Modélisation mathématique de la survie et du développement des maladies : Applications à des données médicales

Tasnime Hamdeni

THÈSE DE DOCTORAT

présentée par

Tasnime HAMDENI

pour l'obtention du

DIPLOME NATIONAL DE DOCTORAT

en Mathématiques Appliquées

Préparée dans le cadre d'une cotutelle entre l'Ecole Nationale d'Ingénieurs de Tunis de l'Université Tunis El Manar et l'Université de Toulon

---

Modélisation mathématique de la survie et du développement des maladies: Applications à des données médicales

---

Soutenue le 01/12/2021 devant le jury composé de :

M. Maher MOAKHER,	PR, Université de Tunis El-Manar	Président
M. Soufiane GASMI,	PR, Université de Tunis	Directeur de Thèse
M. Jean-Marc GINOUX,	MCF-HDR, Université de Toulon	Directeur de Thèse
M. Afif MASMOUDI,	PR, Université de Sfax	Rapporteur
M. Raphaël SERREAU ,	PU-PH, Université Paris Saclay	Rapporteur
M. Franck BETIN,	PR, Université de Picardie Jules Verne	Examinateur
Mme Roomila NAECK,	Expert clinique, Novatech	Membre invité

Thèse préparée au sein du Laboratoire Modélisation Mathématique, Analyse harmonique et Théorie de Potentiel: Optimisation, Modélisation et Aide à la Décision (OMAD)- Tunisie, et le Laboratoire d'Informatique et des Systèmes (LIS)- France

Cette thèse a été préparée au sein de deux laboratoires :

OMAD :
Laboratoire Modélisation Mathématique, Analyse harmonique et Théorie de Potentiel :
Optimisation, Modélisation et Aide à la Décision
Faculté des Sciences de Tunis (FST)
Université Tunis El-Manar
Campus universitaire BP 37
1002 Le Belvédère Tunis
Tunisie

LIS :
Laboratoire d'Informatique et des Systèmes
LIS UMR 7020 CNRS / AMU / UTLN
Université de Toulon – Campus de La Garde – Bat X
CS 60584
83041 Toulon Cedex 9
France

Remerciement

Je tiens à exprimer ma reconnaissance à mes directeurs de thèse M. Soufiane GASMI, Professeur à l'Université de Tunis, et M. Jean-Marc GINOUX, Professeur à l'Université de Toulon, pour leur encadrement, leur disponibilité, et leur aide considérable tout au long de mes années de thèse. Je leur suis très reconnaissante pour les nombreuses discussions et suggestions qui m'ont permis d'améliorer mes connaissances. Leurs remarques et leurs critiques constructives ont énormément contribué à l'amélioration de la qualité de la thèse.

Mes remerciements et ma plus profonde gratitude s'adressent à M. Maher MOAKHER, Professeur à l'Université de Tunis El-Manar, pour me conférer l'honneur d'accepter de présider mon jury de thèse.

J'exprime mon profond respect à M. Afif MASMOUDI, Professeur à l'Université de Sfax, ainsi que Dr Raphaël SERREAU, Chef de service de Médecine Préventive Orléans Métropole, pour avoir accepté d'être rapporteurs et de m'accorder une partie de leur temps précieux pour bien critiquer ce travail. J'exprime, également ma reconnaissance à M. Franck BETIN, Professeur à l'Université de Picardie Jules Verne, qui a accepté d'examiner mes travaux de thèse et de faire partie du jury.

Mes remerciements les plus chaleureux vont au Professeur Mounir SAYADI qui m'a accueilli au sein du laboratoire Signal, Image et Maîtrise de l'Énergie (SIME) dès mon Projet de Fin d'Étude en m'offrant un cadre de travail qui m'a permis d'aboutir à l'objectif que je poursuis. Ses conseils, son intérêt et son suivi continu m'ont aidé à mener cette thèse à son terme. Je voudrais également remercier et exprimer mon profond respect à tous les membres du laboratoire.

Mes remerciements et mes profondes gratitudes s'adressent aussi aux membres du laboratoire Modélisation Mathématique, Analyse harmonique et Théorie de Potentiel : Optimisation, Modélisation et Aide à la Décision (OMAD), en particulier M.Lakhdar RACHDI qui m'a accueilli au sein du laboratoire.

Je voudrais tout particulièrement remercier le Professeur Farhat FNAIECH qui n'a pas cessé de m'encourager et de me soutenir durant toutes les années de thèse et qui a contribué à la fondation de notre collaboration et à la cotutelle de thèse.

