Identification par fouille interactive de biomarqueurs pour l'aide au diagnostic de la sclérose latérale amyotrophique

Arnaud Soulet*, Arnaud Giacometti*,
Ines Akaichi **

*prénom.nom@univ-tours.fr
**prénom.nom@etu.univ-tours.fr

Résumé. l'analyse longitudinale et l'analyse de survie répondent à plusieurs questions de recherche dans le domaine médical comme le suivi de l'évolution de la maladie, la description des trajectoires de patients ou le pronostic de la maladie. Nous avons travaillé dans ce projet sur des données longitudinales des patients ayant la maladie SLA. Nous avons essayé de détailler en première étape les caractéristiques de ces données en décrivant des approches utiles pour leur manipulation. Et en deuxième étape, nous avons abordé la question du pronostic en interrogeant notre base de données cliniques et biologiques de patients ALS pour associer la survie à plusieurs variables cliniques et bilogiques.

1 Introduction

La sclérose latérale amyotrophique (SLA), plus connue sous le nom de maladie de Charcot, est une maladie neurodégénérative qui provoque la paralysie progressive des muscles impliqués dans la motricité volontaire . Elle affecte également la phonation, la déglutition et l'appareil respiratoire.Dans la majorité des cas, ces paralysies conduisent à un décès après une évolution moyenne de 3-5 ans (Swinnen et Robberecht, 2014). Le plus souvent, c'est l'atteinte des muscles respiratoires qui cause le décès. Les options thérapeutiques existantes prolongent la survie de quelques mois seulement(Miller et Moore, 2012). Les patients SLA affichent des patterns très différents de manifestation et de progression de la maladie. Cet hétérogénéité rend l'analyse difficile.
De nombreux challenges restent posés, en particulier pour l'aide à l'identification de sous-groupes de patients présentant des évolutions cliniques similaires, mais aussi pour aider à l'identification d'ensembles de biomarqueurs pour l'aide à la prédiction de l'évolution de la maladie.
Ainsi, le fait de réussir à regrouper les patients atteints de SLA en sous-groupes significatifs sur le plan clinique peut être très utile pour faire progresser le développement de traitements efficaces et obtenir de meilleurs soins pour les patients atteints de SLA.
Dans ce projet en partenariat avec une équipe médicale de l'hôpital universitaire de Tours, deux jeux de données sont principalement utilisés. Le premier est déjà intégré

dans un entrepôt de données qui était déjà mis en place et qui contient des données cliniques sur 1245 patients. Le deuxième contient des données biologiques sur l'ensemble de ces patients.
Dans cet article nous allons décrire en premier temps les spécificités des données reçues et puis nous allons discuter des modèles d'analyses présents dans la littérature pour le traitement de ces données dans le cas d'une analyse longitudinale et de survie. Et en deuxième temps, nous allons présenter des analyses de survie effectués sur des groupes des patients que nous avons construits à l'aide de nos données cliniques et biologiques.

2 Contexte

Suivant Robert et Neta (2019), la compréhension des sous-groupes de patients est liée à la compréhension de la progression et de la survie chez les patients SLA. Ceci est déclaré après leur découverte des sous-populations de patients à partir de leurs durée de vie ainsi que d'autres variables comme le lieu d'apparition des premiers symptômes. Ils ont pu distinguer quatre groupes de patients : les patients à progression lente et les patients à progression rapide, ainsi que les patients dont le taux de progression moyen était soit précoce, soit tardif au début de la période d'observation clinique enregistrée. Les groupes distingués ne sont pas des groupes définitifs.
Dans ce sens, nous avons utilisé les données disponibles pour établir une analyse primaire de survie ainsi qu'une analyse longitudinale pour avoir un aperçu de la progression de la maladie chez les patients SLA inscrits à l'hôpital de Tours.
Dans les sections qui suivent, nous donnons un aperçu des données collectées et comment nous avons pu organisé ces données pour que nous puissions les exploiter après dans nos analyses.

2.1 Entrepôt de données :

Les données cliniques qui sont une partie des données disponibles étaient intégrées dans un entrepôt de données mis en place lors de stage d'été de l'année dernière dans le but de rassembler toutes les données historiques collectées sur les patients SLA inscrits à l'hôpital de Tours et qui viennent de différentes sources.
Récemment, nous avons reçu les données biologiques qui étaient intégrées dans l'entrepôt et puis exploitées avec les données cliniques dans les analyses effectuées dans la deuxième phase du projet transversal.

2.1.1 Description des données

la totalité des données reçues sont décrites dans le tableau suivant :

Description des données
Données sur le patient
Démographiques : date de naissance et sexe.
date et lieu d'apparition des premiers symptômes (bulbaire, spinale ou respiratoire).
date du diagnostic
Date de décès.
Utilisation du riluzole (Oui ou Non) et date de la première utilisation.
Données longitudinales cliniques
Poids et hauteur.
Score ALS-FRS et détails du score.
Mesures respiratoires : FVC (Forced Vital Capacity) et SVC (Slow Vital Capacity).
Données longitudinales biologiques
29 mesures biologiques :
Sodium, Potassium, Chlorures, Albumine, Calcium, Urée, Glucose,
Créatine kinase, Bêta 2 microglobuline, Créatinine, LDH, ASAT, ALAT, Phosphatases alc., Gamma-GT,
Préalbumine, Bilirubine conjuguée, Bilirubine non conjuguée, Bilirubine totale, HDL Cholestérol, LDL Cholestérol, Fer,
Transferrine, Capacité totale de fixation en transferrine, Ferritine, Triglycérides et CRP ;

Tab. 1 : Description des données collectées

La gravité des symptômes est souvent évaluée à l'aide d'une échelle fonctionnelle : ALSFRS-R (ALS Functional Rating Scale).L'échelle ALSFRS est une liste de 12 évaluations de la fonction motrice, respiratoire et bulbaire.
chaque évaluation prend une valeur de 0 à 4, 4 étant la plus élevée (fonction normale) et 0 étant sans fonction. Le score pour les questions individuelles sont ensuite additionnées pour générer le score ALSFRS-R.
Les Mesures respiratoires (FVC et SVC) sont deux mesures communiquées en pourcentages qui renseignent sur la fonction pulmonaire.

