Prédiction du pronostic des patients atteints de muscoviscidose

Dorette Lionelle Nkam Beriye

le cnam

École doctorale Sciences des Métiers de l'Ingénieur

Laboratoire Modélisation, Épidémiologie et Surveillance des Risques Sanitaires

THÈSE DE DOCTORAT

présentée par : **Dorette Lionelle NKAM BERIYE**

soutenue le : **22 décembre 2017**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline : **Santé publique, environnement et société**

Spécialité : **Sécurité sanitaire**

PRÉDICTION DU PRONOSTIC DES PATIENTS ATTEINTS DE MUCOVISCIDOSE

THÈSE dirigée par

Mounia Nacima HOCINE	*Maître de conférences, CNAM, Paris*

PRÉSIDENT

Avner BAR-HEN	*Professeur des Universités, CNAM, Paris*

RAPPORTEURS

Cécile PROUST-LIMA	*Chargée de Recherche, INSERM U1219, Bordeaux*
Yohann FOUCHER	*Maître de conférences, INSERM U1246, Nantes*

EXAMINATEURS

Isabelle SERMET-GAUDELUS	*Professeur des Universités - Praticien Hospitalier, Hôpital Necker, Paris*
Pierre-Régis BURGEL	*Professeur des Universités - Praticien Hospitalier, Hôpital Cochin, Paris*

Résumé : La mucoviscidose est une maladie génétique rare et incurable. Malgré les nombreux progrès réalisés dans la recherche à ce sujet, il reste indispensable d'avoir davantage une meilleure connaissance de la maladie afin de proposer des traitements encore plus adaptés aux patients. La majorité des traitements actuels visent principalement à réduire les symptômes de la maladie sans toutefois la guérir. La transplantation pulmonaire reste le moyen le plus adéquat pour améliorer la qualité de vie et prolonger la vie des patients dont l'état respiratoire s'est considérablement dégradé. Il est donc nécessaire de fournir aux cliniciens des outils d'aide à la décision pour mieux identifier les patients nécessitant une transplantation pulmonaire. Pour ce faire, il est indispensable de connaître d'une part, les facteurs pronostiques de la transplantation pulmonaire et d'autre part, de savoir prédire la survenue de cet événement chez les sujets atteints de mucoviscidose. L'objectif de ce travail de thèse est de développer des outils pronostiques utiles à l'évaluation des choix thérapeutiques liés à la transplantation pulmonaire. Dans un premier temps, nous avons réévalué les facteurs pronostiques de la transplantation pulmonaire ou du décès sans transplantation pulmonaire chez les adultes atteints de mucoviscidose. Suite aux progrès thérapeutiques qui ont conduit à l'amélioration du pronostic au cours des dernières années, ce travail a permis d'identifier les facteurs pronostiques en adéquation avec l'état actuel de la recherche. Un deuxième travail a consisté à développer un modèle conjoint à classes latentes fournissant des prédictions dynamiques pour la transplantation pulmonaire ou le décès sans transplantation pulmonaire. Ce modèle a permis d'identifier trois profils d'évolution de la maladie et également d'actualiser le risque de survenue de la transplantation pulmonaire ou du décès sans transplantation pulmonaire à partir des données longitudinales du marqueur VEMS. Ces modèles pronostiques ont été développés à partir des données du registre français de la mucoviscidose et ont fourni de bonnes capacités prédictives en termes de discrimination et de calibration.

Mots clés : mucoviscidose, transplantation pulmonaire, modèles conjoints, prédiction dynamique

Abstract : Cystic Fibrosis is an incurable inherited disorder. Despite real progress in research, it is essential to always have a better understanding of the disease in order to provide suitable treatments to patients. Current treatments mostly aim to reduce the disease symptoms without curing it. Lung transplantation is proposed to cystic fibrosis patients with terminal respiratory failure with the aim of improving life expectancy and quality of life. It is necessary to guide clinicians in identifying in a good way patients requiring an evaluation for lung transplantation. It is thus important to clearly identify prognostic factors related to lung transplantation and to predict in a good way the occurrence of this event in patients with cystic fibrosis. The aim of this work was to develop prognostic tools to assist clinicians in the evaluation of different therapeutic options related to lung transplantation. First, we reevaluated prognostic factors of lung transplantation or death without lung transplantation in adult with cystic fibrosis. Indeed, therapeutic progress in patients with cystic fibrosis has resulted in improved prognosis over the past decades. We identified prognostic factors related to the current state of research in the cystic fibrosis field. We further developed a joint model with latent classes which provided dynamic predictions for lung transplantation or death without lung transplantation. This model identified three profiles of the evolution of the disease and was able to update the risk of lung transplantation or death without lung transplantation taking into account the evolution of the longitudinal marker FEV_1 which describes the lung function. These prognostic models were developed using the French cystic fibrosis registry and provided good predictive accuracy in terms of discrimination and calibration.

Key words : cystic fibrosis, lung transplantation, joint models, dynamic prediction

À ma famille

Remerciements

À Madame Mounia Hocine

Merci de m'avoir confié ce travail. Dans un premier temps, tout s'est passé au téléphone. Avant même de m'avoir rencontrée, tu m'as fait confiance. Merci pour cette confiance. Merci pour ta gentillesse, tes encouragements, tes conseils, ta disponibilité tout au long de cette thèse. Ça a été une expérience très enrichissante et j'ai appris beaucoup de choses au cours de ces trois années passées avec toi.

À Monsieur William Dab

Merci pour l'accueil au sein de cette équipe MESuRS, dans ce laboratoire où il fait bon de travailler. Merci pour toutes ces délicieuses tartelettes de la maison du chocolat.

À Madame Cécile Proust-Lima

Après ces stages effectués avec toi et vu le sujet sur lequel je travaille, il m'a semblé évident de te proposer d'être rapporteur de cette thèse. Merci d'avoir accepté de donner ton avis sur ce travail, j'en suis très ravie. J'ai beaucoup appris avec toi lors de mes stages. Les connaissances que j'ai acquises m'ont beaucoup aidé à comprendre mon sujet de thèse. Mon nom dans le package *lcmm* et les remerciements dans [Proust-Lima et al. 2017] « c'est la classe » ! J'en suis très reconnaissante. Merci encore.

À Monsieur Yohann Foucher

Merci d'avoir accepté aussi rapidement et spontanément de juger mon travail. Je vous en suis très reconnaissante.

À Madame Isabelle Sermet-Gaudelus

Merci d'avoir accepté de participer à mon jury de thèse. Certes, nous n'avons toujours pas eu le temps pour faire cette collaboration mais j'espère bien qu'elle se fera.

À Monsieur Avner Bar-Hen

Merci de me faire l'honneur de juger mon travail. Je vous en suis très reconnaissante.

Aux collaborateurs

M. Pierre-Régis Burgel, merci pour le temps que vous avez consacré à ce projet. Nous avons eu besoin d'un spécialiste dans le domaine de la mucoviscidose pour faire ce travail, et votre avis nous a été très précieux. Merci pour avoir su guider nos choix et pour vos critiques constructives dans le but d'améliorer ce travail. **M. Aurélien Latouche** et **M. Jérôme Lambert**, merci pour vos conseils, vos suggestions proposées pour améliorer ce travail. **M. Gil Bellis**, merci de m'avoir aidée à comprendre les données du registre et d'avoir toujours répondu à mes questions concernant le registre.

À l'équipe MESuRS

Laura Temime, future directrice du laboratoire MESuRS, merci pour ton aide et tes conseils quand j'en ai eu besoin. A tous ceux et celles qui ont partagé le bureau dit « des doctorants » avec moi, ce fut un plaisir de vous avoir à mes côtés. **Rania** (l'organisatrice des sorties !), **Narimane** (l'Américaine, toujours aussi sérieuse !), **Audrey** (la motarde !), **Jonathan** (tu as arrêté avec les « œufs », Bravo !), **François** et **Kadiatou** (ces stidiens beaucoup trop drôles !), **Karim** (le « jackpot » !) et tous les autres, ça a été un réel plaisir de travailler avec vous. On a beaucoup travaillé, mais surtout on a pas du tout rigolé. Merci pour votre bonne humeur et tous ces moments de joie qu'on a partagés. **Kévin**, **Isabelle**, **Sonia**, **Thomas** ça a été un plaisir de travailler à vos côtés. Sans oublier les nouveaux **Hélène**, **Nathalie**, **Jérôme**, **Tom** et tous les autres **Cnamiens**...

À l'association Vaincre la Mucoviscidose

Merci d'avoir financé cette thèse. Merci à **Anna Ronayette** qui a toujours su répondre à nos questions en rapport avec l'association.

Aux Amis d'ici et d'ailleurs, merci pour tous vos encouragements, merci d'être là.

À ma Famille

Le meilleur pour la fin.... **Dad**, **Mom**, **Brothers** and **Sisters**, merci pour vos encouragements, pour votre soutien. Merci infiniment d'être toujours là pour moi. C'est un plaisir de vous avoir dans ma vie. **Merci à toute ma famille**.

Table des matières

Chapitre 1

Contexte épidémiologique de la mucoviscidose

1.1 La mucoviscidose

1.1.1 Définition et symptômes

La mucoviscidose encore appelée fibrose kystique, est une maladie génétique rare. Elle est le résultat d'une mutation qui survient au niveau du chromosome 7 du gène CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) codant pour la protéine CFTR [Kerem et al. 1989; Rommens et al. 1989]. Cette dernière est impliquée dans le transfert des ions chlorure. Elle est présente sur la membrane des cellules et permet le passage des ions chlorure vers l'intérieur ou vers l'extérieur de la cellule [Riordan et al. 1989]. Les ions chlorure ont un rôle essentiel dans le transport de l'eau dans les tissus des organes. De ce fait, un des rôles principaux de la protéine CFTR est l'hydratation du mucus. Le mucus étant une substance visqueuse qui recouvre l'intérieur de certains organes ou cavités de l'organisme (poumon, intestin, nez...). Il permet entre autres de protéger les organes contre les agents pathogènes. La mutation du gène CFTR se traduit donc par la production de protéines CFTR anormales. Le transfert d'ions et l'hydratation du mucus deviennent alors très difficiles. Les sujets malades ont de ce fait, un mucus visqueux d'où l'appellation mucoviscidose.

La mucoviscidose est une maladie qui affecte principalement les voies respiratoires et les voies digestives. Elle peut néanmoins affecter plusieurs autres organes de l'organisme dans une moindre mesure.

Les symptômes de la mucoviscidose au niveau de l'appareil respiratoire se manifestent par un épaississement du mucus dans les bronches qui entraîne une obstruction des voies respiratoires. L'obstruction des bronches empêche le passage de l'air et rend

difficile la respiration. À cause du manque de fluidité et de l'écoulement difficile du mucus, les bronches deviennent un lieu propice au développement des germes qui sont responsables d'infections pulmonaires. Les poumons représentent les organes les plus affectés par la maladie, qui peut entraîner dans certains cas une insuffisance respiratoire. Celle-ci est responsable d'au moins 80% de décès liés à la mucoviscidose [Lyczak et al. 2002].

Au niveau de l'appareil digestif, la maladie se manifeste par l'obstruction des parois du pancréas suite à l'épaississement du mucus. Or le pancréas est l'organe produisant les enzymes qui permettent la digestion des aliments. À cause du mauvais transfert des enzymes produites par le pancréas vers l'intestin, les aliments ingérés sont mal ou pas digérés. Ceci entraîne entre autres des troubles digestifs pouvant entraîner une malnutrition sévère et des carences en vitamines. De plus, l'atteinte du pancréas peut entraîner une insuffisance en insuline qui peut engendrer un diabète [Kelly and Moran 2013]. Ce diabète lié à la mucoviscidose est cependant différent du diabète classique de type I ou de type II. Il a tendance à se développer avec l'âge chez les malades. Environ 20% d'adolescents et, entre 40% et 50% d'adultes en sont atteints [Moran et al. 2009].

La mucoviscidose peut également affecter d'autres organes dans une moindre mesure. Le foie peut être endommagé chez certains sujets atteints de la maladie et éventuellement évoluer vers une cirrhose hépatique. Cependant, la cirrhose n'apparait que chez environ 5% de malades et en général chez les plus de 15 ans [Wilschanski and Durie 2007]. Les organes génitaux peuvent également être affectés entraînant ainsi une baisse de la fertilité chez les hommes et chez les femmes.

La mucoviscidose se manifeste différemment d'un sujet à l'autre et indépendamment de l'âge. Elle peut se déclarer à la naissance, durant l'enfance mais aussi à l'âge adulte, selon sa sévérité. On observe chez le nouveau-né un arrêt du transit intestinal et une prise de poids plus lente que la normale. Chez l'enfant, la maladie se manifeste entre autres par un retard de croissance, des douleurs abdominales, des bronchites. Plus généralement, les sujets atteints de la mucoviscidose ont des douleurs abdominales, des sinusites répétitives, une respiration laborieuse due à l'obstruction des voies respiratoires, des infections pulmonaires fréquentes, une insuffisance pancréatique. Une stérilité masculine est très souvent observée contrairement à la stérilité féminine. Cependant, les femmes peuvent être sujettes à une infertilité.

1.1.2 Diagnostic

La mucoviscidose est la plus fréquente des maladies génétiques dans les populations de type caucasien. C'est une maladie héréditaire avec un mode de transmission autosomique récessif. C'est-à-dire que deux parents porteurs d'une mutation du gène CFTR mais sains, ont un risque de 25% d'avoir un enfant atteint de la maladie. Ce dernier aura alors les deux gènes mutés de ses parents et est dit homozygote pour le gène CFTR. Le risque est de 50% pour les parents d'avoir un enfant hétérozygote pour le gène CFTR. Tout comme ses parents, ce dernier sera porteur d'une mutation du gène CFTR mais sain, et pourrait éventuellement le transmettre à sa descendante. Enfin, les parents peuvent avoir un enfant ayant leurs gènes non mutés et de ce fait sera non atteint de la maladie. La mucoviscidose est une maladie non associée au chromosome sexuel et affecte donc de la même manière les hommes et les femmes.

La méthode standard, la plus fiable et la plus répandue pour détecter la présence de la maladie est le test de la sueur qui consiste à mesurer la concentration en chlore dans la sueur. Le diagnostic est généralement bien établi. Une concentration de chlore dans la sueur inférieure à 40 mmol/L est normale et par conséquent traduit un test négatif vis-à-vis de la maladie. Un test est considéré comme positif si cette concentration est supérieure à 60 mmol/L [De Boeck et al. 2006]. Cependant, une concentration de chlore dans la sueur comprise entre 40 et 60 mmol/L peut être un indice de la présence de la maladie [Shwachman and Mahmoodian 1967; De Boeck et al. 2006; Farrell et al. 2008]. Dans ce cas, le diagnostic doit être confirmé par un deuxième test de la sueur suivi d'un test génétique pour identifier les mutations du gène présentes chez le malade.

Il est possible de faire dépister les nouveau-nés dès leur naissance. Un nouveau-né susceptible d'être atteint de mucoviscidose a un niveau élevé de la protéine trypsine immuno-réactive (TIR) dans les premiers mois de vie en cas de dysfonctionnement pancréatique. Le dépistage néonatal consiste alors à doser la protéine TIR dans le sang dans les premiers jours après la naissance. Ce dépistage précoce n'est pas un test de diagnostic, il permet principalement d'identifier les nouveau-nés ayant un risque élevé d'être atteints de mucoviscidose. Il se fait chez le nouveau-né avant l'apparition des symptômes, le troisième jour après la naissance. Le dépistage néonatal permet une prise en charge immédiate et adaptée de l'enfant et favorise ainsi une meilleure qualité de vie [Sims et al. 2005]. Ce test possède une bonne capacité à identifier les sujets malades. Cependant, il a une moins bonne spécificité, car sélectionne également

des sujets non atteints de la maladie. En cas de dépistage positif chez le nouveau-né, il est nécessaire que le diagnostic soit confirmé par le test de la sueur et le test génétique [Castellani et al. 2008; Farrell et al. 2008].

Un consensus sur les recommandations pour le diagnostic de la mucoviscidose a récemment été établit par Farrell et al.. Les critères de diagnostic ont été reconsidérés du fait d'une meilleure connaissance actuelle de la maladie. Ce consensus confirme que le diagnostic de la naissance à l'âge adulte se fait par le test de la sueur. Cependant, la dose de chlore en dessous de laquelle le test de la sueur est considéré comme négatif a été réduite à 30 mmol/L pour tous les âges. Cette dose était au préalable fixée à 40 mmol/L pour des sujets de plus de 6 mois. Or, chez certains sujets, des diagnostics se sont révélés positifs avec des concentrations de chlore dans la sueur comprises entre 30 et 39 mmol/L. La dose de chlore supérieure à 60 mmol/L reste inchangée pour un diagnostic positif de la maladie. Il est recommandé de confirmer le diagnostic par la recherche de mutations si la dose de chlore est comprise entre 30 et 59 mmol/L. Ce consensus vise non seulement à améliorer le diagnostic, mais aussi à clarifier et standardiser les critères de diagnostic dans le monde [Farrell et al. 2017].

1.1.3 Facteurs de risque

La mucoviscidose est une maladie qui touche plus de 70000 individus dans le monde [Foundation 2016]. Sa prévalence moyenne dans les pays de l'union européenne est de 0.737/10000, valeur assez proche de la prévalence moyenne aux États-Unis qui vaut 0.707/10000 [Farrell 2008]. Bien que présente dans toutes les populations, la mucoviscidose est plus fréquente dans les populations de type caucasien. Son incidence est d'environ 1 sur 3000 naissances dans la population caucasienne, et est plus faible dans les autres populations. On compte environ une naissance sur 4000 à 10000 chez les latino-américains, environ une naissance sur 15000 à 20000 chez les afro-américains et encore moins chez les indiens d'Amériques [O'Sullivan and Freedman 2009]. La maladie est très rare dans les populations africaines et asiatiques. L'histoire familiale est donc incontestablement un facteur de risque de la mucoviscidose étant donné son caractère génétique.

1.1.4 Évolution de la maladie

La mucoviscidose est une maladie génétique complexe due à une mutation du gène CFTR. Selon la base de données de mutations de la mucoviscidose, on dénombre à ce jour plus de 2000 mutations de ce gène dont la plupart sont des mutations peu fréquentes dans la population de sujets malades. Seules une vingtaine de mutations ont une fréquence supérieure à 0.1% dans le monde. La mutation la plus fréquente est la mutation F508del, présente chez environ 70% de malades sous forme hétérozygote et chez environ 50% de malades sous forme homozygote [Foundation 2016]. La majorité des mutations sont présentes chez les malades issus des populations caucasiennes [Morral et al. 1994]. Néanmoins, il existe des mutations présentes chez des malades issus de populations non-caucasiennes, notamment africaines et asiatiques [Macek et al. 1997]. Cependant, aucune de ces mutations n'atteint la fréquence de F508del.

Les mutations du gène CFTR ont été classées en 6 classes selon le niveau de dysfonctionnement de la protéine (absence de canaux, diminution du nombre de canaux, dysfonctionnement des canaux...) [Zielenski and Tsui 1995]. La classe de mutations 1 correspond à une absence totale ou partielle de la protéine CFTR. Dans la classe 2, celle où on retrouve la mutation la plus fréquente F508del, le processus de maturation de la protéine CFTR est perturbé. Dans la classe 3, on retrouve les mutations qui provoquent une mauvaise régulation des ions chlorure. Ces trois classes 1, 2 et 3 sont généralement des indicateurs d'une expression sévère de la maladie. La classe 4 regroupe les mutations altérant le transport des ions chlorure dans les cellules. Dans la classe 5, les protéines CFTR sont normales mais produites en petite quantité. La classe 6 correspond aux mutations altérant la stabilité de la protéine CFTR. Les classes de mutations 4, 5 et 6 sont associées à une expression modérée de la maladie [Castellani et al. 2008]. Bien que la sévérité de la maladie dépende du type de mutation, il n'y a pas de corrélation directe entre le génotype et le phénotype (manifestations cliniques). En effet d'autres facteurs peuvent avoir un impact sur la progression de la maladie, tels que le contexte génétique de l'individu, la présence d'autres pathologies, la qualité des soins, les conditions de vie, la pollution, la pratique d'une activité physique [Goss et al. 2004; Drumm et al. 2005].

Les infections pulmonaires sont constantes chez les personnes atteintes de mucoviscidose et constituent la cause prédominante de mortalité et de morbidité associée à la maladie [De Boeck 2000; Doring et al. 2004]. Le mucus épais présent dans les bronches favorise le développement d'agents pathogènes. Plusieurs bactéries et virus

s'y installent et jouent un rôle critique dans la progression de la maladie. Le germe le plus fréquent chez les sujets atteints de mucoviscidose est le *Pseudomonas aeruginosa*. Les infections à *Pseudomonas aeruginosa* augmentent avec l'âge et plus de 80% des malades âgés de 19 ans et plus en sont porteurs [Kosorok et al. 2001; Nixon et al. 2001; Lyczak et al. 2002]. Cette bactérie est associée à une augmentation de la morbidité et de la mortalité, ainsi qu'une réduction de la qualité de vie des malades[Pressler et al. 2011]. Elle fait partie des germes difficiles à éradiquer chez les malades car résistante aux antibiotiques[Oliver et al. 2000; Winstanley et al. 2016]. Néanmoins, sa prévalence a tendance à diminuer avec le temps, grâce à l'amélioration du contrôle des infections et des stratégies d'éradication de la bactérie.

À l'instar de *Pseudomonas aeruginosa*, *Staphylococcus aureus* est une bactérie très fréquente chez les malades. Elle est en général le premier pathogène isolé à infecter les poumons et est associée à un déclin rapide de la fonction pulmonaire. Il a été montré que les sujets infectés par *Staphylococcus aureus* ont un taux de déclin plus rapide de 43% de la fonction pulmonaire, comparé aux sujets non infectés [Dasenbrook et al. 2008]. Contrairement à *Pseudomonas aeruginosa*, *Staphylococcus aureus* est surtout présent chez les enfants et les adolescents et a tendance à diminuer avec l'âge. Ces bactéries sont impliquées dans les infections respiratoires chroniques chez les malades [Lyczak et al. 2002; Phillips et al. 2006; Lipuma 2010].

Des bactéries telles *Burkholderia cepacia* (une bactérie complexe d'environ 20 espèces) sont acquises relativement tard et sont assez rares. Elles restent cependant un facteur de mauvais pronostic car, sont résistantes aux antibiotiques, très virulentes et associées à une dégradation rapide de la fonction pulmonaire [De Boeck et al. 2004; Ellaffi et al. 2005].

D'autres pathogènes émergents ont également été associés à des infections pulmonaires sévères, c'est le cas de *Stenotrophomonos maltophilia*, *Achromobacter xylosoxidans*, *Aspergillus fumigatus*, *Non-tuberculous mycobacteria*. Les infections pulmonaires sont problématiques dans l'évolution de la maladie. Elles sont responsables de la baisse de la fonction pulmonaire, de l'augmentation de la mortalité, des complications aigües de la maladie (exacerbations pulmonaires, hémoptysie, pneumothorax).

La mucoviscidose est une maladie qui est à ce jour incurable. Néanmoins, on note depuis quelques décennies de nombreuses avancées dans la recherche. La survie des sujets atteints de mucoviscidose a considérablement augmenté durant les dernières décennies [Reid et al. 2011]. L'espérance de vie qui était de 7 ans il y a une cinquantaine d'année, est estimée à plus de 50 ans pour les sujets nés après 2000 [Dodge et al. 2007]. La médiane de survie quant à elle approche les 40 ans [MacKenzie

et al. 2014]. Grâce à l'innovation thérapeutique, au dépistage néonatal, à la prise en charge immédiate et le suivi constant des malades, la survie des sujets atteints de mucoviscidose a considérablement augmenté au cours des dernières années. Bien que l'état de santé des malades a tendance à s'aggraver avec l'âge, la proportion de malades adultes est désormais supérieure à celle des malades enfants [Bellis et al. 2015; Foundation 2016]. Il y a de moins en moins de décès infantile et on note une augmentation du nombre d'adultes malades. D'ici 2025, 75% des malades seront adultes [Burgel et al. 2015].

L'évolution de la maladie est très hétérogène car, il existe une multitude de mutations ayant des manifestations phénotypiques différentes. Généralement, seules une trentaine de mutations sont recherchées du fait de leur trop grand nombre. Or, la connaissance des mutations est indispensable pour envisager des thérapies ciblées. De plus, la maladie se manifeste différemment d'un sujet à l'autre. Cependant, une bonne prise en charge dès le diagnostic favorise le ralentissement de la progression et les malades vivent désormais plus longtemps avec une meilleure qualité de vie.

1.1.5 Fonction pulmonaire

La mucoviscidose est une maladie chronique et évolutive dont il est nécessaire d'étudier la progression afin d'améliorer la prise en charge des malades. La surveillance clinique se fait en grande partie par l'exploration fonctionnelle respiratoire.

L'exploration fonctionnelle respiratoire regroupe un ensemble d'examens permettant de mesurer la capacité respiratoire. L'examen incontournable de l'exploration fonctionnelle respiratoire est la spirométrie. Elle permet de suivre l'évolution de la fonction respiratoire et ainsi, diagnostiquer, évaluer la sévérité et assurer le suivi de certaines maladies. La spirométrie consiste à mesurer les débits d'air entrant et sortant des poumons au cours d'une expiration normale ou forcée. Cet examen est indolore et est réalisé dès l'âge de 6 ans. Plusieurs paramètres sont mesurés au cours d'une spirométrie, parmi lesquels la capacité vitale forcée (CVF) et le volume maximal expiré la première seconde (VEMS).

Le VEMS est le volume maximal expulsé au cours de la première seconde suivant une inspiration maximale. C'est le paramètre qui reflète clairement la fonction pulmonaire [Taussig et al. 1973; Davies and Alton 2009; VanDevanter et al. 2010]. Il

représente de ce fait une mesure fondamentale pour suivre l'évolution de la maladie. En effet, l'appareil respiratoire est l'organe le plus affecté par la maladie et plusieurs malades décèdent à cause de problèmes respiratoires. Il est donc nécessaire de surveiller régulièrement la fonction pulmonaire des sujets atteints de la mucoviscidose pour mieux étudier la progression et la sévérité de la maladie.

Le VEMS reste à ce jour le meilleur prédicteur de la mortalité chez les patients atteints de mucoviscidose [Kerem et al. 1992]. Il existe différentes équations de régression permettant de calculer le VEMS.

Crapo et al. propose une méthode de calcul du VEMS basée sur une régression linéaire, à partir de la taille et de l'âge de 251 sujets non fumeurs.

Knudson et al. propose une méthode de calcul du VEMS à partir de l'âge et du genre de 697 sujets non fumeurs, tirés de façon aléatoire d'un échantillon représentatif de la population caucasienne de la ville de Tucson en Arizona (États-Unis). De plus Knudson et al. propose des valeurs de référence des paramètres de spirométrie qui tiennent compte de la variabilité inter-sujet et de la distribution non Gaussienne de certains paramètres.

Coultas et al. propose une méthode de calcul des paramètres de spirométrie à partir des données de 576 sujets de la communauté hispanique. Un modèle logarithmique et un modèle de régression linéaire sont utilisés pour prédire les paramètres de spirométrie respectivement pour les sujets âgés de 6 à 18 ans, et pour les sujets âgés de 25 à 80 ans. Pour les sujets âgés de 19 à 24 ans, une interpolation linéaire est faite entre les valeurs obtenues en utilisant les équations de prédiction développées pour les deux autres tranches d'âge. De plus, Coultas et al. évalue d'autres méthodes de calcul basées sur les données de populations caucasiennes, qui sous-estiment les valeurs prédites de spirométrie sur les données des populations hispaniques.

Schwartz et al. propose une méthode de calcul des paramètres de spirométrie à partir des données de 1963 sujets non fumeurs, issus des populations caucasiennes et africaines. Les méthodes de calcul sont des analyses de régression basées sur la taille, l'âge, le genre des sujets.

Wang et al. propose une méthode de calcul des paramètres de spirométrie à partir des données de 11630 et 989 enfants issus des populations caucasiennes et africaines respectivement. Les méthodes de calcul sont basées sur la taille, l'âge, le genre et l'ethnie des sujets âgés de 6 à 18 ans.

Glindmeyer et al. propose une méthode de calcul des paramètres de spirométrie à partir des données de 5042 sujets non fumeurs, sans problèmes respiratoires, issus des populations caucasiennes et africaines. Les méthodes de calcul sont des régressions

polynomiales basées sur la taille, l'âge, le genre des sujets âgés de 18 à 65 ans.

Hankinson et al. propose une méthode de calcul des paramètres de spirométrie plus récente, à partir des données de 7429 sujets non fumeurs, issus des populations caucasiennes, africaines et mexicaines. Les méthodes de calcul sont basées sur la taille, l'âge, le genre et l'ethnie des sujets âgés de 8 à 80 ans. De plus, Hankinson et al. propose des valeurs de référence des paramètres de spirométrie utiles pour la recherche.

Le VEMS est en général exprimé en pourcentage de valeurs prédites et calculé en fonction de l'âge, le genre, la taille, le poids, l'ethnie, selon les formulations. De ce fait, la fonction pulmonaire des sujets atteints de mucoviscidose est comparable à celle de la population de référence, qui est une population saine, sans problèmes respiratoires.

La baisse du VEMS est un indicateur de dégradation de l'état de santé des patients. Elle est associée à la mortalité et à la morbidité chez les sujets atteints de mucoviscidose [Corey and Farewell 1996; Liou et al. 2001; Schluchter et al. 2002]. Plusieurs autres facteurs ont été associés à la baisse de la fonction pulmonaire. On peut citer entre autres un risque élevé de cures d'antibiotiques [Amadori et al. 2009], une insuffisance pancréatique [Corey et al. 1997; Schaedel et al. 2002] la présence de diabète [Schaedel et al. 2002], des infections telles que *Pseudomonas aeruginosa* [Schaedel et al. 2002; Konstan et al. 2007; Olszowiec-Chlebna et al. 2016]. Comme autre facteur associé à la baisse de la fonction pulmonaire, on peut également citer un risque élevé d'exacerbations pulmonaires qui sont des complications aigües dues à des infections respiratoires répétées et nécessitant généralement des cures d'antibiotiques [Konstan et al. 2007; Amadori et al. 2009; Olszowiec-Chlebna et al. 2016]. Le VEMS a tendance à diminuer avec l'âge du fait d'infections et d'inflammations chroniques des voies respiratoires [Flume 2009]. La diminution continue du VEMS est donc un indice d'aggravation de la maladie.

1.1.6 Facteurs pronostiques

La connaissance des facteurs pronostiques de la mucoviscidose est importante pour suivre son évolution. Identifier les facteurs pronostiques de la maladie permet de détecter les malades ayant un pronostic vital engagé afin de leur proposer des traitements adéquats pour améliorer leur état de santé. Plusieurs facteurs sont liés au pronostic des sujets atteints de mucoviscidose, le plus important étant le VEMS qui mesure la fonction pulmonaire. Des modèles pronostiques ont été développés afin

d'identifier les facteurs qui prédisent au mieux la mortalité chez les sujets atteints de mucoviscidose. En plus de la baisse du VEMS qui est associée à la mortalité chez les malades, on peut citer comme autre facteur pronostique, l'âge. Un âge élevé est en général associé à un risque élevé de décès car la maladie a tendance à s'aggraver avec l'âge [Kerem et al. 1992; Liou et al. 2001; Mayer-Hamblett et al. 2002; McCarthy et al. 2013]. Des indicateurs du statut nutritionnel, notamment un faible indice de masse corporel [George et al. 2011; McCarthy et al. 2013] et un faible poids [Aaron et al. 2015] sont également associés à un risque élevé de décès. Comme autres facteurs pronostiques de la maladie, on cite l'insuffisance pancréatique [Liou et al. 2007], le diabète [Aaron et al. 2015; Liou et al. 2001], les infections par *Pseudomonas aeruginosa* [Aaron et al. 2015; Courtney et al. 2007], *Burkholderia cepacia* [Liou et al. 2001; Courtney et al. 2007; Buzzetti et al. 2012] qui sont des germes associés à une dégradation rapide de la fonction pulmonaire. Des indicateurs d'aggravation de la maladie tels que les exacerbations pulmonaires Liou et al. [2001]; Buzzetti et al. [2012], les hospitalisations, les cures d'antibiotiques [Mayer-Hamblett et al. 2002] ont été identifiés comme facteurs pronostiques.

Des modèles pronostiques ont été développés incluant entre autres ces facteurs et permettant d'identifier les sujets à risque élevé de décès au cours d'une période de temps spécifique. À partir de ces modèles, les cliniciens sont à mesure d'identifier aisément les sujets nécessitant une prise en charge particulière.

1.1.7 Traitements

La mucoviscidose est à ce jour une maladie rare et incurable. Néanmoins, des progrès très significatifs ont été accomplis dans la recherche médicale dans ce domaine. La prise en charge thérapeutique fait intervenir une équipe pluridisciplinaire de professionnels de santé. D'une part, la maladie atteint divers organes chez le malade et d'autre part, elle se manifeste différemment d'un malade à l'autre. Pour une meilleure prise en charge, il est donc nécessaire que chaque malade ait un suivi personnalisé par différents professionnels de santé (médecins, diététiciens, infirmiers, assistants sociaux ...).

Bien que la maladie affecte plusieurs organes, les poumons restent les organes les plus endommagés entraînant ainsi de nombreuses difficultés respiratoires chez le malade. L'épais mucus endommage les poumons et conduit progressivement à une diminution de la fonction respiratoire et à des inflammations et infections pulmonaires. Pour

améliorer la qualité de vie des sujets atteints de mucoviscidose, il est indispensable de nettoyer les voies respiratoires notamment en se débarrassant du mucus qui obstrue les bronches. Il existe des médicaments à inhaler grâce à des nébuliseurs, qui permettent de fluidifier le mucus. Le mucus fluidifié sera ensuite expulsé sous forme de crachats. Pour cela, des techniques de nettoyage des bronches telles que la kinésithérapie sont appliquées. Elles consistent généralement à relaxer les muscles et dilater les bronches, ce qui favorise la fluidification et une extraction plus facile du mucus. De cette façon, le malade se débarrasse à la fois du mucus, mais également des germes présents dans les poumons.

Les infections pulmonaires par des agents pathogènes sont fréquentes chez les sujets atteints de mucoviscidose. Elles représentent une grande cause de morbidité et de mortalité chez les malades. Un des principaux challenges dans la prise en charge de la maladie est le traitement des infections pulmonaires [Flume et al. 2007]. Avec des durées plus ou moins longues, ces infections peuvent être sévères, on parle alors d'exacerbations pulmonaires. Dans ce cas, les infections sont traitées par des antibiotiques pris par voie intraveineuse. Ce mode de traitement peut être administré soit à domicile, soit à l'hôpital. Dans ce dernier cas, les malades sont admis à l'hôpital pendant toute la durée de la cure. Il existe également des antibiotiques pris par voie orale ou inhalés selon le type de germe à traiter ou la gravité de l'infection. Leur but principal est d'éradiquer les agents pathogènes présents dans les voies respiratoires. Bien que plusieurs pathogènes peuvent être éradiqués, il en existe quelques-uns qui restent résistants aux antibiotiques.

Tout comme dans les poumons, le mucus s'accumule dans le pancréas et empêche le passage des enzymes vers l'intestin. De ce fait, la digestion des aliments et l'absorption des nutriments devient difficile. Jusqu'à 90% des sujets atteints de mucoviscidose ont une insuffisance pancréatique [Collins 1992]. Un supplément en enzymes pancréatiques est donc nécessaire pour améliorer la digestion, en facilitant l'absorption des matières grasses. De même des suppléments en vitamines et des suppléments de minéraux sont recommandés. De plus, un régime alimentaire adéquat est nécessaire pour éviter les retards de croissance, notamment dans l'enfance, ceci favorise une meilleure qualité de vie. Par ailleurs, il a été montré qu'une amélioration du statut nutritionnel chez les malades améliorerait la fonction pulmonaire [Konstan et al. 2003; Sanders et al. 2015].

Un traitement novateur a été mis en place, il s'agit de la thérapie génique qui vise

à réparer la fonction défectueuse du gène muté. Le principe est d'introduire une copie de l'ADN fonctionnant normalement dans les cellules affectées. Cependant, cette technique s'avère plus compliquée en pratique. Sachant qu'il existe plusieurs types de mutations, les médicaments disponibles pour la thérapie génique ne s'appliquent que pour certaines mutations. Il y a à ce jour deux médicaments approuvés pour la thérapie génique : Ivacaftor (Kalydeco) et Lumacaftor/Ivacaftor (Orkambi). Ivacaftor est prescrit aux malades âgés de 2 ans ou plus avec un type de mutation précis [Davies et al. 2016]. C'est un traitement pris par voie orale qui facilite l'ouverture du canal chlore à la surface de la cellule et permet ainsi le passage des ions vers l'intérieur et l'extérieur de la cellule. Une autorisation européenne de mise sur le marché a été obtenue pour Ivacaftor en 2012. La combinaison Lumacaftor/Ivacaftor quant à elle, est prescrite aux malades âgés de plus de 12 ans et homozygotes F508del, mutation la plus fréquente chez les personnes atteintes de la mucoviscidose. Dans ce cas, Lumacaftor a pour rôle de positionner la protéine défectueuse au bon endroit à la surface de la cellule. Une autorisation européenne de mise sur le marché a été obtenue pour Lumacaftor/Ivacaftor en 2015.

Un essai clinique visant à évaluer l'impact de Ivacaftor sur la fonction pulmonaire chez des malades âgés de 12 ans ou plus a été réalisé. Il en découle que les malades ayant reçu Ivacaftor ont une meilleure fonction pulmonaire (10.6% de plus que les malades du groupe placebo), moins d'exacerbations pulmonaires, une augmentation de poids et une meilleure qualité de vie [Ramsey et al. 2011]. Les mêmes constats ont été faits pour une étude similaire réalisée chez les sujets âgés de 6 à 11 ans [Davies et al. 2013]. De même, une étude de phase III a été réalisée sur une durée de 24 semaines pour évaluer l'impact sur la fonction pulmonaire de Lumacaftor/Ivacaftor chez des malades âgés de 12 ans ou plus. Le taux d'exacerbations pulmonaires était de 30% à 39% plus bas dans le groupe ayant reçu le traitement comparé au groupe ayant reçu le placebo. Le même constat est fait pour le taux d'événements entrainant des hospitalisations ou des cures d'antibiotiques intraveineuses. L'amélioration de la fonction pulmonaire était observée au bout de 15 jours. La différence était de 4.3% à 6.7% ($p<0.001$) entre les deux groupes, avec des valeurs plus faibles dans le groupe placebo [Wainwright et al. 2015].

Plusieurs autres thérapies géniques suivent actuellement leur processus de fabrication pour améliorer le quotidien des malades. La majorité d'entre elles ont un rôle similaire soit à Ivacaftor, soit à Lumacaftor. Cependant, elles peuvent concerner d'autres types de mutations, d'autres tranches d'âge ou alors avoir une posologie

plus légère pour les malades.

Les traitements délivrés aux sujets atteints de la mucoviscidose sont divers et dépendent de l'âge et de la sévérité de la maladie. Ils peuvent être regroupés selon leur approche thérapeutique. Malgré le fait que la maladie affecte principalement les voies respiratoires et digestives, tous les organes peuvent être atteints. Il est important que le mucus soit expulsé. Certains traitements permettent de diminuer l'épaisseur du mucus et de le fluidifier pour une extraction plus facile. Certaines thérapies permettent de dégager les voies respiratoires et donc limiter le développement des bactéries. Les antiinflammatoires et les antibiotiques permettent de lutter contre les infections pulmonaires. Les enzymes facilitent la digestion. En cas d'insuffisance respiratoire, l'oxygénothérapie ou encore la ventilation non invasive aident le malade à respirer avec beaucoup moins d'efforts. Des thérapies ciblées et adaptées pour chaque type de mutation permettent d'améliorer le fonctionnement du gène défectueux. Bien que la maladie soit incurable, on note beaucoup de progrès dans la recherche qui améliorent la fonction respiratoire et la qualité de vie, et prolongent ainsi l'espérance de vie des malades.