Je voudrais également remercier et exprimer mon profond respect à Mme Roomila NAECK, Docteur en Ingénierie Biomédicale, et M. Moez BOUCHOUICHA, Maitre de Conférences à l'Université de Toulon, pour leurs participations constantes et leurs aides permanentes.

Je voudrais également dire un grand Merci à tous les membres du Laboratoire Informatique et Systèmes (LIS) qui m'ont accueilli durant mes stages à l'Université de Toulon dans le cadre de la cotutelle.

Mes sincères remerciements vont à tous les membres du service radiologie de l'Institut Salah Azaiez de Tunis (ISA) pour m'avoir accueilli au sein du service durant mon stage et pour leurs soutiens et le climat agréable qu'ils ont su créer et dont je serai incapable d'oublier. Je pense tout particulièrement à Dr Asma BEN KHEDHER ZIDI, le chef de service Radiologie de l'ISA, et Dr Frederick TSHIBASU, médecin radiologue aux Cliniques Universitaires de Kinshasa à la République Démocratique du Congo, pour m'avoir aidé à collecter et analyser la base de données sur laquelle se base une partie importante de ce travail. J'apprécie également leurs grandes contributions scientifiques enrichissantes dans l'étude effectuée.

Résumé

Cette thèse d'inscrit dans le cadre de la problématique globale portant sur l'analyse des données médicales, en particulier, celles de survie. Le travail ne se limite pas sur une maladie bien déterminée. Au contraire, nous avons élargi notre application à des maladies cancéreuses, neurologiques, infectieuses etc. Nous avons pour but d'apporter aux praticiens hospitaliers (en oncologie, neurologie et virologie) une aide à la décision médicale et une vision globale des malades pris en charge. En prenant appui sur des méthodes rigoureuses, des techniques robustes de biostatistique et de modélisation biomathématique ont été utilisée, pour permettre aux médecins de formuler des hypothèses physiopathologiques, de randomisation et d'essai thérapeutiques. Pour ce faire, nous avons traité des problèmes de modélisation de données médicales en utilisant le concept des distributions défectueuses qui permettent de, non seulement décrire le comportement des données de survie, mais aussi de prédire et quantifier la présence éventuelle d'une fraction survivante. Nous avons travaillé de façon à explorer les distributions bien fondées de la littérature pouvant être utilisée pour la modélisation de données de survie avec un taux de guéri. Ces distributions sont à être utilisées d'une manière compétitive. Des nouvelles distributions très flexibles ont été proposé. Leur surperformance a été démontrée à l'aide des techniques d'inférence statistique et de tests d'ajustement paramétriques et non paramétriques. Les particularités des populations, notamment, lorsque le risque de décès variable au cours du temps est envisagé selon les catégories de patients en cours d'étude, ont été adaptées en utilisant des méthodes de régression linéaire, des méthodes non paramétriques comme Kaplan-Meier, et des méthodes semi-paramétriques comme la régression de Cox à travers des exemples de la réalité. Le phénomène de censure, habituellement rencontré dans l'analyse de survie, a été aussi traité.

Les mots clés — analyse de données de survie, modélisation mathématique, distributions défectueuses, taux de guérison, distribution de Gompertz, famille Marshall-Olkin, survie sans progression, données biomédicales.

Abstract

This thesis lies within the scope of the overall problem of medical data analysis, in particular those of survival. The work is not limited to a specific disease. On the contrary, we have broadened our application to cancerous, neurological, infectious diseases, etc. Our goal is to provide hospital practitioners (in oncology, neurology, and virology) with medical decision support and a global view of the patients' situation. Based on rigorous methods, robust techniques of biostatistics and biomathematical modeling were used, to allow physicians to make assumptions related to patients' pathophysiology, randomization, and therapeutic trials. To that end, we have addressed medical data modeling problems using the concept of defective distributions which, not only describe the behavior of survival data but also predict and quantify the possible presence of a surviving fraction. Well-founded distributions were also explored to be used competitively way. Marked by their high flexibility, new distributions were proposed in this work. The outperformance of these new distributions was proved using statistical inference techniques, parametric and non-parametric goodness-of-fit tests. The particularities of the populations, especially, when the risk of death varying over time is considered according to the categories of patients under study, have been adapted using linear regression methods, non-parametric methods such as Kaplan-Meier, and semi-parametric methods like Cox regression through real-life examples. The censorship phenomenon, often encountered in the analysis of survival, was also studied.

Keywords — survival data analysis, mathematical modeling, defective distributions, survival rate, Gompertz distribution, Marshall-Olkin family, progression-free survival, biomedical data.