Les variations de score ALSFRS ,des mesures respiratoires et de poids sont reconnues comme des indicateurs de pronostic pour la maladie SLARobert et Neta (2019). Une étude récente de a déclaré que l'utilisation de la variation des sous-scores (Bulbaire et Spinale) combinée avec la variation du score total ALSFRS peut garantir un meilleur ajustement dans la prédiction de la survieJames et Tom (2016).

2.1.2 Présentation de DFM

La figure 1 montre le modèle dimensionnel de fait (DFM) de l'entrepôt de données conçu qui est composé de trois tables faits qui stocke un nouveau examen de patient à une date donnée en renseignant soit les informations cliniques soit les informations biologiques.

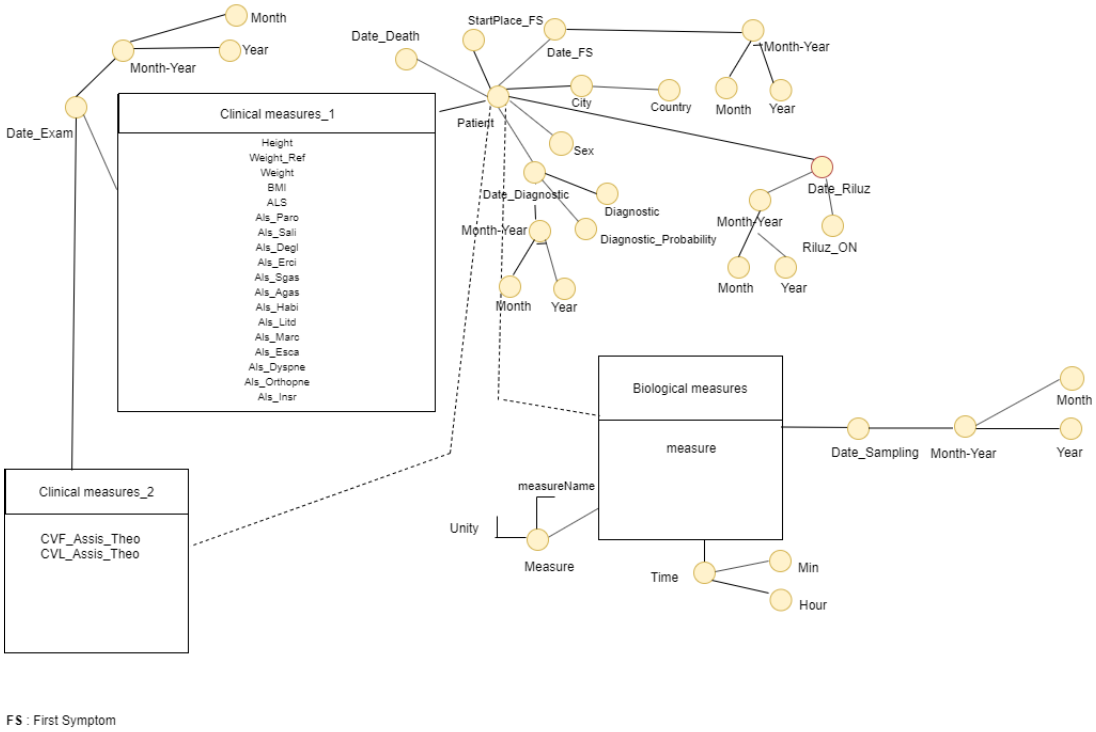

Fig. 1 : Le modèle DFM de l'entrepôt de données.

2.2 Motivation

L'aspect longitudinal des données va nous aider à suivre la progression ainsi que l'évolution de la maladie chez les patients SLA.
Cependant,Ces données ont d'autres spécificités : Elles sont également non alignées et incomplètes.

Données non alignées Lors de l'analyse des données longitudinales, le chercheur doit déterminer si les données sont «alignées» ou «non alignées». En règle générale, les plans d'étude longitudinaux exigent un nombre fixe de mesures répétées sur tous les participants. Lorsque tous les individus possèdent le même nombre de mesures, et que celles-ci ont été prises à intervalle de temps régulier , l'étude est dite "alignée" dans le temps. Bon nombre des premières méthodes statistiques élaborées pour l'analyse longitudinale (e.g : l'analyse de variance des mesures répétées) exigeaient que les données soient alignées.
Toutefois, dans les études longitudinales en sciences de la santé, surtout celles qui comportent des mesures répétées sur une période relativement longue, certaines personnes manquent presque toujours la date prévue de leur visite ou de leur observation.

Par conséquent, la séquence des temps d'observation n'est plus commune à tous les individus. Les données sont dites "non alignées" dans le temps(Xian, 2016).