1.1.8 Transplantation pulmonaire

Selon la société internationale de transplantation cardiaque et pulmonaire, plus de 62000 transplantations pulmonaires ont été réalisées jusqu'en 2016, dont plus de 60000 chez les adultes. La proportion de transplantation pulmonaire chez les patients adultes atteints de mucoviscidose représentait environ 16%. Malgré les nombreuses avancées thérapeutiques, la transplantation pulmonaire est à ce jour le moyen le plus adéquat pour augmenter l'espérance de vie et améliorer la qualité de vie des sujets ayant une fonction respiratoire détériorée [Kugler et al. 2004; Vermeulen et al. 2004; Thabut et al. 2013]. En effet, les malades avec une expression sévère de la maladie auront tendance à développer des bronchiectasies (altération des voies respiratoires) et à avoir un déclin continue de la fonction pulmonaire. Dans ce cas, la transplantation pulmonaire devient la solution optimale pour améliorer la qualité de vie des malades.

Le nombre de transplantations pulmonaires est en constante évolution du fait qu'il y ait de plus en plus de malades adultes et que la maladie a tendance à s'aggraver avec l'âge. Le temps d'attente d'une greffe est variable selon les pays et peut être critique.

En effet, Plusieurs paramètres sont pris en considération pour réaliser une transplantation pulmonaire notamment la disponibilité et la compatibilité d'un greffon, l'avis du malade, les contrindications à la transplantation, le jugement du médecin qui dépend fortement des manifestations cliniques de la maladie. Afin d'optimiser l'utilisation des greffons disponibles et de maximiser les chances de non rejet du greffon, les centres de transplantation doivent déterminer le temps optimal et le malade le plus compatible qui recevra le greffon.

Plusieurs critères sont pris en compte pour une évaluation à la transplantation pulmonaire. On peut citer une dégradation continue de la fonction pulmonaire notamment un VEMS inférieur à 30% de valeurs prédites, mais aussi une diminution de la quantité d'oxygène dans le sang et une augmentation de la concentration en gaz carbonique dans le sang. Une insuffisance respiratoire importante, une augmentation d'exacerbations pulmonaires nécessitant des cures d'antibiotiques ou des hospitalisations, ainsi qu'un déclin rapide du VEMS sont également à considérer. Comme autres critères d'évaluation à la transplantation pulmonaire, on peut citer des symptômes pneumothorax, des hémoptysies (rejet du sang provenant des voies respiratoires par la bouche) récurrentes ou encore des indices de malnutrition ou de perte de poids [Yankaskas and Mallory 1998; Spahr et al. 2007]. D'autres critères tels que l'âge ou encore les infections et colonisations pulmonaires sont également pris en compte. Cependant, il existe des contrindications à la transplantation pulmonaire. Ceux-ci sont spécifiques aux centres de transplantation.

Certes il peut y avoir quelques complications après une transplantation pulmonaire, néanmoins la santé des malades s'améliore considérablement. Notamment il y a une nette amélioration de la fonction pulmonaire avec des valeurs de VEMS en moyenne plus élevées de 20% à 80% de valeurs prédites [Egan et al. 1995]. Il a été montré que la survie un an après la transplantation pulmonaire était de 81%. Elle était de 59% et 38% respectivement 5 ans et 10 ans après la transplantation pulmonaire [Lama et al. 2001]. On note également une amélioration de la qualité de vie, avec une diminution de la prise de traitements, une pratique plus aisée d'activités physiques et sportives.

Malgré la progression dans la recherche, la mise en place de nouveaux traitements qui permettent d'améliorer l'état de santé des malades, la mucoviscidose reste une maladie incurable. De ce fait, la transplantation pulmonaire reste un moyen inévitable pour prolonger la vie des malades. Les sujets éligibles à la transplantation pulmonaire sont les cas les plus graves. Certes la transplantation pulmonaire ne fait

pas disparaître la maladie, les symptômes sont toujours présents et les infections pulmonaires restent un problème majeur. Néanmoins, le malade reçoit des poumons sains, et on note une augmentation non négligeable de la fonction pulmonaire et de la qualité de vie des sujets ayant reçu une transplantation pulmonaire.

1.1.9 La mucoviscidose en France

La mucoviscidose est la maladie rare la plus fréquente en France avec plus de 2 millions de sujets porteurs sains du gène CFTR. On compte environ 7000 sujets atteints de mucoviscidose. Ceci représente environ 90% de couverture de la population des malades en France. Près de 200 enfants atteints de mucoviscidose naissent chaque année ce qui correspond à une incidence moyenne d'une naissance sur 4500.

Le dépistage néonatal systématique existe en France depuis 2002. Il permet une prise en charge immédiate et un suivi constant des malades par des équipes pluridisciplinaires. Ce suivi régulier est assuré par les Centres de Ressources et de Compétences de la Mucoviscidose (CRCM). Ces centres ont été créés en 2002 et sont présents sur tout le territoire français. Ce sont des centres spécialisés et constitués d'équipes de professionnels pour une meilleure prise en charge des malades. On compte 45 CRCM dont 16 CRCM pédiatriques, 12 CRCM adultes et 17 CRCM mixtes. Les malades sont transférés progressivement des centres pédiatriques vers les centres adultes à l'âge de 18 ans (ou entre 17 et 21 ans). Les CRCM assurent un suivi médical régulier et rigoureux, permettant de réduire la gravité et la fréquence des complications chez les malades.

En 1992, l'Observatoire National de la Mucoviscidose a été mis en place, puis a été transformé en Registre Français de la Mucoviscidose (RFM) en 2006. Depuis sa création, il collecte les données des sujets atteints de la maladie. Entre 1992 et 2017, il y a eu une inclusion de plus en plus croissante de malades dans le registre, ce qui lui donne une exhaustivité d'environ 90% des malades en France. Ce taux de couverture est estimé à partir d'un modèle de population stationnaire [Bellis et al. 2007] et également en référence aux données de la caisse nationale de l'assurance maladie des travailleurs salariés (CNAMTS), aux données des régimes spéciaux et aux données des malades non déclarés en affectation de longue durée (ALD) pour la mucoviscidose. Un des objectifs visé par le RFM est l'amélioration de la qualité et de l'exhaustivité des données du registre. En effet, le RFM qui a longtemps été alimenté

exclusivement par les centres de soins possède aujourd'hui un recueil multi-sources. Les données de diagnostic sont désormais complétées par l'association française pour le dépistage et la prévention des handicaps de l'enfant (AFDPHE) ainsi que par la base de données française des mutations CFTR rares (CFTR-France). Les données concernant les décès sont complétées par le centre d'épidémiologie sur les causes médicales de décès (CépiDc).

Le RFM constitue une importante source de données pour la recherche et leur exploitation est indispensable pour une meilleure connaissance de la maladie. Grâce à des études épidémiologiques, il est possible de détecter les déterminants de l'évolution de la mucoviscidose qui reste à ce jour une maladie complexe.

Suite au diagnostic, chaque malade est affecté à un centre de soin où il peut être vu une ou plusieurs fois dans l'année. Une visite par trimestre est recommandée aux malades. Néanmoins, le suivi peut être réalisé selon l'état de santé du malade ou selon les prescriptions du médecin. À partir des données recueillies dans l'année, un bilan annuel est établi via un questionnaire par les centres de soin et permet d'alimenter les données du RFM. Elles concernent entre autres l'état civil, les données démographiques et sociales, les éléments de diagnostic, les données anthropométriques, les données de spirométrie, les traitements, les données bactériologiques, les éléments de mortalité et de morbidité.

Le contrôle qualité des données est réalisé par l'association Vaincre la Mucoviscidose. Des bilans annuels sont réalisés par l'association Vaincre la Mucoviscidose en collaboration avec l'Institut National d'Études Démographiques (Ined). Ces bilans visent à décrire de manière générale l'état de santé des malades. Les caractéristiques démographiques et les éléments de morbidité y sont décrits. L'évolution de la maladie y est décrite, notamment par l'estimation des risques de survenue d'événements tels que le décès ou encore des complications médicales. Néanmoins, le RFM constitue une base de données de qualité mise à la disposition des chercheurs pour la réalisation d'études permettant de faire avancer la recherche.

Le bilan du RFM pour l'année 2015 montre qu'il y a presqu'autant d'hommes (52%) que de femmes (48%). Le nombre d'adultes est en constante évolution. En 1992, les adultes représentaient seulement 18.7% de la population de malades. En 2015, plus de la moitié des malades étaient adultes, soit 53.7% recensés par le registre. L'âge moyen était de 10 ans, 32 ans et 20 ans respectivement dans les CRCM pédiatriques,

adultes et mixtes.

Les données du registre montrent que jusqu'à l'âge de 9 ans, il n'y a pas de différence significative de survie entre les cohortes de naissance. Après cet âge, les cohortes de naissance les plus anciennes (jusqu'en 2001) ont une survie significativement plus faible que les cohortes de naissance les plus récentes (à partir de 2002, année de mise en place du dépistage néonatal). Parmi les nouveaux patients inclus dans le registre en 2015, 60% ont eu un dépistage néonatal. La mutation la plus fréquente en France est la mutation F508del. En 2015, elle était présente chez 83% des malades, suivie de la mutation G542X présente chez seulement 5% des malades. Jusqu'en 2010, pour le bilan annuel des malades, le registre retenait la dernière valeur du VEMS relevée. À partir de 2011, le registre retenait la plus grande valeur du VEMS relevée dans l'année. Le VEMS médian est passé de 77% à 92% chez les sujets âgés de 6 à 19 ans, entre 1995 et 2015. Il est passé de 48% à 72% entre 1995 et 2015 chez les malades âgés de 20 ans et plus. Le VEMS diminue donc avec l'âge et a tendance à s'améliorer au cours du temps, grâce aux progrès dans la recherche.

Tout comme dans la population mondiale des malades, *Staphylococcus aureus* et *Pseudomonas aeruginosa* sont les germes les plus fréquents. Contrairement à *Pseudomonas aeruginosa*, *Staphylococcus aureus* est plus fréquent dans l'enfance et a tendance à diminuer avec l'âge. La fréquence des germes est restée assez constante au cours des dix dernières années dans la population française. En France, les transplantations chez les sujets atteints de mucoviscidose concernent surtout les poumons. D'autres transplantations peuvent être réalisées dans des proportions plus faibles (rein, foie, cœur-poumon, foie-poumon, foie-rein...). En 2015, 86% des transplantations réalisées étaient des transplantations bi-pulmonaires. L'âge moyen des patients greffés était de 30 ans, avec un écart-type de 11 ans. Cinq décès post-greffe ont été enregistrés au cours de cette année. Très peu de décès sont enregistrés sur liste d'attente chaque année. Depuis 2007, moins de 4% de malades décèdent en attente de greffe, chaque année. Concernant la prise en charge thérapeutique, plus de 80% des malades ont une insuffisance pancréatique et reçoivent des extraits pancréatiques. Les malades reçoivent en moyenne 2 cures par an pour une durée moyenne d'environ 38 jours. Peu de malades bénéficient de la thérapie génique. En 2015, seuls 1.9% ont reçu Ivacaftor du fait que ce traitement ne concerne que des mutations peu fréquentes dans la population de malades. La même année 2015, année de mise sur le marché de Lumacaftor/Ivacaftor, 0.8% ont reçu ce traitement.

Le RFM est outil important pour réaliser des études épidémiologiques permet-

tant une meilleure connaissance de la maladie. Il est efficace pour comprendre la progression de la maladie et identifier les facteurs cliniques associés à des évènements d'intérêt comme la transplantation pulmonaire ou le décès. Identifier de tels facteurs permettraient aux cliniciens de mieux identifier les patients nécessitant une prise en charge personnalisée.

1.2 Problématique et objectif de la thèse

L'espérance de vie des sujets atteints de mucoviscidose est en constante évolution grâce aux progrès scientifiques. Néanmoins, la transplantation pulmonaire reste le moyen le plus efficace pour prolonger la vie des malades dont l'état respiratoire s'est considérablement dégradé. Il est donc important d'identifier au mieux les malades éligibles à la transplantation pulmonaire. Dans cette optique, des modèles de prédiction ont été développés et ont permis d'identifier les facteurs associés au décès chez les sujets atteints de mucoviscidose. Parmi les facteurs identifiés, certains sont des contrindications à la transplantation pulmonaire. De plus, dans le contexte actuel de la mucoviscidose, une prédiction du risque de décès sans tenir compte de la transplantation pulmonaire serait une approche biaisée. En effet, ne considérer que les sujets décédés exclurait les sujets à risque très élevé de décéder, qui sont les sujets transplantés. Dans le contexte de la mucoviscidose, soit les malades en phase terminale décèdent, soit ils reçoivent une transplantation pulmonaire. La transplantation pulmonaire permet donc d'éviter le décès et, ne saurait être ignorée dans le contexte de la survie des sujets atteints de mucoviscidose.

Par ailleurs, suite aux progrès thérapeutiques liés à la mucoviscidose, le pronostic des malades s'est nettement amélioré et la mortalité infantile a presque disparue dans les pays développés. Réévaluer les facteurs pronostiques et prédire le risque de survenue du décès ou de transplantation pulmonaire fourniraient des recommandations pour la prise de décision de greffe.

L'objectif principal ce travail de thèse est de proposer des outils d'aide à la décision clinique pour l'identification des sujets nécessitant une évaluation pour une transplantation pulmonaire. Pour cela, des modèles de prédiction du risque de décès ou de transplantation pulmonaire chez les adultes atteints de mucoviscidose ont été développés.

Une étape préalable est d'identifier les facteurs pronostiques adaptés au contexte actuel de la mucoviscidose chez les adultes. Ensuite il est question de prédire le

risque de décès ou de transplantation pulmonaire en tenant compte de l'évolution temporelle du VEMS, principal marqueur de la mucoviscidose qui permet de décrire la fonction pulmonaire. Ces outils d'aide à la décision permettront d'estimer de façon appropriée le risque pour un malade de recevoir une transplantation pulmonaire. Il s'agit d'outils pronostiques qui fourniront des prédictions dynamiques du risque de transplantation pulmonaire ou du décès à partir de l'évolution temporelle du VEMS. Ainsi, chaque sujet aura un risque de recevoir une greffe pulmonaire ou de décéder sur une fenêtre de temps précise, qui sera actualisé après chaque nouvelle valeur du VEMS.

1.3 Plan du mémoire

Dans le chapitre 1, nous présentons le contexte épidémiologique de la mucoviscidose.

Dans le chapitre 2, nous présentons l'état de l'art des méthodes d'analyse de variables binaires, des données longitudinales et des données de survie. Les méthodes de modélisation conjointe pour données longitudinales et temps d'événements y sont ensuite décrites. Puis, l'évaluation des capacités prédictives de ces modèles, ainsi que leur validation sont présentées.

Dans le chapitre 3, nous présentons le premier travail de cette thèse qui consiste à identifier, chez les adultes atteints de mucoviscidose, les facteurs associés à la transplantation pulmonaire ou au décès à 3 ans. Ce travail a été fait à partir d'un modèle de régression logistique dont les résultats sont par la suite comparés à ceux obtenus par un modèle de survie prenant en compte les données censurées.

Dans le chapitre 4, nous présentons le deuxième travail de cette thèse qui consiste à développer un modèle conjoint pour la prédiction dynamique du risque de transplantation pulmonaire ou du décès chez les adultes atteints de mucoviscidose. Les modèles développés sont des modèles conjoints à classes latentes. En plus de fournir des prédictions dynamiques pour la transplantation pulmonaire ou le décès, ils permettent d'identifier des profils différents d'évolution de la mucoviscidose.

Dans le chapitre 5, nous faisons une discussion générale sur les travaux présentés dans cette thèse et donnons une conclusion et quelques perspectives.

Chapitre 2

État de l'art

Dans ce chapitre, nous présentons les différentes méthodes statistiques utilisées pour l'analyse des données. Nous commençons par la méthode d'analyse de variables binaires, suivie des méthodes d'analyse de données longitudinales et des données de survie. Ensuite nous présentons les modèles conjoints pour données longitudinales et temps d'événement. Nous terminons par l'évaluation des capacités prédictives des modèles conjoints et leur validation.

2.1 Modèle de régression logistique

Le pronostic dans le contexte d'une maladie, se rapporte à l'évolution future de la maladie et au devenir du malade vis-à-vis de la maladie. Déterminer le pronostic d'un sujet vis-à-vis d'une maladie, nécessite la connaissance des facteurs pouvant influencer l'évolution de cette maladie. Pour un sujet atteint d'une maladie particulière, on parle de pronostic vital lorsqu'on fait référence au risque qu'a ce sujet d'en décéder. Il est alors possible de s'intéresser directement aux facteurs liés au décès en rapport avec la maladie. Il s'agit d'identifier les facteurs liés à la survenue d'un événement caractérisé par une variable qualitative binaire avec les catégories « décédé » et « vivant ». En statistique le modèle de régression logistique est approprié pour l'étude de telles variables. Il permet d'étudier la relation existante entre une variable qualitative binaire et une ou plusieurs variables qualitatives ou quantitatives.

2.1.1 Spécification du modèle logistique

Dans un modèle logistique, on s'intéresse à la survenue d'un événement sur une période de temps fixée, qui est décrite par une variable à expliquer ayant deux catégories. Cette variable est caractérisée par une variable aléatoire notée Y. Elle est codée 0 pour l'absence de l'événement et 1 pour la présence de l'événement. L'idée du modèle logistique est de modéliser la probabilité de survenue de l'événement d'intérêt, connaissant les valeurs de p variables explicatives, pouvant être quantitatives ou qualitatives qu'on note $X_1, X_2, X_3, \ldots, X_p$.

$Y, X_1, X_2, \ldots, X_p$ sont des variables de la population dont on extrait un échantillon de n individus i. Pour un sujet i, $i = 1, 2, \ldots, n$, on note y_i l'observation de la variable à expliquer Y_i et $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ le vecteur de réalisation des variables explicatives $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$. Les hypothèses du modèle de logistique portent sur les distributions des variables Y_i sachant les x_i. Elles sont indépendantes et suivent une loi de Bernoulli de paramètre π défini par :

$$\begin{aligned} \pi(x_i) = E(Y_i|x_i) &= P(Y_i = 1|x_i) \\ &= \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})} \end{aligned}$$

On modélise la probabilité de survenue d'un événement en fonction des variables explicatives. Cette probabilité est l'espérance de la variable aléatoire $Y_i|x_i$, comprise entre 0 et 1 pour toutes les valeurs prises par les variables explicatives $x_{i1}, x_{i2}, \ldots, x_{ip}$. Pour s'assurer que les valeurs prises par l'espérance de $Y_i|x_i$ soient bien comprises entre 0 et 1, on la modélise par la fonction logistique qui a la forme suivante :

$$f(x) = \frac{exp(x)}{1+exp(x)}$$

De ce fait, pour des valeurs de x pouvant varier de $-\infty$ à $+\infty$, les valeurs prises par la fonction logistique $f(x)$ sont toujours comprises entre 0 et 1.

Il est également possible d'exprimer la relation entre $\pi(x_i)$ et le vecteur de variables explicatives x_i en utilisant la transformation Logit, de la manière suivante :

$$Logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right)$$

En remplaçant la probabilité de survenue de l'événement π par sa valeur, on peut écrire :

$$\begin{aligned} Logit(\pi(x_i)) &= ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) \\ &= \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \end{aligned}$$

$\frac{\pi(x_i)}{1-\pi(x_i)}$ représente une côte qui peut être calculée pour des catégories d'individus selon leurs caractéristiques. Il est alors possible de déterminer des rapports de cotes entre 2 catégories d'individus. Le rapport de cotes est utile pour l'interprétation des paramètres du modèle logistique.

2.1.2 Interprétation des paramètres

Soit un modèle de régression logistique à p variables explicatives définit comme suit :

$$Logit(\pi(x)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

L'interprétation de β_0, ainsi que le calcul de son intervalle de confiance n'ont de sens que si toutes les autres variables peuvent prendre la valeur 0. Dans le cas contraire, seules les interprétations des paramètres associés aux variables explicatives et leurs intervalles de confiance ont un intérêt pratique.

Prenons le cas du paramètre β_1 associé à la variable explicative x_1. On s'intéresse à $exp(\beta_1)$ qui représente le rapport de cotes du risque de survenue de l'événement d'intérêt pour une augmentation d'une unité de la variable x_1 si elle est quantitative, ajusté sur les autres variables explicatives présentes dans le modèle.

Si la variable x_1 est qualitative, $exp(\beta_1)$ représente le rapport de cotes du risque de survenue de l'événement d'intérêt chez les exposés par rapport aux non exposés, ajusté sur les autres variables explicatives présentes dans le modèle. Les exposés sont les sujets pour lesquels la variable x_1 vaut 1 et les non exposés sont ceux pour lesquels la variable x_1 vaut 0. x_1 étant une variable qualitative binaire indiquant l'exposition à un facteur.

On parle d'ajustement lorsqu'on compare des groupes ayant des valeurs différentes pour la variable considérée (x_1) et ayant des valeurs identiques pour toutes les autres variables présentes dans le modèle.

Le rapport de cotes est toujours positif et fournit une information sur la force et le sens de l'association entre la variable à expliquer et les variables explicatives. Les variables sont indépendantes si le rapport de cotes vaut 1. Plus le rapport de

cotes se rapproche de 0 ou de $+\infty$, plus le lien entre les variables est fort. La variable explicative est considérée comme un facteur de risque pour l'événement si le rapport de cotes est supérieur à 1. Cependant, s'il est inférieur à 1, la variable explicative est considérée comme un facteur protecteur. Si l'événement étudié est rare, avec une prévalence inférieure à 10% par exemple, le rapport de cotes est comparable au risque relatif (mesure d'association correspondant au rapport de risques absolus entre les sujets exposés et les sujets non exposés).

2.1.3 Estimation du modèle logistique

L'estimation des paramètres du modèle logistique se fait par la méthode du maximum de vraisemblance. Cette méthode fournit des estimateurs ayant de bonnes propriétés asymptotiques, c'est-à-dire sans biais et avec une faible variance. $Y, X_1, X_2, \ldots, X_p$ sont des variables de la population dont on extrait un échantillon de n individus i. Pour un sujet i, $i = 1, 2, \ldots, n$, on note y_i l'observation de la variable à expliquer Y_i et $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ le vecteur de réalisation des variables explicatives $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$. Soit le modèle de régression logistique définit par :

$$Logit(\pi(x_i)) = ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

On estime le vecteur de paramètres $\beta = (\beta_0, \beta_1, ..., \beta_p)$ par la fonction de vraisemblance correspondant à la probabilité d'observer l'échantillon y_i $(i = 1, \ldots, n)$ sachant les x_i. La fonction de vraisemblance s'écrit :

$$\begin{aligned}
V(\beta) &= \prod_{i=1}^{n}\left[\pi^{y_i}(1-\pi)^{(1-y_i)}\right] \\
&= \prod_{i=1}^{n}\left[\left(\frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}\right)^{y_i}\left(1 - \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}\right)^{(1-y_i)}\right] \\
&= \prod_{i=1}^{n}\left[\left(\frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}\right)^{y_i}\left(\frac{1}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}\right)^{(1-y_i)}\right]
\end{aligned}$$

Le principe du maximum de vraisemblance est de déterminer les valeurs de $\beta = (\beta_0, \beta_1, ..., \beta_p)$ qui maximisent la log-vraisemblance $L(\beta)$.

$$L(\beta) = \sum_{i=1}^{n}\left[y_i\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\right) - ln\left(1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})\right)\right]$$

La fonction log-vraisemblance est maximisée en déterminant les solutions de l'équation qui annule sa dérivée par rapport au paramètre β. On obtient le vecteur

score

$$U(\beta) = \frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \left(y_i - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)$$

avec $j(j = 0, 1, ..., p)$, $x_{i0} = 1$ et $x_i^T \beta = \sum_{j=0}^{p} \beta_j x_{ij}$

Une fois les paramètres $\widehat{\beta}$ du modèle estimés par des méthodes itératives utilisant des algorithmes numériques, il est possible de calculer leur intervalle de confiance. En pratique, on s'intéresse à l'intervalle de confiance du rapport de cotes, donc de $exp(\beta)$. Dans un modèle de régression logistique, l'intervalle de confiance du rapport de cotes à 100(1-α)% associé à la $j^{ième}$ $(j = 1, 2, \ldots, p)$ variable explicative est donné par :

$$IC(RC_j) = exp\left(\widehat{\beta}_j \pm \left|Z_{\frac{\alpha}{2}}\right| \sqrt{\widehat{var}(\widehat{\beta}_j)}\right)$$

Si l'intervalle de confiance exclut la valeur 1, alors il existe une association significative entre la variable dépendante et la variable explicative x_j, après ajustement sur les autres variables explicatives présentes dans le modèle. Si l'intervalle de confiance inclut la valeur 1, il n'y a pas d'association entre les deux variables, tenant compte des autres variables explicatives présentes dans le modèle.

%———————————————————————————

–

2.1.4 Tests d'hypothèse sur les paramètres du modèle

Il est possible de comparer des modèles emboités pour tester l'apport de variables explicatives dans un modèle de régression logistique. On peut tester l'apport global de toutes les variables explicatives d'un modèle logistique. Le but étant de déterminer si toutes les variables prises simultanément sont associées à la survenue de l'événement d'intérêt. On peut également s'intéresser à l'apport d'un sous ensemble de variables dans un modèle de régression logistique. Le but étant de déterminer si en présence des autres variables explicatives incluses dans le modèle, le sous-groupe de variables testées a une influence sur le risque de survenue de l'événement d'intérêt. Enfin, on peut s'intéresser à l'apport spécifique d'une variable dans le modèle de régression en présence d'autres variables explicatives.

La méthode de maximum de vraisemblance est utile pour faire ces tests d'hypothèses sur les paramètres du modèle, notamment par le test du rapport de vraisemblance. En pratique, il est surtout utilisé pour tester l'apport global des variables explicatives ou l'apport d'un sous-groupe de variables explicatives d'un modèle.

D'autres tests équivalents tels que le test du score ou le test de Wald sont également utilisés, notamment pour tester l'apport spécifique d'une variable explicative dans un modèle de régression logistique.

2.1.5 Adéquation du modèle logistique

Calibration

L'adéquation du modèle de régression logistique permet de déterminer la qualité d'ajustement du modèle aux données. Si l'ajustement est correct, alors les valeurs prédites par le modèle sont proches des valeurs observées de la variable à expliquer. Pour tester l'adéquation du modèle, il est donc possible de comparer les probabilités prédites par le modèle et celles observées dans l'échantillon de données. Pour ce faire, on peut utiliser le test d'Hosmer et Lemeshow [Hosmer et al. 1997]. Il s'agit d'une mesure d'adéquation qui consiste à calculer pour chaque observation la probabilité prédite par le modèle, puis de regrouper les probabilités prédites par le modèle en déciles, et de comparer dans chaque classe les effectifs observés et les effectifs théoriques.
Dans chaque groupe, on observe l'écart entre les valeurs observées et les valeurs prédites. Pour un modèle adéquat, il faut que cet écart soit minimal. On le calcule au moyen d'une statistique $\widehat{C}$ définie de la manière suivante :

$$\widehat{C} = \sum_{G=1}^{10} \left[\frac{(O_{1G} - E_{1G})^2}{E_{1G}} + \frac{(O_{0G} - E_{0G})^2}{E_{0G}} \right]$$

On a 10 groupes $G = 1, 2, \ldots, 10$. $O_{1G} = \sum_{i \in G} Y_i$ représente le nombre observé de sujets ayant connu l'événement d'intérêt et $O_{0G} = \sum_{i \in G} (1 - Y_i)$ représente de nombre observé de sujets n'ayant pas connu l'événement. Ces deux valeurs sont comparées respectivement aux nombres prédits par le modèle $E_{1G} = \sum_{i \in G} \widehat{\pi}_i$ et $E_{0G} = \sum_{i \in G} (1 - \widehat{\pi}_i)$, avec

$\widehat{\pi_i} = \widehat{y}_i \frac{exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip})}{1 + exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip})}$ et $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ les valeurs estimées des paramètres du modèle.

Si dans chaque classe ces deux effectifs sont proches alors le modèle est calibré. L'hypothèse nulle testée est H_0 : « le modèle est calibré », « le modèle est adéquat », « le modèle s'ajuste bien aux données », « les probabilités observées sont proches des probabilités théoriques ». Sous H_0, la statistique $\widehat{C}$ suit une loi du χ^2 à 8 degrés de liberté (nombre de groupes – 2).

Discrimination

Dans un contexte de prédiction, il est possible de construire un modèle visant à prédire la probabilité de survenue d'un événement d'intérêt en fonction des variables explicatives. On peut alors évaluer la qualité de prédiction du modèle, c'est-à-dire évaluer la capacité qu'a le modèle à correctement classer les sujets vis-à-vis de l'événement en fonction des probabilités prédites par le modèle.
Pour chaque sujet i, il est possible à partir des probabilités prédites par le modèle $\widehat{\pi}(x_i) = \widehat{y}_i$, de fixer un seuil s entre 0 et 1 (par exemple 0.5%) et de classer le sujet i comme « ayant connu l'événement » si $\widehat{y}_i \geq s$ ou « n'ayant pas connu l'événement » si $\widehat{y}_i \leq s$, avec

$\widehat{\pi_i} = \widehat{y}_i \frac{exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + ... + \widehat{\beta}_p x_{ip})}{1 + exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + ... + \widehat{\beta}_p x_{ip})}$ et $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p$ les valeurs estimées des paramètres du modèle.

Pour chaque seuil s, il est possible de calculer la sensibilité et la spécificité du modèle. La sensibilité correspond à la proportion des sujets prédits comme « ayant connu l'événement » parmi les sujets ayant réellement connu l'événement. La spécificité correspond à la proportion de sujets prédits comme « n'ayant pas connu l'événement » parmi les sujets n'ayant réellement pas connu l'événement.
Pour chaque valeur de s comprise entre 0 et 1, on aura une valeur pour la sensibilité et une valeur pour la spécificité. Il est alors possible de tracer, à partir de toutes ces valeurs, la courbe ROC (Reciever Operating Curve) qui est représentée par la sensibilité en fonction de 1-spécificité. Le critère de discrimination utilisé dans ce cas est l'aire sous la courbe ROC. Plus cette aire est proche de 1, meilleure est la capacité prédictive du modèle.

2.2 Analyse de données longitudinales

Les données longitudinales sont des données recueillies sur des individus à différents instants dans le temps. Ces données représentent un intérêt majeur en épidémiologie, notamment pour étudier l'évolution d'un paramètre au cours du temps ou encore d'identifier les facteurs qui influencent cette évolution. En analyse des données longitudinales, les mesures d'un même individu ne peuvent être considérées comme des réalisations indépendantes. Cette corrélation doit être prise en compte lors de l'analyse des données pour ne pas obtenir des résultats incorrects et avoir une meilleure

précision des effets. Les modèles linéaires mixtes sont des méthodes statistiques adaptés à l'analyse des données longitudinales.

2.2.1 Spécification du modèle linéaire mixte

Les modèles linéaires mixtes sont une extension des modèles linéaires simples. Ils sont constitués d'effets fixes (présents dans le modèle linéaire simple) qui décrivent la variable dépendante d'intérêt au niveau de la population, et d'effets aléatoires qui prennent en compte la variabilité liée aux individus. En effet, dans un suivi longitudinal, les sujets n'ont pas forcément le même nombre de mesures, de plus les mesures ne sont pas toujours prises au même moment. Il existe des différences dans le suivi de deux sujets distincts. Cependant, il existe une corrélation entre les mesures répétées d'un même sujet. Les effets aléatoires présents dans le modèle linéaire mixte permettent de capturer cette corrélation. Ils représentent alors des déviations individuelles par rapport à l'évolution moyenne du paramètre étudié dans la population.

Les modèles linéaires mixtes ont été décrits par Laird and Ware [1982] pour étudier l'évolution d'un paramètre Gaussien mesuré au cours du temps. Dans une population de N sujets ($i = 1, \ldots, N$), notons $Y_i = (Y_i(t_{i1}), Y_i(t_{i2}), ..., Y_i(t_{in_i}))$ le vecteur de taille n_i d'observations mesurées aux temps $t_{i1}, \ldots, t_{in_i}$ pour le sujet i. La formulation générale du modèle linéaire mixte est la suivante :

$$Y_i(t_{ij}) = X_i(t_{ij})^T \beta + Z_i(t_{ij})^T b_i + \epsilon_i(t_{ij})$$

où Y_i est le vecteur de la variable à expliquer pour l'individu i ($i = 1, \ldots, N$) aux différents temps t_{ij} ($j = 1, \ldots, n_i$). $X_i(t_{ij})$ est le vecteur de variables explicatives de dimension p associé au p-vecteur des effets fixes β. $Z_i(t_{ij})$ est un sous-vecteur de $X_i(t_{ij})$, de dimension q ($q < p$) et associé au q-vecteur des effets aléatoires b_i. Les effets aléatoires b_i sont indépendants et identiquement distribués selon une loi normale de moyenne 0 et de matrice de variance-covariance B. $\epsilon_i = (\epsilon_i(t_{i1}), \epsilon_i(t_{i2}), ..., \epsilon_i(t_{in_i}))$ est le vecteur des erreurs de mesures qui sont indépendantes entre elles et indépendantes des effets aléatoires b_i. Les erreurs de mesures sont distribuées selon une loi normale de moyenne 0 et de matrice de variance-covariance $\sigma^2_\epsilon I_{n_i}$, de dimension n_i x n_i. Bien que les erreurs de mesures soient supposées indépendantes, il est également possible de définir l'erreur comme la somme d'une composante autorégressive et d'une erreur de mesure. Dans ce cas, les effets aléatoires représentent une variation individuelle à long terme, l'erreur autorégressive représente une variation individuelle à court

terme, à laquelle on rajoute l'erreur résiduelle souvent liée aux instruments de mesure.

$X_i(t_{ij})^T\beta + Z_i(t_{ij})^T b_i$ représente la vrai valeur du paramètre étudié au cours du temps pour le sujet i, avec $X_i(t_{ij})^T\beta$ l'évolution moyenne du paramètre au cours du temps et $Z_i(t_{ij})^T b_i$ les déviations individuelles du sujet i par rapport à la moyenne.

2.2.2 Estimation des paramètres du modèle

Les paramètres du modèle linéaire mixte sont généralement estimés par la méthode du maximum de vraisemblance, en utilisant la formulation marginale du modèle, à travers des algorithmes itératifs tels que l'algorithme de Newton Raphson. La formulation marginale du modèle s'écrit :

$$Y_i = X_i\beta + e_i$$
$$\text{avec } e_i = Z_i b_i + \epsilon_i \text{ et } e_i \sim N\left(0, V_i = \sigma_\epsilon^2 I_{n_i} + Z_i B Z_i^T\right)$$

soit θ le vecteur des paramètres de variance-covariance de V_i (variance-covariance des effets aléatoires et des erreurs de mesure). La vraisemblance s'écrit comme le produit des densités marginales de la manière suivante :

$$V(\beta,\theta) = \prod_{i=1}^{N} f(Y_i)$$
$$V(\beta,\theta) = \prod_{i=1}^{N} (2\pi)^{-n_i/2}|V_i|^{-1/2} exp\left[-\frac{1}{2}\left(Y_i - X_i\beta\right)^T V_i^{-1}\left(Y_i - X_i\beta\right)\right]$$

Plus généralement, c'est la log-vraisemblance qui est maximisée, elle s'écrit :

$$L(\beta,\theta) = -\tfrac{1}{2}\sum_{i=1}^{N} n_i ln(2\pi) + ln|V_i| + \left(Y_i - X_i\beta\right)^T V_i^{-1}\left(Y_i - X_i\beta\right)$$

Avec $|V_i|$ le déterminant de V_i.

Les paramètres qui maximisent la log-vraisemblance sont les solutions de l'équation du score suivante :

$$\frac{\partial L(\beta,\theta)}{\partial\beta} = \sum_{i=1}^{N} X_i^T V_i^{-1}\left(Y_i - X_i\beta\right) = 0$$

Lorsque les paramètres de variance θ sont connus, l'estimateur des paramètres de régression est donné par :

$$\widehat{\beta} = \left(\sum_{i=1}^{N} X_i^T V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^T V_i^{-1} Y_i\right)$$

Si les paramètres de variance θ sont inconnus, on remplace β par $\widehat{\beta}$ dans $L(\beta, \theta)$ et on maximise la vraisemblance obtenue sur θ.

Si l'intérêt porte sur les paramètres de variance, il est possible de maximiser la vraisemblance restreinte pour estimer les paramètres θ [Harville 1977]. Notamment lorsque les échantillons sont de petites tailles ou le nombre d'effets fixes est élevé. En effet, dans les échantillons de petites tailles, l'estimation du maximum de vraisemblance de θ peut être biaisé car ne prend pas en compte la perte de degrés de liberté induite par l'estimation des effets fixes. Toute fois les deux approches donnent des estimations proches dans le cas d'échantillons de grande taille.

2.2.3 Prédictions et résidus

En plus de l'estimation des paramètres du modèle marginal, il est possible d'estimer les effets aléatoires utiles pour les prédictions individuelles. Les effets aléatoires peuvent être estimés par des estimateurs bayésiens empiriques (Best Linear Unbiased Predictor) qui sont donnés par :

$$\begin{aligned} \widehat{b_i} = E(b_i|Y_i) &= E(b_i) + cov(b_i, Y_i) var(Y_i)^{-1} \left[Y_i - E(Y_i)\right] \\ &= B Z_i^T V_i^{-1} (Y_i - X_i \beta) \end{aligned}$$

On peut obtenir des prédictions marginales qui représentent l'espérance de Y_i pour les sujets ayant les mêmes caractéristiques X_i que le sujet i : $\widehat{E}(Y_i) = X_i \widehat{\beta}$. On peut également obtenir les prédictions individuelles sachant les effets aléatoires qui représentent l'espérance spécifique au sujet i et prenant en compte les variations individuelles à long terme : $\widehat{Y}_i = \widehat{E}(Y_i|b_i) = X_i \widehat{\beta} + Z_i \widehat{b}_i$.

De même, on peut définir à partir des prédictions, les résidus marginaux $Y_i - X_i \widehat{\beta}$ et les résidus spécifiques aux sujets $Y_i - X_i \widehat{\beta} - Z_i \widehat{b}_i$.

2.2.4 Évaluation du modèle

L'évaluation des hypothèses du modèle linéaire mixte peut se faire par différentes méthodes. Les principales hypothèses du modèle linéaire mixte sont la normalité de la distribution des effets aléatoires, la normalité et l'homoscédasticité des erreurs de mesure. L'évaluation de la normalité des résidus peut se faire par le graphe

Quantile-Quantile des résidus de Cholesky. Il est possible de repérer des observations extrêmes ou d'évaluer la bonne spécification d'une variable explicative en représentant le graphe des résidus divisés par l'écart-type en fonction du temps ou en fonction de la variable considérée. Le graphe des valeurs prédites marginales ou conditionnelles en fonction des valeurs observées est également utile pour l'évaluation du modèle linéaire mixte.

2.2.5 Classification des données manquantes

Il est fréquent d'être confronté au problème de données manquantes dans l'analyse des données longitudinales. Les données peuvent être manquantes par intermittence, lorsque les données sont manquantes à une visite mais non manquantes à la visite suivante. Les données peuvent également être manquantes de façon monotone c'est-à-dire de façon définitive. Ceci correspond en général aux sorties d'étude, aux perdus de vue. Une classification du mécanisme des données manquantes a été proposée par Little and Rubin [1987, 2002].