Table des matières

Publications

Articles de journaux scientifiques

- **Hamdeni, T.**, Gasmi, S. (2020). The Marshall–Olkin generalized defective Gompertz distribution for surviving fraction modeling. *Communications in Statistics-Simulation and Computation*, https ://doi.org/10.1080/03610918.2020.1804937, (Impact Factor : 1.118).
- **Hamdeni, T.**, Gasmi, S. (2020). A proportional-hazards model for survival analysis and long-term survivors modeling : Application to amyotrophic lateral sclerosis data. *Journal of Applied Statistics*, https ://doi.org/10.1080/02664763.2020.1830954, (Impact Factor : 1.404).
- Tshibasu, F.T., Eba, G.N., Bushaba, F.N., Mbarki, W., **Hamdeni, T.**, Sayadi M., Mbenza,B.L., Muamba, J.M. (2021). Radiosurgical Occurrence of Lumbar Disc Herniation Operated in Kinshasa / DRC. *International Journal of Medical Imaging*, https ://doi : 10.11648/j.ijmi.20210903.11, 9(3) : 130-140, (Impact Score : 3.61)
- **Hamdeni, T.**, Gasmi, S., Ginoux, J. M. (2020). Cure rate and statistical Inference for the Marshall-Olkin Generalized Defective Gompertz Distribution under Censored Survival Data. *Journal of Interdisciplinary Mathematics*, (Article soumis).
- **Hamdeni, T.**, Gasmi, S. (2020). Survival Analysis and Cure Rate Modeling of the ongoing pandemic COVID-19, *International Journal of Biostatistics* (Article soumis).
- **Hamdeni, T.**, Mouelhi, A., Zidi, A., Tshibasu, F., Gasmi, S. , Naeck, R., Bouchouicha, M., Ginoux, J. M. and Fnaiech, F. New Trends in CT Scan for Pulmonary Nodule Detection : Overview.
- **Hamdeni T.**, Gasmi, S. , Ginoux J. M., Guillet P. (MD), Dr. Escarguel B. (MD), Naeck R., Fnaiech F., Frizzi S., Bouchouicha M. Overview of Ordinary Differential Equations for Tumor Growth Modeling

Communications publiées dans des conférences internationales

- **Hamdeni, T.**, Naeck, R., Bouchouicha, M., Fnaiech, F, Zidi, A., Tshibasu, F., and Ginoux, J. M. (2018) Overview and Definitions on Lung Cancer Diagnosis. *IEEE 4th Middle East Conference on Biomedical Engineering (MECBME), 28-30 March 2018.*
- **Hamdeni, T.**, Gasmi, S. (2018) Parameter Estimation of Generalized Gompertz Distribution Under Type-I Censoring. *Euro-Mediterranen Conferences on Mathematical Reliability (ECMR), 4-6 July 2018.*
- **Hamdeni, T.**, Gasmi, S. (2019) La distribution de Gompertz généralisée et défectueuse pour la modélisation du taux de guérison de quelques données de survie. *New Trends in Analysis and Probability (NTAP'19), 23-26 September 2019.*
- **Hamdeni, T.**, Ginoux, J. M., Fnaiech, F. Application of segmentation methods to estimate lung tumor size by Response Evaluation Criteria In Solid Tumor (RECIST). *3rd IEEE CIS Summer School on Computational Intelligence (Poster). November 17-20, 2016.*

Liste des figures

Liste des tableaux

Notations

Conventions

- Toute variable aléatoire, sera notée par une lettre majuscule.
- Toute réalisation de variable aléatoire sera notée par la lettre minuscule correspondant à la lettre désignant cette variable aléatoire (exemple : si T est une variable aléatoire, alors une réalisation sera notée t).
- Pour toute variable aléatoire T_i avec $i = 0$, la réalisation vaut zéro ($t_0 = 0$).
- Nous utiliserons souvent dans le rapport l'acronyme (v.a.) pour désigner une variable aléatoire.

Acronymes

CDF	:	Fonction de répartition.
PDF	:	Fonction de densité de probabilité.
MO	:	Marshall-Olkin.
GD	:	Distribution de Gompertz.
MGD	:	Distribution de Gompertz modifiée.
GGD	:	Distribution de Gompertz généralisée.
MGGD	:	Distribution de Gompertz modifiée et généralisée.
MO-GD	:	Distribution de Gompertz généralisée selon Marshall-Olkin.
GDGD	:	Distribution de Gomepertz généralisée et défectueuse.
MO-GDGD	:	Distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin.