Données non complètes Dans l'analyse des données longitudinales, les observations manquantes sont presque toujours présentes. Dans les études biomédicales, les sujets peuvent abandonner le suivi pour diverses raisons (santé, décès ...).
Si les données longitudinales à informations complètes sont très rares, une structure de données complète et parfaite ne correspond pas à la réalité de la dynamique physiologique et sociale. Par conséquent, l'une des caractéristiques les plus uniques des données longitudinales est la présence d'un grand nombre d'observations manquantes.
L'élimination de tous les cas pour lesquels il manque des observations rendra les résultats d'analyse moins fiables, d'autant plus que ceux qui ont été retirés d'une analyse longitudinale peuvent présenter des caractéristiques individuelles différentes de celles des cas pour lesquels l'information est complète.
Pour les petits échantillons, l'élimination de ceux dont les données sont incomplètes peut réduire la précision des estimations des paramètres en raison de la perte d'informations importantes. En effet, la modélisation par régression à partir de données longitudinales incomplètes est la marque distinctive de l'analyse moderne des données longitudinales(Xian, 2016).

Rappelons l'objectif de ce travail qui est l'analyse de survie ainsi la progression de la maladie.La non complétude et la non alignement des données doivent être pris en compte lors de ces différentes analyses.
Dans la section suivante, Nous allons commencer par présenter un état de l'art des méthodes utilisés et qui prennent en compte ces spécificités et puis montrer un aperçu des courbes d'analyse de survie et d'analyse longitudinale.

3 État de l'art

3.1 Analyse de Survie

Définition : L'analyse de la survie est un terme générique pour toute analyse de la survenue au cours du temps d'un événement, comme par exemple le décès.Ce type d'analyse est largement utilisé en épidémiologie clinique. Il permet la description de la survie d'un groupe de patients mais aussi la comparaison de la survie de deux ou plusieurs groupes afin d'étudier les facteurs pronostiques, c'est-à-dire les facteurs susceptibles d'expliquer la survenue du décès (ou d'un autre événement) au cours du temps(Alberti et al., 2005).

La survie à partir du diagnostic varie considérablement. Plusieurs facteurs pronostiques sont connus, y compris le site d'apparition (bulbe ou spinale), l'âge au moment de l'apparition des symptômes, le délai entre l'apparition et le diagnostic et le recours au riluzole et à la ventilation non invasive (Knibb et al., 2016).
L'un des challenges lors d'une analyse de survie est l'absence de la date de décès, on parle dans ce cas d'une censure à droite qui est une forme d'incomplétude des données.

Méthodes de modélisation : Le modèle standard des risques proportionnels de Cox (standard Cox proportional hazards model), couramment utilisé pour explorer la dépendance entre les caractéristiques cliniques et la survie, ignore ces cas censurés(Robert et Neta, 2019).
Cependant, la courbe kaplan Meier traite ces cas mais ne peut pas intégrer d'autres variables prédictives pour la prédiction de la survie (Knibb et al., 2016).
Une nouvelle méthode qui prend en compte à la fois des données censurées et intègre d'autres variables prédictives est proposé dans Robert et Neta (2019) appelée Guan-Rank.
Dans la section suivante,nous allons montrer notre analyse de survie établie sur deux jeux de données dont le premier contient des données censurées et le deuxième contient que des données dont la date de décès est toujours fournie.
Nous voulons montrer par cette comparaison l'effet d'éliminer les données censurées sur les résultats obtenus et l'importance d'utiliser des modèles qui peuvent traiter ces cas de censure.

3.1.1 Application

Description de jeu de données : Nous avons pris un jeu de données extrait de note entrepôt de données avec 897 patients dont les dates de premier symptômes sont disponibles.Le jeu de données comprend 4 attributs :

- l'identifiant du patient.
- un attribut Booléen qui renseigne sur le pronostic vital du patient.1 s'il est décédé,0 sinon.
- Un attribut qui renseigne sur la durée de survie du patient calculée à partir de la date de décès et la date d'apparition de premier symptôme.Si la date de décès est inconnu ,on calcule la durée de survie à partir de date de dernier examen clinique qui peut représenter dans notre cas la date des dernières nouvelles sur le patient.
- Un attribut qui renseigne sur le site d'apparition de premier symptôme. Cet attribut sert à regrouper les patients et sera utilisé pour montrer La variation de durée de survie entre les différents groupes.

Statistiques descriptifs : Dans le tableau 1, nous montrons quelques statistiques sur le jeu de données utilisé.

n=897	
Site d'apparition des premiers symptomes (%)	
Bulbaire	69
Spinale	31
Décédés (%)	
Oui	56
Non	44
Durée de survie(mois, médiane, min ; max)	26 [3 ;175]

TAB. 2 : statistiques descriptives

La courbe de survie : A partir de ces données, nous avons crée deux courbes de survie avec l'estimateur Kaplan Meier pour deux groupes de patients : Des patients dont le site d'apparition des premiers symptômes est Bulbaire et des patients dont le site d'apparition des premiers symptômes est spinale(Figure 2).