Soit $Y = (Y^o, Y^m)$ le vecteur des réponses avec Y^o les réponses observées et Y^m les réponses non observées. Soit R une variable indicatrice de l'observation pour laquelle $R(t) = 1$ si $Y(t)$ est observé et $R(t) = 0$ si $Y(t)$ n'est pas observé. Les données manquantes sont monotones lorsque $P(R(t) = 0|R(s) = 0, s < t) = 1$, dans le cas contraire, elles sont intermittentes. La classification des données manquantes est faite en 3 classes :

— Dans le cas où R et Y sont indépendants, les données sont dites manquantes complètement aléatoirement (Missing Completely At Random, MCAR). Dans ce cas, la probabilité des réponses dépend des variables explicatives observées.
$P(R|Y^o, Y^m, X) = P(R|X)$

— Dans le cas où R dépend des réponses observées Y^o et pas des réponses manquantes Y^m, les données sont dites manquantes aléatoirement (Missing At Randon, MAR).
$P(R|Y^o, Y^m, X) = P(R|Y^o, X)$

Lorsque les paramètres θ du modèle pour l'espérance de Y sont distincts des paramètres ω du processus d'observation de R et que les données sont manquantes aléatoirement, alors les données manquantes sont dites ignorables. Dans ce cas, les estimateurs issus de la méthode du maximum de vraisemblance ne sont pas biaisés. θ peut alors être estimé sans biais par maximisation de la

log-vraisemblance $L(\theta|Y^o)$ calculée sur les réponses observées. La vraisemblance des données observées est définie comme suit :

$f(Y^o, R|\theta, \omega) = \int f(Y^o, Y^m|\theta) f(R|Y^o, Y^m, \omega) dY^m$

Or, si les données sont manquantes aléatoirement, on a :

$$\begin{aligned} f(Y^o, R|\theta, \omega) &= f(R|Y^o, \omega) \int f(Y^o, Y^m|\theta) dY^m \\ &= f(Y^o|\theta) f(R|Y^o, \omega) \end{aligned}$$

Et la log-vraisemblance conjointe vaut : $L(\theta, \omega|Y^o, R) = L(\theta|Y^o) + L(\omega|R, Y^o)$

— Dans le cas où R dépend des réponses non observées Y^m, les données sont dites manquantes non aléatoirement (Missing Not At Random) ou informatives. $P(R|Y^o, Y^m, X) = P(R|Y^m, X)$

Dans le cas des données manquantes non aléatoires, il est nécessaire de maximiser la log-vraisemblance conjointe $L(Y^o, R)$ sur les données observées sur les processus Y et R. Pour cela, des méthodes telles que les modèles de mélange (Pattern Mixture Models) [Michiels et al. 2002; Little 1993] et les modèles de sélection [Diggle and Kenward 1994] ont été proposés. Un exemple de modèle de sélection lorsque les données manquantes sont monotones est le modèle conjoint pour données longitudinales et délai de survie.

Les modèles linéaires mixtes permettent de décrire l'évolution d'un paramètre gaussien en fonction du temps. Cependant, il existe des données répétées dans le temps dont la distribution est éloignée de la distribution normale supposée dans le modèle linéaire mixte. Des extensions du modèle mixte, notamment les modèles linéaires généralisés mixtes ont été proposées pour étudier des variables dépendantes binaires, ordinales ou de comptage. Des modèles mixtes à classes latentes ont également été proposés pour tenir compte de l'hétérogénéité souvent existante dans les populations étudiées [Verbeke and Lesaffre 1996].

2.3 Analyse des données de survie

Dans l'analyse des données de survie, la variable aléatoire étudiée est le délai de survenue d'un événement. Si l'événement considéré est le décès par exemple, on peut parler de durée de survie. On s'intéresse donc au temps écoulé entre une date d'origine et la date de survenue de l'événement considéré.

Le choix de la date d'origine dépend de l'événement étudié et de l'objectif visé par l'étude. Il peut être propre au sujet, la naissance par exemple si le risque de survenue de l'événement dépend de l'âge, dans ce cas le délai considéré est l'âge. La date d'origine peut être une date de mise sous traitement si on étudie le délai jusqu'à l'apparition d'un symptôme, ou une date de transplantation si on étudie le délai jusqu'au décès après greffe ou le délai jusqu'à l'échec de la greffe. La date d'origine peut également être commune à tous les sujets si on s'intéresse à un phénomène ayant affecté plusieurs individus à la fois (guerre, tremblement de terre...).

2.3.1 Définitions

Notons T la variable aléatoire positive et continue qui représente le délai entre la date d'origine et la date de survenue de l'événement considéré. La loi de probabilité de T est définie par plusieurs fonctions.

— La fonction de densité de probabilité qui donne la probabilité que l'événement survienne entre le temps t et le temps $t + \Delta_t$

$$f(t) = \lim_{\Delta_t \to 0^+} \frac{P(t \leq T < t + \Delta_t)}{\Delta_t}$$

— La fonction de répartition qui donne la probabilité que l'événement survienne entre 0 et t :

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

— La fonction de survie qui donne la probabilité de n'avoir pas subi l'événement jusqu'au temps t :

$$S(t) = P(T > t) = 1 - F(t)$$

— La fonction de risque instantané qui donne la probabilité que l'événement survienne entre le temps t et le temps $t + \Delta_t$ sachant que le sujet était encore

à risque au temps t :

$$\lambda(t) = \lim_{\Delta_t \to 0^+} \frac{P(t \leq T < t+\Delta_t | T \geq t)}{\Delta_t} = \frac{f(t)}{S(t)}$$

— La fonction de risque cumulée qui est l'intégrale de la fonction de risque instantané :

$$\Lambda(t) = \int_o^t \lambda(u)du = -ln(S(t))$$

En analyse des données de survie, tous les événements ne sont pas toujours observés. Ces observations incomplètes sont des conséquences de la censure et de la troncature.

— On parle de censure à droite si le sujet n'a pas subi l'événement à sa dernière visite. Ce cas intervient par exemple lorsque le sujet est perdu de vue ou tout simplement s'il ne subira pas l'événement jusqu'à la fin de son suivi. Pour une durée de survie censurée à droite, on définit une variable aléatoire de la censure à droite C et une variable indicatrice de l'événement δ.
Chaque sujet est défini par le couple $(\tilde{T}, \delta)$ avec $\tilde{T} = min(T, C)$ et $\delta = 0$ si T>C et $\delta = 1$ sinon.

— On parle de censure à gauche si le sujet a déjà subi l'événement avant le début de son suivi, sans que la date de survenue de l'événement ne soit connue avec exactitude. Pour éviter le problème de censure à gauche, en général on ne considère que les sujets n'ayant pas encore subi l'événement dans l'étude. Dans ce cas, on a $\tilde{T} = max(T, C)$ et $\delta = 0$ si T<C et $\delta = 1$ sinon.

— On parle de censure par intervalle lorsque le sujet subit l'événement entre deux dates connues, mais le temps exact de survenue de l'événement n'est pas connu. Ce cas intervient par exemple lorsque l'événement est connu suite à un diagnostic. Le sujet est vu à une visite et a un résultat négatif vis-à-vis de l'événement suite au diagnostic. Cependant, à la visite suivante le résultat vis-à-vis de l'événement est positif suite au diagnostic.

— En analyse des données de survie, en plus de la censure, la troncature engendre également des données incomplètes. On parle de troncature à gauche si le sujet n'est observable que si sa durée de survie est supérieure à une certaine valeur. Soit T_0 une variable indépendante de T. T est tronquée à gauche si elle n'est

observable qu'à condition que $T > T_0$. T est tronquée à droite si elle n'est observable qu'à condition que $T < T_0$. T est tronquée par intervalles si elle est à la fois tronquée à gauche et tronquée à droite.

Dans notre étude, nous nous intéressons à la survenue du décès ou de la transplantation pulmonaire chez les sujets âgés de 18 ans ou plus. Les données de l'étude ne concernent que les sujets vivants et non transplantés à l'inclusion. Ces deux événements étant indiqués par des dates connues, les problèmes de censure à gauche et par intervalles ne se posent pas.

2.3.2 Modèles de régression pour l'analyse des données de survie

L'analyse des données de survie se fait par des modèles de régression qui prennent en compte des variables explicatives pouvant avoir un impact sur la durée de survie. Les principales approches utilisées pour cela sont les approches paramétriques et semi-paramétriques.

L'approche paramétrique fait l'hypothèse que la distribution des durées de survie appartient à une famille de lois paramétriques. Parmi les modèles avec une approche paramétrique il y a le modèle exponentiel qui suppose que la fonction de risque instantané est constante au cours du temps. sa densité de probabilité s'écrit $f(t, \gamma) = \gamma e^{-\gamma t}$ et sa fonction de répartition vaut $F(t, \gamma) = 1 - e^{-\gamma t}$.

Comme autre modèle paramétrique, on note le modèle de Weibull qui dépend de deux paramètres positifs γ et ρ. Sa densité de probabilité s'écrit $f(t, \gamma, \rho) = \rho\gamma^{\rho}t^{\rho-1}e^{-(\gamma t)^{\rho}}$ et sa fonction de répartition vaut $F(t, \gamma, \rho) = 1 - e^{-(\gamma t)^{\rho}}$. Pour ces modèles, si $\rho > 1$ alors la fonction de risque est croissante, si $0 < \rho < 1$ alors la fonction de risque est décroissante et si $\rho = 1$ on retrouve la loi exponentielle.

La distribution de Weibull et la distribution exponentielle sont les plus utilisées en analyse de données de survie. Néanmoins il existe d'autres approches paramétriques telles que Gompertz, Gamma. Dans l'approche paramétrique, les paramètres du modèle sont estimés en maximisant la vraisemblance des observations par des méthodes itératives telles que l'algorithme de Newton-Raphson. Dans le cas des données censurées à droite, la vraisemblance de n sujets s'écrit :

$$V = \prod_{i=1}^{n} f(\tilde{T}_i)^{\delta_i} S(\tilde{T}_i)^{1-\delta_i} = \prod_{i=1}^{n} \lambda(\tilde{T}_i)_i^{\delta} S(\tilde{T}_i)$$

Pour l'analyse des données de survie, le modèle à risques proportionnels proposé par Cox [1972] est très largement utilisé. Il s'agit d'un modèle avec une approche

semi-paramétrique qui ne fait pas d'hypothèse sur la fonction de risque. L'estimation des paramètres qui sont les coefficients associés aux variables explicatives du modèle se fait par maximisation de la vraisemblance partielle. Soit $Z = (Z_1, Z_2, ..., Z_p)^T$ un vecteur de variables explicatives de taille p et $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ un vecteur de coefficients de régression. La fonction de risque associée à la survenue d'un événement s'écrit :

$$\alpha(t, Z, \beta) = \alpha_0(t) exp(\beta^T Z)$$

où $\alpha_0(t)$ est la fonction de risque de base qui est la fonction de risque pour les sujets dont toutes les variables explicatives sont nulles, si cela a un sens.

La vraisemblance partielle ne dépend pas de la fonction de risque de base mais uniquement des paramètres qui sont les coefficients de régression. Elle correspond au produit des probabilités conditionnelles que le sujet i subisse l'événement au temps t_i sachant qu'il est à risque à ce temps et qu'il n'y a qu'un seul événement à ce temps. Elle s'écrit :

$$VP(\beta, Z) = \prod_{i=1}^{k} \frac{exp(\beta^T Z_i)}{\sum\limits_{l:\tilde{T}_l \geq t_i} exp(\beta^T Z_{(l)})}$$

La vraisemblance partielle est maximisée par une méthode itérative telle que l'algorithme de Newton-Raphson. Cette écriture de la vraisemblance nécessite qu'il n'y ait qu'un seul événement par date d'événement. Cependant, des extensions de la vraisemblance partielle ont été proposées dans le cas où plusieurs sujets auraient un événement au même temps [Breslow 1974; Efron 1977].

Cependant, ce modèle n'est pas adéquat lorsque les temps d'événement sont censurés par intervalles.

2.3.3 Variables dépendantes du temps

Dans le modèle de Cox, il est possible de prendre en compte, en plus des variables définies au moment de l'inclusion, des variables dépendantes du temps et donc recueillies au cours du suivi. On distingue deux types de variables dépendantes du temps : les variables exogènes et les variables endogènes. Les variables exogènes sont celles dont le recueil ne dépend pas de la survenue de l'événement (la température extérieure, les données de pollution par exemple). Les variables endogènes sont celles dont les valeurs sont associées à une modification du risque de survenue de l'événement. Elles sont relevées chez le sujet à condition qu'il soit encore en vie. Ce sont des

variables mesurées avec erreur. Lors de la mesure d'une variable endogène, si celle-ci est faite à partir d'un appareil il y a une erreur de mesure liée à l'appareil. De plus, il y a une erreur de mesure liée au sujet. En effet, deux mesures successives de la variable endogène faite sur un sujet ne seront pas les mêmes du fait des variations biologiques du sujet. Dans le cas de la mucoviscidose, le VEMS est une variable endogène mesurée régulièrement chez les sujets pour évaluer leur fonction pulmonaire grâce à la spirométrie.

La prise en compte des variables dépendantes du temps est très délicate notamment en ce qui concerne les variables endogènes. En effet, l'introduction des variables endogènes dans le modèle de Cox modifie l'interprétation des fonctions de survie et de densité. De ce fait, la vraisemblance n'est plus adaptée aux données. Plus particulièrement, le calcul de la vraisemblance partielle nécessite la connaissance des valeurs des variables explicatives des sujets à risque à tous les temps d'événement. Dans le cas des variables endogènes, les valeurs sont recueillies lors des visites et supposées constantes entre les visites. Seulement, en pratique, cette hypothèse parait peu raisonnable. De plus, les variables endogènes sont mesurées avec des erreurs qui sont typiquement des erreurs liées au sujet et au mécanisme de recueil de données. Les données recueillies chez le sujet le sont généralement à des temps précis, lors des visites par exemple. Elles ne sont pas relevées de manière continue dans le temps. Pour ces raisons, il n'est pas approprié d'inclure des variables endogènes dépendantes du temps dans un modèle de Cox.

2.3.4 Adéquation du modèle

L'adéquation du modèle de Cox se fait par l'évaluation de la log-linéarité et l'hypothèse de proportionnalité des risques.

Il est possible de vérifier la proportionnalité des risques par des méthodes graphiques. Par exemple en traçant les courbes $ln(-ln(\widehat{S}(t)))$ pour chaque catégorie de la variable considérée et en vérifiant que l'écart entre les courbes, correspondant à $\widehat{\beta}$ reste constant. Un test d'interaction entre une explicative et une fonction de temps, ainsi que les résidus de Schoenfeld, les résidus de martingale [Lin et al. 1993] et les résidus de Cox-snell [Kalbfleisch and Prentice 2002] sont également utiles pour évaluer l'hypothèse de proportionnalité des risques.

Cependant, pour certaines variables cette hypothèse de proportionnalité des risques n'est pas vérifiée. Dans ce cas, il est possible de faire une stratification sur

ces variables. Ainsi, la fonction de risque sera supposée différente pour les catégories de la variable considérée, mais l'effet des variables explicatives restera le même pour toutes ces catégories. La fonction de risque de base sera différente d'une catégorie à l'autre, mais au sein de chaque catégorie de la variable de stratification, tous les sujets auront la même fonction de risque de base. Il est également possible de faire des stratifications sur plusieurs variables. La vraisemblance partielle reste un produit de vraisemblances partielles sur toutes les catégories.

L'hypothèse de log-linéarité pour une variable quantitative peut être vérifiée en la transformant en classes et en vérifiant que les risques relatifs obtenus avec la variable quantitative sont contenus dans les intervalles de confiances des risques relatifs obtenus avec la variable en classes.

2.3.5 Modèles à risques compétitifs

En analyse des données de survie, on peut être intéressé par la survenue de plusieurs événements plutôt que d'un seul événement [Putter et al. 2007]. Dans le cas où la survenue d'un événement empêche la survenue des autres événements, on parle de risques compétitifs. Dans ce contexte, on est en présence de K différents événements, D représente le type d'événement (D vaut 0 si le sujet est censuré) et on s'intéresse au temps T de survenue du premier événement. Tout comme dans le contexte où on se limite à la survenue d'un seul événement, il est possible de définir des fonctions de risque en présence de plusieurs événements. On parle de fonctions de risque instantané cause-spécifique. Pour l'événement k, elle est définie de la manière suivante :

$$\alpha_k(t) = \lim_{\Delta_t \to 0^+} \frac{P(t \leq T \leq t+\Delta_t, D=k | T \geq t)}{\Delta_t}$$

Il s'agit de la fonction de risque instantané au temps t pour l'événement k en présence des autres événements. Elle représente la probabilité que l'événement k survienne entre le temps t et le temps $t + \Delta_t$ chez un sujet, sachant que le sujet est encore à risque au temps t.

Il est également possible de définir la fonction de risque globale. Similaire à la fonction de risque instantané dans le cas d'un seul événement, elle ne tient pas compte du type d'événement associé au délai. Elle représente la probabilité qu'un événement survienne entre le temps t et le temps $t + \Delta_t$ chez un sujet, sachant que le sujet est encore à risque au temps t.

$$\alpha(t) = \lim_{\Delta_t \to 0^+} \frac{P(t \leq T \leq t+\Delta_t | T \geq t)}{\Delta_t} \text{ avec } \alpha(t) = \sum_{k=1}^{K} \alpha_k(t)$$

A partir de cette fonction, on peut déterminer la fonction de survie sans événement qui est donnée par :

$$S(t) = exp\left(-\sum_{k=1}^{K}\left(\int_0^t \alpha_k(s)ds\right)\right)$$

Il est également possible de définir les fonctions d'incidences cumulées qui représentent la probabilité que l'événement k survienne avant le temps t en présence des autres événements. La fonction d'incidence cumulée représente une alternative à l'estimateur de Kaplan-Meier qui est biaisé dans le contexte des risques compétitifs. La fonction d'incidence cumulée s'écrit :

$$\begin{aligned} I_k(t) = P(T \leq t, D = k) &= \int_0^t \alpha_k(u)S(u)du \\ &= \int_0^t \alpha_k(u)exp\left(-\int_0^u \sum_{k=1}^{K} \alpha_k(s)ds\right)du \end{aligned}$$

Il est possible d'utiliser des modèles à risques proportionnels basés sur les risques cause-spécifiques pour étudier l'effet de variables explicatives sur la survenue d'un événement. Le modèle s'écrit de la manière suivante :

$$\alpha_k(t|X) = \alpha_{0k}(t)exp(\beta_k X^T)$$

avec k l'événement d'intérêt, α_{0k} le risque cause-spécifique de base, $\beta = (\beta_{1k}, \beta_{2k}, ..., \beta_{pk})^T$ le vecteur de covariables. Dans ce modèle, tous les événements autres que k sont considérés comme des censures.

Plus particulièrement on peut écrire l'adaptation du modèle de Cox en présence de plusieurs événements :

$$\begin{aligned} &\alpha_k(t|X) = \alpha_1(t)exp(b_k + \beta_k X^T + \gamma X^T) \\ &\text{avec } \gamma = (\gamma_1, \gamma_2, ..., \gamma_p)^T \text{ et } b_1 = \beta_1 = 0 \end{aligned}$$

$\alpha_1(t)$ est le risque de base pour les sujets ayant subi l'événement $k = 1$ et pour lesquels les covariables sont nulles, si cela a un sens. La fonction de risque de base pour tout événement autre que k est donnée par $\alpha_k(t) = \alpha_1(t)exp(b_k)$.

$exp(\gamma_1)$ représente le taux relatif d'accroissement d'une unité de la variable explicative X_1, si elle est quantitative pour l'événement $k = 1$.

$exp(\gamma_1 + \beta_{1k})$ représente le taux relatif d'accroissement d'une unité de X_1, si elle est quantitative pour tout événement autre que $k = 1$. Il est donc possible de tester si l'effet de la variable X_1 est identique pour les événements 1 et k avec $H_0 : \beta_{1k} = 0$.

Il est par ailleurs possible de faire un modèle stratifié sur le type événement où le risque de base α_{0k} est spécifique à l'événement k.

$$\alpha_k(t|X) = \alpha_{0k}(t)exp(\beta_k X^T + \gamma X^T)$$

2.4 Modèles conjoints pour données longitudinales et temps d'événements

L'analyse des données longitudinales et l'analyse des temps d'événements ont été décrites précédemment et séparément. Dans chacune de ces analyses, au moins un problème a été soulevé. Nous avons vu dans l'analyse des données longitudinales que les estimations par maximum de vraisemblance étaient biaisées en présence de données manquantes non aléatoires. Elles sont non biaisées uniquement en présence de données manquantes aléatoirement. Dans ce cas, on suppose que l'évolution du marqueur étudié n'est pas modifiée par la survenue de l'événement, après ajustement sur les données précédentes. Cependant, dans des contextes comme celui de la mucoviscidose, la transplantation pulmonaire entraîne une augmentation du VEMS et donc, change complètement la dynamique du VEMS. Il est possible dans ce cas de ne considérer que les mesures prises avant la transplantation pulmonaire pour étudier l'évolution du VEMS. De ce fait, les données prises après la transplantation pulmonaire sont considérées comme des données manquantes non aléatoires.

Concernant l'analyse des données de survie, nous avons vu qu'il n'est pas approprié d'inclure des variables endogènes dépendantes du temps, mesurées avec erreur dans un modèle de Cox. En effet, la vraisemblance partielle nécessite la connaissance des valeurs des variables explicatives des sujets à risque à tous les temps d'événement. L'estimation du paramètre associé à la variable endogène dépendante du temps dans un modèle de Cox sera biaisée si on suppose que les valeurs de la variable restent constantes entre deux mesures, notamment lorsque les intervalles entre les mesures sont grands [Prentice 1982]. Les modèles conjoints pour données longitudinales et temps d'événements sont une alternative pour remédier à ces problèmes.

Les modèles conjoints ont été développés dans les années 90 [Faucett and Thomas 1996; Wulfsohn and Tsiatis 1997; Henderson et al. 2000]. Ils permettent d'étudier simultanément le risque de survenue d'un événement et l'évolution d'un marqueur longitudinal. Ils combinent donc deux sous-modèles à savoir un modèle mixte pour l'évolution du marqueur et un modèle de survie pour les temps d'événements, qui sont liés à travers une structure latente. Les modèles conjoints sont particulièrement utiles pour régler les problèmes cités précédemment, mais aussi lorsqu'on souhaite étudier le lien existant entre l'évolution d'un marqueur longitudinal et la survenue d'un événement. Ils sont également utiles lorsqu'on souhaite prédire de façon dynamique le risque de survenue d'un événement à partir de l'évolution d'un marqueur longitudinal. Il existe deux principaux types de modèles conjoints : le modèle conjoint à effets aléatoires partagés et le modèle conjoint à classes latentes.

2.4.1 Modèles conjoints à effets aléatoires partagés

Spécification

Soit $Y(t)$ un marqueur longitudinal Gaussien défini au temps t. $Y_i(t_{ij})$ représente la variable Y mesurée au temps t_{ij} pour le sujet $i(i = 1, \ldots, N)$. On note T_i le temps de survenue de l'événement d'intérêt et C_i le temps de censure à droite pour le sujet i. Comme précédemment, on définit le couple $(\tilde{T}_i, \delta_i)$ avec $\tilde{T}_i = min(T_i, C_i)$ et $\delta_i = 0$ si $T_i > C_i$ et $\delta_i = 1$ sinon. Le modèle conjoint à effets aléatoires partagés est une combinaison d'un modèle linéaire mixte qui décrit l'évolution du marqueur longitudinal $Y(t)$ et d'un modèle de survie, généralement un modèle à risques proportionnels qui décrit le risque d'événement. Il est défini comme suit :

$$\begin{aligned} Y_i(t_{ij}) &= X_{1i}(t_{ij})^T\beta + Z_i(t_{ij})^T b_i + \epsilon_i(t_{ij}) \\ &= \tilde{Y}_i(t_{ij}) + \epsilon_i(t_{ij}) \\ \alpha_i(t) &= \alpha_0(t) exp(X_{2i}^T\gamma + g_i(b_i, t)^T\eta) \end{aligned}$$

où Y_i est le vecteur de la variable à expliquer pour l'individu $i(i = 1, \ldots, N)$ aux différents temps t_{ij} $(j = 1, \ldots, n_i)$. $X_{1i}(t_{ij})$ est le vecteur de variables explicatives dépendantes du temps, associé au vecteur des effets fixes β. $Z_i(t_{ij})$ est un sous-vecteur de $X_{1i}(t_{ij})$, de variables également dépendantes du temps, associé au vecteur des effets aléatoires b_i. Les effets aléatoires b_i sont indépendants et identiquement distribués selon une loi normale de moyenne 0 et de matrice de variance-covariance B. $\epsilon_i = (\epsilon_i(t_{i1}), \epsilon_i(t_{i2}), \ldots, \epsilon_i(t_{in_i}))$ est le vecteur des erreurs de mesures qui sont

indépendantes entre elles et indépendantes des effets aléatoires b_i. Les erreurs de mesures sont distribuées selon une loi normale de moyenne 0 et de matrice de variance-covariance $\sigma_\epsilon^2 I_{n_i}$, de dimension n_i x n_i.

$X_{1i}(t_{ij})^T\beta$ représente l'évolution moyenne du paramètre au cours du temps et $Z_i(t_{ij})^T b_i$ les déviations individuelles du sujet i par rapport à la moyenne. $\tilde{Y}_i(t_{ij}) = E\left[Y_i(t_{ij}|b_i\right] = X_{1i}(t_{ij})^T\beta + Z_i(t_{ij})^T b_i$ est l'espérance conditionnelle qui représente la vrai valeur du paramètre étudié au cours du temps pour le sujet i, sans les erreurs.

X_{2i} est le vecteur de variables explicatives associé au vecteur des paramètres γ qui mesure l'effet des variables sur le risque de survenue de l'événement. $\alpha_0(t)$ est la fonction de risque de base qui peut être paramétrique et modélisée par des distributions telles que Weibull, Gamma. Des distributions plus souples telles que des fonctions splines [Rosenberg 1995] ou des fonctions constantes par morceaux peuvent également être utilisées.

Le sous-modèle de survie et le sous-modèle mixte sont liés par la fonction $g_i(b_i, t)$ et η est le paramètre qui quantifie cette association. Comme dans un modèle de Cox classique, η s'interprète en terme de rapport de risques. La fonction $g_i(b_i, t)$ peut être spécifiée de manières différentes selon l'objectif de l'étude.

— $g_i(b_i, t) = \tilde{Y}_i(t) = E\left[Y_i(t)|b_i\right] = X_{1i}(t)^T\beta + Z_i(t)^T b_i$
Cette spécification suppose que le risque de survenue de l'événement au temps t dépend de la valeur courante du marqueur longitudinal au même temps t, sans prise en compte de l'erreur de mesure. Dans ce cas, pour une augmentation d'une unité du niveau courant de $\tilde{Y}_i(t)$ au temps t, η quantifie l'augmentation ou la diminution du risque relatif de la survenue de l'événement au même temps t.

— $g_i(b_i, t) = \tilde{Y}_i(max(t - c, 0))$
Cette spécification suppose que le risque de survenue de l'événement au temps t dépend de la valeur courante du marqueur longitudinal au temps $t - c$, c étant un temps de décalage. Dans ce cas, pour une augmentation d'une unité du niveau courant de $\tilde{Y}_i(t)$ au temps $t - c$, η quantifie l'augmentation ou la diminution du risque relatif de la survenue de l'événement au temps t.

— $g_i(b_i, t) = \tilde{Y}_i(t)'$
Cette spécification suppose que le risque de survenue de l'événement au temps

t dépend de la pente courante du marqueur longitudinal au même temps t. Dans ce cas, pour une augmentation d'une unité de la pente du marqueur longitudinal au temps t, η quantifie l'augmentation ou la diminution du risque relatif de la survenue de l'événement au même temps t.

— $g_i(b_i, t) = \tilde{Y}_i(t) + \tilde{Y}_i(t)'$
Cette spécification suppose que le risque de survenue de l'événement au temps t dépend à la fois du niveau courant et de la pente courante du marqueur longitudinal au même temps t [Ye et al. 2008]. Dans ce cas, η_1 est le paramètre associé au niveau courant $\tilde{Y}_i(t)$ et η_2 est le paramètre associé à la pente courante $\tilde{Y}_i(t)'$. Ces deux paramètres s'interprètent comme précédemment, toutes choses égales par ailleurs. Notons que dans cette spécification, il est également possible d'introduire un temps de décalage c.

— $g_i(b_i, t) = \int_0^t \tilde{Y}_i(s)ds$
Cette spécification suppose que le risque de survenue de l'événement au temps t dépend de toute la trajectoire du marqueur longitudinal qui est représentée par l'aire sous la trajectoire du marqueur longitudinal jusqu'au temps t. Cette spécification suppose des poids égaux pour les valeurs du marqueur longitudinal. Il est néanmoins possible de rajouter dans le calcul de l'intégrale une fonction de poids qui donnera des poids plus faibles aux valeurs les plus anciennes. Dans ce cas, pour une augmentation d'une unité de l'aire sous la trajectoire du marqueur, η quantifie l'augmentation ou la diminution du risque relatif de la survenue de l'événement au temps t.

— $g_i(b_i, t) = b_i$
Cette spécification suppose que le risque de survenue de l'événement au temps t dépend des déviations individuelles sur le niveau initial et sur la pente du marqueur longitudinal. Dans ce cas, les sujets ayant des valeurs faibles/élevées du niveau initial du marqueur ou les sujets ayant une importante diminution/augmentation de la pente du marqueur seront plus/moins vulnérables à la survenue de l'événement.

Estimation

L'estimation des paramètres du modèle conjoint se fait par maximum de vraisemblance conjointe. Soit $\theta = (\theta_e, \theta_y, \theta_b)$ le vecteur contenant tous les paramètres du modèle conjoint avec θ_e le vecteur des paramètres du sous-modèle de survie, θ_y le

vecteur des paramètres du sous-modèle longitudinal et θ_b le vecteur des paramètres de la matrice de variance-covariance des effets aléatoires. À partir de l'hypothèse d'indépendance conditionnelle entre le marqueur longitudinal Y et le temps d'événement T, la log-vraisemblance s'écrit :

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{N} lnV(Y_i, T_i, \delta_i) \\ &= \sum_{i=1}^{N} ln \int V(Y_i, T_i, \delta_i | b_i) f_{b_i}(b_i) db_i \\ &= \sum_{i=1}^{N} ln \int f_{Y_i}(Y_i | b_i, \theta_y) S_i(T_i | b_i, \theta_e) \alpha_i(T_i | b_i, \theta_e)^{\delta_i} f_{b_i}(b_i, \theta_b) db_i \end{aligned}$$

La maximisation de la log-vraisemblance peut se faire par les algorithmes standards tels que l'algorithme EM (Expectation-Maximization) [Dempster et al. 1977], l'algorithme de Newton-Raphson ou Quasi-Newton. Dans le cas de l'algorithme EM, les effets aléatoires sont considérés comme des données manquantes. Un algorithme qui combine l'algorithme EM et l'algorithme Quasi-Newton est disponible dans le package *JM* du logiciel R. la maximisation de la log-vraisemblance se fait au départ par l'algorithme EM pour un nombre d'itérations fixé et si la convergence n'est pas atteinte, l'algorithme Quasi-Newton prend le relai.

Les intégrales sur les effets aléatoires de la log-vraisemblance n'ont généralement pas de solution analytique. Le calcul de ces intégrales se fait par des approches numériques telles que la quadrature de Gauss ou par la méthode MCMC (Markov Chain Monte Carlo) [Wulfsohn and Tsiatis 1997; Henderson et al. 2000; Song et al. 2002]. Il est également possible d'utiliser une approximation de Laplace lorsque le modèle inclut beaucoup d'effets aléatoires [Ye et al. 2008; Rizopoulos et al. 2009].

L'estimation des paramètres du modèle conjoint peut également être faite par une approche Bayésienne [Brown and Ibrahim 2003; Guo and Carlin 2004; Rizopoulos 2016].

Évaluation

Il est nécessaire d'évaluer l'ajustement de chacun des sous-modèles. L'évaluation se fait généralement à travers les prédictions et les résidus conditionnels et marginaux des deux sous-modèles. Dans le sous-modèle longitudinal, il est possible de vérifier les hypothèses d'homoscédasticité et de normalité des erreurs à partir des résidus conditionnels. Les résidus marginaux permettent de valider les hypothèses sur la va-

riance intra-sujet. Dans le sous modèle de survie, les résidus de martingale permettent d'identifier les sujets ayant une mauvaise prédiction par le modèle. L'adéquation du sous-modèle de survie peut également être testée par les résidus de Cox-Snell en comparant leur distribution à une distribution exponentielle de paramètre 1.

2.4.2 Modèles conjoints à classes latentes

Tout comme les modèles conjoints à effets aléatoires partagés, les modèles conjoints à classes latentes permettent de modéliser simultanément un marqueur longitudinal et un temps d'événement. La principale différence entre ces deux types de modèles est la structure de lien entre le sous-modèle de survie et le sous-modèle longitudinal. Dans le cas des modèles à effets aléatoires partagés, le lien entre les deux sous-modèles est continu et se fait par les effets aléatoires. Ces modèles supposent que toute la population est homogène et possède une évolution moyenne du marqueur longitudinal considéré et les sujets présentent des déviations individuelles par rapport à cette évolution moyenne. Cependant, l'hypothèse d'homogénéité peut paraître assez forte notamment dans l'étude des maladies. Prenons le cas de la mucoviscidose qui est une maladie complexe, avec plus de 2000 mutations qui ont des effets plus ou moins graves chez les malades. La maladie se manifeste différemment d'un malade à l'autre, même lorsqu'il s'agit de malades de la même fratrie. Il serait peu raisonnable de supposer une évolution similaire de la maladie chez tous les malades. Dans ce cas, on peut utiliser les modèles conjoints à classes latentes.

Les modèles conjoints à classes latentes sont plus récents et moins utilisés que les modèles conjoints à effets aléatoires partagés [Lin et al. 2002; Proust-Lima et al. 2014]. Ils permettent de modéliser conjointement la survenue d'un événement et l'évolution d'un marqueur longitudinal. Ces modèles supposent que la population est hétérogène et peut être divisée en sous-groupes d'individus, homogènes vis-à-vis du marqueur longitudinal et de l'événement considéré. Dans ces modèles la corrélation entre le sous-modèle longitudinal et le sous-modèle de survie est capturée par une structure latente discrète qui correspond à la variable aléatoire définissant les sous-groupes d'individus ou classes latentes.

Spécification

Dans une population de N sujets ($i = 1, \ldots, N$), notons Y le marqueur longitudinal étudié avec $Y_i = (Y_i(t_{i1}), Y_i(t_{i2}), \ldots, Y_i(t_{in_i}))$ le vecteur de taille n_i d'observations

mesurées aux temps $t_{i1}, \ldots, t_{in_i}$ pour le sujet i. Notons T_i le temps de survenue de l'événement considéré, C_i variable aléatoire de la censure à droite, δ_i la variable indicatrice de l'événement. On définit le couple $(\tilde{T}_i, \delta_i)$ avec $\tilde{T}_i = min(T_i, C_i)$ et $\delta_i = 0$ si $T_i > C_i$ et $\delta_i = 1$ sinon. On suppose que la population de N sujets est constituée au total de G sous-populations homogènes g ($g = 1, \ldots, G$) et on note c_i la variable latente discrète qui vaut g si le sujet i appartient à la classe g.
Le modèle conjoint à classes latentes est constitué de trois sous modèles : un modèle mixte qui décrit l'évolution du marqueur longitudinal, un modèle de survie le plus souvent un modèle à risques proportionnels qui modélise le risque de survenue de l'événement et un modèle logistique multinomial qui donne la probabilité d'appartenance aux classes pour les sujets.

La probabilité pour le sujet i d'appartenir à la classe g est donnée par :

$$\pi_{ig} = P(c_i = g|X_{0i}) = \frac{e^{\tau_{0g} + X_{0i}^T \tau_{1g}}}{\sum_{l=1}^{G} e^{\tau_{0l} + X_{0i}^T \tau_{1l}}}$$

Ce modèle est rendu identifiable lorsqu'une classe est choisie comme classe de référence, ici la classe G. On a alors $\tau_{0G} = 0$ et $\tau_{1G} = 0$. τ_{0g} représente l'intercept pour la classe g et τ_{1g} est le vecteur de paramètres associé au vecteur de covariables X_{0i} et spécifiques à la classe g. La probabilité d'appartenir à la classe g par rapport à la classe G qui est la classe de référence pour l'augmentation d'une unité de X_{0i} est donnée par $exp(\tau_{1G})$.

La trajectoire du marqueur longitudinal pour le sujet i dans la classe g est donnée par :

$$Y_i(t_{ij})|_{c_i=g} = X_{2i}(t_{ij})^T \beta + X_{3i}(t_{ij})^T \delta_g + Z_i(t_{ij})^T b_{ig} + \epsilon_i(t_{ij})$$

où $X_{2i}(t_{ij})^T$ représente le vecteur de covariables associé au vecteur β d'effets fixes commun à toutes les classes. $X_{3i}(t_{ij})^T$ représente le vecteur de covariables associé au vecteur δ_g d'effets fixes spécifiques à chaque classe. $Z_i(t_{ij})^T$ est le vecteur de covariables associé au vecteur $b_{ig} = b_i|_{c_i=g} \sim N(\mu_g, B)$ d'effets aléatoires spécifiques à chaque classe. B est la matrice de variance-covariance des effets aléatoires qui peut être spécifique à chaque classe ($B = Bg$) ou commune à toutes les classes. $\epsilon_i = (\epsilon_i(t_{i1}), \epsilon_i(t_{i2}), ..., \epsilon_i(t_{in_i}))^T \sim N(0, \sigma^2 I_{n_i})$ represente le vecteur des erreurs de mesure pour le sujet i, où $\sigma^2 I_{n_i}$ est une matrice diagonale de variance-covariance des erreurs indépendantes. Selon la spécification du modèle, les erreurs peuvent également

être corrélées selon différentes structures de corrélation.

Le risque de survenue de l'événement pour le sujet i, dans la classe g est donné par un modèle à risques proportionnels :

$$\alpha_i(t|_{c_i=g}) = \alpha_{0g}(t;\zeta_g)exp(X_{4i}(t)^T\lambda_g + X_{5i}(t)^T\nu)$$

où α_{0g} représente la fonction de risque de base spécifique à la classe g et décrite par le vecteur de paramètres ζ_g. Une distribution paramétrique peut être appliquée sur la fonction de risque de base (Weibull, Gamma, spline...). $X_{4i}(t)^T$ représente le vecteur de covariables associé au vecteur λ_g d'effets spécifiques à chaque classe. $X_{5i}(t)^T$ représente le vecteur de covariables associé au vecteur ν d'effets commun à toutes les classes.

Les deux processus sont supposés indépendants conditionnellement à la classe latente. Chaque classe est définie par un risque d'événement et une trajectoire moyenne du marqueur longitudinal.

Estimation

Dans le cas des modèles conjoints à classes latentes, l'estimation des paramètres se fait par la méthode du maximum de vraisemblance pour un nombre de classes G fixé. À partir de l'hypothèse d'indépendance entre les deux processus conditionnellement aux classes latentes, la log-vraisemblance s'écrit :

$$L(\theta_G) = \sum_{i=1}^{N} log\left(\sum_{g=1}^{G} \pi_{ig} f_{Y_i|c_i}(Y_i|c_i = g)\alpha_i(T_i|c_i = g)^{\delta_i} S_i(T_i|c_i = g)\right)$$

où θ_G est le vecteur de tous les paramètres du modèle conjoint. π_{ig} est la probabilité d'appartenir à la classe g pour le sujet i. $f_{Y_i|c_i}(Y_i|c_i = g)$ est la densité pour le marqueur longitudinal qui est distribuée selon une loi normale de moyenne $X_{2i}\beta + X_{3i}\delta_g + Z_i\mu_g$ et de matrice de variance-covariance $Z_iBZ_i^T + \sigma^2 I_{n_i}$. $\alpha_i(T_i|c_i = g)^{\delta_i} S_i(T_i|c_i = g)$ représente la densité du temps d'événement pour données censurées avec $\alpha_i(T_i|c_i = g)$ la fonction de risque instantané et $S_i(T_i|c_i = g)$ la fonction de survie.

Contrairement au modèle conjoint à effets aléatoires partagés, la log-vraisemblance du modèle conjoint à classes latentes ne comporte pas d'intégrale et a donc une expression analytique. Elle peut être maximisée par un algorithme itératif tel que Marquardt et les variances des paramètres sont obtenues en inversant la matrice

Hessienne, tel que dans la fonction *Jointlcmm* du package *lcmm* sous le logiciel R qui permet d'estimer ces modèles [Proust-Lima et al. 2017].

