé.
Nous avons également considéré une méta-analyse sur données individuels (IPD) pour quantifier le bénéfice de l'ajout de la chimiothérapie à la radiothérapie sur chaque risque concurrent pour les patients avec un carcinome nasopharyngien. Les recommandations pour l'analyse des risques concurrents dans le cadre d'essais cliniques randomisés sont bien établies. Étonnamment, aucune recommendation n'a encore été proposée pour l'anlayse d'une méta-analyse IPD avec les risque concurrents. Pour combler cette lacune, ce travail a détaillé la manière de traiter l'hétérogénéité entre les essais par un modèle de régression stratifié pour les risques concurrents et il souligne que les mesures standardes d'hétérogénéité pour évaluer l'incohérence peuvent facilement être utilisées. Nous avons aussi proposé une approche landmark pour la fonction d'incidence cumulée pour étudier l'impact du temps de suivi sur l'effet du traitement.
L'hypothèse d'une taille de grappe non informative était faite dans les deux analyses. On dit que la taille de grappe est informative lorsque la variable réponse dépend de la taille de grappe conditionnellement à un ensemble de variables explicatives. Intuitivement, une méta-analyse répondrait à cette hypothèse. Cependant, la taille de grappe non informative est généralement supposée, même si elle peut être fausse dans certaines situations, ce qui conduit à des résultats incorrects. La taille des grappes informatives (ICS) est un problème difficile et sa présence a un impact sur le choix de la méthodologie. Nous avons discuté plus en détail de l'interprétation des résultats et des quantités qui peuvent être estimées et dans quelles conditions. Nous avons proposé un test pour l'ICS avec des données en grappes censurées. À notre connaissance, il s'agit du premier test sur le contexte de l'analyse de survie. Une étude de simulation a été réalisée pour évaluer la puissance du test et quelques exemples sont fournis à titre d'illustration. L'implémentation de chacun de ces développements est disponible sur https://github.com/AMeddis.

Title: Inference and validation of prognostic marker for correlated survival data with application to cancer

Keywords: clustered survival data, ROC curve, meta-analysis, center effect, informative cluster size

Abstract: Clustered data are characterized by correlations between observations belonging to the same cluster. Here, we discuss some extension to clustered data in different contexts. Initially, we considered a meta-analysis for breast cancer to assess how well the circulating tumor cells discriminate patients with same tumor stage regarding the risk of death. Although the time dependent ROC curve has been widely used for biomarker's discrimination, there is no methodology that can handle clustered censored data. We proposed an estimator for the covariate-specific time dependent ROC curve and AUC when clustered failure times are detected considering a shared frailty model to account for the cluster effect. A simulation study was conducted showing negligible bias for the proposed estimator and a nonparametric one based on inverse probability censoring weighting, while a semiparametric estimator, ignoring the clustering, is markedly biased.
We further considered an IPD meta-analysis with competing risks to assess the benefit of the addition of chemotherapy to radiotherapy on each competing endpoint for patients with nasopharyngeal carcinoma. Recommendations for the analysis of competing risks in the context of randomized clinical trials are well established. Surprisingly, no formal guidelines have been yet proposed to conduct an IPD meta-analysis with competing risks. To fill this gap, this work detailed: how to handle the heterogeneity between trials via a stratified regression model for competing risks and it highlights that the usual metrics of inconsistency to assess heterogeneity can readily be employed. We further proposed a landmark approach for the cumulative incidence function to investigate the impact of follow up on the treatment effect.
The assumption of non informative cluster size was made in both the analyses. The cluster size is said to be informative when the outcome depends on the size of the cluster conditional on a set of covariates. Intuitively, a meta-analysis would meet this assumption. However, non informative cluster size is commonly assumed even though it may be not true in some situations. Informative cluster size (ICS) is a challenging problem and its presence has an impact on the choice of the correct methodology. We discussed more in details interpretation of results and which quantities can be estimated under which conditions. We proposed a test for ICS with censored clustered data. To our knowledge, this is the first test on the context of survival analysis. A simulation study was conducted to assess the power of the test and some illustrative examples were provided.
The implementation of each of these developments are available at https://github.com/AMeddis.

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