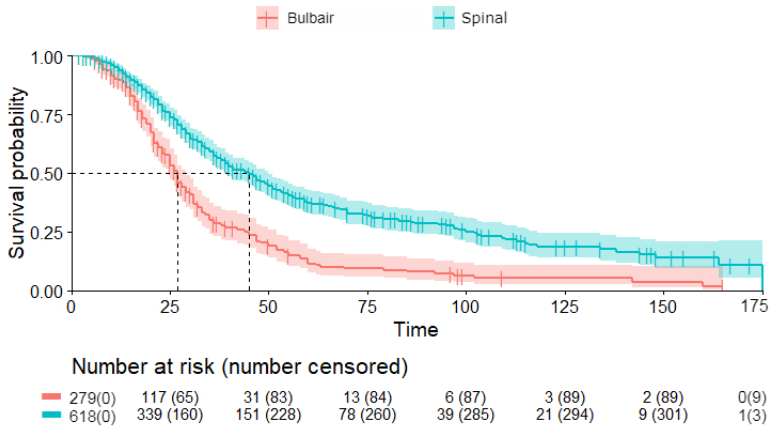

FIG. 2 : Kaplan Meier sans données censurées.

Interprétation : En observant les deux estimateurs de survie , nous remarquons que les patients du premier groupe (Bulbaire) ont une durée de survie médiane de 27 mois.Par contre,pour les patients du deuxième groupe (Spinal),Nous observons une durée de survie médiane de 43 mois.
Nous avons pu observer aussi que la durée de survie du premier groupe décroît plus vite que les patients du deuxième groupe. Nos résultats appuient sur les constatations de Knibb et al. (2016) mais ils aident aussi à les placer dans un contexte plus utilisable et testable dans le futur pour la construction des sous-groupes de patients en supposant que le site d'apparition des premiers symptômes peut sert à regrouper les patients.

Cette courbe nous montre aussi les patients censurés.L'estimateur Kaplan Meier suppose que les patients censurés ont les mêmes chances de survie que ceux qui continuent d'être suivis.
Par convention, les lignes verticales indiquent les données censurées,leurs valeurs x correspondantes correspondant à l'heure à laquelle la censure s'est produite.
Pour les observations censurées, le temps de survie est considéré comme au moins aussi long que la durée de l'étude. Nous ne pouvons pas exclure ces sujets, sinon la taille de l'échantillon de l'étude pourrait devenir petite et les résultats peuvent devenir biaisés. Ceci est montré dans la figure 3. Nous avons essayé de modéliser la courbe de survie en prenant en compte que les patients dont la date de décès est connue. Nous avons observé une grande différence dans la durée médiane de survie pour les deux groupes bulbaire et spinale qui sont respectivement 23 et 27 mois.

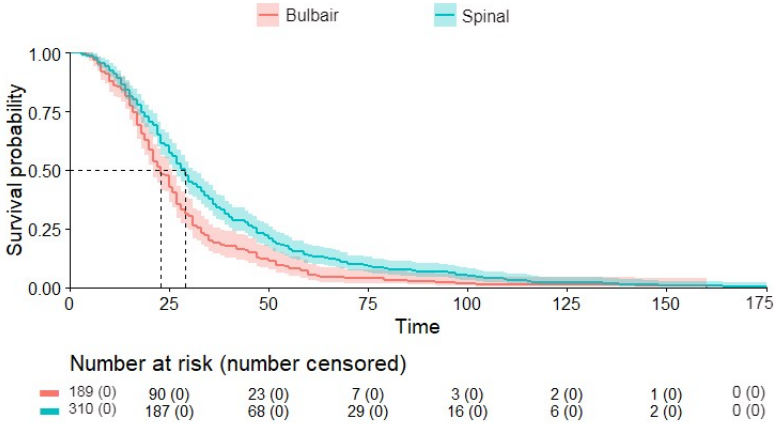

FIG. 3 : Kaplan Meier avec des données non censurées.

Grâce à cette analyse de survie ,Nous avons pu montrer l'importance de traiter les données censurées qui est une forme d'incomplétude des données. Cependant, d'autres variables indépendantes peuvent être incluses dans le modèle de survie pour aider à expliquer la variabilité des résultats de survie. Ceci est impossible avec la courbe Kaplan Meier mais possible avec d'autres modèles comme nous avons déjà mentionné.

3.2 Analyse longitudinale :

Définition des données longitudinales : Les données longitudinales sont le résultat du recueil des valeurs de plusieurs variables sur un échantillon d'individus à différents moments. Plusieurs objectifs font l'objet d'une analyse longitudinale, On cite :

— Étudier l'évolution au cours du temps afin d'étudier le changement dans le temps des caractéristiques des individus et/ou de comparer des évolutions entre individus.

— Observer la survenue et la date d'un événement (E.g :Analyse de survie).
— Étudier la variabilité inter ou intra-individuelle.

Méthodes de modélisation : Pour l'analyse des données longitudinales,nous faisons recours soit à des méthodes exploratoires ou à des méthodes de modélisation. Dans cette partie,Nous allons donner un aperçu sur des méthodes de modélisations qui sont divisés en deux :les méthodes classiques et les méthodes nouvelles.

Les méthodes classiques : Les approches ANOVA et MANOVA :Construction de Modèles linéaires générals :ANOVA (uni-variée) ou MANOVA(multivariée) à mesures répétées.
Les deux mettent l'accent sur la comparaison des moyennes de groupe mais ni l'une ni l'autre n'informe sur les tendances propres au sujet au fil du temps.
Les approches ANOVA sont limitées dans le traitement des données irrégulières et manquantes. Les mesures répétées ANOVA exige que tous les participants soient mesurés au même nombre de points dans le temps, et MANOVA exige des données entièrement complètes. L'application des méthodes ANOVA aux données pour lesquelles il manque des observations donne des estimations paramétriques biaisées(TanyaP et Karen, 2017).