La présence de maxima locaux est assez fréquente dans l'estimation des paramètres du modèle conjoint à classes latentes. Il est recommandé d'estimer le modèle plusieurs fois en partant de valeurs initiales différentes pour s'assurer de la convergence du modèle vers le maximum global. Comme précisé précédemment, le modèle est estimé pour un nombre de classes latentes fixé. Pour déterminer le nombre de classes latentes optimal, on utilise entre autres le critère BIC (Bayesian Information Criterion). Dans ce cas, on retiendra le modèle ayant le plus petit BIC.

Classification *a posteriori*

Il est possible à partir des modèles conjoints à classes latentes de calculer les probabilités *a posteriori* d'appartenance aux classes pour les sujets. À partir de la formule de Bayes, on a l'expression de la probabilité *a posteriori* d'appartenance à la classe g pour le sujet i sachant les données longitudinales et le temps d'événement :

$$\begin{aligned}\widehat{\pi_{ig}}^{Y,T} &= P(c_i = g|Y_i|T_i, \delta_i, \widehat{\theta_G}) \\ &= \frac{\widehat{\pi_{ig}} f_{Y_i|c_i}(Y_i|c_i = g, \widehat{\theta_G}) \alpha_i(T_i|c_i = g, \widehat{\theta_G})^{\delta_i} S_i(T_i|c_i = g, \widehat{\theta_G})}{\sum_{l=1}^{G} \widehat{\pi_{il}} f_{Y_i|c_i}(Y_i|c_i = l, \widehat{\theta_G}) \alpha_i(T_i|c_i = l, \widehat{\theta_G})^{\delta_i} S_i(T_i|c_i = l, \widehat{\theta_G})}\end{aligned}$$

Avec $\widehat{\pi}_{ig}$ la probabilité d'appartenir à la classe g pour le sujet i calculée à partir des valeurs estimées des paramètres $\widehat{\theta}_G$. Cette probabilité est utilisée pour évaluer la qualité d'ajustement du modèle conjoint.

Le sujet est classé *a posteriori* dans la classe pour laquelle cette probabilité est la plus grande.

Évaluation du modèle

L'évaluation du modèle conjoint à classes latentes en similaire à celle du modèle conjoint à effets aléatoires partagés en ce qui concerne le sous-modèle de survie et le sous-modèle longitudinal. Il est possible d'obtenir les prédictions marginales et les prédictions conditionnelles aux effets aléatoires qui sont dans ce cas spécifiques à la classe latente g. En plus, il est possible de moyenner les prédictions sur toutes les classes latentes afin d'obtenir des prédictions individuelles par sujet. De même,

on peut obtenir des prédictions spécifiques aux classes en calculant la moyenne des prédictions sur tous les sujets dans chaque classe latente [Proust-Lima et al. 2014]. Comme dans le cas des modèles conjoints à effets aléatoires, les résidus marginaux et spécifiques aux sujets peuvent être obtenus. Dans les modèles conjoints à classes latentes, il est important de bien déterminer le nombre de classes latentes. En plus du critère BIC, la classification *a posteriori* permet de guider ce choix en évaluation la capacité discriminante du modèle. Une bonne classification correspond au cas où chaque sujet a une probabilité élevée (proche de 1) d'appartenir à une classe et des probabilités faibles (proches de 0) d'appartenir aux autres classes.

2.4.3 Extension des modèles conjoints

Pour étudier l'évolution d'une maladie par exemple, on est amené à faire un recueil des données sur les sujets. Généralement, plusieurs paramètres sont recueillis au cours du temps dont des paramètres ayant un caractère évolutif. Plutôt que de s'intéresser uniquement au lien entre un marqueur longitudinal et la survenue d'un événement, on pourrait être s'intéresser à l'analyse conjointe de plusieurs marqueurs longitudinaux et la survenue d'un ou de plusieurs événements. Des extensions de modèles conjoints ont été faits dans cette optique.
Proust-Lima et al. [2016] propose un modèle conjoint à classes latentes pour étudier le déclin cognitif à partir de deux tests de mémoire, associé aux risques de démence et de décès. Sur le même sujet, Rouanet et al. [2016] propose un modèle conjoint à classes latentes dans un contexte de risques semi-compétitifs censurés par intervalles. Han et al. [2007] propose un modèle conjoint à classes latentes pour événements récurrents dans l'analyse des crises d'épilepsie.
Cependant, les extensions de modèles conjoints concernent surtout les modèles conjoints à effets aléatoires partagés. Andrinopoulou et al. [2014, 2017] propose un modèle conjoint à effets aléatoires partagés pour étudier simultanément deux marqueurs longitudinaux qui décrivent le fonctionnement de la valve aortique, et la survenue du décès ou de la ré-opération chez des sujets ayant des problèmes de valve aortique. Comme autres extensions de modèles conjoints, on peut citer les modèles conjoints pour plusieurs marqueurs longitudinaux et un temps d'événement [Brown et al. 2005; Rizopoulos 2011], les modèles conjoints pour un marqueur longitudinal et plusieurs temps d'événements [Elashoff et al. 2008; Yu and Ghosh 2010; Huang et al. 2011], les modèles conjoints pour événements récurrents [Liu and Huang 2009; Mauguen et al. 2013; Krol et al. 2016], les modèles conjoints incorporant un processus multi-états [Ferrer et al. 2016].

2.4.4 Prédiction dynamique

Les modèles conjoints permettent de prédire le risque de survenue d'un événement sur une fenêtre de temps précise. Ces prédictions ont la particularité d'être dynamiques, c'est-à-dire évolutives dans le temps. En effet, toutes les données répétées du marqueur sont utilisées pour prédire l'événement. Ainsi, à chaque nouvelle valeur du marqueur, le risque d'événement est mis à jour[Proust-Lima and Taylor 2009; Rizopoulos 2011].

Les prédictions dynamiques sont importantes dans le suivi des patients atteints d'une maladie. Elles sont en adéquation avec le principe de la médecine personnalisée qui est de plus en plus appliquée. Elles serviraient notamment de guide pour les cliniciens dans la prise de décision liée aux traitements des malades. Dans le contexte de la mucoviscidose par exemple, les critères d'identification des patients à la transplantation pulmonaire sont assez variés et diffèrent d'un centre de transplantation à un autre. Alors, fournir un outil pronostique aux cliniciens leur permettraient de mieux identifier les malades éligibles à la transplantation pulmonaire. Grâce aux prédictions dynamiques du risque de décès ou de transplantation pulmonaire, le clinicien aurait plus d'informations sur le devenir du malade et pourrait adapter le traitement si nécessaire.

Le principe des prédictions dynamiques issues des modèles conjoints consiste à fixer un temps de prédiction s, et d'utiliser toute l'information disponible jusqu'au temps s, notamment les mesures répétées du marqueur longitudinal et les variables explicatives pour prédire la survenue de l'événement d'intérêt entre le temps s et le temps $s + t$. Ici, s représente le temps de prédiction et t représente l'horizon de prédiction. Lorsque le temps s augmente, il y a plus d'information sur le marqueur longitudinal et le risque de survenue de l'événement d'intérêt est mis à jour.

Soit $s \geq 0$ et $t \geq 0$. Pour un sujet i, notons T_i le temps de survenue de l'événement à prédire, $Y_i^{(s)} = \{Y_i(t_{ij}),\, j = 1, \ldots, n_i \text{ tels que } t_{ij} \leq s\}$ le vecteur de données observées du marqueur longitudinal jusqu'au temps s et $X_i^{(s)} = \{X_{0i}, X_{2i}, X_{1i}(t_{ij}),\, j = 1, \ldots, n_i$ tels que $t_{ij} \leq s\}$ le vecteur de données observées des variables explicatives jusqu'au temps s et θ le vecteur des paramètres du modèle. À partir du modèle conjoint à classes latentes, la probabilité de survenue de l'événement d'intérêt entre le temps s et le temps $s + t$ s'écrit :

$$P(T_i \leq s+t|T_i \geq s, Y_i^{(s)}, X_i^{(s)}, \theta) = \sum_{g=1}^{G} P(T_i \leq s+t|T_i \geq s, c_i = g, X_i^{(s)}, \theta) \, P(c_i = g|T_i \geq s, Y_i^{(s)}, X_i^{(s)}, \theta)$$

À partir du modèle conjoint à effets aléatoires, la probabilité de survenue de l'événement d'intérêt entre le temps s et le temps $s+t$ s'écrit comme précédemment, seulement la somme sur les classes est remplacée par l'intégrale sur les effets aléatoires :

$$P(T_i \leq s+t|T_i \geq s, Y_i^{(s)}, X_i^{(s)}, \theta) = \int P(T_i \leq s+t|T_i \geq s, b_i = u, X_i^{(s)}, \theta) \, f_{b_i}(u|T_i \geq s, Y_i^{(s)}, X_i^{(s)}, \theta)du$$

À partir des estimations des paramètres du modèle conjoint $\widehat{\theta}$ et de leurs variances $\widehat{V(\theta)}$, la probabilité de survenue de l'événement peut être calculée aussi bien pour les sujets inclus dans l'échantillon d'estimation que pour les sujets n'y étant pas inclus.

Une approche différente des modèles conjoints pour faire de la prédiction dynamique est l'approche « landmark ». À partir des mesures collectées jusqu'à un temps s de prédiction, l'approche « landmark » consiste à estimer un modèle de survie classique chez les sujets encore à risque jusqu'à ce temps s. Ce principe utilise alors la valeur du marqueur longitudinal au temps s pour prédire la probabilité de survenue de l'événement [van Houwelingen 2007; Parast et al. 2012].

2.4.5 Évaluation des capacités pronostiques

Les capacités pronostiques d'un modèle de prédiction peuvent être évaluées en termes de discrimination et de calibration. La calibration permet d'évaluer la capacité du modèle à prédire en comparant la survie observée et la survie prédite [Graf et al. 1999; Schemper and Henderson 2000; Gerds and Schumacher 2006]. La discrimination permet d'évaluer la capacité du modèle à discriminer les sujets qui sont plus susceptibles de subir l'événement de ceux qui ne le sont pas [Harrell et al. 1982; Heagerty et al. 2000; Antolini et al. 2005].

La discrimination

Un modèle ayant une bonne capacité à discriminer fournit des probabilités de survenue de l'événement plus élevées pour un sujet plus susceptible de subir l'événement comparé à un sujet moins susceptible de subir l'événement.
Pour évaluer la discrimination on peut utiliser l'AUC décrit dans le cadre des modèles de régression logistique. L'AUC peut être adapté dans le contexte de la prédiction dynamique.

Soit s un temps de prédiction et t un horizon de prédiction. On note $\pi_i(s,t)$ les prédictions pour le sujet $i(i = 1, \ldots, n)$ au temps de prédiction s et à l'horizon t. $\pi_i(s,t) = 0$ pour les sujets qui ne sont plus à risque au temps de prédiction s. L'AUC dynamique adapté aux modèles conjoints peut être défini de la manière suivante :

$$AUC(s,t) = P(\pi_i(s,t) > \pi_j(s,t) | D_i(s,t) = 1, D_j(s,t) = 0, T_i > s, T_j > s)$$

Où D représente l'événement d'intérêt, $D_i(s,t) = 1$ si le sujet i a subi l'événement dans l'intervalle de temps $]s, s+t]$ et $D_i(s,t) = 0$ si le sujet i n'a pas subi l'événement au temps $s+t$. Dans le cas des risques compétitifs $D_i(s,t) = 0$ si le sujet i a subi un événement autre que l'événement d'intérêt dans l'intervalle $]s, s+t]$. Le sujet i est donc considéré comme un cas si $D_i(s,t) = 1$ et comme un contrôle si $D_i(s,t) = 0$ [Zheng et al. 2012; Blanche et al. 2013].

Le score de Brier

Le score de Brier (BS) ou erreur de prédiction quadratique permet d'évaluer à la fois la calibration et la discrimination d'un modèle de prédiction. Ce score permet de comparer la survie observée et la survie prédite. Ce paramètre est adapté au contexte de la prédiction dynamique et est défini par :

$$\begin{aligned} BS(s,t) &= E\left[(D(s,t) - \pi(s,t))^2 | T > s\right] \\ &= E\left[\left(E\left[D(s,t)|H(s)\right] - \pi(s,t)\right)^2 | T > s\right] \\ &+ E\left[\left(D(s,t) - E\left[D(s,t)|H(s)\right]\right)^2 | T > s\right] \end{aligned}$$

Ici, $H(s)$ représente l'historique du sujet jusqu'au temps s, c'est-à-dire les mesures répétées du marqueur et les covariables à l'inclusion.

Le terme $E\left[\left(E\left[D(s,t)|H(s)\right] - \pi(s,t)\right)^2 | T > s\right]$ représente la calibration et permet de mesurer l'écart entre les prédictions et le « vrai » risque de l'événement

$E\left[D(s,t)|H(s)\right]$ dans l'intervalle de temps $]s, s+t]$ sachant l'historique du sujet.

Le terme $E\left[\left(D(s,t) - E\left[D(s,t)|H(s)\right]\right)^2 |T > s\right]$ permet d'évaluer la capacité à discriminer du modèle à partir des mesures répétées du marqueur jusqu'au temps de prédiction s.

Contrairement à l'AUC qui fournit de meilleurs capacités lorsqu'il est proche de 1, le BS qui est une erreur de mesure fournit de bonnes capacités lorsqu'il est proche de 0.

Le statut d'un sujet censuré dans l'intervalle de temps $]s, s+t]$ est inconnu. Pour résoudre ce problème, l'estimateur Inverse Probability of Censoring Weighting (IPCW) a été proposé pour l'AUC et le score de Brier [Blanche et al. 2013, 2015]. Le principe de cette approche est d'affecter des poids aux sujets en fonction de leur probabilité d'être des cas ou des contrôles. On obtient alors un AUC et un score de Brier qui valent :

$$\widehat{AUC}(s,t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}_{(\pi_i(s,t)>\pi_j(s,t))}\tilde{D}_i(s,t)(1-\tilde{D}_j(s,t))\widehat{W}_i(s,t)\widehat{W}_j(s,t)}{\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{D}_i(s,t)(1-\tilde{D}_j(s,t))\widehat{W}_i(s,t)\widehat{W}_j(s,t)}$$

$$\widehat{BS}(s,t) = \frac{1}{N}\sum_{i=1}^{N} W_i(s,t)(\tilde{D}_i(s,t) - \pi_i(s,t))^2$$

Où N représente le nombre de sujets à risque au temps de prédiction s. $\tilde{D}_i(s,t) = \mathbb{1}_{(s<\tilde{T}_i\leq s+t, \tilde{\eta}_i=1)}$ qui vaut 1 si le sujet i a subi l'événement dans l'intervalle de temps $]s, s+t]$ et 0 sinon. $\pi_i(s,t)$ est la survie prédite au temps $s+t$ sachant que le sujet est encore à risque au temps de prédiction s et tenant compte des mesures répétées du marqueur jusqu'au temps s. Les poids sont représentés par :

$$\widehat{W}_i(s,t) = \frac{\mathbb{1}_{(\tilde{T}_i>s+t)}}{\widehat{G}(s+t|s)} + \frac{\mathbb{1}_{(s<\tilde{T}_i\leq s+t)}\Delta_i}{\widehat{G}(\tilde{T}_i|s)}$$

où $\widehat{G}(u)$ est l'estimateur de Kaplan-Meier pour la fonction de survie au temps de censure u [Graf et al. 1999]. $\widehat{G}(\tilde{T}_i|s)$ représente la probabilité d'avoir subi l'événement au temps de suivi $\tilde{T}_i$ sachant que le sujet i était à risque au temps s.

2.4.6 Validation

Il est nécessaire de valider les modèles de prédiction sur des données autres que celles ayant servi à leur développement. Le but étant d'évaluer les performances prédictives du modèle développé, c'est-à-dire évaluer la capacité discriminante et la

calibration du modèle. En effet, les capacités prédictives d'un modèle peuvent être surestimées sur les données ayant servi au développement [Lee et al. 2012]. Dans ce cas, le modèle a tendance à trop bien prédire sur les données ayant servi au développement.

La validation d'un modèle de prédiction peut être faite en utilisant des données extérieures, on parle de validation externe. Dans ce cas, les données utilisées sont différentes des données de développement soit temporellement, soit géographiquement. Le modèle sera alors développé sur un premier échantillon de données et les performances prédictives seront évaluées sur cet autre échantillon de données.

Il est également possible de séparer au préalable, aléatoirement, les données en deux échantillons. Un premier échantillon d'apprentissage qui servira à développer le modèle. Un second échantillon de test qui servira à évaluer les capacités prédictives du modèle [May et al. 2004]. Des techniques de rééchantillonnage qui consistent à utiliser des sous-ensembles aléatoires des données pour la validation peuvent également être utilisées. C'est le cas de la validation croisée, des méthodes de bootstrap [Efron and Tibshirani 1997; Altman and Royston 2000].

Chapitre 3

Facteurs pronostiques de la transplantation pulmonaire

La mucoviscidose est une maladie rare dont l'espérance de vie est en constante évolution. La médiane de vie des sujets atteints de mucoviscidose qui était inférieure à 10 ans dans les années 50, est actuellement estimée à 40 ans [MacKenzie et al. 2014; Knapp et al. 2016]. Il y a donc une amélioration constante de la prise en charge de la maladie grâce aux progrès scientifiques. Ceci a entraîné une diminution du décès chez les enfants atteints de mucoviscidose. Cette maladie qui était au préalable considérée comme une maladie infantile est désormais considérée comme une maladie chronique. Les décès infantiles ont presque disparu dans les pays développés [Urquhart et al. 2013]. Plus de 50% des malades sont adultes et cette proportion est estimée à 75% pour l'année 2025 [Burgel et al. 2015].

La mucoviscidose étant une maladie incurable à ce jour, il est indispensable que les progrès s'intensifient et s'améliorent dans la recherche à ce sujet. Pour ce faire, il est nécessaire d'avoir une bonne connaissance de la maladie notamment des facteurs de risque associés au décès chez les malades. Une meilleure connaissance des facteurs pronostiques de la mucoviscidose permettrait d'identifier les malades ayant un risque élevé de décéder prématurément et d'adapter leur prise en charge en conséquence. De nombreux traitements permettent d'améliorer la santé des malades, néanmoins la transplantation pulmonaire reste le moyen le plus efficace pour prolonger la vie des malades dont l'état respiratoire s'est considérablement dégradé. Les transplantations pulmonaires sont surtout réalisées chez les adultes du fait de la diminution du risque de décès chez les enfants malades, entre autres. Ainsi, des modèles pronostiques ont été développés dans le contexte de la mucoviscidose afin d'identifier les malades éligibles à la transplantation pulmonaire.

En 1992, Kerem et al. a identifié le VEMS qui décrit la fonction pulmonaire, comme le principal facteur lié à la mortalité chez les sujets atteints de mucoviscidose. Kerem et al. a alors identifié le seuil de 30% de VEMS en dessous duquel les patients seraient considérés comme éligibles à la transplantation pulmonaire. Ce seuil a par la suite été adopté comme un des critères majeurs d'identification des malades à la transplantation pulmonaire. Cependant, depuis 1992, de nombreux progrès ont été réalisés dans le domaine de la mucoviscidose. Le pronostic des malades s'est considérablement amélioré et il a été montré une amélioration de la survie des malades avec un VEMS inférieur à 30% [George et al. 2011]. De ce fait, le critère du VEMS inférieur à 30% de valeurs prédites ne devrait pas être considéré comme seul critère d'identification des malades à la transplantation pulmonaire.

D'autres modèles pronostiques ont été développés et ont permis d'identifier des facteurs autres que le VEMS liés à la mortalité chez les sujets atteints de mucoviscidose. Parmi les facteurs identifiés, on peut citer entre autres les colonisations aux les germes (*Burkholderia cepacia* [Courtney et al. 2004, 2007], *Pseudomonas aeruginosa* [Courtney et al. 2007; Aaron et al. 2015], *Staphylococcus aureus*[Buzzetti et al. 2012]...), les complications de la maladie telles que les symptômes pneumothorax, les hémoptysies, les exacerbations pulmonaires [Mayer-Hamblett et al. 2002], les maladies associées à la mucoviscidose comme le diabète [Aaron et al. 2015], l'indice de masse corporelle (IMC) [Courtney et al. 2007], l'insuffisance pancréatique [Aaron et al. 2015]... Cependant, ces modèles pronostiques considèrent généralement le décès comme seul événement d'intérêt pour l'identification des malades à la transplantation pulmonaire. Ce choix du décès seul, sans tenir compte de la transplantation pulmonaire pourrait être dû au fait que ces modèles ont été développés à un moment où la transplantation pulmonaire était peu développée. Aujourd'hui, il y a de plus en plus de transplantations pulmonaires. De plus, les transplantations pulmonaires sont réalisées pour sauver la vie des malades dont l'état respiratoire s'est considérablement dégradée. L'hypothèse sous-jacente est que si ces sujets n'avaient pas reçu de greffe, ils seraient sans doute décédés. De ce fait, le choix d'étudier le décès seul comme événement d'intérêt ne saurait être réaliste.

Dans ce travail, nous avons fait deux choix essentiels pour l'identification des facteurs pronostiques dans le contexte de la mucoviscidose. Pour les raisons évoquées précédemment, nous avons choisi d'une part, de nous intéresser à l'événement composite transplantation pulmonaire ou décès dans le but d'identifier les malades éligibles

à la transplantation pulmonaire. D'autre part, nous avons choisi de nous limiter aux données des malades adultes. Ces choix nous semblent pertinents et mieux adaptés au contexte actuel de la mucoviscidose notamment en France. En effet, plus de la moitié des malades recensés dans le registre français de la mucoviscidose sont des adultes. Tous les patients transplantés sont au préalable inscrits sur liste d'attente et l'âge moyen d'inscription sur liste d'attente est de 30 ans avec un écart-type de 11 ans. Les transplantations pulmonaires sont en général réalisées chez les adultes, avec un âge moyen de 30 ans et un écart-type de 11 ans pour l'année 2015. En 2015, l'âge moyen et l'âge médian au décès des sujets atteints de mucoviscidose en France étaient respectivement de 34 ans et de 31 ans. Par ailleurs, en France on note très peu de décès sur liste d'attente, soit moins de 4% par an depuis 2007 [Bellis et al. 2015]. Il est donc plus judicieux de s'intéresser à la transplantation pulmonaire ou au décès comme événement pour identifier les sujets éligibles à la transplantation pulmonaire chez les adultes.

Nous avons développé un modèle pronostique identifiant les facteurs associés à la transplantation pulmonaire ou au décès chez les adultes atteints de mucoviscidose à partir des données du registre français de la mucoviscidose. La fenêtre de prédiction choisie était de 3 ans, et ce choix a été fait en concertation avec des cliniciens. Pour un modèle pronostique dont le but est d'identifier les patients éligibles à la transplantation pulmonaire, il est préférable de ne pas considérer une fenêtre de prédiction qui soit très courte, ni très longue. La plupart des modèles pronostiques développés à ce jour ne font pas de distinction entre les données d'enfants et les données d'adultes. Pour ces modèles, les fenêtres de prédictions vont de 1 à 7 ans. Aaron et al. a proposé un modèle pour prédire le risque de décès à 1 an chez les sujets atteints de mucoviscidose dont une des limites est la surestimation de ce risque. Mayer-Hamblett et al. a développé un modèle de prédiction de la mortalité à 2 ans chez les sujets atteints de mucoviscidose dont le pouvoir prédictif était moins bon que le critère du VEMS inférieur à 30% de valeurs prédites. Pour une fenêtre de prédiction très courte, on peut être confronté au problème de manque de greffon, notamment dans les pays où il y a peu de transplantations pulmonaires. Par ailleurs, faire un modèle pronostique à long terme chez les adultes atteints d'une maladie dont la médiane de vie est de 40 ans ne semble pas pertinent d'un point de vue clinique.

En 2001, Liou et al. a proposé un modèle de prédiction à 5 ans qui a permis d'identifier des facteurs liés au décès chez les sujets atteints de mucoviscidose. Ce modèle a été développé chez les enfants et les adultes sans tenir compte de la transplantation pulmonaire. Quinze ans après le développement de ce modèle, la recherche a évolué

dans le domaine de la mucoviscidose. Certains traitements inexistants à cette époque ont permis d'améliorer le pronostic des malades. Ce modèle pourrait ne plus être adapté au contexte actuel de la mucoviscidose. Par ailleurs ce modèle a été validé en 2011 par Buzzetti et al. sur les données du registre italien de la mucoviscidose. Des différences significatives ont été notées entre les décès observés et les décès prédits. Les auteurs supposent que les mauvaises prédictions seraient dues à l'amélioration dans les traitements et la prise en charge de la maladie.

Il est donc nécessaire de réévaluer les facteurs associés au décès ou à la transplantation pulmonaire chez les adultes atteints de mucoviscidose. Ce travail a été réalisé à partir des données du registre français de la mucoviscidose et a donné lieu à une publication dans Journal of Cystic Fibrosis.

3.1 Manuscrit publié dans Journal of Cystic Fibrosis

ELSEVIER
Journal of Cystic Fibrosis xx (2017) xxx – xxx
Journal of Cystic Fibrosis

Original Article

A 3-year prognostic score for adults with cystic fibrosis

L. Nkam [a,*], J. Lambert [b], A. Latouche [c], G. Bellis [d], PR. Burgel [e,f,1], M.N. Hocine [a,1]

[a] Laboratoire Modélisation, Epidémiologie et Surveillance des Risques Sanitaires, Conservatoire National des Arts et Métiers, Paris, France
[b] Hôpital Saint-Louis, Service de Biostatistique et Information Médicale, Assistance Publique-Hôpitaux de Paris, Paris, France
[c] Centre d'Etudes et de Recherche en Informatique et Communications, Conservatoire National des Arts et Métiers, Paris, France
[d] Institut National d'Etudes Démographiques, Paris, France
[e] Assistance Publique-Hôpitaux de Paris, Hôpital Cochin, Service de Pneumologie, Paris, France
[f] Université Paris Descartes, Sorbonne Paris Cité, Paris, France

Received 20 September 2016; revised 2 March 2017; accepted 2 March 2017
Available online xxxx

Abstract

Background: Therapeutic progress in patients with cystic fibrosis (CF) has resulted in improved prognosis over the past decades. We aim to reevaluate prognostic factors of CF and provide a prognostic score to predict the risk of death or lung transplantation (LT) within a 3-year period in adult patients.
Methods: We developed a logistic model using data from the French CF Registry and combined the coefficients into a prognostic score. The discriminative abilities of the model and the prognostic score were assessed by c-statistic. The prognostic score was validated using a 10-fold cross-validation.
Results: The risk of death or LT within 3 years was related to eight characteristics. The development and the validation provided excellent results for the prognostic score; the c-statistic was 0.91 and 0.90 respectively.
Conclusion: The score developed to predict 3-year death or LT in adults with CF might be useful for clinicians to identify patients requiring specialized evaluation for LT.

Keywords: Cystic fibrosis; Prognostic factors; Registry data; Logistic model; Prognostic score

1. Introduction

Cystic fibrosis (CF) is a multiorgan disease that affects primarily the lungs, causing diffuse bronchiectasis which often leads to progressive respiratory insufficiency and premature death [9]. Lung transplantation (LT) is proposed to CF patients with terminal respiratory failure with the aim of improving life expectancy and quality of life [25]. Although criteria for referring patients for LT have been proposed [14], the optimal timing for referring CF patients for transplantation remains difficult to establish in an individual patient. A recent study in France has shown that respiratory death in CF patients often occurs due to late or no referral for LT [19], suggesting the need to develop novel strategies for referring patients at high risk of death for transplant evaluation.

In the past 25 years, several statistical models have been developed to identify prognostic factors in CF patients. In their seminal study, Kerem et al. identified forced expiratory volume in 1 s (FEV_1) as the main prognostic factor and suggested that patients with an FEV_1 value less than 30% should be considered for LT [16]. Subsequent studies identified multiple other factors related to death in patients with CF including older age [17,21],

* Corresponding author at: Laboratory 'Modélisation, Epidémiologie et Surveillance des Risques Sanitaires', Conservatoire National des Arts et Métiers, 292 Rue Saint-Martin, 75003 Paris, France.
E-mail address: lionelle.nkam@cnam.fr (L. Nkam).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jcf.2017.03.004

female gender [16,17], lower body mass index (BMI) [21], pancreatic insufficiency [1,17], diabetes [1,17], *Pseudomonas aeruginosa* colonization [1,6], *Burkholderia cepacia* colonization [4,6,17], *Staphylococcus aureus* colonization [7], massive hemoptysis [11], pneumothorax [10] and high number of pulmonary exacerbations [4,17,21]. Although several attempts at developing prognostic scores have been performed in CF patients [1,4,13,17,20,21], it has proven difficult to develop a score that better predicts death than $FEV_1 < 30\%$ predicted [20]. Further, prognosis has dramatically improved over the past decades due to advance in integrated care provided by multidisciplinary teams in CF centers [2,8,12,18,23]. As a result, prognostic factors have changed over time and studies performed using data obtained in previous decades may not be appropriate for current evaluation of CF patients. For example, George et al. showed an important improvement in the survival of patients whose FEV_1 has fallen below 30% of predicted value. Consequently, they suggested that the threshold of 30% predicted for FEV_1 should no longer be considered in isolation as an indication for LT [12]. Moreover, pediatric mortality in patients with CF has almost disappeared in developed countries due to improvement in patients care by multidisciplinary teams [26].

The aim of the present study was to develop a 3-year predictive model that provides a prognostic score to better predict the risk of death or LT in adult patients with CF.

2. Materials and methods

2.1. The French Cystic Fibrosis Registry

We used data from the French Cystic Fibrosis Registry (French CF Registry). This registry contains longitudinal data on more than 8000 patients since 1992, which represents approximately 90% of all CF patients in France [3]. Each patient is assigned to a center specialized in CF, where his/her health status is regularly monitored. A numeric code is assigned to each patient to link information between specialized centers and the French CF Registry. This registry records annual health-check data for each subject including vital status, therapeutic management, anthropometry, spirometry, morbidity factors, consultations and hospitalizations, arterial blood gas, microbiological tests, pregnancy and paternity, and transplantations and sociodemographic data [3].

2.2. Patients and data collection

The period of the study was 2010–2013. Patients alive and aged 18 years or older on 31st December 2010 and for whom vital status was known on 31st December 2013 were included in the study. Patients who received a lung transplant before 2010 and patients lost to follow-up between 2011 and 2013 were excluded from the study. Forty-two covariates (listed in Table 1 and Supplementary Table 1) considered as potential predictors and records of the year 2010 were extracted to predict the outcome, defined as death or LT before the end of 2013. Fig. 1 presents the selection scheme of patients who were included in the study.

2.3. Missing data imputation

In 2010, only 12% of patients had complete information for all the 42 potential predictors. However, the percentage of missing data represented only 4%, as illustrated in Table 1. To deal with missing data in the covariates, a multiple imputation by chained equations was used [27]. We assumed that data were missing at random that is, the probability of missingness depends on the values of the observed covariates.

2.4. Model development

The characteristics of patients according to the outcome were compared using chi-square test or Fisher's exact test for categorical variables, and the Mann–Whitney test for continuous variables. We developed a multivariable logistic regression model to predict the outcome of interest, defined as death or LT by the end of 2013. Covariates that were significantly associated to the outcome in 2013 (p value <0.25) with univariate analysis were considered for the multivariable logistic regression model. A forward stepwise selection process was used to select the subset of variables independently associated with the outcome. The predictors retained in the final model were combined into a prognostic score to easily estimate the individual risk of death or LT within 3 years. To this end, continuous predictors were transformed into categorical variables according to clinically relevant thresholds. The contribution of each predictor to the prognostic score was proportional to its regression coefficient. To help the clinician to easily obtain the prognostic score and the risk of death or LT in a 3-year period using the patient characteristics, a nomogram was provided.

2.5. Model performances

Performances of the developed model and the prognostic score were investigated in terms of discrimination and calibration. Discrimination assesses how well the model can distinguish patients with the outcome of interest and patients without. This was evaluated using the c-statistic, also known as the area under the receiver operating characteristic curve [5]. The calibration compares the observed proportion of events against the predicted probabilities. It was evaluated using the Hosmer-Lemeshow test [15]. These performances were tested for both the developed model and the prognostic score, on each imputed dataset.

2.6. Model internal validation

To avoid overestimation of the model performances, we performed an internal validation using a 10-fold cross-validation. Overestimation happens when the model performs well on the data used for development but not on test data. Cross-validation can help detect overestimate models and helps to assess how well the model fits new observations. We randomly partitioned the initial dataset into 10 subsamples, fitted the model on nine of the subsamples and evaluated its performances on the other. We repeated this ten times, leaving out each subsample once. The performance of the prognostic score was evaluated using the

Please cite this article as: Nkam L, et al, A 3-year prognostic score for adults with cystic fibrosis, J Cyst Fibros (2017), http://dx.doi.org/10.1016/j.jcf.2017.03.004

Table 1
Characteristics of 2096 adult CF patients in 2010 according to vital status on December 31st 2013.

Patient characteristics	Missing data (%)	Total n = 2096 (%)	Alive without LT n = 1828 (%)	Death or LT n = 268 (%)	p value
Gender, Male	0	1101 (52.5)	960 (52.5)	141 (52.6)	1
CFTR genotype	0				<0.001
Classes 1–3		1373 (65.5)	1165 (63.7)	208 (77.6)	
Classes 4/5		254 (12.1)	238 (13.0)	16 (6.0)	
Classes unknown		469 (22.4)	425 (23.3)	44 (16.4)	
Airway colonization	0	2013 (96.0)	1747 (95.6)	266 (99.3)	
Achromobacter xylosoxidans	0	133 (6.6)	98 (5.6)	35 (13.2)	<0.001
Aspergillus fumigatus	0	641 (31.8)	556 (31.8)	85 (32.0)	0.02
Burkholderia cepacia	0	69 (3.4)	51 (2.9)	18 (6.8)	<0.001
Non-tuberculous mycobacteria	0	95 (4.7)	79 (4.5)	16 (6.0)	0.009
Pseudomonas aeruginosa	0	1319 (65.5)	1102 (63.1)	217 (81.6)	<0.001
Staphylococcus aureus	0	1290 (64.1)	1136 (65.0)	154 (57.9)	0.001
Stenotrophomonas maltophilia	0	203 (10.1)	166 (9.5)	37 (13.9)	0.001
Comorbidities	0	2007 (95.8)	1741 (95.2)	266 (99.3)	0.004
Cirrhosis	0	98 (4.9)	77 (4.4)	21 (7.9)	<0.001
Insulin-treated diabetes	0	343 (17.1)	256 (14.7)	87 (32.7)	<0.001
Pancreatic insufficiency	0	1758 (87.6)	1504 (86.4)	254 (95.5)	<0.001
Allergic bronchopulmonary aspergillosis	12 (0.6)	378 (18.9)	302 (17.5)	76 (28.7)	<0.001
Hemoptysis	9 (0.4)	204 (10.2)	152 (8.8)	52 (19.5)	<0.001
Pneumothorax	9 (0.4)	29 (1.5)	19 (1.1)	10 (3.8)	<0.001
Depression	30 (1.4)	147 (7.3)	114 (6.6)	33 (12.6)	<0.001
FEV_1, % predicted [a]	144 (6.9)	58.3 (39.4–79.8)	63.3 (44.9–82.3)	29.1 (22.2–36.3)	<0.001
FVC, % predicted [a]	150 (7.2)	79.0 (62.2–95.2)	82.4 (67.0–97.7)	50.3 (38.6–60.6)	<0.001
Age (years) [a]	0	25.5 (21–32.3)	25 (21–32)	26.5 (22–33)	0.02
Height (cm^2) [a]	84 (4.0)	167 (160–173)	167 (160–173)	165.9 (160–172)	0.26
Weight (kg) [a]	63 (3.0)	56 (50–63)	57 (51–64)	51 (46–57)	<0.001
BMI (kg/m^2) [a]	92 (4.4)	20.2 (18.5–22.1)	20.4 (18.9–22.3)	18.4 (17.3–20.0)	<0.001
Number of IV antibiotics courses/year [a]	22 (1.0)	1 (0–2)	1 (0–2)	3 (2–5)	<0.001
Number of IV antibiotics days/year [a]	76 (3.6)	14 (0–30)	0 (0–28)	45 (21–75)	<0.001
Number of days of hospitalization/year [a]	327 (15.6)	0 (0–23)	0 (0–1)	2 (1–3)	<0.001
Azithromycin	10 (0.5)	1254 (59.8)	1056 (58.1)	198 (74.2)	<0.001
Non-invasive ventilation	11 (0.5)	127 (6.1)	51 (2.8)	76 (28.5)	<0.001
Long-term oxygen therapy	11 (0.5)	230 (11.0)	97 (5.3)	133 (49.8)	<0.001
Oral corticosteroids	13 (0.6)	171 (8.2)	125 (6.9)	46 (17.3)	<0.001
Inhaled therapies	0	1873 (89.4)	1618 (88.5)	255 (95.1)	0.001
Inhaled antibiotics	0	1122 (59.9)	928 (57.4)	194 (76.1)	<0.001
Inhaled corticosteroids	0	1012 (54.0)	868 (53.6)	144 (56.5)	0.003

FEV_1%: forced expiratory volume in 1 s as percentage of predicted values.
FVC %: forced vital capacity as percentage of predicted values.
BMI: body mass index.
IV: intravenous.
CFTR: cystic fibrosis transmembrane conductance regulator.
LT: lung transplantation.
[a] Continuous variables, median (interquartile range).

c-statistic. The process above was repeated 10 times on each imputed datasets and we averaged all the values of c-statistic to produce a single estimation of it.

2.7. Sensitivity analysis

A total of 184 patients lost to follow-up have been excluded from the analysis. In order to provide the lack of impact of this exclusion on the model, we reanalyzed the data including these patients after imputation of their outcome.

The two outcomes (death without LT and occurrence of LT) may be predicted by different covariates. We further investigated if the use of the separate outcomes altered the interpretation, by reanalyzing our data using each individual component.

All statistical analyses were performed using R software version 3.3.1.

3. Results

3.1. Data description

There were 2096 patients in the French CF Registry who satisfied the study criteria. A total of 268 (13%) died or received LT within the 3-year follow-up period, including 55 deaths without LT and 213 transplantations. Patient characteristics according to death or LT at 3 years are provided in Table 1. Significant differences in baseline characteristics were observed between patients who died or received LT during the 3-year follow-up period and those who

Please cite this article as: Nkam L, et al, A 3-year prognostic score for adults with cystic fibrosis, J Cyst Fibros (2017), http://dx.doi.org/10.1016/j.jcf.2017.03.004

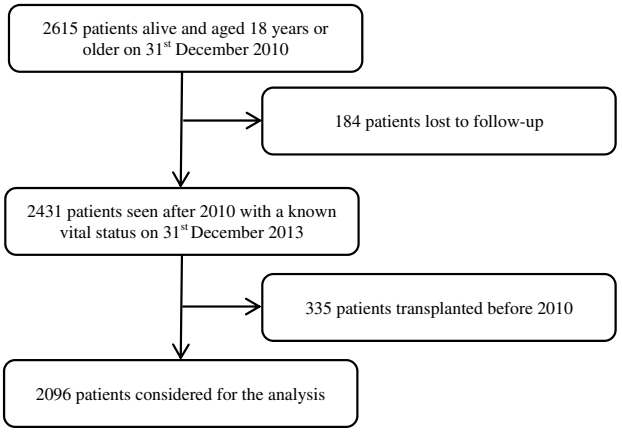

Fig. 1. Patients selection in the French CF Registry on 31st December 2010.

remained alive without LT. The median (interquartile range) age of patients at death or LT was 26.5 (22–33) years. Patients who died or received LT had lower BMI, FEV_1 and forced vital capacity (FVC), and had higher rates of airway microbial colonization than patients who remained alive without LT.

Patients who died or received LT were more likely to have been diagnosed with ABPA, hemoptysis, pneumothorax, insulin-treated diabetes, liver cirrhosis and depression. Treatment burden was also higher in these patients with higher rates of long-term oxygen therapy, non-invasive ventilation, oral corticosteroids, azithromycin and higher number of days with intravenous (IV) antibiotics and hospitalization.

3.2. Model development

The results of the multivariable logistic regression obtained after forward selection for variables associated with death or LT are shown in Table 2. The predictive model included 3 predictors directly related to the patient clinical characteristics in particular: FEV_1, BMI and *Burkholderia cepacia* colonization. It also identified 5 therapeutic variables, namely oral corticosteroids, long-term oxygen therapy, non-invasive ventilation, number of IV antibiotics courses per year and number of days of hospitalization per year. The risk of death or LT was higher for patients with lower values of FEV_1, BMI or with *Burkholderia cepacia* colonization.

Table 2
Logistic regression model for prediction of within 3-year death or lung transplantation in adults with CF.

	Odds ratio (95% CI)	p value
FEV_1, % predicted	0.94 (0.92–0.95)	<0.001
BMI (kg/m^2)	0.87 (0.81–0.93)	<0.001
Burkholderia cepacia colonization		0.007
Test negative	1	
Test positive	3.15 (1.55–6.41)	
No test	1.17 (0.23–5.93)	
Number of intravenous antibiotics courses/year	1.16 (1.07–1.26)	<0.001
Number of days of hospitalization/year	1.17 (1.06–1.28)	0.001
Oral corticosteroids	2.05 (1.25–3.35)	0.004
Long-term oxygen therapy	2.81 (1.83–4.32)	<0.001
Non-invasive ventilation	1.74 (1.01–3.00)	0.04

Furthermore, the risk of death or LT within 3 years increased with number of days of hospitalization per year, number of IV antibiotics courses per year, the use of oral corticosteroids, long-term oxygen therapy and non-invasive ventilation. The Hosmer–Lemeshow test for this model showed good agreement between the predicted and observed values on 95% of the imputed datasets. The average c-statistic of the model across the imputed datasets was 0.91 (95% CI: 0.89–0.93). The model was correctly specified and had an excellent ability to distinguish individuals who experienced the outcome, and those who did not.

Furthermore, we considered the rate of decline in FEV_1 before 2010 as a potential predictor. For each subject, this was estimated by using a linear mixed model with a random intercept and linear slope. The rate of decline of FEV_1 before study entry was not a better predictor than the value of FEV_1 at the study entry. Thus, it was not considered further.

3.3. Prognostic score

We developed a prognostic score, which was a weighted score calculated from the 8 predictors (Table 2). The parameters of the model were reestimated after transforming continuous variables into categorical variables according to clinically relevant thresholds. A value proportional to each estimated parameter was added to the score (Table 3).

Fig. 2 illustrates a nomogram which facilitates the use of the predictive model and the calculation of the prognostic score for each CF patient, given his or her characteristics. The nomogram provides a score value according to the scale given at the top of the figure. All the scores calculated for the eight predictors are then summed to obtain the prognostic score of the patient. The risk of death or LT corresponding to the prognostic score is given by the scale at the bottom of the figure.

A logistic regression using only the prognostic score as predictor of death or LT provided an odds ratio of 2.75 (95%CI 2.47–3.06, p value $<10^{-3}$). The risk of death or LT was greater for higher scores. The fitting of the prognostic score provided an average discriminant ability of 0.91 (95% CI: 0.89–0.92) which indicated good discriminative ability, and thus supports the use of this score to predict death or LT in a 3-year period. The Hosmer–Lemeshow test indicated that there were no significant differences between the prognostic score's predictions and the observed values on all the imputed datasets. These results were confirmed with the cross-validation of the prognostic score, which provided a c-statistic of 0.90 (95% CI: 0.88–0.93).

3.4. Classes of score

Based on risk of death or LT, we classified patients into three groups with low, intermediate and high risk (see Supplementary Fig. 1 for detailed risk of death or LT at each level of the score). From the score 0 to 1.5, the percentages of death or LT were lower than 2%. From the score 2 to 3.5, the percentages of death or LT were ranged between 7% and 15%. Finally, for the scores higher or equal to 4, the percentages of death or LT were ranged between 33% and 100%. Using the above distribution of death or LT in each value of the score, we created three risk groups. In the

Please cite this article as: Nkam L, et al, A 3-year prognostic score for adults with cystic fibrosis, J Cyst Fibros (2017), http://dx.doi.org/10.1016/j.jcf.2017.03.004

Table 3
Risk score for prediction of within 3-year death or lung transplantation in adults with CF.

	Coefficient	Odds ratio (95% CI)	p-value	Score
FEV_1, % predicted			$<10^{-3}$	
> = 60	0	1		0
[30–60]	1.60	4.97 (2.53–9.74)		1.5
<30	2.99	19.91 (9.79–40.49)		3
BMI (kg/m^2)			$<10^{-3}$	
> = 18.5	0	1		0
[16–18.5]	0.70	2.01 (1.41–2.86)		0.5
<16	1.21	3.36 (1.62–6.97)		1
Burkholderia cepacia colonization			0.01	
Test negative	0	1		0
Test positive	1.08	2.96 (1.45–6.03)		1
No test	-0.10	0.90 (0.19–4.35)		0
Number of intravenous antibiotics courses/year			$<10^{-3}$	
0	0	1		0
[1-2]	0.61	1.84 (1.07–3.18)		0.5
>2	1.37	3.92 (2.29–6.72)		1
Hospitalization (yes vs no)	0.33	1.39 (0.95–2.03)	0.09	0.5
Oral corticosteroids (yes vs no)	0.80	2.20 (1.35–3.58)	10^{-3}	1
Long-term oxygen therapy (yes vs no)	1.12	3.08 (2–4.73)	$<10^{-3}$	1
Non-invasive ventilation (yes vs no)	0.69	1.99 (1.16–3.42)	0.01	1

first group (score ≤ 1.5), the score identified a group at very low risk of death or LT (1%) at 3 years. The second group (score [2–3.5]) corresponded to patients at moderate risk of death or LT (10%) at 3 years. Lastly, the highest group (score ≥ 4) identified patients with a very high risk of death or LT (55%) at 3 years (Fig. 3).

3.5. Comparison between the prognostic score and its components

We compared the diagnostic accuracy of the prognostic score with the one provided by each of its components. The discriminant ability of the prognostic score (c-statistic 0.91, 95% CI: 0.89–0.92) was significantly higher than the one provided by each of its components. In particular, it was significantly higher than the discriminant ability obtained with the criterion of FEV_1 < 30% for LT eligibility (c-statistic 0.74, 95% CI: 0.71–0.77) (see Supplementary Table 2 for details).

3.6. Sensitivity analyses

The model obtained after imputation of the outcome of patients lost to follow-up included the same predictors as the model excluding patients lost to follow-up. The characteristics at baseline of these lost to follow-up patients lead us to believe that they were patients with a low risk of death or LT (Supplementary Table 3). Using our prognostic score, we obtained their median

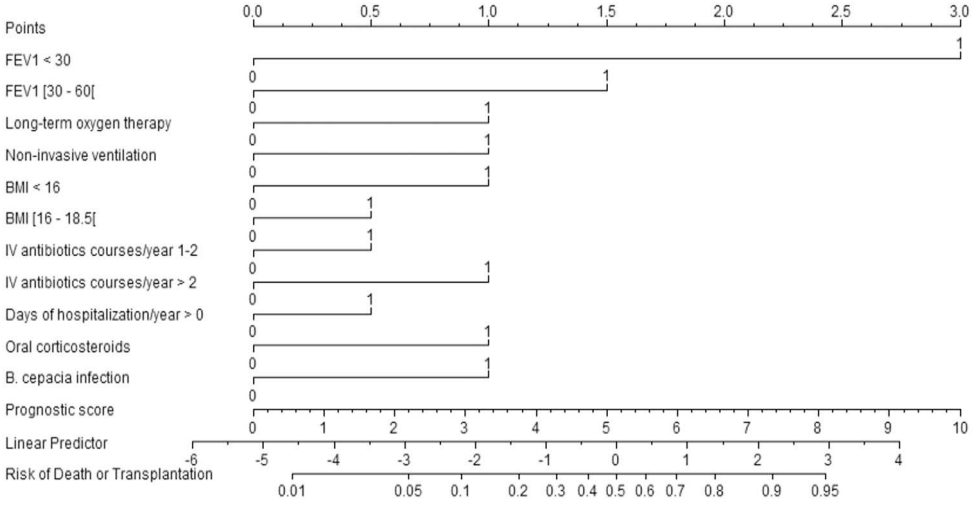

Fig. 2. Nomogram designed to estimate risk of death or lung transplantation. To display the application of the nomogram, we calculated the score and the corresponding risk of death or LT of a given patient. The patient is an 18 years old female gender with a FEV_1 at inclusion of 54% predicted (1.5 points) and a BMI at inclusion of 20.57 kg/m^2 (0 point). She had 2 IV antibiotics courses in the year 2010 (0.5 points). She had no oral corticosteroids (0 point), no long-term oxygen therapy (0 point), no non-invasive ventilation (0 point), no colonization with *Burkholderia cepacia* (0 point) and no days of hospitalization in the year 2010 (0 point). These clinical characteristics correspond to a total of 2 points and a risk of death or LT of 0.04. Thus, the expected risk of death or LT within the next 3 years is estimated at 4%.

Please cite this article as: Nkam L, et al, A 3-year prognostic score for adults with cystic fibrosis, J Cyst Fibros (2017), http://dx.doi.org/10.1016/j.jcf.2017.03.004

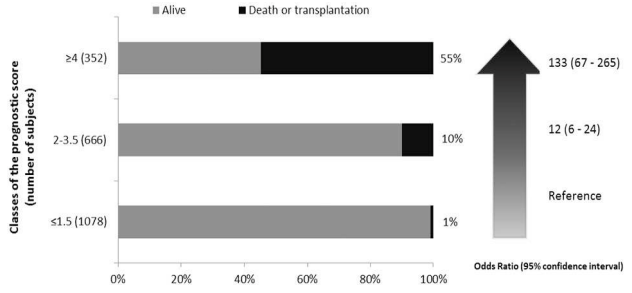

Fig. 3. Three-year risk of death or lung transplantation according to proportion of events in each score.

(interquartile range) score: 0.5 (0–1.5), indicating a low risk of death or LT at 3 years.

Next, we examined the impact of having chosen a composite outcome (occurrence of LT or death without LT) versus analyzing them as separate outcomes. The results of these analyses are presented in Tables 4 and 5 Supplementary materials respectively. Factors associated with LT were remarkably comparable with those obtained for the composite outcome. Factors associated with death without LT likely reflected some of the relative contraindications to transplant (for example cirrhosis, older age, *Burkholderia cepacia* colonization), but no treatment-associated variables.

4. Discussion

We used the most recent data from the French CF Registry to identify current prognostic factors in adults with CF. Eight risk factors were associated with death or LT within 3 years for adults: FEV_1, BMI, *Burkholderia cepacia* colonization, number of IV antibiotics courses per year, number of days of hospitalization per year, use of long-term oxygen therapy, non-invasive ventilation and use of oral corticosteroids. These factors were combined in a single prognostic score which showed good performance in terms of calibration and discrimination. This score allowed the identification of three groups of patients with markedly different risks of poor outcomes: from the lowest to the highest group, there was more than a 50-fold increase in the risk of death or LT. Importantly, the developed score was better at predicting the risk of death or LT than its individual components (especially $FEV_1 < 30\%$), confirming the multidimensional nature of disease severity in CF. These findings confirm data by George et al. who suggested that $FEV_1 < 30\%$ predicted is not by itself sufficient for the identification of patients at risk of poor outcome (i.e., LT or death without LT) in the modern era of CF care [12]. It extends these data by identifying important variables to poor outcomes in adults CF patients and by combining them in a clinically usable score.

Our goal was to examine prognostic factors associated with poor outcomes in adult CF patients, leading to the choice of a combined outcome defined as death without LT or occurrence of LT. Although older studies, which were performed at a time when LT was less developed, have focused on death without considering LT, the choice of studying a combined outcome of death without LT or occurrence of LT was also used in previous study [21]. As sensitivity analyses, we reanalyzed our data using the occurrence of LT and death without LT as separate outcomes. The findings obtained when considering the occurrence of LT as the primary outcome were quite close to those obtained when considering the composite outcome, largely reflecting the fact that 80% of the patients in the combined outcome underwent a LT. Focusing only on death without LT would have biased our analyses towards patients with relative contraindication to transplant and considerably reduced the usefulness of our score, which is intended to be used in all adult patients with CF. Thus, the use of the composite outcome of death without LT or the occurence of LT appeared appropriate in the present era when LT has become the standard of care in CF patients with refractory respiratory insufficiency.

The present study has several strengths. Analyses were performed using data from the French CF Registry that covers around 90% of French CF patients. Our model was developed using a large variety of covariates considered as potential predictors of poor outcome in patients with CF. The continuous variables included in the prognostic score were categorized using predetermined (based on clinical knowledge) cut-offs, which appeared easier for use in daily practice. The study focused on adults with CF because death from CF in children is now very rare in developed countries [26], and because over 95% of LT in France are performed in adults. Missing covariates data, a common issue in observational studies, were handled using the multiple imputation approach, which helped to minimize data loss, avoid biased estimates, and provide better estimations [24]. We also recognize limitations to the present study. Some covariates, such as professional status and marital status, available in the French CF Registry were not included in the models because of a high percentage of missing data. Additionally, variables not collected in the French CF Registry such as physical activity and exercise capacity, pulmonary hemodynamics [22,28] may also be important in assessing the prognosis of CF adults. On the other hand, these variables may not be routinely obtained in all CF patients, and thus may not be appropriate for inclusion in a score designed for referring patients for specialized evaluation for LT. Patients who received combined lung-liver transplantation during the follow-up were considered as having an outcome, whereas patients who received isolated liver transplantation during the follow-up were not considered as having a poor outcome. Because isolated liver transplantation was performed in a very small number of patients (as less than 5 patients per year received liver or lung/liver transplantation), this choice was unlikely to alter our conclusions. The 184 patients who were lost to follow-up (i.e., those with an unknown outcome at the end of the study) were excluded from the analyses. Analyses of the clinical characteristics of these patients suggested that, most patients lost to follow-up were those with milder disease and were at very low risk of death or LT. Additionally, sensitivity analyses suggested that the exclusion of these patients had only limited impact on our results. The vital status of patients corresponded to the one at the end of the year, and the use of annual data did not exclude the possibility that the prognosis could have changed considerably between the last visit and the end of the year. Lastly, the prognostic score included variables related to the patient's status (FEV_1, BMI, *Burkholderia*

Please cite this article as: Nkam L, et al, A 3-year prognostic score for adults with cystic fibrosis, J Cyst Fibros (2017), http://dx.doi.org/10.1016/j.jcf.2017.03.004

cepacia) and also variables related to therapeutic interventions (hospitalization, IV antibiotics, oral corticosteroids, non-invasive ventilation, long-term oxygen therapy) that rely on physician choices. Future studies should aim address the validity of this score in other countries with different healthcare systems. Additionally, the value of assessing longitudinal changes (e.g., related to treatment modifications) in the score in individual patients should be evaluated in the future.

The prognostic score was built after reevaluation of prognostic factors in adult patients with CF, to predict the risk of death or LT in a 3-year period. This score showed good performance and was significantly better in terms of discrimination than its components, including the criterion of FEV_1 lower than 30% proposed for LT eligibility. If validated in other settings, this score could be a useful tool in the future for selecting patients requiring an evaluation for LT.

Conflicts of interest

None

Acknowledgments

This work was supported by a research grant from the *Vaincre La Mucoviscidose* Association.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jcf.2017.03.004.

References

[1] Aaron SD, Stephenson AL, Cameron DW, Whitmore GA. A statistical model to predict one-year risk of death in patients with cystic fibrosis. J Clin Epidemiol 2015;68:1336–45.
[2] Bellis G, Cazes MH, Parant A, Gaimard M, Travers C, Le Roux E, et al. Cystic fibrosis mortality trends in France. J Cyst Fibros 2007;6:179–86.
[3] Bellis G, Dehillotte C, Lemonnier L. Registre Français De La Mucoviscidose. Bilan Des Données 2015. Vaincre la Mucoviscidose et Institut national d'études démographiques; 2015.
[4] Buzzetti R, Alicandro G, Minicucci L, Notarnicola S, Furnari ML, Giordano G, et al. Validation of a predictive survival model in Italian patients with cystic fibrosis. J Cyst Fibros 2012;11:24–9.
[5] Metz CF. Basic principles of roc analysis. Semin Nucl Med 1978;8:283–98.
[6] Courtney JM, Bradley J, McCaughan J, O'Connor TM, Shortt C, Bredin CP, et al. Predictors of mortality in adults with cystic fibrosis. Pediatr Pulmonol 2007;42:525–32.
[7] Dasenbrook EC, Konstan MW, VanDevanter DR. Association between the introduction of a new cystic fibrosis inhaled antibiotic class and change in prevalence of patients receiving multiple inhaled antibiotic classes. J Cyst Fibros 2015;14:370–5.
[8] Dodge JA, Lewis PA, Stanton M, Wilsher J. Cystic fibrosis mortality and survival in the Uk: 1947–2003. Eur Respir J 2007;29:522–6.
[9] Elborn JS. Cystic fibrosis. Lancet 2016.
[10] Flume PA, Strange C, Ye X, Ebeling M, Hulsey T, Clark LL. Pneumothorax in cystic fibrosis. Chest 2005;128:720–8.
[11] Flume PA, Yankaskas JR, Ebeling M, Hulsey T, Clark LL. Massive hemoptysis in cystic fibrosis. Chest 2005;128:729–38.
[12] George PM, Banya W, Pareek N, Bilton D, Cullinan P, Hodson ME, et al. Improved survival at low lung function in cystic fibrosis: cohort study from 1990 to 2007. BMJ 2011;342:d1008.
[13] Hayllar KM, Williams SG, Wise AE, Pouria S, Lombard M, Hodson ME, et al. A prognostic model for the prediction of survival in cystic fibrosis. Thorax 1997;52:313–7.
[14] Hirche TO, Knoop C, Hebestreit H, Shimmin D, Sole A, Elborn JS, et al. Practical guidelines: lung transplantation in patients with cystic fibrosis. Pulm Med 2014;2014:621342.
[15] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16:965–80.
[16] Kerem E, Reisman J, Corey M, Canny GJ, Levison H. Prediction of mortality in patients with cystic fibrosis. N Engl J Med 1992;326:1187–91.
[17] Liou TG, Adler FR, Fitzsimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-year survivorship model of cystic fibrosis. Am J Epidemiol 2001;153:345–52.
[18] MacKenzie T, Gifford AH, Sabadosa KA, Quinton HB, Knapp EA, Goss CH, et al. Longevity of patients with cystic fibrosis in 2000 to 2010 and beyond: survival analysis of the Cystic Fibrosis Foundation patient registry. Ann Intern Med 2014;161:233–41.
[19] Martin C, Hamard C, Kanaan R, Boussaud V, Grenet D, Abely M, et al. Causes of death in French cystic fibrosis patients: the need for improvement in transplantation referral strategies! J Cyst Fibros 2016;15:204–12.
[20] Mayer-Hamblett N, Rosenfeld M, Emerson J, Goss CH, Aitken ML. Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality. Am J Respir Crit Care Med 2002;166:1550–5.
[21] McCarthy C, Dimitrov BD, Meurling IJ, Gunaratnam C, McElvaney NG. The Cf-able score: a novel clinical prediction rule for prognosis in patients with cystic fibrosis. Chest 2013;143:1358–64.
[22] Schneiderman JE, Wilkes DL, Atenafu EG, Nguyen T, Wells GD, Alarie N, et al. Longitudinal relationship between physical activity and lung health in patients with cystic fibrosis. Eur Respir J 2014;43:817–23.
[23] Stephenson AL, Tom M, Berthiaume Y, Singer LG, Aaron SD, Whitmore GA, et al. A contemporary survival analysis of individuals with cystic fibrosis: a cohort study. Eur Respir J 2015;45:670–9.
[24] Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 2007;60:979.
[25] Thabut G, Christie JD, Mal H, Fournier M, Brugiere O, Leseche G, et al. Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation. Am J Respir Crit Care Med 2013;187:1335–40.
[26] Urquhart DS, Thia LP, Francis J, Prasad SA, Dawson C, Wallis C, et al. Deaths in childhood from cystic fibrosis: 10-year analysis from two London specialist centres. Arch Dis Child 2013;98:123–7.
[27] van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67.
[28] Venuta F, Rendina EA, Rocca GD, De Giacomo T, Pugliese F, Ciccone AM, et al. Pulmonary hemodynamics contribute to indicate priority for lung transplantation in patients with cystic fibrosis. J Thorac Cardiovasc Surg 2000;119:682–9.

Supplementary materials

Table 1: Additional covariates to those in table 1, considered as potential predictors, French CF Registry, 2010

Intravenous antibiotic courses in the current year (yes/no)
Genotype (Severe, Moderate, Unknown)
Cancer
Number of day admissions
Total number of days of hospitalization
Rate of decline of FEV_1
Number of outpatient visits
PaO2 blood gas
PaCO2 blood gas
SpO2 blood gas
SaO2 blood gas

Table 2: Discriminative abilities of the prognostic score, its components and FEV_1 <30% criterion

	c-statistic (95%CI)
Score	0.91 (0.89 - 0.92)
Individual components of the score	
FEV_1, % predicted	0.84 (0.82 - 0.86)
Number of intravenous antibiotics courses/year	0.78 (0.75 - 0.81)
Long term oxygen therapy	0.72 (0.69 - 0.75)
Hospitalization	0.67 (0.64 - 0.70)
BMI (kg/m^2)	0.67 (0.64 - 0.70)
Non-invasive ventilation	0.63 (0.60 - 0.66)
Oral corticosteroids	0.55 (0.53 - 0.58)
Burkholderia cepacia colonization	0.54 (0.52 - 0.55)
FEV_1 <30% criterion	0.74 (0.71 - 0.77)

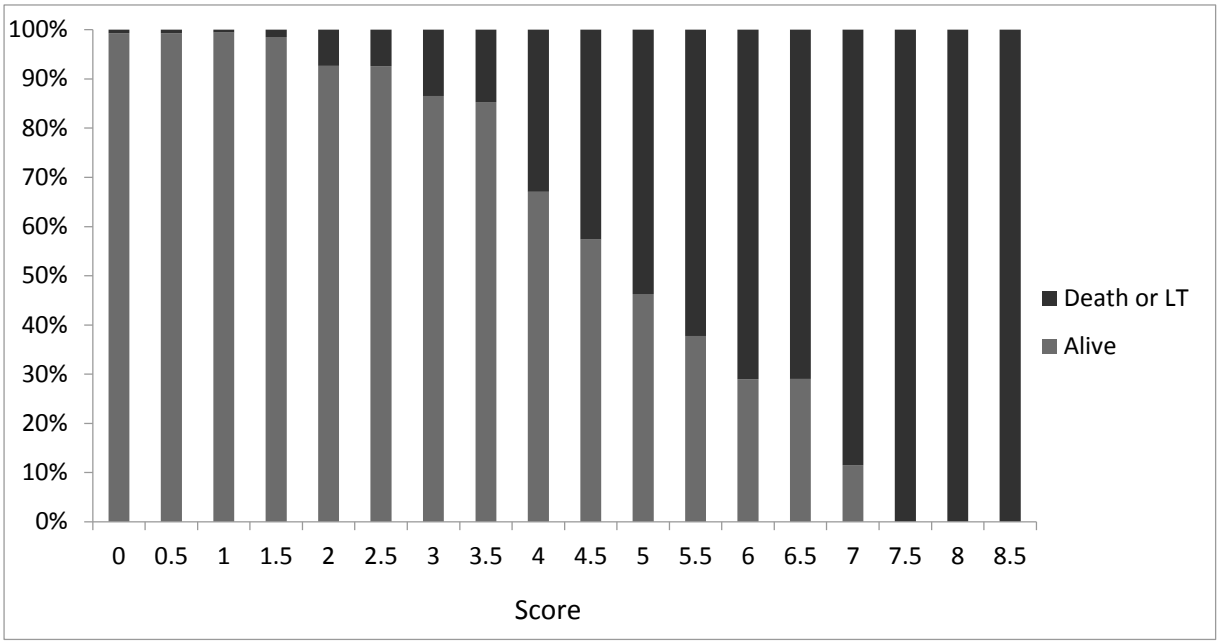

Figure 1 : Distribution of death or lung transplantation (LT) in each level of the prognostic score

Score	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	Total
Alive	430	274	165	200	190	188	135	87	55	39	25	17	11	9	3	0	0	0	**1828**
Death or LT	3	2	1	3	15	15	21	15	27	29	29	28	27	22	23	3	4	1	**268**
Total	433	276	166	203	205	203	156	102	82	68	54	45	38	31	26	3	4	1	**2096**
Percentage of Death or LT	0.7	0.7	0.6	1.5	7.3	7.4	13.5	14.7	32.9	42.6	53.7	62.2	71.1	71	88.5	100	100	100	**12.8**

Table 3: Characteristics of the patients lost to follow-up and those included in the study in 2010

Patient characteristics	Total n = 2,096 (%)	Patients lost to follow-up n = 184 (%)	p value
Gender, Male	1101 (52.5)	106 (57.6)	0.2125
CFTR genotype			<0.001
Classes 1-3	1,373 (65.5)	90 (48.9)	
Classes 4/5	254 (12.1)	27 (14.7)	
Classes unknown	469 (22.4)	67 (36.4)	
Airway colonization	2,013 (96.0)	127 (69)	
Pseudomonas aeruginosa	1,319 (65.5)	59 (46.5)	<0.001
Staphylococcus aureus	1,290 (64.1)	61 (48)	<0.001
Comorbidities	2,007 (95.8)	149 (81)	<0.001
Pancreatic insufficiency	1,758 (87.6)	109 (73.2)	<0.001
Allergic bronchopulmonary aspergillosis	378 (18.9)	9 (6.1)	<0.001
Hemoptysis	204 (10.2)	8 (5.5)	<0.001
FEV_1, % predicted*	58.3 (39.4 – 79.8)	82.7 (60.3 – 98.4)	<0.001
FVC, % predicted*	79.0 (62.2 – 95.2)	90.7 (73.8 – 100.8)	<0.001
Age*	25.5 (21 – 32.3)	26 (22 – 36)	0.1053
BMI (kg/m^2)*	20.2 (18.5 – 22.1)	20.8 (18.7 – 23.5)	0.3052
Number of IV antibiotics courses/year*	1 (0 – 2)	0 (0 – 0)	<0.001
Number of IV antibiotics days/year*	14 (0 – 30)	0 (0 – 0)	<0.001
Number of days of hospitalization/year*	0 (0 – 23)	0 (0 – 0.25)	<0.001
Azithromycin	1,254 (59.8)	64 (35.4)	<0.001
Non-invasive ventilation	127 (6.1)	5 (2.7)	0.0898
Long-term oxygen therapy	230 (11.0)	7 (4)	0.0047
Inhaled therapies	1,873 (89.4)	109 (59.2)	<0.001
Inhaled antibiotics	1122 (59.9)	53 (48.6)	<0.001
Inhaled corticosteroids	1012 (54.0)	61 (56)	<0.001

* Continuous variables, median (interquartile range)
FEV_1%: forced expiratory volume in one second as percentage of predicted values
FVC %: forced vital capacity as percentage of predicted values
BMI: body mass index
IV: intravenous
CFTR: cystic fibrosis transmembrane conductance regulator

Table 4: Logistic regression model for prediction of within 3-year lung transplantation in adults with CF

	Odds Ratio (95% CI)	p value
FEV_1, % predicted	0.92 (0.91 - 0.94)	<0.001
BMI (kg/m^2)	0.89 (0.82 - 0.96)	0.001
Pancreatic insufficiency		0.020
No	1	
Yes	3.93 (1.28 – 12.11)	
Unknown	0.75 (0.07 – 8.64)	
Number of intravenous antibiotics courses/year	1.17 (1.07 - 1.27)	<0.001
Number of days of hospitalization/year	1.21 (1.10 - 1.33)	<0.001
Oral corticosteroids	1.98 (1.16 – 3.37)	0.017
Long-term oxygen therapy	2.90 (1.90 - 4.43)	<0.001

Number of patients: 2096
Number of events (lung transplantation): 213

Table 5: Logistic regression model for prediction of within 3-year death without lung transplantation in adults with CF

	Odds Ratio (95% CI)	p value
FEV_1, % predicted	0.97 (0.95 - 0.99)	<0.001
BMI (kg/m^2)	0.84 (0.74 - 0.96)	0.011
Age*	1.97 (1.92 – 2.02)	
Burkholderia cepacia colonization		<0.001
Test negative	1	
Test positive	6.26 (2.67 – 14.68)	
Cirrhosis		0.021
No	1	
Yes	3.40 (1.41 – 8.21)	
Unknown	0.59 (0.07 - 4.70)	

***Odds ratio for an increase of 10 years**
Number of patients: 2096
Number of events (death without lung transplantation): 55

3.2 Compléments

Dans ce travail, l'identification des facteurs associés au décès sans transplantation pulmonaire ou à la transplantation pulmonaire a été réalisée via une sélection ascendante, à partir d'une quarantaine de variables considérées comme prédicteurs potentiels. Les variables étaient introduites dans le modèle selon la valeur de leur p value. Une imputation multiple a été appliquée aux données pour éviter une perte d'information. Pour cela, 20 jeux de données et 50 itérations ont été considérés. Pour chaque variable, la convergence des imputations a été vérifiée en comparant les densités de probabilité des données observées avec celles des données imputées. Les performances prédictives du modèle développé ont été évaluées grâce à l'AUC pour la discrimination et le test d'Hosmer-Lemeshow pour la calibration. Un programme permettant de moyenner l'AUC sur les tables de données imputées a été appliqué. Les données de registre français de la mucoviscidose étaient figées jusqu'en 2013. Pour cette raison, nous avons considéré les données de l'année 2010 pour la prédiction à 3 ans.

À partir des données récentes du registre français de la mucoviscidose, ce travail a permis d'identifier les facteurs pronostiques actuels de la maladie. Parmi les facteurs pronostiques identifiés, on en retrouve qui ont déjà été identifiés comme étant associés au décès chez les sujets atteints de mucoviscidose. Il s'agit du VEMS, de l'IMC, de la colonisation à *Burkholderia cepacia*, du nombre de cures d'antibiotiques dans l'année précédente et du nombre de jours d'hospitalisation dans l'année précédente. Ceci permet de confirmer les résultats de la littérature. Par ailleurs, ce modèle a également permis d'identifier des paramètres qui, jusque-là n'avaient pas été identifiés dans les modèles pronostiques, à notre connaissance. Ce sont la ventilation nasale, l'oxygénothérapie et la prise de corticoïdes oraux. Ce sont des traitements particulièrement attribués aux malades ayant une insuffisance respiratoire. En effet, la ventilation nasale est généralement prescrite aux malades ayant une insuffisance respiratoire sévère et qui sont amenés à recevoir une greffe ou alors qui sont en évaluation pour l'obtention d'une greffe pulmonaire [Madden et al. 2002]. La ventilation nasale constitue souvent un « pont » vers la transplantation pulmonaire [Hodson et al. 1991] et permet de stabiliser le déclin de la fonction pulmonaire chez les malades ayant une fonction respiratoire dégradée [Fauroux et al. 2008]. De même, l'oxygénothérapie est prescrite aux sujets ayant une insuffisance respiratoire sévère. Les malades sous oxygénothérapie sont des malades éligibles à la transplantation pulmonaire [Orens et al. 2006]. Les corticoïdes oraux sont prescrits dans le but d'améliorer la fonction pulmonaire et la qualité de vie des malades et également pour

réduire les exacerbations pulmonaires [Flume et al. 2007; Mogayzel et al. 2013].

Nous avons identifié les facteurs associés au décès sans transplantation pulmonaire ou à la transplantation pulmonaire comme événement composite. Nous avons effectué des analyses supplémentaires dans le but de comparer les résultats obtenus en utilisant l'événement composite avec ceux obtenus en utilisant chacun des événements séparément. Les résultats obtenus en utilisant uniquement la transplantation pulmonaire comme événement d'intérêt étaient très proches des résultats obtenus en utilisant l'événement composite. Ceci est vraisemblablement dû au fait que 80% des sujets ayant eu l'événement composite étaient des sujets transplantés. Cependant, en utilisant uniquement le décès comme événement d'intérêt, l'âge apparaît comme associé au risque de décès. Certes la maladie a tendance à s'aggraver avec l'âge, mais un âge avancé est souvent une contrindication à la transplantation pulmonaire [Flume et al. 2007; Hook and Lederer 2012; Mogayzel et al. 2013]. De plus, il n' y a pas d'association entre le décès et les facteurs associés à une éventuelle sévérité de la maladie tels que les hospitalisations, les cures, l'oxygénothérapie.

Pour l'identification des facteurs pronostiques, nous avons utilisé un modèle de régression logistique. Nous aurions pu utiliser un modèle de survie et s'intéresser au délai écoulé jusqu'à la survenue de l'événement. Seulement les sujets ont des visites annuelles et la fenêtre de prédiction considérée est de 3 ans. La régression logistique nous a semblé mieux appropriée dans ce contexte, sachant qu'on est sur une courte période d'exposition et le temps écoulé jusqu'à la survenue de l'événement a peu d'intérêt. Un modèle de survie aurait été utile si le temps de suivi était plus long.

Certes le modèle logistique est relativement simple à interpréter pour les cliniciens, mais son usage pour modéliser les probabilités de décès pourrait donner des prédictions biaisées en présence de suivi incomplet [Szczesniak et al. 2017]. Pour gérer le problème de perdus de vue dans notre étude, nous avons opté pour une imputation du statut des sujets perdus de vue. Les résultats obtenus avec ou sans perdus de vue étaient similaires. L'étude a tout de même était réalisée à partir d'un modèle à risques proportionnels et nous avons obtenu les résultats similaires à ceux obtenus en utilisant le modèle logistique. Les rapports de cotes et les risques relatifs étaient similaires et l'effet des facteurs pronostiques sur le risque d'événement évoluaient de la même manière avec les deux approches. Cependant, avec le modèle à risques proportionnels, la ventilation nasale n'était pas associée au risque de survenue de l'événement.

De plus, nous avons considéré le cas des risques compétitifs. En termes de variables

TABLE 3.1 – *Modèle de Cox pour l'identification des facteurs associés au décès sans transplantation pulmonaire ou à la transplantation pulmonaire chez les sujets adultes atteints de mucoviscidose, N= 2280*

	Risque relatif (95% CI)	**p value**
VEMS, % valeurs prédites	0.95 (0.94 - 0.96)	$< 10^{-3}$
IMC (kg/m^2)	0.88 (0.83 - 0.93)	$< 10^{-3}$
Burkholderia cepacia		0.004
Test négatif	1	
Test positif	2.48 (1.51 – 4.07)	
Pas de test	0.65 (0.16 – 2.71)	
Nombre de cures intraveineuse par année	1.12 (1.07 - 1.17)	$< 10^{-3}$
Nombre de jours d'hospitalisation par année	1.11 (1.05 - 1.18)	$< 10^{-3}$
Corticoïdes oraux	1.62 (1.16 – 2.27)	0.004
Oxygénothérapie	2.13 (1.56 – 2.91)	$< 10^{-3}$
Ventilation nasale	1.28 (0.91 – 1.78)	0.1

associées à chacune des causes, nous avons obtenus des résultats similaires aux régressions logistiques en considérant les événements séparément. Ci-dessous, nous avons les résultats du modèle pour la transplantation pulmonaire et pour le décès sans transplantation pulmonaire.

TABLE 3.2 – *Résultat du modèle de risque cause-spécifique pour la transplantation pulmonaire chez les sujets adultes atteints de mucoviscidose*

	Risque relatif (95% CI)	**p value**
VEMS, % valeurs prédites	0.94 (0.93 - 0.95)	$< 10^{-3}$
IMC (kg/m^2)	0.88 (0.82 - 0.95)	$< 10^{-3}$
Burkholderia cepacia		0.37
Test négatif	1	
Test positif	1.69 (0.85 – 3.38)	
Pas de test	1.01 (0.24 – 4.24)	
Nombre de cures intraveineuse par année	1.11 (1.06 - 1.17)	$< 10^{-3}$
Nombre de jours d'hospitalisation par année	1.11 (1.03 - 1.19)	0.003
Corticoïdes oraux	1.66 (1.12 – 2.45)	0.01
Oxygénothérapie	2.30 (1.60 – 3.30)	$< 10^{-3}$
Ventilation nasale	1.16 (0.79 – 1.71)	0.45

TABLE 3.3 – *Résultat du modèle de risque cause-spécifique pour le décès sans transplantation pulmonaire chez les sujets adultes atteints de mucoviscidose*

	Risque relatif (95% CI)	**p value**
VEMS, % valeurs prédites	0.96 (0.94 - 0.97)	$< 10^{-3}$
IMC (kg/m^2)	0.88 (0.80 - 0.98)	0.02
Burkholderia cepacia		$< 10^{-3}$
Test négatif	1	
Test positif	4.69 (2.25 – 9.75)	
Pas de test	0 (0 - 0)	
Nombre de cures intraveineuse par année	1.16 (1.04 - 1.28)	0.005
Nombre de jours d'hospitalisation par année	1.10 (0.98 - 1.24)	0.11
Corticoïdes oraux	1.51 (0.77 – 2.93)	0.22
Oxygénothérapie	1.64 (0.88 – 3.06)	0.12
Ventilation nasale	1.66 (0.87 – 3.19)	0.12

Chapitre 4

Profils d'évolution de la maladie et prédictions dynamiques

La mucoviscidose est une maladie rare qui est incurable à ce jour. Le moyen le plus adéquat actuellement pour prolonger la vie des patients dont l'état respiratoire s'est considérablement dégradé est la transplantation pulmonaire. Une bonne connaissance des facteurs associés au décès est indispensable pour améliorer l'identification des patients pour la transplantation pulmonaire. En effet, les critères d'identification des patients pour la transplantation pulmonaire varient d'un centre de transplantation à l'autre. De plus, le pronostic des patients atteints de mucoviscidose s'est considérablement amélioré durant les précédentes années. Des thérapies nouvelles ont été mises en place pour améliorer les conditions de vie des malades. Dans le contexte actuel de la mucoviscidose, il nous a paru intéressant de réévaluer les facteurs pronostiques associés à la transplantation pulmonaire ou au décès sans transplantation pulmonaire. Parmi les facteurs identifiés, on retrouve le VEMS qui décrit la fonction pulmonaire. Ce paramètre est indispensable pour le suivi des malades car il est un bon indicateur de la progression de la maladie.

À partir du modèle logistique développé dans l'article précédent, il est tout à fait possible de déterminer le risque qu'a un malade de décéder ou de recevoir une greffe sur une période de 3 ans. Ce risque est constant et calculé à partir des valeurs initiales du VEMS, entre autres. Cependant, il est possible de fournir, à partir des modèles conjoints, des risques pour le décès sans transplantation pulmonaire ou la transplantation pulmonaire qui évoluent dans le temps. Grâce à la modélisation conjointe de l'évolution du VEMS et de la survenue du décès sans transplantation pulmonaire ou de la transplantation pulmonaire, on obtient des prédictions individuelles dynamiques du risque. Dans ce cas, le risque prédit est mis à jour après chaque nouvelle valeur du VEMS. Cette approche permettra de mieux évaluer l'impact du VEMS sur le

risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire.

L'objectif de ce travail est de développer un outil pronostique fournissant des prédictions dynamiques du risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire à partir des mesures répétées du VEMS. Le modèle développé est un modèle conjoint à classes latentes qui permet d'identifier différents profils d'évolution de la maladie.

4.1 Manuscrit en cours

Dynamic prediction of survival in adults with Cystic Fibrosis

L. Nkam[1]PhD, PR. Burgel[2,3]MD PhD, M.N. Hocine[1]PhD
[1]Laboratoire Modélisation, Epidémiologie et Surveillance des Risques Sanitaires, Conservatoire National des Arts et Métiers, Paris, France
[2]Assistance Publique Hôpitaux de Paris, Hôpital Cochin, Service de Pneumologie, Paris, France
[3]Université Paris Descartes, Sorbonne Paris Cité, Paris, France

Correspondence to : L. Nkam, Conservatoire National des Arts et Métiers, Laboratoire Modélisation, Epidémiologie et Surveillance des Risques Sanitaires, 75003 Paris, France, lionelle.nkam@cnam.fr, +331 53 01 80 69

Abstract

Background Cystic Fibrosis (CF) is a rare genetic disease that particularly affects the lungs and leads to progressive loss in lung function and premature death. Lung transplantation (LT) is a potentially life-saving treatment option for CF patients with terminal respiratory failure. However, determining the optimal timing for LT referral has proven challenging due to heterogeneity in CF evolution and difficulties in predicting prognosis. The aim of this work was to predict dynamically the prognosis in adults with CF taking into account the temporal evolution of forced expiratory volume in 1 second (FEV_1), and to identify different profiles of disease progression.