Les méthodes nouvelles : Le modèles Linéaire à effets mixtes :Est une extension du modèle linéaire qui prend en compte la variabilité liée aux individus. Ce modèle est composé d'une partie fixe et d'une partie aléatoire. La partie fixe est identique pour chaque individu et représente l'effet population. La partie aléatoire est propre à chacun des individus et traduit la variabilité liée à chaque sujet(TanyaP et Karen, 2017).
Suivant TanyaP et Karen (2017),les modèles à effets mixtes peuvent tolérer des données déséquilibrées , c.-à-d.le nombre des mesures peut varier d'un individu ou d'un groupe à l'autre. Autrement dit, s'il y a des données manquantes dans les réponses, l'inférence pour les modèles d'effets mixtes peut se faire de la manière habituelle comme s'il n'y avait pas de réponses manquantes.

3.2.1 Application :

Nous avons fait une analyse exploratoire sur les jeux de données longitudinales afin de voir l'aspect de progression de la maladie sur une période d'étude de 60 mois. Rappelons que la progression de la maladie est liée principalement à la variation du score ALS.

Description de jeu de données : En se basant sur les données cliniques dans l'entrepôt de données,nous avons préparé un jeu de données indiquant des détails du score ALS dans des moments différents pour un nombre de patients. Les patients pris en compte sont les patients avec plus qu'une date d'examen et qui ont la valeur de site d'apparition de la maladie connue. Le nombre de patients est égale à 720 avec un total de 5112 essais cliniques. Le jeux de données comportent respectivement :
— l'identifient du patient
— Le score ALS.

— Une colonne pour le site d'apparition des premiers symptômes(Bulbaire ou spinale).
— Les détails du score ALS.
— Une colonne Delta donné en mois depuis le début de l'essai. Delta est une variable que nous avons calculé à partir de la date d'examen.Delta est au début de l'essai clinique est égale à zéro.

Statistiques Descriptifs : Dans le tableau 2, Nous montrons une synthèse de nos données.

n=790	
Site d'apparition des premiers symptomes (%)	
Bulbaire	68
Spinale	32
Score ALS (moyenne ± Ecart-type)	
Mois 0	36 ± 0.96
Mois 60	28 ± 1.53

TAB. 3 : Statistiques descriptives

Dans la figure 4, nous montrons un nuage de points qui met en évidence la progression de la maladie durant la période de 60 mois chez les patients choisis.

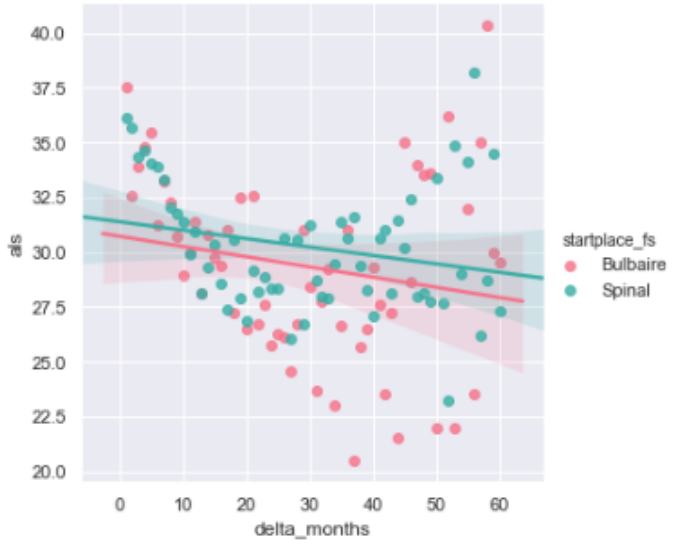

FIG. 4 : La variation de score ALS au cours temps pour l'ensemble des patients étudiés.

La courbe a pu montrer des cas spéciaux dont la valeur du score ALS reste élevée après les 60 mois.Ceci atteste les constatations de Robert et Neta (2019),ils ont observé la présence des sous-groupes de patients qui ont comme spécificité la progression lente de la maladie. Une autre constatation est mise en avant pour l'évolution la valeur ALSFRS chez les groupes de patients avec le site bulbaire qui décroît plus vite que l'autre groupe.

4 Implémentation :

Rappelons l'objectif derrière ce projet est l'identification des-sous groupes de patients dont la progression de la maladie est similaire ainsi l'identification des indicateurs ou biomarqueurs qui servent à la prédiction de l'évolution de la maladie chez les patients SLA.
Dans cette phase du projet, Nous avons focalisé sur l'analyse de survie des patients SLA. En effet, en l'appliquant sur différents groupes de patients, nous avons voulu mettre en évidence les indicateurs les plus probables qui servent à distinguer des sous-groupes de patients SLA en fonction de leurs survies.
Nous avons essayé de modéliser plusieurs courbes de survies afin de trouver les hypothèses les plus probables pour distinguer un sous-groupe d'un autre. Plusieurs attributs cliniques et biologiques étaient utilisés pour la création des groupes et qui vont être détaillés dans la sous-section 2.
Tous ces tests étaient automatisés à l'aide d'une application (sous python) dédiée aux médecins. Cet outil accède à l'entrepôt des données, effectue tous les tests possibles de création de groupes puis retourne les meilleurs résultats significatifs pouvant impacter la survie.
L'interface graphique de l'application est encore sous développement mais nous avons arrivé à montrer quelques tests présentés des les sections suivantes.

4.1 Stratégie d'analyse

Estimer la probabilité de survie d'une cohorte de patients implique de connaître, pour chaque sujet, la date de début et la date de fin de son suivi.
Cette dernière peut correspondre à la date de décès du sujet ou, si le patient n'est pas décédé, à la date de fin d'observation. Celle-ci est soit la date de fin d'étude, soit la date à laquelle le patient est perdu de vue (date de dernier examen dans notre jeu de données). Si un sujet n'est pas décédé à la fin de la période d'observation, le temps de survie de ce sujet est alors dit censuré à droite (Voir section 2).
Dans notre cas, la date de début est la date d'apparition des premiers symptômes puisque cette date peut refléter le plus adéquatement l'échelle de temps naturelle de la SLA.Rooney et Byrne (2013)
La période d'observation est la période de 5 ans après la date d'apparition des premiers symptômes. Les patients dont la date des premières symptômes est inconnue sont négligés de l'étude. Après la préparation des données de survie, nous avons construit les groupes de patients pour permettre à la comparaison de durée de survie de chaque cohorte.

4.2 Création de groupes

Pour la création des groupes, nous avons utilisé les informations cliniques et biologiques des patients. De différents groupes étaient construits en se basant sur l'attribut sexe et site d'apparition de la maladie et d'autres attributs calculés comme l'âge au moment de diagnostic, le délai de diagnostic et l'évolution de certaines mesures biologiques.
Les patients dont la valeur de site d'apparition de la maladie est inconnue sont exclus de l'étude. L'âge au moment du diagnostic a été classé en quatre groupes se rapprochant des quartiles de la distribution : moins de 55 ans ; 55 à <65 ans ; 65 à <75 ans ; et 75 ans ou plus.
le délai de diagnostic a été classé en 3 groupes : <12 mois ; 12 à <24 et 24 ou plus.
Pour l'évolution des mesures biologiques, nous avons modélisé pour chaque patient l'évolution de chaque mesure à l'aide d'un modèle de régression linéaire. Seulement les patients ayant au minimum 2 examens biologiques différents sont pris en compte. Cette modélisation sert à capturer la variation (la pente de la courbe)de la mesure pour un patient donné. Mais aussi à remédier à la non alignement des données en fixant la période dont ces mesures sont prises (Dans notre cas, nous intéressons aux périodes de 3, 6 ,9 et 12 mois après la date d'apparition des premiers symptômes).
Ensuite, Une modélisation de la distribution des pentes de chaque mesure donnée est effectuée et qui a servi pour la création des groupes de mesures à l'aide de la méthode "Jenks optimization method". Cette méthode de clustering qui est aussi une méthode de discrétisation de données numériques conçue pour déterminer la meilleure disposition des valeurs dans différentes classes. En d'autres termes, elle cherche à réduire la variance au sein des classes et à maximiser la variance entre les classes. Pour simplification, nous avons choisis de repérer seulement deux groupes à l'aide de cette méthode.

4.3 La courbe Kaplan Meier

estimation de survie Rappelons l'une des spécificités de nos données de survie est l'absence de la date de décès pour certains patients à la fin de la période d'observation. L'estimateur Kaplan Meier prend en compte ces cas de censure contrairement au modèle de régression Cox (Voir section 3).
Cette méthode représente en fonction du temps le taux de survie, c'est à dire la proportion des sujets initialement inclus dans l'essai toujours vivants au temps t. C'est la probabilité de survivre au moins jusqu'au temps t. Elle utilise les information des individus censurés jusqu'au moment où le patient est censuré. Ainsi elle maximise l'utilisation de l'information disponible sur le temps écoulé avant évènement(Manish et Pardeep, 2010).

Comparaison de survies Puisque nous voulons générer des hypothèses sur la meilleure répartition des groupes ayant une influence sur la survie, nous avons besoin d'un test statistique qui nous permet de comparer de différentes courbes de survies pour différents groupes afin de conclure sur la pertinence des cohortes choisis. L'un des tests statistiques les plus utilisés est le test "LogRank".
Ce test est utilisé pour tester l'hypothèse nulle selon laquelle il n'y a pas de différence

entre les populations ou cohortes(Rooney et Byrne, 2013).
Les différences entre les cohortes sont considérées comme statistiquement significatives pour les valeurs de p du "LogRank" inférieures à 0,05.

4.4 Résultats

4.4.1 Résultats des groupes cliniques :

Le tableau ci-dessous montre les résultats d'analyse de survie des groupes de patients en se basant sur la répartition des attributs démographiques comme l'âge et le sexe et d'autres attributs comme le site d'apparition de la maladie et le délai de diagnostic.

Groupes	**Nombre de patients**	**Médiane de survie**	**Test LogRank**
Tous les cas avec date de diagnostique et date des premiers symptomes connues : - 228 censurés	595	31	
Groupes d'age :			
- <55	117	45	
- 55-65	138	33	p = 0.00023
- 65-75	194	28	
- >75	146	29	
Groupes de délai de diagnostic :			
- <12	395	95	
- 12-24	134	42	p <0.0001
- >24	66	25	
Tous les cas avec date des premiers symptômes connue : - 395 censurés	897	36	
Sex :			
- Male	504	37	p = 0.044
- Femelle	393	34	
Site d'apparition de la maladie :			
	618	27	p <0.0001
- Bulbaire	279	45	
- Spinale			

Tab. 4 : Résultats de la comparaison des groupes de survie liés aux attributs cliniques

La durée médiane de survie de 31 mois à partir du date d'apparition des premiers symptômes est similaire à celle rapportée dans la littérature (Kim et Crystal, 2013). L'âge avancé au date de diagnostic s'est avéré être un facteur prédictif solide d'une survie plus courte (Jordan et Fagliano, 2015). Dans cette cohorte, la durée médiane de survie des personnes âgées de moins de 55 ans était plus longue que ceux âgés de 75 ans ou plus.