Methods We developed a joint latent class model (JLCM) and a landmark approach to provide dynamic predictions of the risk of death without LT or the occurrence of LT in adults with CF, using the temporal evolution of FEV_1. In addition, the JLCM allowed to classify patients into prognostic classes considering jointly the prognosis and FEV_1 evolution over time.

Results We used data from the French CF Registry, which included 1806 CF adults, among whom 5.7% died without LT and 17.9% received LT between 2007 and 2013. The dynamic models provided better predictive performances for patient prognosis compared to standard survival analyses, with prediction errors lower than 0.10 and discrimination values up to 0.91. Three classes of patients with poor, moderate and good prognosis were identified. These classes represented respectively 34%, 43%, and 23% of patients; among them 52%, 14%, and 0.4% died without LT or received LT during follow-up.

Interpretations Dynamic predictions models, which use longitudinal evolution of FEV_1 at each encounter, represent an improvement in predicting CF prognosis. These models may represent useful tools for the identification of CF patients requiring referral for LT and for implementing risk-adapted treatment.

Funding Association Vaincre la Mucoviscidose

Key words: cystic fibrosis, prognosis, lung transplantation, dynamic prediction

Introduction

Cystic fibrosis (CF) is a complex genetic disorder that particularly affects the lungs and the digestive tract. The disease which often leads to progressive decline in lung function and respiratory failure, remains the most common cause of death in patients with CF (1). Lung transplantation (LT) is proposed to patients with terminal respiratory failure with the aim of improving survival and quality of life (2, 3). Although criteria for referring CF patients for LT have been proposed (4), determining the optimal timing for LT referral remains a challenge (5). Thus, developing tools that would help physicians to determine the optimal timing for LT referral appears important.

Forced expiratory volume in one second (FEV_1, usually expressed as percent of predicted -pp- values), a measurement of lung function, is considered one of the most important predictor of death in patients with CF (6). Most statistical models that have been developed to identify prognostic factors in patients with CF relied on cross-sectional analyses, focusing on a single value of FEV_1 at inclusion to estimate the risk of death in patients with CF (7-11). An important drawback with this approach is that CF patients have currently improved survival at low lung function as compared to the past decades (12). In addition, longitudinal variations in FEV_1 may affect patient prognosis. Therefore, survival prediction needs to be updated and analyzing repeated measurements of FEV_1 may provide a better description of the progression of CF lung disease with the potential to provide more accurate information on the optimal timing for LT referral. Such endogenous covariate is measured with error and is related to the risk of event. Therefore, it cannot be considered as a time-varying covariate in standard survival analysis(13). The use of statistical approaches such as joint models (14-16) or landmark concept (17) represent interesting alternatives. Joint models couple a survival model, which describes the occurrence of the clinical event of interest (e.g., death without LT or the occurrence of LT), and a model that describes the evolution of a longitudinal marker (e.g., FEV_1). The landmark approach is a way of handling time-dependent covariates in survival models. The aim of the present study was to assess the use of the repeated measurements of FEV_1 in predicting the risk of death without LT or the occurrence of LT in adults with CF. This study also aimed to identify classes of patients with different prognosis profiles.

Methods

Subjects and data collection

This study was based on longitudinal follow-up of patients included in the French CF Registry. This registry contains data on more than 8000 subjects over the period 1992-2016, which represents at least 90% of all CF patients in France. Data are collected annually and the registry contains reliable information on vital status, therapeutic management, anthropometry, spirometry, hospitalizations, microbiological tests, transplantation, morbidity and sociodemographic data (18).

The developed models considered eight variables that were associated to an increased risk of "death without LT or occurrence of LT", the composite event of interest, in a recent analysis of the French CF registry, which aimed to reevaluate prognostic factors in adults with CF (11). These variables were $ppFEV_1$, body mass index (BMI), Burkholderia cepacia colonization, number of intravenous (IV) antibiotics courses per year, number of days of hospitalization per year, use of long-term oxygen therapy or non-invasive ventilation and use of oral corticosteroids $\geq$ 3 months/year.

We used data collected between December 31st 2007 and December 31st 2013. Because pediatric mortality has almost disappeared in CF patients in developed countries (19), we restricted the analyses to the adult population (age$\geq$18 years). However, delayed entry occurred and was handled for patients followed after 18 years. Patients who received LT before the study period and those with one measure of FEV_1 were excluded. Because LT considerably modifies FEV_1 values, measurements acquired after LT were not considered. Figure S1 in the Supplementary Appendix presents the selection scheme of patients included in the study.

Statistical approaches

We used two distinct statistical approaches that provided individual dynamic predictions of death without LT or LT, taking into account the temporal evolution of FEV_1in adults with CF.

The first approach is the joint model which describes jointly the evolution of the main marker of the disease (e.g., FEV_1), and the occurrence of the event of interest (e.g., death without LT or the occurrence of LT). At a given time of prediction, patient's characteristics at the beginning of the study and longitudinal measures of the marker are used together to predict the risk of the event of interest within a prediction horizon.
We used a joint latent class model (JLCM), a particular case of joint models which is made up of three submodels. The first submodel described the risk of death without LT or occurrence of LT according to age and adjusted for the eight afore-mentioned variables, using a proportional hazard model. Weibull baseline risk function was assumed for the survival model. The second submodel described the evolution of $ppFEV_1$ over time using a linear mixed model, and adjusted for the initial level of BMI and the initial number of IV antibiotics courses per year, a marker for severe pulmonary exacerbation. In the JLCM, these two submodels are linked through a discrete latent variable which define the underlying population subgroups. Indeed, JLCM assume that the population is heterogeneous and can be divided into a finite number of homogeneous subgroups called latent classes. Each class provides a homogeneous mean profile of the marker evolution and a homogeneous risk of the event of interest. For a given patient, the probability to belong to a specific class is given using a multinomial logistic model, a third JLCM submodel (20, 21).We did not consider explanatory covariates in this multinomial logistic submodel.
The optimal number of classes defining the CF population, corresponded to the model with the lowest Bayesian Information Criterion (BIC) (22). Each patient was classified in the class where its mean posterior probability was the highest. We compared the characteristics of subjects at the beginning of the study according to the identified classes, using a chi square test for categorical covariates and using ANOVA for continuous covariates.

The second approach is landmarking, which provides dynamic predictions of the event of interest by considering endogenous time-dependent covariates in a standard survival model. The principle is to consider the subset of subjects at risk at a given time of prediction and the value of the time-dependent covariate (FEV_1) as a time fixed covariate in a survival model to predict the risk of event within a horizon of prediction (17). As the time of prediction increases, the subset of subjects at risk, as well as the available information on patients are updated, leading to individual dynamic predictions. In this approach, the risk of death without LT or the occurrence of LT was described according to age and adjusted for the eight afore-mentioned variables, using a proportional hazard model.

Model predictive performances
The predictive performances of the developed models were evaluated in terms of discrimination and calibration (23).The discrimination assesses how well the model can predict a high risk of event occurrence in subjects who are more likely to experience it, compared to those who are less likely to experience it. The discriminative ability of models was assessed by the time-dependent area under the receiver operating curve (AUC). The closer is this measure to 1, the better is the discriminative ability (23). The calibration measures the predictive ability of models. It was assessed by the time-dependent Brier Score (BS) which compares at a given time point, the predicted and the observed survival status. Low values of the BS indicate a good predictive performance of the model (23). The AUC and the BS were provided at different times of prediction in a prediction horizon of 3 years.

To assess the ability of the JLCM to provide an accurate classification of patients into the identified classes, posterior probabilities were evaluated. They were obtained from the estimated parameters of the probability to belong to a specific class mentioned previously. In addition, the weighted mean subject-specific predictions from the mixed model were compared with the weighted mean observed values of FEV_1 according to age in each latent class to evaluate the goodness-of-fit of the JLCM.

Additional analyses
In order to evaluate the impact of using longitudinal measures of FEV_1 on prognosis prediction, we compared the predictive performances of longitudinal models (JCLM and landmark) to those of two proportional hazard models (C1 and C2) that used only data collected at baseline. Model C1adjusted for FEV_1 at baseline only and model C2 adjusted for the eight afore-mentioned covariates at baseline.
Because LT is considered the standard of care to improve longevity and quality of life for CF patient with terminal respiratory failure, our main analysis was based on a composite outcome (death without LT or occurrence of LT), as discussed in details in our previous report (11). As a sensitivity analysis, we reanalyzed the

data using death without LT as the primary outcome and compared the predictive performances of this analysis to our main analysis.

Results

Data description

Data included 1806 adult patients from the French CF Registry who were followed between December 31st, 2007 and December 31st, 2013. A total of 102 (5.7%) patients died without LT and 324 (17.9%) patients received LT within this follow-up period. The median (interquartile range) number of FEV_1 measurements was 7 (5 - 7). Table 1 presents the baseline characteristics of patients who died without LT or received LT compared to those of patients who remained alive without LT. The population contained 53% males and 66% of subjects with a severe cystic fibrosis transmembrane conductance regulator (CFTR) genotype (classes of CFTR mutation I – III). Subjects were often colonized with Staphylococcus aureus and/or Pseudomonas aeruginosa and were mostly pancreatic insufficient. Most of the subjects received inhaled antibiotics (60%) and azithromycin (57%). The mean±SD age was 27±9 years and $ppFEV_1$ was 60±26%.

Table 1 shows that except for gender and age at baseline, all characteristics were significantly different between subjects who died without LT or received a LT compared with subjects who did not. Subjects who died without LT or received a LT had lower $ppFEV_1$ and BMI at baseline and were more likely to have comorbidities and to be colonized with pathogens. They were also more likely to receive azithromycin, non-invasive ventilation, long-term oxygen therapy, corticosteroids, inhaled therapies, and they had received more days of IV courses in the previous year.

Table 1: Baseline sociodemographic and clinical characteristics of the study patients

Patient characteristics	All n = 1806 (%)	Alive without LT n = 1380 (%)	Death without LT or occurrence of LT n = 426 (%)	p value[c]
Gender, Male	958 (53.0)	736 (53.3)	222 (52.1)	0.7
CFTR genotype[a]				$<10^{-3}$
Classes of mutation I-III	1190 (65.9)	861 (62.4)	329 (77.2)	
At least one class of mutation IV/V	202 (11.2)	178 (12.9)	24 (5.7)	
Classes of mutation unknown	414 (22.9)	341(24.7)	73 (17.1)	
Airwaycolonization	1728 (95.7)	1310 (94.9)	418 (98.1)	
Achromobacter xylosoxidans	109 (6.3)	68 (5.2)	41 (9.8)	$<10^{-3}$
Aspergillus fumigatus	492 (28.5)	357 (27.6)	135 (32.3)	0.002
Burkholderia cepacia	56 (3.2)	29 (2.2)	27 (6.5)	$<10^{-3}$
Non-tuberculous mycobacteria	46 (2.7)	39 (3.0)	7 (1.7)	0.006
Pseudomonas aeruginosa	1114 (64.5)	783 (59.8)	331 (79.2)	$<10^{-3}$
Staphylococcus aureus	1046 (60.5)	828 (63.2)	218 (52.2)	$<10^{-3}$
Stenotrophomonas maltophilia	137 (7.9)	94 (7.2)	43 (10.3)	0.002
Comorbidities	1613 (89.3)	1202 (87.1)	411 (96.5)	
Cirrhosis	84 (5.2)	57 (4.7)	27 (6.6)	$<10^{-3}$
Insulin-treated diabetes	264 (16.4)	166 (13.8)	98 (23.8)	$<10^{-3}$
Pancreatic insufficiency	1334 (82.7)	979 (81.4)	355 (86.4)	$<10^{-3}$
Treated aspergillosis	285 (17.7)	179 (14.9)	106 (25.8)	$<10^{-3}$
Hemoptysis	153 (9.5)	86 (7.2)	67 (16.3)	$<10^{-3}$
Pneumothorax	22 (1.4)	8 (0.7)	14 (3.4)	$<10^{-3}$
FEV_1, % predicted[b*]	59.5 (25.9)	67.4 (23.9)	35.5 (15.1)	$<10^{-3}$
Age (years) [b]	27.5 (9.0)	27.4 (9.0)	28.0 (8.8)	0.2
Age at the end of the follow-up (years)[d]	32.5 (9.0)	32.9 (9.0)	31.1 (9.0)	$<10^{-3}$
BMI (kg/m^2) [b]	20.5 (3.1)	20.9 (3.2)	18.9 (2.3)	$<10^{-3}$
Number of IV antibiotics courses/year[b]	1.5 (1.9)	1.0 (1.4)	3.1 (2.3)	$<10^{-3}$
Number of IV antibiotics days/year[b]	22.1 (30.0)	14.7 (20.6)	46.4 (40.8)	$<10^{-3}$
Number of days of hospitalization/year[b]	0.7 (1.4)	0.5 (1.0)	1.6 (2.1)	$<10^{-3}$
Azithromycin	1036 (57.4)	719 (52.1)	317 (74.4)	$<10^{-3}$
Non-invasive ventilation	122 (6.8)	37 (2.7)	85 (20.0)	$<10^{-3}$
Long-term oxygen therapy	179 (9.9)	39 (2.8)	140 (32.9)	$<10^{-3}$
Oral corticosteroids	93 (5.1)	53 (3.8)	40 (9.4)	$<10^{-3}$
Inhaled therapies	1524 (84.4)	1128 (81.7)	396 (92.9)	
Inhaled antibiotics	914 (60.0)	623 (55.2)	291 (73.5)	$<10^{-3}$
Inhaled corticosteroids	715 (46.9)	521 (46.2)	194 (49.0)	$<10^{-3}$

[a] A classification of the main CFTR mutations is given in the Supplementary Appendix, Table S1

[b] Continuous variables, mean (standard deviation)

[c] Comparison of proportions in Alive without LT and LT or death without LT using chi-square test or Fisher's exact test for categorical variables, and the Student test for continuous variables.

[d] Age a death for those who died without LT, age at LT for those who received LT, age at the end of the follow-up for those who were alive without LT

* 7% of missing data

Dynamic prediction of patient prognosis

Two third (N=1204) of the patients were randomly selected for the development of the models and one third (N=602) for their validation. Table 2 shows that the estimated parameters were quite similar using the longitudinal predictive models (JLCM and landmark approach) and the cross-sectional survival models (C1 and C2). All models show that the risk of death without LT or occurrence of LT increase with the number of intravenous antibiotics courses per year. The risk of event was also higher for subjects with long-term oxygen therapy, those having oral corticosteroids and those with lower values of BMI.

An illustration of dynamic predictions of death without LT or occurrence of LT using the developed JLCM is given below for one patient A (Figure 1). For this patient, predictions are estimated at four times of prediction 26, 27, 28 and 30 years for a horizon of prediction of 3 years (top left, top right, bottom left, bottom right of Figure 3, respectively).

At the first time point 26 years, the characteristics at the beginning of the study and the first two measurements of FEV_1 are used to estimate the risk of death without LT or occurrence of LT for this patient. The probability to die without LT or to receive LT within a prediction horizon of 3 years was 0.07. As time of prediction increased, the available number of FEV_1 measurements also increased, which lead to an updated risk of death without LT or occurrence of LT. However, at time 30 years, there was a steady decrease in FEV_1 values, down to 20% predicted. Consequently, the probability to die or to receive a lung transplant within 3 years increased up to 0.25 for this patient.

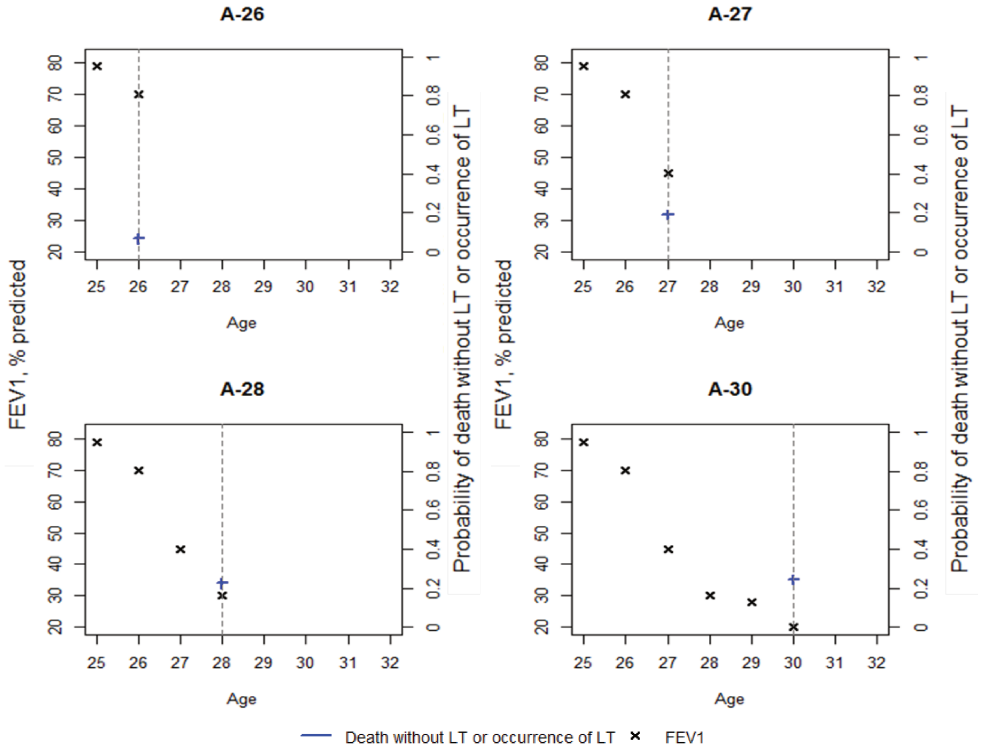

Figure 1: Example of dynamic predictions for a horizon of 3 years
Baseline characteristics for Subject A: man of 25 years old having azithromycin, long-term oxygen therapy, inhaled antibiotics and inhaled corticosteroids as treatments, BMI of 15 kg/m², 2 intravenous antibiotics and 1 day of hospitalization in the year, Pseudomonas aeruginosa colonization

Table 2 : Developed models

Variables	Joint model*			Cox landmark			C^1			C^2		
	parameter	SE**	p value	parameter	SE	p value	parameter	SE	p value	parameter	SE	p value
Survival model												
FEV_1, % predicted	0.09	0.01	$<10^{-3}$	-0.07	0.01	$<10^{-3}$	-0.08	0.01	$<10^{-3}$	-0.05	0.01	$<10^{-3}$
BMI (kg/m^2)	-0.22	0.03	$<10^{-3}$	-0.13	0.03	$<10^{-3}$				-0.12	0.03	$<10^{-3}$
Number of intravenous antibiotics courses/year	0.44	0.04	$<10^{-3}$	0.16	0.03	$<10^{-3}$				0.17	0.03	$<10^{-3}$
Number of days of hospitalization/year	0.03	0.05	0.57	0.03	0.03	0.35				0.03	0.04	0.43
Non-invasive ventilation	-0.15	0.22	0.5	-0.16	0.19	0.39				-0.08	0.92	0.68
Long-term oxygen therapy	0.99	0.2	$<10^{-3}$	0.44	0.17	0.01				0.59	0.17	$<10^{-3}$
Oral corticosteroids	0.48	0.23	0.03	0.51	0.22	0.02				0.69	0.21	$<10^{-3}$
Burkholderiacepacia												
No Test	0.44	0.37	0.24	0.20	0.74	0.79				-0.11	0.74	0.87
Test Positive	-0.35	0.26	0.18	0.89	0.77	0.25				0.49	0.28	0.09

C^1 Proportional hazard model, covariate at baseline: FEV_1
C^2 Proportional hazard model, covariates at baseline: FEV_1, BMI, Burkholderia cepacia colonization, number of intravenous antibiotics courses per year, number of days of hospitalization per year, use of long-term oxygen therapy, non-invasive ventilation and use of oral corticosteroids
*Survival part of the joint model
**SE: standard error

Profiles of CF evolution

The developed JLCM suggested that patients can optimally be classified into 3 classes, according to the BIC criterion. Figure 2 shows the class-specific mean trajectories of $ppFEV_1$ over time and the class-specific survival curves for death without LT or occurrence of LT.

Class 1 included 381 (34%) subjects and was characterized by low $ppFEV_1$values at baseline and during the follow-up. In class 1, there was a very high risk of death without LT or LT and a decreasing survival curve from the age of 20 years. This class had the worst prognosis for death without LT or LT with 52% of subjects having this combined event. At the age of 30 years, 31% of subjects in class 1 already experienced death without LT or occurrence of LT. Class 2 was composed of 481 (43%) subjects, which represented the larger group in the population. This class was characterized by a moderate decline in$ppFEV_1$ with age. Patients in class 2 had a moderate risk of death without LT or LT, with 14% of subjects experiencing death without LT or LT during follow-up. Class 3 included 255 (23%) subjects and represented the class with the highest values of $ppFEV_1$ at baseline and during the follow-up. The survival curve in this class was the highest, close to one in all ages. The prognosis was better in class 3 with only 0.4% of subjects who experienced death without LT or LT (Figure S2).

The characteristics of subjects according to the identified classes are given in Table S2. Except for the use of oral corticosteroids and gender, all characteristics were different between latent classes. Major differences at the beginning of the study between the classes were the mean of $ppFEV_1$which was clearly lower in class 1 (36%) compared to class 2 (60.6%) and class 3 (92.8%).The percentage of subjects with pneumothorax, insulin-treated diabetes, allergic bronchopulmonary aspergillosis, Pseudomonas aeruginosa, Stenotrophomonas maltophilia, azithromycin and inhaled antibiotics or respiratory support (long-term oxygen therapy and/or non-invasive ventilation) were higher in class 1.

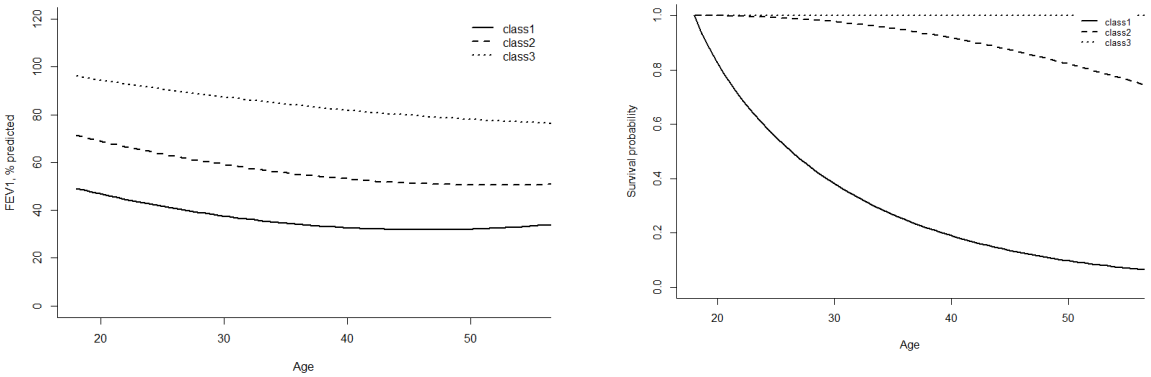

Figure 2: Class-specific mean predicted trajectories of FEV_1 (left) and the associated class-specific survival curves for LT or death without LT (right). The curves are plotted for the mean of covariates which are: initial level of FEV_1, initial level of BMI, initial number of intravenous antibiotics courses per year, initial number of days of hospitalization, oral corticosteroids, long-term oxygen therapy, non-invasive ventilation, Burkholderia cepacia colonization

Figure S3 shows how much the observed and fitted values of FEV_1 were comparable in all classes. The developed JLCM provided a good discrimination between classes with a posterior probability of belonging to each class 0.92, 0.87, 0.92 for class 1, class 2 and class 3, respectively. This means that subjects who had a final classification in class 1 had a mean probability to belong to class 1 equal to 0.92. Similarly, they had mean probability of 0.07 and 0.01 to belong to class 2 and class 3, respectively (Table S3). We concluded that the developed JLCM was reliable in classifying subjects in the class where they were more likely to belong to.

Predictive performances of the models

Table 3 provides the predictive performances for the combined event death without LT or occurrence of LT, according to the developed models at the times of prediction20, 22, 24, 26, 28, 30, 32 and 34 years and the horizon 3 years. The developed models provided good abilities to predict the event with high values of AUC and low values of BS. This shows the models ability to well predict the event and to provide a higher probability of dying without LT or of receiving LT within 3 years for a subject who is more likely to experience this event, than for a subject who is less likely to experience it. The discrimination (AUC) provided by model C2 is slightly better than the one provided by model C1, at all times of prediction. The discrimination provided by the JLCM is better in younger ages than that provided by C1 and C2.The predictive performances provided by the dynamic prognostic model landmark are especially high between the times of prediction 26 and 32 years (0.86<AUC<0.91). These times are associated with the median age for transplantation which is 30 years old, interquartile range (26-36).

Table 3: Predictive performances for LT or death without LT according to the developed models

Time of prediction (Age) / Models	JLCM		Landmark approach		C1[1]		C2[2]	
	AUC	BS	AUC	BS	AUC	BS	AUC	BS
20	0.81	0.02	0.77	0.01	0.65	0.01	0.68	0.01
22	0.86	0.04	0.76	0.02	0.67	0.04	0.68	0.04
24	0.85	0.04	0.80	0.02	0.73	0.04	0.74	0.04
26	0.76	0.05	0.86	0.03	0.79	0.06	0.82	0.06
28	0.83	0.06	0.88	0.03	0.79	0.06	0.81	0.06
30	0.75	0.04	0.91	0.02	0.76	0.04	0.83	0.04
32	0.74	0.08	0.86	0.04	0.77	0.07	0.81	0.07
34	0.89	0.04	0.83	0.02	0.78	0.04	0.80	0.04

[1] Proportional hazard model, covariate at baseline: FEV_1
[2] Proportional hazard model, covariates at baseline: FEV_1, BMI, Burkholderia cepacia colonization, number of intravenous (IV) antibiotics courses per year, number of days of hospitalization per year, use of long-term oxygen therapy, non-invasive ventilation and use of oral corticosteroids.

Sensitivity analysis

We compared the predictive abilities provided by the JLCM using death without LT as the primary event to those provided by the JLCM using the composite event (Figure S4).Both models provided good predictive abilities. The discrimination was better when considering death without LT before 26 years. However, events mostly occurred in subjects after 26 years old, where the discrimination was better using the composite event death without LT or occurrence of LT (Figure S5).

Discussion

We developed a JLCM and a survival model using the landmark approach to predict dynamically the prognosis of CF adults in relation with the temporal evolution of $ppFEV_1$. The JLCM allowed the identification of 3 profiles of disease progression in adults with CF and subjects were classified at the end of the study into 3 classes according to the observed prognosis. Class 1 (34% of patients) represented a group of patients with poor prognosis, characterized by low values of $ppFEV_1$during the follow-up and the highest rates of death without LT or LT over the study period. Class 2 (43% of patients) represented a group of patients characterized by intermediate rates of death with LT or occurrence of LT, less severe evolution with moderate decline in FEV_1. Finally, class 3 (23% of patients) represented patients with the highest values of FEV_1 during the follow-up and low rates of death without LT or LT. The developed JLCM also allowed individual dynamic predictions of the risk of death without LT or occurrence of LT using repeated measurements of FEV_1, with good predictive performances.

An important aspect of the present study is that we used JCLM for the identification of classes of patients with differences in lung function decline and prognosis. Previous studies have applied joint models using data obtained in patients with CF to characterize the evolution of FEV_1 over time with survival (24, 25), or with time to gastrostomy tube initiation(26). Joint models have also been used to assess the survival benefit of LT in adults with CF (3). However, none have focused on identifying classes of evolution of the disease. Previous studies pointed out the heterogeneity in the severity of CF (27, 28) and a recent study identified three phenotypes of rapid decline in lung function (early, late and middle) during adolescence and adulthood (29). Identifying different profiles of evolution of CF, as was described in the present study, could be useful to develop specific therapeutic intervention in patients belonging to the class at elevated risk for LT.

$ppFEV_1$ is an important marker, which is used to assess the severity of CF lung disease, and contributes to detect patients eligible for LT (4). It is then important to consider all repeated measures of FEV_1, to better assess the dependency between this marker and the occurrence of death without LT or LT. Thus, an important aim of the present study was to compare models that considered longitudinal evolution of FEV_1 (i.e., JCLM and landmark) to more conventional Cox models (i.e., C1 and C2 models) that relied on a single FEV_1 value obtained at baseline.
At all chosen times and prediction horizons, we showed that all the developed models provided good ability to predict death without LT or occurrence of LT. Interestingly, we showed that using the repeated measures of FEV_1 was better in terms of discrimination and calibration to predict the risk of death without LT or occurrence of LT than using the single value of FEV_1 at baseline, at all times of prediction with the landmark approach and in young ages with the JLCM. Indeed, an important motivation behind using dynamic prediction models is to provide individualized dynamic predictions for death without LT or occurrence of LT, using the repeated measures of FEV_1.

We particularly aimed to develop a prognostic model with good predictive performances. To do this, many approaches were considered in order to identify covariates related to the evolution of FEV_1 over time, and to the occurrence of the poor outcome. We particularly focused on backward selection, forward selection and a third approach which included the eight covariates identified as related to the poor outcome in our previous work. We finally retained the model using this last approach with good predictive performances.
When considering death without LT as the outcome of interest, the JLCM provided good predictive performances. However, the model performed less well in terms of discrimination that the one using the

composite outcome, especially around the median age of LT (Figure S4). Moreover the confidence intervals of the AUC were quite large (results not shown) due to the small number of deaths.
Although we developed prognostic models with good predictive performances, they should be validated on external data.

To conclude, we developed a JLCM that allowed us to identify three profiles of evolution of CF. This model provided dynamic predictions for death without LT or occurrence of LT in adults with CF, as well as the survival model using the landmark approach. These dynamic predictions are updated when there is more information on the repeated values of FEV_1. Furthermore, the dynamic prediction models provided good predictive performances. As LT is proposed to patients with terminal respiratory failure in order to avoid death, improve longevity and quality of life (2, 3), these model can be a useful tools for clinicians in identifying patients requiring an evaluation for LT. Especially, it may be possible to provide an online tool to easily calculate the dynamic predictions for a given patient. Thus, at each visit, the baseline characteristics and the new value of FEV_1 will be provided to calculate the risk of event.

References

1. O'Sullivan BP, Freedman SD. Cystic fibrosis. Lancet. 2009;373(9678):1891-904.
2. Charman SC, Sharples LD, McNeil KD, Wallwork J. Assessment of survival benefit after lung transplantation by patient diagnosis. The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation. 2002;21(2):226-32.
3. Thabut G, Christie JD, Mal H, Fournier M, Brugiere O, Leseche G, et al. Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation. American journal of respiratory and critical care medicine. 2013;187(12):1335-40.
4. Hirche TO, Knoop C, Hebestreit H, Shimmin D, Sole A, Elborn JS, et al. Practical guidelines: lung transplantation in patients with cystic fibrosis. Pulmonary medicine. 2014;2014:621342.
5. Martin C, Hamard C, Kanaan R, Boussaud V, Grenet D, Abely M, et al. Causes of death in French cystic fibrosis patients: The need for improvement in transplantation referral strategies! Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society. 2016;15(2):204-12.
6. Kerem E, Reisman J, Corey M, Canny GJ, Levison H. Prediction of mortality in patients with cystic fibrosis. The New England journal of medicine. 1992;326(18):1187-91.
7. Aaron SD, Stephenson AL, Cameron DW, Whitmore GA. A statistical model to predict one-year risk of death in patients with cystic fibrosis. Journal of clinical epidemiology. 2015;68(11):1336-45.
8. Liou TG, Adler FR, Fitzsimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-year survivorship model of cystic fibrosis. American journal of epidemiology. 2001;153(4):345-52.
9. Mayer-Hamblett N, Rosenfeld M, Emerson J, Goss CH, Aitken ML. Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality. American journal of respiratory and critical care medicine. 2002;166(12 Pt 1):1550-5.
10. McCarthy C, Dimitrov BD, Meurling IJ, Gunaratnam C, McElvaney NG. The CF-ABLE score: a novel clinical prediction rule for prognosis in patients with cystic fibrosis. Chest. 2013;143(5):1358-64.
11. Nkam L, Lambert J, Latouche A, Bellis G, Burgel PR, Hocine MN. A 3-year prognostic score for adults with cystic fibrosis. Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society. 2017.
12. George PM, Banya W, Pareek N, Bilton D, Cullinan P, Hodson ME, et al. Improved survival at low lung function in cystic fibrosis: cohort study from 1990 to 2007. Bmj. 2011;342:d1008.
13. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. NJ: Wiley-Interscience; 2002.
14. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. Statistics in medicine. 1996;15(15):1663-85.
15. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997;53(1):330-9.
16. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics. 2000;1(4):465-80.
17. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. Scand J Stat. 2007;34(1):70-85.
18. Bellis G, Dehillotte C, Lemonnier L. Registre français de la mucoviscidose. Bilan des données 2015. Vaincre la Mucoviscidose et Institut national d'études démographiques. 2015.
19. Urquhart DS, Thia LP, Francis J, Prasad SA, Dawson C, Wallis C, et al. Deaths in childhood from cystic fibrosis: 10-year analysis from two London specialist centres. Archives of disease in childhood. 2013;98(2):123-7.

20. Proust-Lima C, Philipps V, Liquet B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. J Stat Softw. 2017;78(2):1-56.
21. Proust-Lima C, Sene M, Taylor JM, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: a review. Statistical methods in medical research. 2014;23(1):74-90.
22. Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. Psychological methods. 2003;8(3):338-63.
23. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statistics in medicine. 2013;32(30):5381-97.
24. Schluchter MD, Konstan MW, Davis PB. Jointly modelling the relationship between survival and pulmonary function in cystic fibrosis patients. Statistics in medicine. 2002;21(9):1271-87.
25. Piccorelli AV, Schluchter MD. Jointly modeling the relationship between longitudinal and survival data subject to left truncation with applications to cystic fibrosis. Statistics in medicine. 2012;31(29):3931-45.
26. Szczesniak R, Su W, Clancy JP. Dynamics of Disease Progression and Gastrostomy Tube Placement in Children and Adolescents with Cystic Fibrosis: Application of Joint Models for Longitudinal and Time-to-Event Data. Internal medicine review. 2016;2(9).
27. Szczesniak RD, McPhail GL, Duan LL, Macaluso M, Amin RS, Clancy JP. A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. Annals of epidemiology. 2013;23(12):771-7.
28. Moss A, Juarez-Colunga E, Nathoo F, Wagner B, Sagel S. A comparison of change point models with application to longitudinal lung function measurements in children with cystic fibrosis. Statistics in medicine. 2016;35(12):2058-73.
29. Szczesniak RD, Li D, Su W, Brokamp C, Pestian J, Seid M, et al. Phenotypes of Rapid Cystic Fibrosis Lung Disease Progression during Adolescence and Young Adulthood. American journal of respiratory and critical care medicine. 2017;196(4):471-8.