Pour le délai de diagnostic, nous observons que plus le délai est long plus la durée de survie diminue. Ceci est confirmé dans (Kim et Crystal, 2013). Nous remarquons une différence bien significative pour les groupes de site d'apparition de la maladie. Les patients dont le site d'apparition de la maladie est bulbaire présentent une médiane de survie beaucoup plus petit que ceux dont le site est spinale (Kim et Crystal, 2013).

Dans cette étude, les femelles avaient une survie plus courte que les mâles avec un p<0.044 qui n'est pas très significatif. Bien que cette comparaison ne soit pas statistiquement significatives dans certaines études (Kim et Crystal, 2013). D'autres ont défini une différence de survie selon le sexe (Jordan et Fagliano, 2015).

4.4.2 Résultats des groupes biologiques :

Les résultats de la comparaison de survie chez les groupes de patients générés à partir de la variation des mesures biologiques sur trois périodes différentes (6, 9 et 12 mois) sont montrés dans le tableau ci-dessous.

Période	Mesures	Groupe 1/2	Nombre de patients (Groupe 1/2)	Médiane de survie (Groupe 1/2)	Test LogRank
6 mois	Albumine	[-8, -1.25[[-1.25, 7.5]	37 223	17 29	p = 1.848680e-03
	Fer	[-8.75, -1.11[[-1.11,4]	69 154	25 30	p = 2.290289e-02
	Transferrine	[-0.31, -0.01[[-0.01, 0.16]	69 71	25 34	p = 8.316766e-04
	Capacité totale de fixation de la transferrine en fer	[-7.67, -0.4[[-0.4, 4]	61 79	23 31	p = 2.536694e-03
9 mois	Albumine	[-8, -1.43[[-1.43, 7.5]	27 269	17 28	p = 2.226759e-07
	HDL Cholestérol	[-0.19, -0.02[[-0.02, 0.15]	123 215	26 33	p = 3.947743e-05
	Fer	[-13, -0.7[[-0.7, 8]	106 195	25 31	p = 2.798993e-02
	Transferrine	[-0.31, -0.02[[-0.02, 0.15]	96 145	26 34	p = 3.932839e-04
	Capacité totale de fixation de la transferrine en fer	[-7.67, -1[[-1, 4]	66 174	23 33	p = 6.196649e-04
12 mois	Albumine	[-8, -1.43[[-1.43, 7.5]	28 293	17 28	p = 0.000003
	Préalbumine	[-0.06, -0.005[[-0.005, 0.05]	82 147	25 34	p = 0.000053
	HDL Cholesterol	[-0.17, -0.02[[-0.02, 0.21]	108 253	25 32	p = 0.000083
	Fer	[-5, -0.57[[-0.57, 4]	111 214	25 31	p = 0.001974
	Transferrine	[-0.31, -0.03[[-0.03, 0.215]	77 193	25 32	p = 0.013812
	Capacité totale de fixation de la transferrine en fer	[-7.6, -0.63[[-0.63, 5.5]	96 174	27 33	p = 0.004653

TAB. 5 : Résultats de la comparaison des groupes de survie liés à la variation des mesures biologiques

Nous avons observé une différence significative dans la durée de survie entre les groupes des mesures suivantes : Albumine, pré-albumine, fer, transferrine, cholestérol et la capacité totale de fixation de la transferrine(TSC) en fer.

D'autres mesures étaient identifiées comme la biliribune, la créatine Kinase et la Ferritine mais sur des groupes de patients dont la taille de l'échantillon est très petite. Ainsi, nous avons mis une condition sur les tailles des groupes à étudier qui doivent être supérieur à 20 patients.

Nous observons pour toutes les mesures (Tableau 5) que la durée médiane de survie des groupes "2" augmente en fonction de l'augmentation des taux des mesures. des taux élevés en fer, transferrine et TSC chez les patients ALS s'est avéré être liés à une durée de survie plus courte que ceux avec des taux plus bas Nadjar et Paul (2012). Ces derniers suggèrent que les patients ALS peuvent avoir un stockage accru en fer qui peut impacter la survie. Une concentration élevé en albumine est un indicateur d'inflammation dans les muscles. Chiò et Calvo (2014) ont montré que cette mesure peut estimer la sévérité de la maladie ainsi affecter la durée de la survie. En plus, elle peut être un potentiel biomarqueur fiable qui peut être utiliser pour définir le pronostic des patients. un taux de cholestérol plus élevé a été corrélé avec une survie améliorée de 26 à 33 dans la période de 6 mois et de 25 à 32 dans la période de 12 mois. Une corrélation entre le taux élevé de HDL cholestérol et l'amélioration de survie était présenté en (Ahmed et Highton-Williamson, 2018). Les variations des groupes de mesures montrées dans le tableau 5 peuvent être des potentiels biomarqueurs qui servaient à définir le pronostic des patients SLA en ayant cette capacité à affecter significativement la survie dans les cas où les taux sont élevés. les intervalles des groupes pour la même mesure peuvent changer avec le changement de la période et/ou le changement de la taille l'échantillon ce qui nous laisse penser à l'obligation d'élaborer plus d'expérimentation pour avoir plus de précision dans l'identification des seuils de variation qui peuvent servir à reconnaître certains groupes de patients en liaison à la vitesse de la progression de la maladie.

D'autres mesures qui nécessitent plus d'investigation comme la bilirubine (Joanna et Stelmasiak, 2003). Le découpage en suivant la méthode "Natural Jenks" a identifié deux groupes de patients composés respectivement de 400 et 2 patients avec un pvalue < 0.0001. Ce découpage n'est pas pris en compte, et nous pensons que d'autres arrangements possibles des groupes de variations de bilirubine en optimisant le nombre de groupes peuvent retourner de meilleurs résultats.

5 Conclusion

Le travail d'état de l'art présenté dans cet article a porté sur l'analyse longitudinale et l'analyse de survie. Nous avons commencé par présenter les spécificités de nos données collectées puis de décrire brièvement quelques méthodes traditionnelles pour l'exploitation des données longitudinales et enfin d'argumenter le choix d'utilisation du modèle linéaire à effets mixte pour ce type d'analyse. En plus, nous avons défini l'analyse de survie et puis nous avons présenté quelques modèles dans la littérature qui prennent en compte la censure des données. Nous avons montré quelques analyses primaires faites sur les données pour mesurer la qualité de nos données collectées en vérifiant quelques constatations repérés dans Robert et Neta (2019) et Knibb et al. (2016).

Rappelons toujours que l'un des grands challenge de la maladie SLA est l'hétérogénéité dans la progression de la maladie chez les patients, d'où la nécessité de l'identification

des groupes de patients présentant des évolutions similaires.
A l'aide des analyses de survies effectuées sur de différents groupes de patients, nous avons pu identifier des mesures biologiques qui peuvent affecter la survie des malades comme l'Albumine, la pré-albumine, le fer, la transferrine, le cholestérol et la capacité totale de fixation de la transferrine(TSC) en fer. Des taux élevés de ces mesures est un indicateur de dis-fonctionnement de certains organes des patients. Ceci était aussi repéré dans Nadjar et Paul (2012),Chiò et Calvo (2014)et Ahmed et Highton-Williamson (2018) mais pas confirmé jusqu'au moment présent.
le sexe féminin, l'apparition des premiers symptômes au niveau de site bulbaire, un âge plus avancé au moment du diagnostic et un délai de diagnostic plutôt élevé sont liés à une durée plus courte de la maladie (Jordan et Fagliano, 2015).

Décrire le modèle clinique et les facteurs pronostiques de la SLA est d'une grande importance pour les patients, leurs familles et leurs médecins afin d'élaborer un plan de traitement médical utile et personnalisé. Dans ce projet, nous avons essayé de présenter quelques hypothèses de stratification des patients SLA que nous avons penser être pertinentes pour impacter la survie et qui étaient aussi repérés dans la littérature mais qui nécessitent plus d'investigation dans le futue.

Références

Ahmed, R. et E. Highton-Williamson (2018). Lipid metabolism and survival across the frontotemporal dementia-amyotrophic lateral sclerosis spectrum: Relationships to eating behavior and cognition. *J Alzheimers Dis.*

Alberti, C., J.-F. Timsit, et S. Chevret (2005). Analyse de survie : comment gérer les données censurées ? *Revue des Maladies Respiratoires.*

Chiò, A. et A. Calvo (2014). Albumin and creatinine to assess severity of als. *Neurology.*

James, R. et B. Tom (2016). What does the alsfrs-r really measure? a longitudinal and survival analysis of functional dimension subscores in amyotrophic lateral sclerosis. *Neurol Neurosurg Psychiatry.*

Joanna, I. et Z. Stelmasiak (2003). Serum bilirubin concentration in patients with amyotrophic lateral sclerosis. *Clinical Neurology and Neurosurgery.*

Jordan, b. et J. Fagliano (2015). Effects of demographic factors on survival time after a diagnosis of amyotrophic lateral sclerosis. *Neuroepidemiology.*

Kim, T. et K. Crystal (2013). Prognosis and epidemiology of amyotrophic lateral sclerosis. *Neurol Clin Pract.*

Knibb, J., N. Keren, A. Kulka, et al (2016). A clinical tool for predicting survival in als. *Journal of Neurology 87.*

Manish, K. et K. Pardeep (2010). Understanding survival analysis: Kaplan-meier estimate. *Int J Ayurveda Res.*

Miller, R. G., M. J. D. L. M. et D. H. Moore (2012). Riluzole for amyotrophic lateral sclerosis (als)/motor neuron disease (mnd). *Cochrane Database Syst Rev.*

Nadjar, Y. et G. Paul (2012). Elevated serum ferritin is associated with reduced survival in amyotrophic lateral sclerosis. *PLOS One*.

Robert, K. et Z. Neta (2019). Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Scientific Reports 9*.

Rooney, J. et S. Byrne (2013). Survival analysis of irish amyotrophic lateral sclerosis patients diagnosed from 1995–2010. *PLOS*.

Swinnen, B. et W. Robberecht (2014). The phenotypic variability of amyotrophic lateral sclerosis. *Nature Reviews Neurology 10*, 661–670.

TanyaP, G. et M. Karen (2017). Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model. *Curr Neurol Neurosci Rep 14*.

Xian, L. (2016). *Methods and Applications of Longitudinal Data Analysis*.

Summary

the longitudinal analysis and the survival analysis answer several research questions in the medical field such as monitoring the evolution of the disease, describing patient trajectories or the prognosis of the disease. We worked in this project on longitudinal data from patients with ALS disease. We tried to detail in the first step the characteristics of these data by describing useful approaches for their manipulation. And in the second step, we tackled the question of prognosis by questioning our clinical and biological database of ALS patients to associate survival with several clinical and bilogical variables.