Supplementary appendix

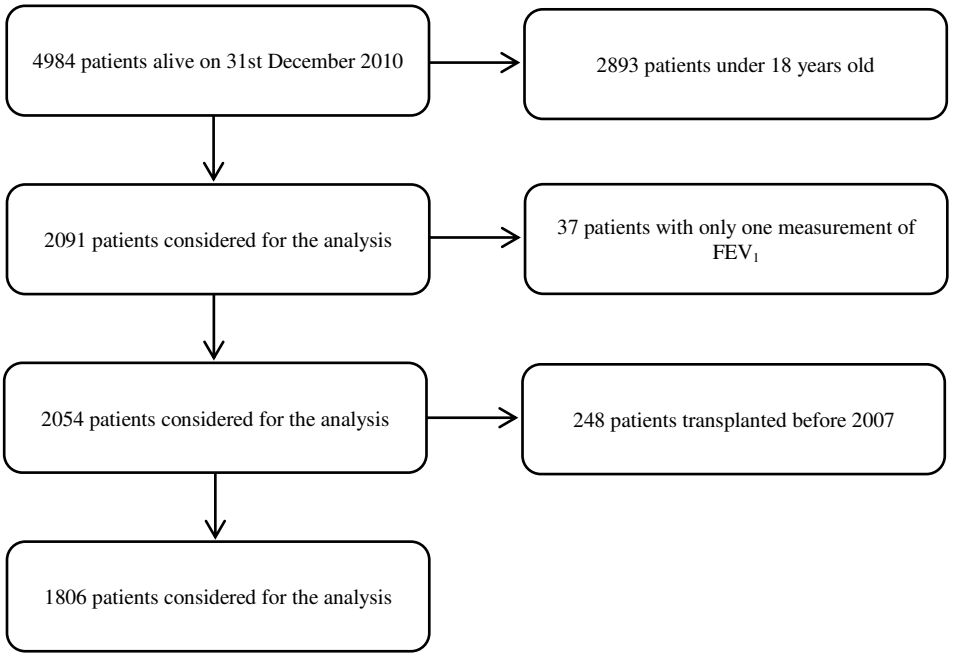

Figure S1: Patients CONSORT flow chart

Table S1: Classification of the main CFTR mutations (i.e., with frequencies≥0.3% in the 2015 French CF Registry)

Class of mutation I	Class of mutation II	Class of mutation III	Class of mutation IV	Class of mutation V
W1282X	F508del	G551D	D1152H	3849+10kbC>T
W846X	I507del	G1244E	R117H	A445E
R553X	N1303K	S1255P	R117C	2789+5G>A
R1162X	L206W	G1349D	R334W	3120+1G>A
R1066C	G85E	S945L	R347H	
G542X	S549N	G551S	R347P	
E60X		R560T	R352Q	
E585X			S1251N	
711+1G>T				
621+1G>T				
394delTT				
3659delC				
2183AA>G				
1811+1.6kbAG				
1078delT				
1717-1G>A				

Table S2: Characteristics of CF patients within the identified classes of prognosis

Patient characteristics*	Class 1 n = 381 (34%)	Class 2 n = 481 (43%)	Class 3 n = 255 (23%)
Gender, Male	214 (56.2)	258 (51.2)	121 (47.5)
CFTR genotype			
Classes of mutationI-III	276 (72.4)	345 (71.7)	136 (53.3)
At least one class of mutation IV/V	40 (10.5)	44 (9.2)	40 (15.7)
Classes of mutation unknown	65 17.1)	92 (19.1)	79 (31.0)
Airway colonization	378 (99.2)	469 (97.5)	232 (91.0)
Achromobacter xylosoxidans	32 (8.5)	34 (7.2)	7 (3.0)
Aspergillus fumigatus	121 (32.0)	136 (29.0)	53 (22.8)
Burkholderiacepacia	16 (4.2)	14 (3.0)	12 (5.2)
Non-tuberculous mycobacteria	7 (1.9)	13 (2.8)	7 (3.0)
Pseudomonas aeruginosa	281 (74.3)	306 (65.2)	104 (44.8)
Staphylococcus aureus	223 (59.0)	284 (60.6)	150 (64.7)
Stenotrophomonasmaltophilia	38 (10.1)	28 (6.0)	14 (6.0)
Comorbidities	368 (96.6)	444 (92.3)	204 (80.0)
Cirrhosis	17 (4.6)	31 (7.0)	5 (2.5)
Insulin-treated diabetes	74 (20.1)	70 (15.8)	18 (8.8)
Pancreatic insufficiency	312 (84.8)	368 (82.9)	153 (75.0)
Allergic bronchopulmonary aspergillosis	79 (21.5)	79 (17.8)	30 (14.7)
Hemoptysis	35 (9.5)	40 (9.0)	13 (6.4)
Pneumothorax	8 (2.2)	3 (0.7)	0 (0)
FEV_1, % predicted**	36.0 (12.3)	60.6 (17.3)	92.8 (16.5)
Age (years)**	27.6 (9.1)	26.8 (8.0)	28.3 (10.3)
Age at the end of the follow-up (years)***	33.0 (9.6)	32.0 (7.8)	33.9 (10.2)
BMI (kg/m^2)**	20.1 (3.7)	20.2 (2.7)	21.4 (2.7)
Number of IV antibiotics courses/year**	1.9 (1.8)	1.6 (2.0)	0.8 (1.4)
Number of IV antibiotics days/year**	21.0 (27.6)	23.7 (33.2)	10.3 (18.8)
Number of days of hospitalization/year**	1.0 (1.6)	0.7 (1.4)	0.4 (1.2)
Azithromycin	266 (69.8)	282 (58.6)	96 (37.6)
Non-invasive ventilation	49 (12.9)	24 (5.0)	6 (2.3)
Long-term oxygen therapy	79 (20.7)	37 (7.7)	4 (1.6)
Oral corticosteroids	25 (6.6)	27 (5.6)	10 (3.9)
Inhaled therapies	352 (92.4)	420 (87.3)	190 (74.5)
Inhaled antibiotics	242 (68.8)	251 (59.8)	79 (41.6)
Inhaled corticosteroids	172 (48.9)	200 (47.6)	97 (51.1)

* Two third of patients (1204) was used to develop the model, 1117 corresponded to patients without missing data in the longitudinal marker FEV_1

** Continuous variables, mean (sd)

*** Age a death for those who died without LT, age at LT for those who received LT, age at the end of the follow-up for those who were alive without LT

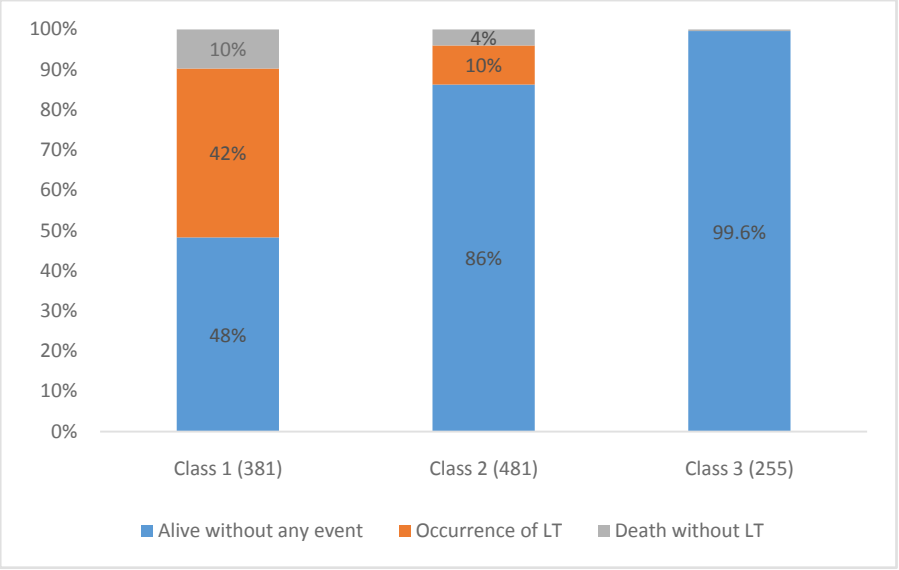

Figure S2: Proportion of death without transplantation and occurrence of lung transplantation in each class

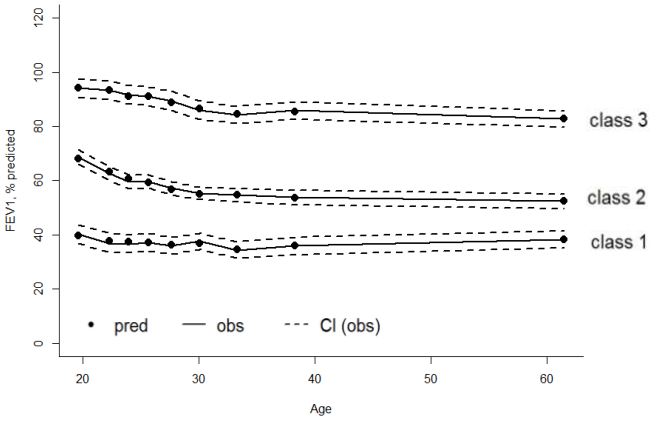

Figure S3: Comparison between the weighted subject-specific predictions and the weighted mean observed FEV_1 according to age, to evaluate the joint latent class model fit

Table S3: Mean probabilities classification of subjects in each class

Final classification	Mean posterior probability in each class		
	Class 1	Class 2	Class 3
Class 1	0.92	0.07	0.01
Class 2	0.07	0.87	0.06
Class 3	0.00	0.08	0.92

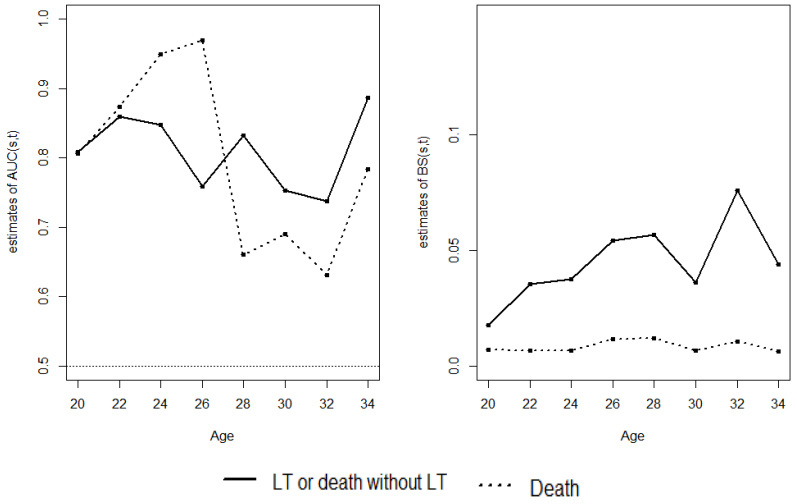

Figure S4: Comparison between the time-dependent AUC and Brier score provided by the joint latent class model using the events death without transplant and the composite event death without transplant or occurrence of lung transplantation

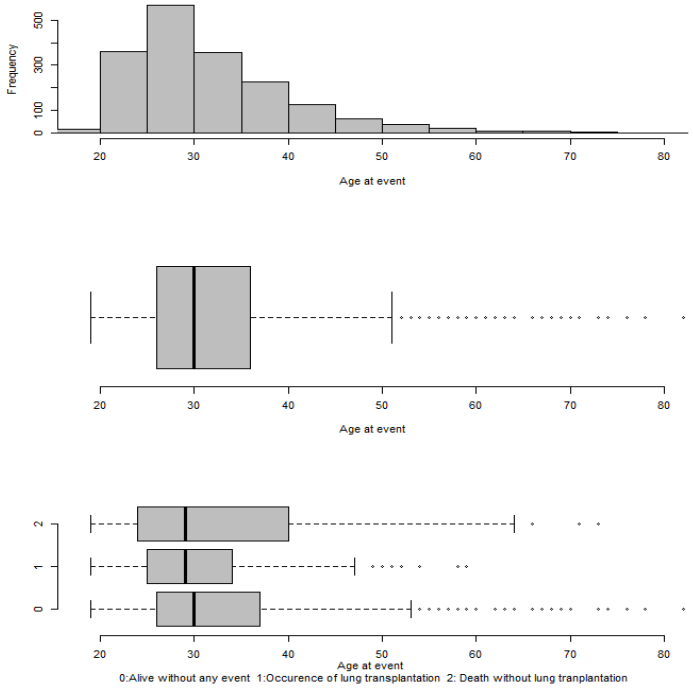

Figure S5: Distribution of age death without transplant or/and lung transplant

Table S4 : Full developedmodels

Variables	Joint model			Cox landmark			C^1			C^2		
	parameter	SE*	p value	parameter	SE	p value	parameter	SE	p value	parameter	SE	p value
Survival model												
FEV_1, % predicted	0.09	0.01	$<10^{-3}$	-0.07	0.01	$<10^{-3}$	-0.08	0.01	$<10^{-3}$	-0.05	0.01	$<10^{-3}$
BMI (kg/m^2)	-0.22	0.03	$<10^{-3}$	-0.13	0.03	$<10^{-3}$				-0.12	0.03	$<10^{-3}$
Number of intravenous antibiotics courses/year	0.44	0.04	$<10^{-3}$	0.16	0.03	$<10^{-3}$				0.17	0.03	$<10^{-3}$
Number of days of hospitalization/year	0.03	0.05	0.57	0.03	0.03	0.35				0.03	0.04	0.43
Non-invasive ventilation	-0.15	0.22	0.5	-0.16	0.19	0.39				-0.08	0.92	0.68
Long-term oxygen therapy	0.99	0.2	$<10^{-3}$	0.44	0.17	0.01				0.59	0.17	$<10^{-3}$
Oral corticosteroids	0.48	0.23	0.03	0.51	0.22	0.02				0.69	0.21	$<10^{-3}$
Burkholderia cepacia												
No Test	0.44	0.37	0.24	0.20	0.74	0.79				-0.11	0.74	0.87
Test Positive	-0.35	0.26	0.18	0.89	0.77	0.25				0.49	0.28	0.09
LT +/-sqrt(Weibull1) class 1	0.28	0.08	$<10^{-3}$									
LT +/-sqrt(Weibull1) class 1	0.95	0.07	$<10^{-3}$									
LT +/-sqrt(Weibull1) class 2	0.12	0.02	$<10^{-3}$									
LT +/-sqrt(Weibull1) class 2	1.48	0.14	$<10^{-3}$									
LT +/-sqrt(Weibull1) class 3	0.01	0.03	0.73									
LT +/-sqrt(Weibull1) class 3	1.29	0.65	0.05									
Longitudinal mixed effects model												
Intercept class 1	8.21	3.51	0.02									
Intercept class 2	30.34	3.51	$<10^{-3}$									
Intercept class 3	55.16	3.75	$<10^{-3}$									
Linearslope class 1	-1.22	0.21	$<10^{-3}$									
Linearslope class 2	-1.22	0.18	$<10^{-3}$									
Linearslope class 3	-0.82	0.23	$<10^{-3}$									
Quadraticslope class 1	0.02	0.01	$<10^{-3}$									
Quadraticslope class 2	0.02	0.01	$<10^{-3}$									
Quadraticslope class 3	0.01	0.01	0.22									
BMI (kg/m^2)	2.14	0.17	$<10^{-3}$									
Number of intravenous antibiotics courses/year	-4.05	0.24	$<10^{-3}$									
Random intercept variance	165.48	20.25										
Random linear slope variance	5.25	0.52										
Random quadratic slope variance	0.003	0.0004										
Random intercept and linear slope covariance	-22.03	2.92	$<10^{-3}$									
Random intercept and quadratic slope covariance	0.39	0.07	$<10^{-3}$									
Random linear and quadratic slope covariance	-0.13	0.02	$<10^{-3}$									
Measurement error	7.28	0.08										
Class-membership model												
Intercept class 1	2.88	0.25	$<10^{-3}$									
Intercept class 2	1.08	0.12	$<10^{-3}$									

C^1 Proportional hazard model, covariate at baseline: FEV_1

C^2 Proportional hazard model, covariates at baseline: FEV_1, BMI, Burkholderia cepacia colonization, number of intravenous antibiotics courses per year, number of days of hospitalization per year, use of long-term oxygen therapy, non-invasive ventilation and use of oral corticosteroids

SE: standard error

4.2 Compléments

Nous avons développé un modèle conjoint à classes latentes en considérant les risques de survenue du décès sans transplantation pulmonaire et de la transplantation pulmonaire comme des risques compétitifs. La classification obtenue à partir de ce modèle est très similaire à celle obtenue en considérant l'événement composite. La classe 1 était constituée de 33% de sujets, la classe 2 comptait 43% de sujets et la classes 3 comptait 24% des sujets. Les probabilités moyennes *a posteriori* d'appartenir aux classes 1, 2 et 3 pour les sujets classés dans ces groupes étaient respectivement de 0.96, 0.92 et 0.96. Outre les proportions similaires, les deux modèles classaient avec concordance les sujets dans les classes. Les deux modèles classaient 30%, 36% et 21% des sujets respectivement dans la classe 1, 2 et 3. Seuls 13% des sujets étaient classés différemment par les deux modèles. Cependant, les différences de classification concernaient principalement les classes 2 et 3. Un seul sujet, soit 0.08% a été classé dans la classe 3 dans le modèle avec l'événement composite et dans la classe 1 avec le modèle considérant les risques compétitifs.

Les courbes d'évolution moyenne du VEMS et les courbes d'incidence cumulée pour la transplantation pulmonaire et pour le décès sans transplantation pulmonaire sont représentées par les figures 4.1, 4.2 et 4.3 respectivement. Ces courbes sont tracées pour un sujet présentant les caractéristiques moyennes de l'ensemble des sujets de l'étude. À l'inclusion, ce sujet a 60% de VEMS, un IMC de $21 kg/m^2$ et une cure d'antibiotiques par voie intraveineuse dans l'année. Il n'a pas été hospitalisé dans l'année et n'a pas reçu de traitements tels que la ventilation nasale, l'oxygénothérapie, les corticoïdes. Il n'a pas été colonisé par *Burkholderia cepacia*.

Les courbes d'évolution moyenne du VEMS sont similaires à celles obtenues en considérant l'événement composite. Les courbes d'incidence cumulée pour le décès sans transplantation pulmonaire sont assez proches dans les classes 2 et 3. Dans la classe 3, le risque de décès sans transplantation pulmonaire reste proche de zéro jusqu'à 40 ans et proche de zéro jusqu'à 55 ans pour la transplantation pulmonaire. Dans la classe 2, ces deux risques augmentent légèrement dès l'âge de 22 ans mais reste relativement faible pour le décès sans transplantation pulmonaire. Dans la classe 1, le risque de décès sans transplantation pulmonaire et le risque de transplantation pulmonaire augmentent dès l'âge de 21 ans et atteignent respectivement 78% et 22% avant l'âge de 25 ans.

Tout comme en considérant l'événement composite, la classe 1 est une classe de sujets ayant un mauvais pronostic pour la mucoviscidose. Dans cette classe, on a 41% de transplantation pulmonaire et 7% de décès sans transplantation pulmonaire. La

classe 2 est constituée de sujets ayant une évolution modérée de la maladie avec 11% de transplantation pulmonaire et 5% de décès sans transplantation pulmonaire. La classe 3 apparaît comme une classe de sujets ayant un bon pronostic de la maladie, avec 0.4% de transplantation pulmonaire et 2.6% de décès sans transplantation pulmonaire.

La figure 4.4 présente des prédictions dynamiques du risque de décès sans transplantation pulmonaire (rouge) et du risque de transplantation pulmonaire (bleu). Ces prédictions sont calculées pour un sujet ayant des valeurs de VEMS décroissantes au cours du temps. À l'inclusion, ce sujet a 60% de VEMS, un IMC de $18 kg/m^2$. Il a reçu deux cures d'antibiotiques par voie intraveineuse et a été hospitalisé une journée dans l'année. Il reçoit de l'oxygénothérapie et est colonisé par *Pseudomonas aeruginosa*.

Les prédictions sont obtenues aux âges 26, 27, 28 et 30 ans, à un horizon de 3 ans. Plus l'âge de prédiction augmente, plus on a de l'information sur le VEMS et les risques de décès sans transplantation pulmonaire et de transplantation pulmonaire sont mis à jour. Pour ce sujet, les risques de décès sans transplantation pulmonaire et de transplantation pulmonaire augmentent lorsque le VEMS diminue.

Nous avons appliqué sur nos données le modèle conjoint à effets aléatoires partagés qui a la particularité de quantifier l'association entre les deux processus. Deux spécifications de l'association ont été considérées : une association entre le risque de survenue du décès sans transplantation pulmonaire ou de la transplantation pulmonaire et le niveau courant du VEMS, une association entre le risque de survenue du décès sans transplantation pulmonaire ou de la transplantation pulmonaire et le niveau courant et la pente courante du VEMS.

Pour ces modèles, le risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire était fortement associé à l'évolution du VEMS. Le risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire augmentait de l'ordre de 1.3 pour une diminution de 1% du niveau courant du VEMS. Le même constat était fait pour la diminution de la pente du VEMS. Ces modèles conjoints à effets aléatoires partagés ont fourni de bonnes capacités de discrimination (supérieures 0.8). Néanmoins le modèle conjoint à classes latentes a été retenu dans le but d'identifier les différents profils d'évolution de la maladie du fait de l'hétérogénéité de la population des malades.

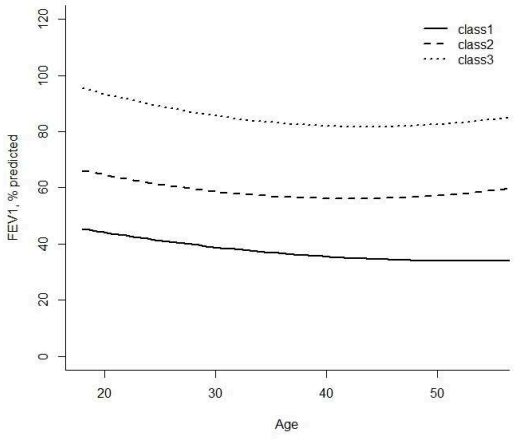

FIGURE 4.1 – Courbes d'évolution moyenne du VEMS

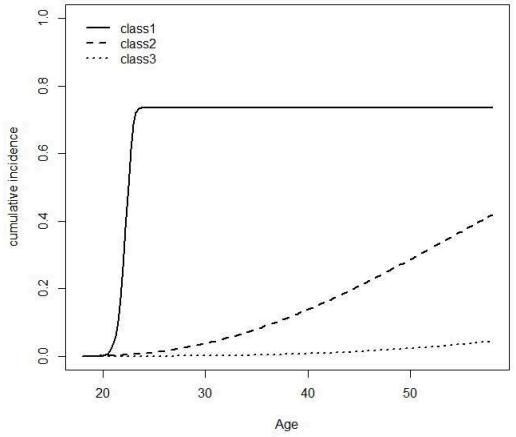

FIGURE 4.2 – Courbes d'incidence cumulée pour la transplantation pulmonaire

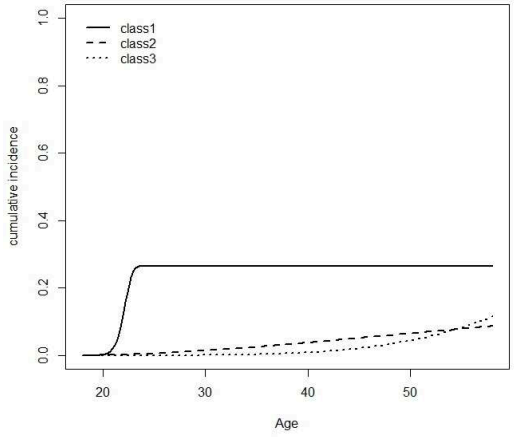

FIGURE 4.3 – Courbes d'incidence cumulée pour le décès sans transplantation pulmonaire

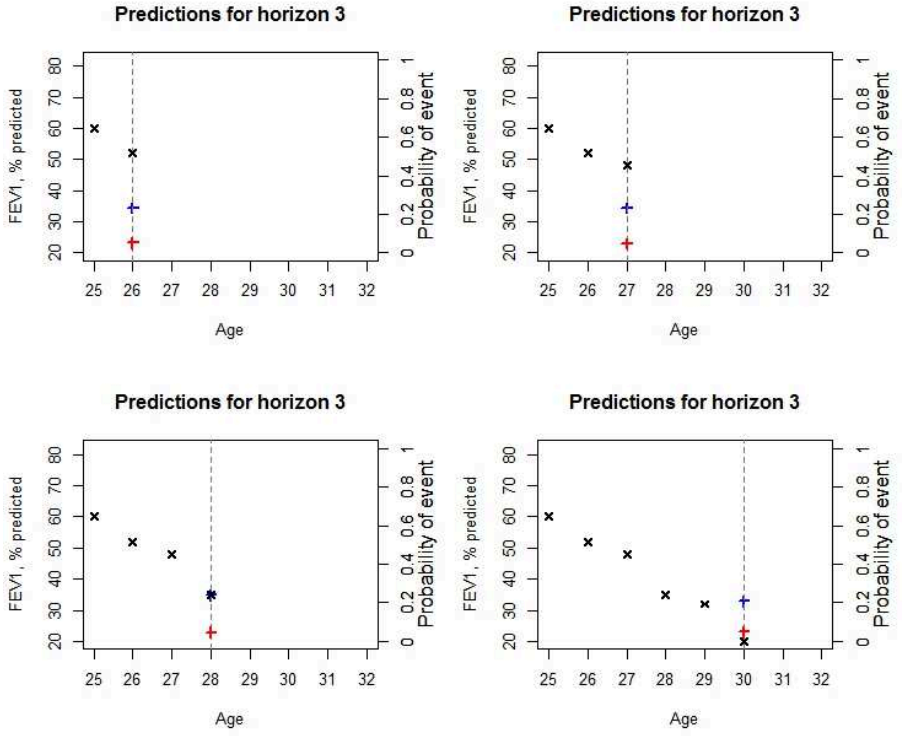

FIGURE 4.4 – Prédictions dynamiques du risque de décès sans transplantation pulmonaire et de la transplantation pulmonaire

Chapitre 5

Discussion, conclusion et perspectives

5.1 Discussion

Choix de la population adulte

Le but de ce travail était de développer des modèles pronostiques afin d'identifier les adultes atteints de mucoviscidose éligibles à la transplantation pulmonaire. Pour diverses raisons, ce travail s'est basé sur les données des sujets adultes du registre français de la mucoviscidose. D'une part, cette maladie au préalable décrite comme une maladie infantile est désormais considérée comme une maladie chronique. En effet, la mortalité infantile liée à la mucoviscidose a presque disparu dans les pays développés [Urquhart et al. 2013]. D'autre part, le nombre d'adultes atteints de mucoviscidose est de plus en plus croissant [Burgel et al. 2015]. La proportion d'adultes est désormais supérieure à la proportion d'enfants atteints de mucoviscidose dans le monde. De plus, les transplantations pulmonaires sont en majorité réalisées chez les adultes [Bellis et al. 2015].

Choix de l'événement composite

Pour l'identification des sujets éligibles à la transplantation pulmonaire, nous avons considéré comme événement d'intérêt un mauvais pronostic du patient, soit le décès sans transplantation pulmonaire ou la survenue de la transplantation pulmonaire. Une approche classique est de considérer uniquement le décès comme événement d'intérêt. Dans ce cas, les facteurs associés au décès permettront au clinicien d'identifier les sujets ayant un risque élevé de décéder et d'évaluer par la suite leur éligibilité à la

transplantation pulmonaire. Cette approche est particulièrement utile lorsqu'on n'a pas assez d'information sur les sujets transplantés et seules les informations sur les sujets décédés nous permettent d'identifier les facteurs qui pourront guider l'avis du clinicien vis-à-vis de la transplantation pulmonaire. Cependant, la transplantation pulmonaire vise à éviter le décès chez les sujets ayant un état respiratoire nettement dégradé. De ce fait, ne considérer que les sujets décédés exclurait les sujets à risque élevé de décéder.

En France, on compte moins de décès sans greffe que de transplantations pulmonaires par année [Bellis et al. 2015]. Les sujets ayant un mauvais pronostic sont donc majoritairement des sujets greffés. De plus, on note peu de décès sur liste d'attente et une récente étude a montré que la moitié des sujets décédés sans greffe auraient été éligibles à la transplantation pulmonaire [Martin et al. 2016]. Selon le registre français de la mucoviscidose, depuis 2009 le ratio du nombre de sujets transplantés par rapport au nombre de sujets inscrits sur liste d'attente varie de 0.9 à 1.2. Le nombre de greffes réalisées dans l'année est donc très proche voire supérieur au nombre d'inscrits sur liste d'attente. Les sujets qui reçoivent une transplantation pulmonaire sont ceux ayant une fonction respiratoire très faible et un état de santé nettement dégradé. Ces sujets ont généralement recours à des traitements supplémentaires pour lutter contre l'insuffisance respiratoire. C'est le cas par exemple de l'oxygénothérapie ou de la ventilation non invasive. La transplantation pulmonaire vise donc à améliorer la qualité de vie et surtout à prolonger la vie des malades dont l'état respiratoire s'est considérablement dégradé [Thabut et al. 2013]. On fait alors l'hypothèse que si ces sujets n'avaient pas été transplantés, ils seraient sans doute décédés. De ce fait, on considère le décès et la transplantation pulmonaire comme des événements associés à un mauvais pronostic chez les sujets atteints de mucoviscidose.

Cette hypothèse peut paraître forte du fait que la décision de la transplantation pulmonaire n'est pas aléatoire, elle est prise par le clinicien selon certains critères. Cela d'autant plus que les critères d'identification à la transplantation pulmonaire peuvent varier selon les centres de transplantation ou selon les pays. Néanmoins, cette approche a déjà été utilisée dans le domaine de la mucoviscidose dans des études récentes [McCarthy et al. 2013; Waters et al. 2013] et est également utilisée dans d'autres contextes médicaux [Raymond et al. 2002; Towbin et al. 2006]. Nous avons tout de même réalisé des études complémentaires comparant les résultats obtenus avec l'événement composite et les résultats obtenus en considérant chacun des événements séparément.

Résumé des travaux réalisés

Dans le premier article, nous avons réévalué les facteurs associés au décès sans transplantation pulmonaire ou à la transplantation pulmonaire chez les adultes atteints de mucoviscidose. Dans ce travail, nous avons identifié trois caractéristiques liées au patient qui sont associées à un risque accru de décès sans transplantation pulmonaire ou de transplantation pulmonaire. Il s'agit d'un VEMS faible, d'un IMC faible et de la colonisation à *Burkholderia cepacia*. Ces facteurs ont été identifiés dans des études antérieures comme associés au décès chez les sujets atteints de mucoviscidose. Nous avons également identifié cinq facteurs qui caractérisent la prise en charge thérapeutique. Il s'agit du nombre de cures intraveineuse dans l'année, du nombre de jours d'hospitalisation dans l'année, de la prise de corticoïdes par voie orale, de la ventilation non invasive et de l'oxygénothérapie. Contrairement aux autres facteurs, la ventilation non invasive et l'oxygénothérapie n'ont pas été au préalable identifiées dans des modèles pronostiques, comme facteurs associés au décès chez les sujets atteints de mucoviscidose. Ces résultats pourraient être expliqués par la prise en charge thérapeutique propre au système de soin français, bien que ces traitements soient également administrés dans d'autres pays. Néanmoins en France, la ventilation non invasive et l'oxygénothérapie sont administrées aux sujets ayant un état respiratoire dégradé indépendamment du fait qu'ils soient sur liste d'attente pour la transplantation pulmonaire ou pas.

À partir de ces huit facteurs identifiés, nous avons construit un score pronostique. De plus, nous avons défini trois classes de scores pronostiques pour lesquels les sujets avaient un risque faible (score allant de 0 à 1.5), modéré (score allant de 2 à 3.5) et élevé (score supérieur ou égal à 4) de décéder ou de recevoir une transplantation pulmonaire. Un nomogramme a été fourni pour faciliter le calcul du score pronostique ainsi que du risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire correspondant. Ce nomogramme, facile d'utilisation, peut servir d'outil pronostique aux cliniciens pour identifier les malades ayant un risque élevé de décéder ou de recevoir une greffe pulmonaire.

Dans une analyse complémentaire, nous avons identifié les facteurs associés au risque de décès sans transplantation pulmonaire uniquement. Parmi les huit facteurs associés au risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire, seuls le VEMS et l'IMC étaient associés au risque de décès. Les indicateurs d'aggravation de la maladie (tels que les cures d'antibiotiques, les hospitalisations) et les traitements (tels que la ventilation non invasive, l'oxygénothérapie) n'ont pas été identifiés comme associés au risque de décès chez les adultes atteints de mucoviscidose.

En revanche, la cirrhose et un âge élevé, qui sont par ailleurs des contrindications à la transplantation pulmonaire, étaient associés à un risque élevé de décéder.

Dans le deuxième article, nous avons développé un modèle conjoint à classes latentes dans le but de fournir des prédictions dynamiques pour le risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire chez les adultes atteints de mucoviscidose. Le modèle conjoint développé a permis d'avoir une meilleure connaissance de l'évolution de la maladie en modélisant simultanément l'évolution du VEMS, principal marqueur de la fonction pulmonaire, et le risque de survenue du décès sans transplantation pulmonaire ou de la transplantation pulmonaire. De plus, il a permis d'identifier trois profils d'évolution de la maladie. En effet, la mucoviscidose est une maladie complexe qui évolue différemment d'un sujet à l'autre. Nous avons identifié un groupe de sujets ayant une évolution sévère de la maladie, un groupe ayant une évolution modérée de la maladie et un groupe ayant une évolution légère de la maladie. Ces résultats sont en adéquation avec les trois classes de scores que nous avons défini dans le premier article.

À partir du modèle conjoint développé, il est possible de prédire pour un sujet inclus dans l'étude ou pas, le risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire qui évolue en fonction des valeurs du VEMS. Ainsi, le risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire est actualisé après chaque nouvelle valeur du VEMS. Un tel outil présente un intérêt majeur en pratique car il permet d'identifier les sujets ayant un risque élevé de décéder ou de recevoir une greffe pulmonaire. L'évaluation d'un outil prédictif est nécessaire pour s'assurer de sa capacité à bien prédire et à bien discriminer les sujets vis-à-vis de l'événement d'intérêt. L'outil prédictif développé a fourni de bonnes capacités prédictives en termes de discrimination et de calibration.

Nous avons réalisé une analyse en considérant le risque de survenue du décès sans transplantation pulmonaire et le risque de transplantation pulmonaire comme des risques compétitifs. Pour une prédiction à trois ans, les performances prédictives du modèle considérant l'événement composite étaient proches de celles fournies en considérant la transplantation pulmonaire dans l'approche par risques compétitifs, à tous les temps de prédiction choisis. Les intervalles de confiances étaient assez larges du fait du nombre d'événement réduit dans les données de validation.

Dans ce deuxième article, nous nous sommes surtout focalisés sur l'identification de différents profils d'évolution de la maladie, les prédictions dynamiques du risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire et les ca-

pacités prédictives du modèle développé. De ce fait, nous avons privilégié les modèles conjoints à classes latentes bien que moins populaires que les modèles conjoints à effets aléatoires partagés. Cependant, les modèles conjoints à effets aléatoires partagés ont été appliqués sur nos données. Ces modèles ont l'avantage de quantifier le lien existant entre le processus longitudinal et le processus de survie. Appliqué sur nos données, le modèle conjoint à effets aléatoires partagés a montré un lien significatif entre le risque de survenue du décès sans transplantation pulmonaire ou de la transplantation pulmonaire et le niveau courant, ainsi que la pente courante du VEMS. De ce fait, à un instant t donné, le risque de survenue du décès sans transplantation pulmonaire ou de la transplantation pulmonaire est supérieur pour un sujet ayant un niveau ou une pente de VEMS faible comparé à un sujet pour lequel ces valeurs sont plus élevées.

La mucoviscidose étant une maladie complexe, nous nous sommes intéressés à l'évolution de la maladie chez les sujets homozygotes F508del pour le gène CFTR. Dans une analyse supplémentaire, nous avons recherché les différents profils d'évolution de la maladie chez ces sujets ayant le gène le plus fréquent dans la population des malades. Les profils identifiés étaient similaires à ceux obtenus à partir de toutes les données de l'échantillon. Les résultats obtenus ne sont pas présentés dans ce document.

Avantages et limites

Les modèles conjoints reposent en grande partie sur les marqueurs longitudinaux. Dans le contexte de la mucoviscidose, le VEMS est le marqueur décrivant au mieux la fonction pulmonaire. Ce paramètre reflète clairement l'évolution de l'état de santé des sujets atteints de mucoviscidose. Il existe plusieurs méthodes de calcul du VEMS. Celle utilisée dans le registre français de la mucoviscidose a été développée en 1983 par Knudson et al.. Elle est calculée à partir des données de 697 caucasiens en fonction de l'âge et du genre. Une méthode plus récente a été proposée en 1999 par Hankinson et al.. Les données de plus de 7000 américains de type caucasien, africain et mexicain ont été utilisées pour le calcul du VEMS. Cette méthode basée sur l'âge, la taille, le genre et l'ethnie est plus utilisée que la méthode proposée par Knudson et al. pour décrire l'évolution de la maladie. Les deux méthodes ont été comparées par Hankinson et al. et les prédictions des valeurs du VEMS obtenues sur les données de sujets mâles de type caucasien étaient très similaires. Cependant, elles étaient légèrement différentes pour les données des sujets issus d'autres ethnies, avec des valeurs prédites du VEMS légèrement meilleures à partir de la méthode

de Hankinson et al.. Contrairement aux États-Unis, la population de malades en France est presqu'exclusivement caucasienne. Bien que relativement ancienne, la méthode de calcul du VEMS utilisée par le registre français de la mucoviscidose pourrait toujours être adaptée au contexte français. Cependant, il serait plus judicieux que les méthodes de calcul du VEMS soient standardisées dans tous les registres.

Ce travail nous a permis de valoriser les données du registre français de la mucoviscidose. En effet, peu d'études de cette nature ont été réalisées en considérant toutes les données du registre. Nous nous sommes limités aux données les plus récentes, et n'avons considéré que les données recueillies après la création officielle du registre en 2006. Les données du registre constituent un réel avantage pour le développement des modèles prédictifs car elles sont centralisées et harmonisées. Ainsi, un effet centre serait peu problématique. Cependant, certaines variables recueillies dans le registre n'ont pas été considérées dans le développement des modèles du fait d'un trop grand nombre de données manquantes, telles que le statut professionnel, le statut marital. De même, certaines variables qui pourraient être utiles pour un score pronostique, telles que les variables liées à l'activité physique n'ont pas été prises en compte, car non recueillies dans le registre. Le registre français de la mucoviscidose recueille les données du bilan annuel des malades, bien que certains malades soient suivis plusieurs fois dans l'année dans leur centre référent. Or, un suivi plus régulier serait plus avantageux, car il y aurait plus d'information disponible pour étudier l'évolution de la maladie.

Dans le modèle conjoint développé, nous nous sommes intéressés à l'âge de survenue de l'événement. Le choix de l'âge comme axe de temps nous a semblé plus pertinent que le temps écoulé depuis l'inclusion dans l'étude, du fait que l'âge soit un critère important dans l'identification des sujets à la transplantation pulmonaire. En effet, pour cette maladie chronique, le risque de mauvais pronostic dépend fortement de l'âge. De plus, les critères de sélection correspondaient principalement au fait d'être adulte et non transplanté au préalable. Le choix du temps écoulé depuis l'entrée dans l'étude aurait plus de sens si on s'intéressait par exemple au délai écoulé entre la greffe pulmonaire et l'échec de la greffe ou le décès. Le suivi considéré dans le modèle conjoint n'était pas très long, car nous n'avons considéré que les données les plus récentes et mieux renseignées. Néanmoins le modèle a fourni de bonnes capacités prédictives.

Pour le modèle conjoint développé, nous nous sommes peu attardés sur la sélection des variables associées à chacun des processus. Le but étant de fournir un outil

prédictif ayant de bonnes capacités, nous nous sommes surtout intéressés aux capacités prédictives du modèle. Cependant, plusieurs modèles ont été testés et leurs capacités prédictives étaient globalement similaires. Le modèle conjoint développé a été comparé en termes de capacité discriminante à un modèle à risque proportionnels ne prenant en compte que la valeur initiale du VEMS. Les capacités discriminantes fournies par le modèle conjoint étaient meilleures. Néanmoins le modèle conjoint reste un modèle plus complexe que les modèles à risques proportionnels. Il est plus récent, éventuellement moins compréhensible pour les cliniciens qu'un modèle logistique ou un modèle de survie. De plus, les modèles conjoints ont un temps de calcul relativement long et présentent souvent des problèmes de convergence.

5.2 Conclusion

Dans ce travail de thèse, nous avons réévalué les facteurs associés au décès sans transplantation pulmonaire ou à la transplantation pulmonaire chez les adultes atteints de mucoviscidose. Nous avons pu confirmer les résultats de la littérature en identifiant des facteurs déjà connus comme associés à un mauvais pronostic chez les sujets atteints de mucoviscidose. À notre connaissance, nous avons pu mettre en évidence dans un modèle pronostique pour la première fois, un lien entre le risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire et l'usage de la ventilation nasale et l'oxygénothérapie. Un score pronostic et un nomogramme ont été développés pour faciliter la mise en œuvre en pratique du modèle développé. Ce modèle a été développé et validé à partir des données du registre français de la mucoviscidose.

Le VEMS étant le marqueur qui décrit au mieux la fonction pulmonaire, une modélisation conjointe de l'évolution temporelle de ce marqueur et de la survie des malades a été réalisée. Cette approche repose sur un modèle à classes latentes qui permet de fournir des prédictions dynamiques du risque de l'événement considéré et d'identifier des profils d'évolution de la maladie. Le modèle développé a confirmé les résultats du modèle précédent, en identifiant trois profils distincts d'évolution de la maladie. Il permet d'obtenir des prédictions dynamiques du risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire avec de bonnes capacités prédictives.

Ces deux outils pronostiques permettent d'identifier les sujets adultes ayant un risque élevé de décéder ou de recevoir une greffe pulmonaire avec de bonnes capacités prédictives. L'outil développé dans le premier article a l'avantage d'être facile d'utilisation et directement applicable par le clinicien. L'outil développé dans

le deuxième article a l'avantage de considérer toute l'information disponible pour le VEMS pour prédire le risque de décès sans transplantation pulmonaire ou de transplantation pulmonaire.

5.3 Perspectives

Extension des modèles conjoints

Plusieurs extensions des modèles conjoints pour données longitudinales et temps d'événement sont disponibles. C'est le cas par exemple de la modélisation simultanée de plusieurs marqueurs longitudinaux et du risque de survenue d'un ou plusieurs événements [Andrinopoulou et al. 2014, 2017; Rizopoulos and Ghosh 2011; Proust-Lima et al. 2016]. Dans le contexte de la mucoviscidose, il est envisageable de considérer l'évolution d'autres marqueurs longitudinaux tels que le nombre de cures intraveineuse par année du sujet, qui est un marqueur d'aggravation de la maladie. Ce paramètre apporterait plus d'information pour mieux modéliser l'évolution de la maladie.

Il est possible d'appliquer les modèles conjoints dans un contexte multi-états [Dantan et al. 2011; Ferrer et al. 2016]. Dans ce cas, on évaluerait les intensités de transition entre les états « vivant sans transplantation pulmonaire vers transplantation pulmonaire », « vivant sans transplantation pulmonaire vers décès », « transplantation pulmonaire vers décès ».

Évaluation de la transplantation pulmonaire

Le bénéfice de la transplantation pulmonaire peut être mesuré par un calcul du score de propension. Cette approche est une alternative à un essai randomisé non réalisable dans ce contexte et dont le but serait d'évaluer le lien causal entre une greffe pulmonaire et le risque de décès. En effet, la décision de transplantation ne pouvant pas être prise aléatoirement d'un sujet à un autre, seules les données de suivi chez les patients transplantés et non transplantés sont disponibles. L'utilisation des données du registre est une opportunité pour calculer ce score et tester ses propriétés.

Validation externe des modèles développés

Les modèles développés sont des outils de prédictions permettant d'identifier les malades adultes ayant un mauvais pronostic pour une évaluation à la transplantation pulmonaire. Ils ont été développés à partir des données du registre français de la

mucoviscidose et ont fourni de bonnes capacités prédictives. Ce sont des données centralisées, récentes et de bonne qualité. Néanmoins le système de soin est propre à la France et certaines variables sont calculées différemment dans d'autres registres. Pour confirmer leurs bonnes performances prédictives, il serait nécessaire qu'ils soient validés sur des données d'un autre registre. Une validation du score pronostique à 3 ans est envisagée en utilisant les données du registre canadien de la mucoviscidose. Si les performances prédictives sont confirmées sur d'autres données, les modèles développés seront d'un intérêt majeur pour l'identification des sujets adultes atteints de mucoviscidose à une évaluation pour la transplantation pulmonaire.

Bibliographie

S. D. Aaron, A. L. Stephenson, D. W. Cameron, and G. A. Whitmore. A statistical model to predict one-year risk of death in patients with cystic fibrosis. *J Clin Epidemiol*, 68(11) :1336–45, 2015. 10, 57, 58

D. G. Altman and P. Royston. What do we mean by validating a prognostic model ? *Statistics in Medicine*, 19(4) :453–473, 2000. 54

A. Amadori, A. Antonelli, I. Balteri, A. Schreiber, M. Bugiani, and V. De Rose. Recurrent exacerbations affect fev(1) decline in adult patients with cystic fibrosis. *Respir Med*, 103(3) :407–13, 2009. 9

E. R. Andrinopoulou, D. Rizopoulos, J. J. M. Takkenberg, and E. Lesaffre. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine*, 33(18) :3167–3178, 2014. 49, 104

E. R. Andrinopoulou, D. Rizopoulos, J. J. M. Takkenberg, and E. Lesaffre. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical Methods in Medical Research*, 26(4) :1787–1801, 2017. 49, 104

L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24) :3927–3944, 2005. 51

G. Bellis, M. H. Cazes, A. Parant, M. Gaimard, C. Travers, E. Le Roux, S. Ravilly, and G. Rault. Cystic fibrosis mortality trends in france. *J Cyst Fibros*, 6(3) : 179–86, 2007. 15

G. Bellis, C. Dehillotte, and L. Lemonnier. Registre français de la mucoviscidose. bilan des données 2015. *Vaincre la Mucoviscidose et Institut national d'études démographiques*, 2015. 7, 58, 97, 98

P. Blanche, J. F. Dartigues, and H. Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30) :5381–5397, 2013. 52, 53

P. Blanche, C. Proust-Lima, L. Loubere, C. Berr, J. F. Dartigues, and H. Jacqmin-Gadda. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1) :102–113, 2015. 53

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1) :89–99, 1974. 36

E. R. Brown and J. G. Ibrahim. A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59(2) :221–228, 2003. 44

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

P. R. Burgel, G. Bellis, H. V. Olesen, L. Viviani, A. Zolin, F. Blasi, J. S. Elborn, and E. E. T. F. o. P. o. C. f. A. w. C. F. i. Europe. Future trends in cystic fibrosis demography in 34 european countries. *Eur Respir J*, 46(1) :133–41, 2015. 7, 56, 97

R. Buzzetti, G. Alicandro, L. Minicucci, S. Notarnicola, M. L. Furnari, G. Giordano, V. Lucidi, E. Montemitro, V. Raia, G. Magazzu, G. Vieni, S. Quattrucci, A. Ferrazza, R. Gagliardini, N. Cirilli, D. Salvatore, and C. Colombo. Validation of a predictive survival model in italian patients with cystic fibrosis. *J Cyst Fibros*, 11(1) :24–9, 2012. 10, 57, 59

C. Castellani, H. Cuppens, J. Macek, M., J. J. Cassiman, E. Kerem, P. Durie, E. Tullis, B. M. Assael, C. Bombieri, A. Brown, T. Casals, M. Claustres, G. R. Cutting, E. Dequeker, J. Dodge, I. Doull, P. Farrell, C. Ferec, E. Girodon, M. Johannesson, B. Kerem, M. Knowles, A. Munck, P. F. Pignatti, D. Radojkovic, P. Rizzotti, M. Schwarz, M. Stuhrmann, M. Tzetis, J. Zielenski, and J. S. Elborn. Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. *J Cyst Fibros*, 7(3) :179–96, 2008. 4, 5

F. S. Collins. Cystic fibrosis : molecular biology and therapeutic implications. *Science*, 256(5058) :774–779, 1992. 11

M. Corey and V. Farewell. Determinants of mortality from cystic fibrosis in canada, 1970-1989. *Am J Epidemiol*, 143(10) :1007–17, 1996. 9

M. Corey, L. Edwards, H. Levison, and M. Knowles. Longitudinal analysis of pulmonary function decline in patients with cystic fibrosis. *J Pediatr*, 131(6) : 809–14, 1997. 9

D. B. Coultas, C. A. Howard, B. J. Skipper, and J. M. Samet. Spirometric prediction equations for hispanic children and adults in new mexico. *Am Rev Respir Dis*, 138 (6) :1386–92, 1988. 8

J. M. Courtney, K. E. Dunbar, A. McDowell, J. E. Moore, T. J. Warke, M. Stevenson, and J. S. Elborn. Clinical outcome of burkholderia cepacia complex infection in cystic fibrosis adults. *J Cyst Fibros*, 3(2) :93–8, 2004. 57

J. M. Courtney, J. Bradley, J. McCaughan, T. M. O'Connor, C. Shortt, C. P. Bredin, I. Bradbury, and J. S. Elborn. Predictors of mortality in adults with cystic fibrosis. *Pediatr Pulmonol*, 42(6) :525–32, 2007. 10, 57

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B-Methodological*, 34 :187–220, 1972. 35

R. O. Crapo, A. H. Morris, and R. M. Gardner. Reference spirometric values using techniques and equipment that meet ats recommendations. *Am Rev Respir Dis*, 123(6) :659–64, 1981. 8

E. Dantan, P. Joly, J. F. Dartigues, and H. Jacqmin-Gadda. Joint model with latent state for longitudinal and multistate data. *Biostatistics*, 12(4) :723–36, 2011. 104

E. C. Dasenbrook, C. A. Merlo, M. Diener-West, N. Lechtzin, and M. P. Boyle. Persistent methicillin-resistant staphylococcus aureus and rate of fev1 decline in cystic fibrosis. *Am J Respir Crit Care Med*, 178(8) :814–21, 2008. 6

J. C. Davies and E. W. Alton. Monitoring respiratory disease severity in cystic fibrosis. *Respir Care*, 54(5) :606–17, 2009. 7

J. C. Davies, C. E. Wainwright, G. J. Canny, M. A. Chilvers, M. S. Howenstine, A. Munck, J. G. Mainz, S. Rodriguez, H. Li, K. Yen, C. L. Ordonez, R. Ahrens, and V. X. S. Group. Efficacy and safety of ivacaftor in patients aged 6 to 11 years with cystic fibrosis with a g551d mutation. *Am J Respir Crit Care Med*, 187(11) : 1219–25, 2013. 12

J. C. Davies, S. Cunningham, W. T. Harris, A. Lapey, W. E. Regelmann, G. S. Sawicki, K. W. Southern, S. Robertson, Y. Green, J. Cooke, M. Rosenfeld, and K. S. Group. Safety, pharmacokinetics, and pharmacodynamics of ivacaftor in patients aged 2-5 years with cystic fibrosis and a cftr gating mutation (kiwi) : an open-label, single-arm study. *Lancet Respir Med*, 4(2) :107–15, 2016. 12

K. De Boeck. Improving standards of clinical care in cystic fibrosis. *Eur Respir J*, 16 (4) :585–7, 2000. 5

K. De Boeck, A. Malfroot, L. Van Schil, P. Lebecque, C. Knoop, J. R. Govan, C. Doherty, S. Laevens, P. Vandamme, and G. Belgian Burkholderia cepacia Study. Epidemiology of burkholderia cepacia complex colonisation in cystic fibrosis patients. *Eur Respir J*, 23(6) :851–6, 2004. 6

K. De Boeck, M. Wilschanski, C. Castellani, C. Taylor, H. Cuppens, J. Dodge, M. Sinaasappel, and G. Diagnostic Working. Cystic fibrosis : terminology and diagnostic algorithms. *Thorax*, 61(7) :627–35, 2006. 3

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1) :1–38, 1977. 44

P. Diggle and M. G. Kenward. Informative drop-out in longitudinal data-analysis. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 43(1) :49–93, 1994. 32

J. A. Dodge, P. A. Lewis, M. Stanton, and J. Wilsher. Cystic fibrosis mortality and survival in the uk : 1947-2003. *Eur Respir J*, 29(3) :522–6, 2007. 6

G. Doring, N. Hoiby, and G. Consensus Study. Early intervention and prevention of lung disease in cystic fibrosis : a european consensus. *J Cyst Fibros*, 3(2) :67–91, 2004. 5

M. L. Drumm, M. W. Konstan, M. D. Schluchter, A. Handler, R. Pace, F. Zou, M. Zariwala, D. Fargo, A. Xu, J. M. Dunn, R. J. Darrah, R. Dorfman, A. J. Sandford, M. Corey, J. Zielenski, P. Durie, K. Goddard, J. R. Yankaskas, F. A. Wright, M. R. Knowles, and G. Gene Modifier Study. Genetic modifiers of lung disease in cystic fibrosis. *N Engl J Med*, 353(14) :1443–53, 2005. 5

B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72 :557–565, 1977. 36

B. Efron and R. Tibshirani. Improvements on cross-validation : The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438) :548–560, 1997. 54

T. M. Egan, F. C. Detterbeck, M. R. Mill, L. J. Paradowski, R. P. Lackner, W. D. Ogden, J. R. Yankaskas, J. H. Westerman, J. T. Thompson, M. A. Weiner, and et al. Improved results of lung transplantation for patients with cystic fibrosis. *J Thorac Cardiovasc Surg*, 109(2) :224–34 ; discussion 234–5, 1995. 14

R. M. Elashoff, G. Li, and N. Li. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3) :762–771, 2008. 49

M. Ellaffi, C. Vinsonneau, J. Coste, D. Hubert, P. R. Burgel, J. F. Dhainaut, and D. Dusser. One-year outcome after severe pulmonary exacerbation in adults with cystic fibrosis. *Am J Respir Crit Care Med*, 171(2) :158–64, 2005. 6

P. M. Farrell. The prevalence of cystic fibrosis in the european union. *J Cyst Fibros*, 7(5) :450–3, 2008. 4

P. M. Farrell, B. J. Rosenstein, T. B. White, F. J. Accurso, C. Castellani, G. R. Cutting, P. R. Durie, V. A. Legrys, J. Massie, R. B. Parad, M. J. Rock, r. Campbell, P. W., and F. Cystic Fibrosis. Guidelines for diagnosis of cystic fibrosis in newborns through older adults : Cystic fibrosis foundation consensus report. *J Pediatr*, 153 (2) :S4–S14, 2008. 3, 4

P. M. Farrell, T. B. White, C. L. Ren, S. E. Hempstead, F. Accurso, N. Derichs, M. Howenstine, S. A. McColley, M. Rock, M. Rosenfeld, I. Sermet-Gaudelus, K. W. Southern, B. C. Marshall, and P. R. Sosnay. Diagnosis of cystic fibrosis : Consensus guidelines from the cystic fibrosis foundation. *J Pediatr*, 181S :S4–S15 e1, 2017. 4

C. L. Faucett and D. C. Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates : A gibbs sampling approach. *Statistics in Medicine*, 15(15) :1663–1685, 1996. 41

B. Fauroux, E. Le Roux, S. Ravilly, G. Bellis, and A. Clement. Long-term noninvasive ventilation in patients with cystic fibrosis. *Respiration*, 76(2) :168–174, 2008. 71

L. Ferrer, V. Rondeau, J. Dignam, T. Pickles, H. Jacqmin-Gadda, and C. Proust-Lima. Joint modelling of longitudinal and multi-state processes : application to clinical progressions in prostate cancer. *Statistics in Medicine*, 35(22) :3933–3948, 2016. 49, 104

P. A. Flume. Pulmonary complications of cystic fibrosis. *Respiratory Care*, 54(5) : 618–625, 2009. 9

P. A. Flume, B. P. O'Sullivan, K. A. Robinson, C. H. Goss, P. J. Mogayzel, D. B. Willey-Courand, J. Bujan, J. Finder, M. Lesters, L. Quittell, R. Rosenblatt, R. L. Vender, L. Hlazle, K. Sabadosa, and B. Marshall. Cystic fibrosis pulmonary guidelines - chronic medications for maintenance of lung health. *American Journal of Respiratory and Critical Care Medicine*, 176(10) :957–969, 2007. 11, 72

C. F. Foundation. Patient registry : 2016 annual data report. 2016. 4, 5, 7

P. M. George, W. Banya, N. Pareek, D. Bilton, P. Cullinan, M. E. Hodson, and N. J. Simmonds. Improved survival at low lung function in cystic fibrosis : cohort study from 1990 to 2007. *BMJ*, 342 :d1008, 2011. 10, 57

T. A. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biom J*, 48(6) :1029–40, 2006. 51

H. W. Glindmeyer, J. J. Lefante, C. McColloster, R. N. Jones, and H. Weill. Blue-collar normative spirometric values for caucasian and african-american men and women aged 18 to 65. *Am J Respir Crit Care Med*, 151(2 Pt 1) :412–22, 1995. 8

C. H. Goss, S. A. Newsom, J. S. Schildcrout, L. Sheppard, and J. D. Kaufman. Effect of ambient air pollution on pulmonary exacerbations and lung function in cystic fibrosis. *Am J Respir Crit Care Med*, 169(7) :816–21, 2004. 5

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18 (17-18) :2529–2545, 1999. 51, 53

X. Guo and B. P. Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *American Statistician*, 58(1) :16–24, 2004. 44

J. Han, E. H. Slate, and E. A. Pena. Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in Medicine*, 26(29) :5285–5302, 2007. 49

J. L. Hankinson, J. R. Odencrantz, and K. B. Fedan. Spirometric reference values from a sample of the general u.s. population. *Am J Respir Crit Care Med*, 159(1) : 179–87, 1999. 9, 101, 102

F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Jama-Journal of the American Medical Association*, 247 (18) :2543–2546, 1982. 51

D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358) : 320–338, 1977. 30

P. J. Heagerty, T. Lumley, and M. S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2) :337–344, 2000. 51

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4) :465–80, 2000. 41, 44

M. E. Hodson, B. P. Madden, M. H. Steven, V. T. Tsang, and M. H. Yacoub. Non-invasive mechanical ventilation for cystic fibrosis patients–a potential bridge to transplantation. *Eur Respir J*, 4(5) :524–7, 1991. 71

J. L. Hook and D. J. Lederer. Selecting lung transplant candidates : where do current guidelines fall short ? *Expert Review of Respiratory Medicine*, 6(1) :51–61, 2012. 72

D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*, 16(9) :965–80, 1997. 26

X. Huang, G. Li, R. M. Elashoff, and J. X. Pan. A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime Data Analysis*, 17(1) :80–100, 2011. 49

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. Wiley-Interscience, NJ, 2002. 37

A. Kelly and A. Moran. Update on cystic fibrosis-related diabetes. *J Cyst Fibros*, 12 (4) :318–31, 2013. 2

B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui. Identification of the cystic fibrosis gene : genetic analysis. *Science*, 245(4922) :1073–80, 1989. 1

E. Kerem, J. Reisman, M. Corey, G. J. Canny, and H. Levison. Prediction of mortality in patients with cystic fibrosis. *N Engl J Med*, 326(18) :1187–91, 1992. 8, 10, 57

E. A. Knapp, A. K. Fink, C. H. Goss, A. Sewall, J. Ostrenga, C. Dowd, A. Elbert, K. M. Petren, and B. C. Marshall. The cystic fibrosis foundation patient registry. design and methods of a national observational disease registry. *Ann Am Thorac Soc*, 13(7) :1173–9, 2016. 56

R. J. Knudson, M. D. Lebowitz, C. J. Holberg, and B. Burrows. Changes in the normal maximal expiratory flow-volume curve with growth and aging. *Am Rev Respir Dis*, 127(6) :725–34, 1983. 8, 101

M. W. Konstan, S. M. Butler, M. E. Wohl, M. Stoddard, R. Matousek, J. S. Wagener, C. A. Johnson, W. J. Morgan, Investigators, and F. Coordinators of the Epidemiologic Study of Cystic. Growth and nutritional indexes in early life predict pulmonary function in cystic fibrosis. *J Pediatr*, 142(6) :624–30, 2003. 11

M. W. Konstan, W. J. Morgan, S. M. Butler, D. J. Pasta, M. L. Craib, S. J. Silva, D. C. Stokes, M. E. Wohl, J. S. Wagener, W. E. Regelmann, C. A. Johnson, G. Scientific Advisory, I. the, and F. Coordinators of the Epidemiologic Study of Cystic. Risk factors for rate of decline in forced expiratory volume in one second in children and adolescents with cystic fibrosis. *J Pediatr*, 151(2) :134–9, 139 e1, 2007. 9

M. R. Kosorok, L. Zeng, S. E. West, M. J. Rock, M. L. Splaingard, A. Laxova, C. G. Green, J. Collins, and P. M. Farrell. Acceleration of lung disease in children with cystic fibrosis after pseudomonas aeruginosa acquisition. *Pediatr Pulmonol*, 32(4) : 277–87, 2001. 6

A. Krol, L. Ferrer, J. P. Pignon, C. Proust-Lima, M. Ducreux, O. Bouche, S. Michiels, and V. Rondeau. Joint model for left-censored longitudinal data, recurrent events and terminal event : Predictive abilities of tumor burden for cancer evolution with application to the ffcd 2000-05 trial. *Biometrics*, 72(3) :907–916, 2016. 49

C. Kugler, M. Strueber, U. Tegtbur, J. Niedermeyer, and A. Haverich. Quality of life 1 year after lung transplantation. *Prog Transplant*, 14(4) :331–6, 2004. 13

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4) :963–74, 1982. 28

R. Lama, A. Alvarez, F. Santos, J. Algar, J. L. Aranda, C. Baamonde, and A. Salvatierra. Long-term results of lung transplantation for cystic fibrosis. *Transplant Proc*, 33(1-2) :1624–5, 2001. 14

M. Lee, K. A. Cronin, M. H. Gail, and E. J. Feuer. Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data. *Statistics in Medicine*, 31(5) :489–500, 2012. 54

D. Y. Lin, L. J. Wei, and Z. Ying. Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3) :557–572, 1993. 37

H. Q. Lin, B. W. Turnbull, C. E. McCulloch, and E. H. Slate. Latent class models for joint analysis of longitudinal biomarker and event process data : Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457) :53–65, 2002. 45

T. G. Liou, F. R. Adler, S. C. Fitzsimmons, B. C. Cahill, J. R. Hibbs, and B. C. Marshall. Predictive 5-year survivorship model of cystic fibrosis. *Am J Epidemiol*, 153(4) :345–52, 2001. 9, 10, 58

T. G. Liou, F. R. Adler, D. R. Cox, and B. C. Cahill. Lung transplantation and survival in children with cystic fibrosis. *N Engl J Med*, 357(21) :2143–52, 2007. 10

J. J. Lipuma. The changing microbial epidemiology in cystic fibrosis. *Clin Microbiol Rev*, 23(2) :299–323, 2010. 6

R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987. 31

R. Little and D. Rubin. *Statistical Analysis with Missing Data*. 2nd edition. Wiley, New York, 2002. 31

R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421) :125–134, 1993. 32

L. Liu and X. L. Huang. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 58 :65–81, 2009. 49

J. B. Lyczak, C. L. Cannon, and G. B. Pier. Lung infections associated with cystic fibrosis. *Clin Microbiol Rev*, 15(2) :194–222, 2002. 2, 6

J. Macek, M., A. Mackova, A. Hamosh, B. C. Hilman, R. F. Selden, G. Lucotte, K. J. Friedman, M. R. Knowles, B. J. Rosenstein, and G. R. Cutting. Identification of common cystic fibrosis mutations in african-americans with cystic fibrosis increases the detection rate to 75 *Am J Hum Genet*, 60(5) :1122–7, 1997. 5

T. MacKenzie, A. H. Gifford, K. A. Sabadosa, H. B. Quinton, E. A. Knapp, C. H. Goss, and B. C. Marshall. Longevity of patients with cystic fibrosis in 2000 to 2010 and beyond : survival analysis of the cystic fibrosis foundation patient registry. *Ann Intern Med*, 161(4) :233–41, 2014. 6, 56

B. P. Madden, H. Kariyawasam, A. J. Siddiqi, A. Machin, J. A. Pryor, and M. E. Hodson. Noninvasive ventilation in cystic fibrosis patients with acute or chronic respiratory failure. *Eur Respir J*, 19(2) :310–3, 2002. 71

C. Martin, C. Hamard, R. Kanaan, V. Boussaud, D. Grenet, M. Abely, D. Hubert, A. Munck, L. Lemonnier, and P. R. Burgel. Causes of death in french cystic fibrosis patients : The need for improvement in transplantation referral strategies ! *J Cyst Fibros*, 15(2) :204–12, 2016. 98

A. Mauguen, B. Rachet, S. Mathoulin-Pelissier, G. MacGrogan, A. Laurent, and V. Rondeau. Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Statistics in Medicine*, 32(30) :5366–5380, 2013. 49

M. May, P. Royston, M. Egger, A. C. Justice, and J. A. C. Sterne. Development and validation of a prognostic model for survival time data : application to prognosis of hiv positive patients treated with antiretroviral therapy. *Statistics in Medicine*, 23 (15) :2375–2398, 2004. 54

N. Mayer-Hamblett, M. Rosenfeld, J. Emerson, C. H. Goss, and M. L. Aitken. Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality. *Am J Respir Crit Care Med*, 166(12 Pt 1) :1550–5, 2002. 10, 57, 58

C. McCarthy, B. D. Dimitrov, I. J. Meurling, C. Gunaratnam, and N. G. McElvaney. The cf-able score : a novel clinical prediction rule for prognosis in patients with cystic fibrosis. *Chest*, 143(5) :1358–64, 2013. 10, 98

B. Michiels, G. Molenberghs, L. Bijnens, T. Vangeneugden, and H. Thijs. Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Stat Med*, 21(8) :1023–41, 2002. 32

P. J. Mogayzel, E. T. Naureckas, K. A. Robinson, G. Mueller, D. Hadjiliadis, J. B. Hoag, L. Lubsch, L. Hazle, K. Sabadosa, B. Marshall, and P. C. P. Guidel. Cystic fibrosis pulmonary guidelines chronic medications for maintenance of lung health. *American Journal of Respiratory and Critical Care Medicine*, 187(7) :680–689, 2013. 72

A. Moran, J. Dunitz, B. Nathan, A. Saeed, B. Holme, and W. Thomas. Cystic fibrosis-related diabetes : current trends in prevalence, incidence, and mortality. *Diabetes Care*, 32(9) :1626–31, 2009. 2

N. Morral, J. Bertranpetit, X. Estivill, V. Nunes, T. Casals, J. Gimenez, A. Reis, R. Varon-Mateeva, J. Macek, M., L. Kalaydjieva, and et al. The origin of the major cystic fibrosis mutation (delta f508) in european populations. *Nat Genet*, 7(2) : 169–75, 1994. 5

G. M. Nixon, D. S. Armstrong, R. Carzino, J. B. Carlin, A. Olinsky, C. F. Robertson, and K. Grimwood. Clinical outcome after early pseudomonas aeruginosa infection in cystic fibrosis. *J Pediatr*, 138(5) :699–704, 2001. 6

A. Oliver, R. Canton, P. Campo, F. Baquero, and J. Blazquez. High frequency of hypermutable pseudomonas aeruginosa in cystic fibrosis lung infection. *Science*, 288(5469) :1251–4, 2000. 6

M. Olszowiec-Chlebna, A. Koniarek-Maniecka, W. Stelmach, K. Smejda, J. Jerzynska, P. Majak, M. Bialas, and I. Stelmach. Predictors of deterioration of lung function in polish children with cystic fibrosis. *Arch Med Sci*, 12(2) :402–7, 2016. 9

J. B. Orens, M. Estenne, S. Arcasoy, J. V. Conte, P. Corris, J. J. Egan, T. Egan, S. Keshavjee, C. Knoop, R. Kotloff, F. J. Martinez, S. Nathan, S. Palmer, A. Patterson, L. Singer, G. Snell, S. Studer, J. L. Vachiery, and A. R. Glanville. International guidelines for the selection of lung transplant candidates : 2006 update - a consensus report from the pulmonary scientific council of the international society for heart and lung transplantation. *Journal of Heart and Lung Transplantation*, 25(7) : 745–755, 2006. 71

B. P. O'Sullivan and S. D. Freedman. Cystic fibrosis. *Lancet*, 373(9678) :1891–904, 2009. 4

L. Parast, S. C. Cheng, and T. X. Cai. Landmark prediction of long-term survival incorporating short-term event time information. *Journal of the American Statistical Association*, 107(500) :1492–1501, 2012. 51

J. R. Phillips, T. J. Tripp, W. E. Regelmann, P. M. Schlievert, and O. D. Wangensteen. Staphylococcal alpha-toxin causes increased tracheal epithelial permeability. *Pediatr Pulmonol*, 41(12) :1146–52, 2006. 6

R. L. Prentice. Covariate measurement errors and parameter-estimation in a failure time regression-model. *Biometrika*, 69(2) :331–342, 1982. 40

T. Pressler, C. Bohmova, S. Conway, S. Dumcius, L. Hjelte, N. Hoiby, H. Kollberg, B. Tummler, and V. Vavrova. Chronic pseudomonas aeruginosa infection definition : Eurocarecf working group report. *J Cyst Fibros*, 10 Suppl 2 :S75–8, 2011. 6

C. Proust-Lima and J. M. G. Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment psa : a joint modeling approach. *Biostatistics*, 10(3) :535–549, 2009. 50

C. Proust-Lima, M. Sene, J. M. Taylor, and H. Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data : a review. *Stat Methods Med Res*, 23(1) :74–90, 2014. 45, 49

C. Proust-Lima, J. F. Dartigues, and H. Jacqmin-Gadda. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death : a latent process and latent class approach. *Statistics in Medicine*, 35(3) :382–398, 2016. 49, 104

C. Proust-Lima, V. Philipps, and B. Liquet. Estimation of extended mixed models using latent classes and latent processes : The r package lcmm. *Journal of Statistical Software*, 78(2) :1–56, 2017. IV, 48

H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics : competing risks and multi-state models. *Stat Med*, 26(11) :2389–430, 2007. 38

B. W. Ramsey, J. Davies, N. G. McElvaney, E. Tullis, S. C. Bell, P. Drevinek, M. Griese, E. F. McKone, C. E. Wainwright, M. W. Konstan, R. Moss, F. Ratjen, I. Sermet-Gaudelus, S. M. Rowe, Q. Dong, S. Rodriguez, K. Yen, C. Ordonez, J. S. Elborn, and V. X. S. Group. A cftr potentiator in patients with cystic fibrosis and the g551d mutation. *N Engl J Med*, 365(18) :1663–72, 2011. 12

R. J. Raymond, A. L. Hinderliter, P. W. Willis, D. Ralph, E. J. Caldwell, W. Williams, N. A. Ettinger, N. S. Hill, W. R. Summer, B. de Boisblanc, T. Schwartz, G. Koch, L. M. Clayton, M. M. Jobsis, J. W. Crow, and W. Long. Echocardiographic predictors of adverse outcomes in primary pulmonary hypertension. *J Am Coll Cardiol*, 39(7) :1214–9, 2002. 98

D. W. Reid, C. L. Blizzard, D. M. Shugg, C. Flowers, C. Cash, and H. M. Greville. Changes in cystic fibrosis mortality in australia, 1979-2005. *Med J Aust*, 195(7) : 392–5, 2011. 6

J. R. Riordan, J. M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J. L. Chou, and et al. Identification of the cystic fibrosis gene : cloning and characterization of complementary dna. *Science*, 245 (4922) :1066–73, 1989. 1

D. Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3) :819–829, 2011. 49, 50

D. Rizopoulos. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7) :1–46, 2016. 44

D. Rizopoulos and P. Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30 (12) :1366–1380, 2011. 104

D. Rizopoulos, G. Verbeke, and E. Lesaffre. Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 71 :637–654, 2009. 44

J. M. Rommens, M. C. Iannuzzi, B. Kerem, M. L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J. L. Cole, D. Kennedy, N. Hidaka, and et al. Identification of the cystic fibrosis gene : chromosome walking and jumping. *Science*, 245(4922) : 1059–65, 1989. 1

P. S. Rosenberg. Hazard function estimation using b-splines. *Biometrics*, 51(3) : 874–87, 1995. 42

A. Rouanet, P. Joly, J. F. Dartigues, C. Proust-Lima, and H. Jacqmin-Gadda. Joint latent class model for longitudinal data and interval-censored semi-competing events : Application to dementia. *Biometrics*, 72(4) :1123–1135, 2016. 49

D. B. Sanders, A. Fink, N. Mayer-Hamblett, M. S. Schechter, G. S. Sawicki, M. Rosenfeld, P. A. Flume, and W. J. Morgan. Early life growth trajectories in cystic fibrosis are associated with pulmonary function at age 6 years. *J Pediatr*, 167(5) : 1081–8 e1, 2015. 11

C. Schaedel, I. de Monestrol, L. Hjelte, M. Johannesson, R. Kornfalt, A. Lindblad, B. Strandvik, L. Wahlgren, and L. Holmberg. Predictors of deterioration of lung function in cystic fibrosis. *Pediatr Pulmonol*, 33(6) :483–91, 2002. 9

M. Schemper and R. Henderson. Predictive accuracy and explained variation in cox regression. *Biometrics*, 56(1) :249–55, 2000. 51

M. D. Schluchter, M. W. Konstan, and P. B. Davis. Jointly modelling the relationship between survival and pulmonary function in cystic fibrosis patients. *Stat Med*, 21 (9) :1271–87, 2002. 9

J. D. Schwartz, S. A. Katz, R. W. Fegley, and M. S. Tockman. Analysis of spirometric data from a national sample of healthy 6- to 24-year-olds (nhanes ii). *Am Rev Respir Dis*, 138(6) :1405–14, 1988. 8

H. Shwachman and A. Mahmoodian. Pilocarpine iontophoresis sweat testing results of seven years' experience. *Bibl Paediatr*, 86 :158–82, 1967. 3

E. J. Sims, J. McCormick, G. Mehta, A. Mehta, and U. K. C. F. D. Steering Committee of the. Neonatal screening for cystic fibrosis is beneficial even in the context of modern treatment. *J Pediatr*, 147(3 Suppl) :S42–6, 2005. 3

X. Song, M. Davidian, and A. A. Tsiatis. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58(4) :742–53, 2002. 44

J. E. Spahr, R. B. Love, M. Francois, K. Radford, and K. C. Meyer. Lung transplantation for cystic fibrosis : current concepts and one center's experience. *J Cyst Fibros*, 6(5) :334–50, 2007. 14

R. Szczesniak, S. L. Heltshe, S. Stanojevic, and N. Mayer-Hamblett. Use of fev1 in cystic fibrosis epidemiologic studies and clinical trials : A statistical perspective for the clinical researcher. *J Cyst Fibros*, 16(3) :318–326, 2017. 72

L. M. Taussig, J. Kattwinkel, W. T. Friedewald, and P. A. Di Sant'Agnese. A new prognostic score and clinical evaluation system for cystic fibrosis. *J Pediatr*, 82(3) : 380–90, 1973. 7

G. Thabut, J. D. Christie, H. Mal, M. Fournier, O. Brugiere, G. Leseche, Y. Castier, and D. Rizopoulos. Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation. *Am J Respir Crit Care Med*, 187(12) :1335–40, 2013. 13, 98

J. A. Towbin, A. M. Lowe, S. D. Colan, L. A. Sleeper, E. J. Orav, S. Clunie, J. Messere, G. F. Cox, P. R. Lurie, D. Hsu, C. Canter, J. D. Wilkinson, and S. E. Lipshultz. Incidence, causes, and outcomes of dilated cardiomyopathy in children. *JAMA*, 296(15) :1867–76, 2006. 98

D. S. Urquhart, L. P. Thia, J. Francis, S. A. Prasad, C. Dawson, C. Wallis, and I. M. Balfour-Lynn. Deaths in childhood from cystic fibrosis : 10-year analysis from two london specialist centres. *Arch Dis Child*, 98(2) :123–7, 2013. 56, 97

H. C. van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1) :70–85, 2007. 51

D. R. VanDevanter, J. S. Wagener, D. J. Pasta, E. Elkin, J. R. Jacobs, W. J. Morgan, and M. W. Konstan. Pulmonary outcome prediction (pop) tools for cystic fibrosis patients. *Pediatr Pulmonol*, 45(12) :1156–66, 2010. 7

G. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91 (433) :217–221, 1996. 32

K. M. Vermeulen, W. van der Bij, M. E. Erasmus, E. J. Duiverman, G. H. Koeter, and E. M. TenVergert. Improved quality of life after lung transplantation in individuals with cystic fibrosis. *Pediatr Pulmonol*, 37(5) :419–26, 2004. 13

C. E. Wainwright, J. S. Elborn, B. W. Ramsey, G. Marigowda, X. Huang, M. Cipolli, C. Colombo, J. C. Davies, K. De Boeck, P. A. Flume, M. W. Konstan, S. A.

McColley, K. McCoy, E. F. McKone, A. Munck, F. Ratjen, S. M. Rowe, D. Waltz, M. P. Boyle, T. S. Group, and T. S. Group. Lumacaftor-ivacaftor in patients with cystic fibrosis homozygous for phe508del cftr. *N Engl J Med*, 373(3) :220–31, 2015. 12

X. Wang, D. W. Dockery, D. Wypij, M. E. Fay, and J. Ferris, B. G. Pulmonary function between 6 and 18 years of age. *Pediatr Pulmonol*, 15(2) :75–88, 1993. 8

V. Waters, E. G. Atenafu, A. Lu, Y. Yau, E. Tullis, and F. Ratjen. Chronic stenotrophomonas maltophilia infection and mortality or lung transplantation in cystic fibrosis patients. *J Cyst Fibros*, 12(5) :482–6, 2013. 98

M. Wilschanski and P. R. Durie. Patterns of gi disease in adulthood associated with mutations in the cftr gene. *Gut*, 56(8) :1153–63, 2007. 2

C. Winstanley, S. O'Brien, and M. A. Brockhurst. Pseudomonas aeruginosa evolutionary adaptation and diversification in cystic fibrosis chronic lung infections. *Trends Microbiol*, 24(5) :327–37, 2016. 6

M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1) :330–339, 1997. 41, 44

J. R. Yankaskas and J. Mallory, G. B. Lung transplantation in cystic fibrosis : consensus conference statement. *Chest*, 113(1) :217–26, 1998. 14

W. Ye, X. H. Lin, and J. M. G. Taylor. A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and Its Interface*, 1(1) :33–45, 2008. 43, 44

B. B. Yu and P. Ghosh. Joint modeling for cognitive trajectory and risk of dementia in the presence of death. *Biometrics*, 66(1) :294–300, 2010. 49

Y. Y. Zheng, T. X. Cai, Y. Y. Jin, and Z. D. Feng. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics*, 68(2) :388–396, 2012. 52

J. Zielenski and L. C. Tsui. Cystic fibrosis : Genotypic and phenotypic variations. *Annual Review of Genetics*, 29 :777–807, 1995. 5

Valorisations scientifiques

Publications :

— Nkam L. *et al.*, A 3-year prognostic score for adults with cystic fibrosis, Journal of Cystic Fibrosis. 2017.

— Nkam L. *et al.*, Dynamic prediction of prognosis in adults with Cystic Fibrosis : a joint latent class model. Travail en préparation.

Communications orales :

— Nkam L. *et al.*, Joint model for longitudinal marker and time-to-event in presence of competing risks : application on cystic fibrosis. International Society for Clinical Biostatistics. Student Day. Juillet 2017. Vigo, Espagne.

— Nkam L. *et al.*, A 3-year prognostic score for adults with cystic fibrosis. European Cystic Fibrosis Society Conference. Juin 2017. Seville, Espagne. **Invitée**.

— Nkam L. *et al.*, Joint model for longitudinal marker (FEV_1) and time-to-event in presence of competing risks (death without lung transplantation and lung transplantation) in adults with cystic fibrosis. European Cystic Fibrosis Society Conference. Juin 2017. Seville, Espagne.

— Nkam L. *et al.*, A 3-year predictive model of French patients with cystic fibrosis. European Cystic Fibrosis Society Conference. Juin 2016. Bâle, Suisse. **Young Investigator Award**.

— Nkam L. *et al.*, Prédiction de la survie à 5 ans des patients atteints de mucoviscidose en France. Colloque Français des Jeunes Chercheurs Mucoviscidose. Février 2016. Paris, France.

Communications affichées :

— Nkam L. *et al.*, Joint model for longitudinal marker and time-to-event in presence of competing risks : application on cystic fibrosis. International Society for Clinical Biostatistics. Juillet 2017. Vigo, Espagne.

— Nkam L. *et al.*, Dynamic prediction of death or lung transplantation for patients with cystic fibrosis using joint model for longitudinal and time-to-event data. Colloque Français des Jeunes Chercheurs Mucoviscidose. Février 2017. Paris, France.

— Nkam L. *et al.*, Dynamic prediction of death or lung transplantation for patients with cystic fibrosis using joint model for longitudinal and time-to-event data . North American Cystic Fibrosis Conference. Octobre 2016. Orlando, États-Unis.

— Nkam L. *et al.*, Prédiction du pronostic des patients atteints de mucoviscidose. Colloque Français des Jeunes Chercheurs Mucoviscidose. Février 2015. Paris, France.

Dorette Lionelle Nkam Beriye

Prédiction du pronostic des patients atteints de mucoviscidose

Résumé : La mucoviscidose est une maladie génétique rare et incurable. Malgré les nombreux progrès réalisés dans la recherche à ce sujet, il reste indispensable d'avoir davantage une meilleure connaissance de la maladie afin de proposer des traitements encore plus adaptés aux patients. La majorité des traitements actuels visent principalement à réduire les symptômes de la maladie sans toutefois la guérir. La transplantation pulmonaire reste le moyen le plus adéquat pour améliorer la qualité de vie et prolonger la vie des patients dont l'état respiratoire s'est considérablement dégradé. Il est donc nécessaire de fournir aux cliniciens des outils d'aide à la décision pour mieux identifier les patients nécessitant une transplantation pulmonaire. Pour ce faire, il est indispensable de connaître d'une part, les facteurs pronostiques de la transplantation pulmonaire et d'autre part, de savoir prédire la survenue de cet événement chez les sujets atteints de mucoviscidose. L'objectif de ce travail de thèse est de développer des outils pronostiques utiles à l'évaluation des choix thérapeutiques liés à la transplantation pulmonaire. Dans un premier temps, nous avons réévalué les facteurs pronostiques de la transplantation pulmonaire ou du décès sans transplantation pulmonaire chez les adultes atteints de mucoviscidose. Suite aux progrès thérapeutiques qui ont conduit à l'amélioration du pronostic au cours des dernières années, ce travail a permis d'identifier les facteurs pronostiques en adéquation avec l'état actuel de la recherche. Un deuxième travail a consisté à développer un modèle conjoint à classes latentes fournissant des prédictions dynamiques pour la transplantation pulmonaire ou le décès sans transplantation pulmonaire. Ce modèle a permis d'identifier trois profils d'évolution de la maladie et également d'actualiser le risque de survenue de la transplantation pulmonaire ou du décès sans transplantation pulmonaire à partir des données longitudinales du marqueur VEMS. Ces modèles pronostiques ont été développés à partir des données du registre français de la mucoviscidose et ont fourni de bonnes capacités prédictives en termes de discrimination et de calibration.

Mots clés : mucoviscidose, transplantation pulmonaire, modèles conjoints, prédiction dynamique

Abstract : Cystic Fibrosis is an incurable inherited disorder. Despite real progress in research, it is essential to always have a better understanding of the disease in order to provide suitable treatments to patients. Current treatments mostly aim to reduce the disease symptoms without curing it. Lung transplantation is proposed to cystic fibrosis patients with terminal respiratory failure with the aim of improving life expectancy and quality of life. It is necessary to guide clinicians in identifying in a good way patients requiring an evaluation for lung transplantation. It is thus important to clearly identify prognostic factors related to lung transplantation and to predict in a good way the occurrence of this event in patients with cystic fibrosis. The aim of this work was to develop prognostic tools to assist clinicians in the evaluation of different therapeutic options related to lung transplantation. First, we reevaluated prognostic factors of lung transplantation or death without lung transplantation in adult with cystic fibrosis. Indeed, therapeutic progress in patients with cystic fibrosis has resulted in improved prognosis over the past decades. We identified prognostic factors related to the current state of research in the cystic fibrosis field. We further developed a joint model with latent classes which provided dynamic predictions for lung transplantation or death without lung transplantation. This model identified three profiles of the evolution of the disease and was able to update the risk of lung transplantation or death without lung transplantation taking into account the evolution of the longitudinal marker FEV_1 which describes the lung function. These prognostic models were developed using the French cystic fibrosis registry and provided good predictive accuracy in terms of discrimination and calibration.

Key words : cystic fibrosis, lung transplantation, joint models, dynamic prediction