MLE	:	Estimateur du maximum de vraisemblance.
AMLE	:	Moyenne des estimations du maximum de vraisemblance.
MSE	:	Erreur quadratique moyenne .
KS	:	Kolmogorov-Smirnov.
CvM	:	Cramér-von Mises .
AIC	:	le critère d'information Akaike
AICc	:	la version corrigée du critère d'information Akaike.
BIC	:	le critère d'information bayésien.
CAIC	:	le critère d'information Akaike cohérent.
HQIC	:	le critère d'information d'Hannan-Quinn .
CI	:	Intervalle de confiance .
ALS	:	La sclérose latérale amyotrophique.
PRO-ACT	:	Pooled Resource Open-Access ALS Clinical Trials.
COVID-19	:	La maladie de Coronavirus.
CL	:	Niveau de censure.
PE	:	Erreurs de prédiction.
SE	:	Erreur standard.
HR	:	Taux de hasard.
OS	:	Survie globale.
PFS	:	Survie sans progression.
DFS	:	Survie sans maladie.
TTP	:	Temps jusqu'à progression.
ISA	:	Institut Salah Azaiez.
OMS	:	Organisation mondiale de la santé.
TDM	:	Tomodensitométrie.
DICOM	:	Digital Imaging and Communica-tions in Medicine.
TEP	:	Tomographie par émission de positions.
IRM	:	L'imagerie par résonance magnétique.
CR	:	Réponse compléte.
PR	:	Réponse partielle.
PD	:	Progression tumorale.
SD	:	Stabilité tumorale.
RECIST	:	Les critères d'évaluation de la réponse dans les tumeurs solides.
DL	:	Les diamètres les plus longs.

Introduction générale

De nos jours, le décès par des maladies mortelles, comme le cancer ou certaines maladies neurologiques par exemple, n'est pas inévitable, surtout si les cas sont détectés à un stade précoce. Selon l'OMS, 50% des patients diagnostiqués avec un cancer survivent à la maladie. L'augmentation de la probabilité de survie est due à l'effet positif de traitements efficaces, à la fonction majeure du système immunitaire et au développement rapide, technologique et clinique, de la recherche médicale.

Axes de recherche et limites des travaux précédents

Il s'agit d'une étude mathématique du développement des maladies. Ce thème est abordé de plus qu'un point de vue : modélisation mathématique, inférence statistique, analyse de survie, prédiction du taux de guérison, évaluation de l'évolution des tumeurs au cours du temps (quand il s'agit d'un cancer), analyse de l'effet des variables indépendantes sur la survie des patients, analyse de la sensibilité des différents paramètres cliniques …

Autant que nous sachions, la majorité de méthodes d'analyse de survie supposent que les sujets atteints de ce type de maladie sont tous susceptibles de mourir ou de récidiver de la maladie après avoir été traités. En termes mathématiques, le taux de guérison est généralement considéré comme égal à zéro. Ou bien, dans certains travaux, on considère un taux de guérison non nul (strictement positif), négligeant le fait qu'un patient peut mourir de la maladie ou que la maladie peut réapparaître après le traitement.

Motivation, Objectifs et problématique de la thèse

Parmi les objectifs de cette thèse, la prise en charge des déficits mentionnées ci-dessus. Pour le côté modélisation, nous avons pour but de créer des modèles mathématiques extrêmement flexibles qui tiennent compte d'une panoplie de contraintes. Le travail proposé vise, en premier lieu, à l'utilisation des méthodes d'inférence statistique pour ajuster les modèles à plusieurs types de bases de données récentes, riches et massives. Les banques de données utilisées seront créées

à partir des bases de données primaires, collectées rigoureusement des hôpitaux Tunisiens et une diversité de données secondaires, sélectionnées minutieusement pour représenter une variété de taille d'échantillon, de courbes de survie et de risque, de différentes maladies cancéreuses et non cancéreuses.

Nous visons par ce travail à aider la communauté médicale à donner la procédure thérapeutique appropriée, à prévoir si un traitement est suffisamment efficace pour être utilisé à plus grande échelle et à fournir des soins plus efficaces au patient. Une méthodologie moderne et contemporaine sera suivie pour analyser la survie des sujets étudiés et prédire le taux de guérison et l'efficacité des traitements avec le minimum de coût possible (de point de vue complexité et nombre de paramètres) et le maximum de gain possible. Les effets des variables indépendantes naturelles et provoquées seront étudiés et quantifiés à l'aide de méthodes paramétriques, semi-paramétriques et non paramétriques. L'existence d'une variété de modèles mathématiques permet une description plus adéquate des phénomènes naturels en attribuant le modèle le plus approprié. Autrement dit, les distributions sont censées être utilisées de manière compétitive.

Organisation du rapport de thèse

Cette thèse est organisée comme suit : dans **le premier chapitre**, nous introduisons les notions générales de l'analyse de survie, les outils et les méthodes statistiques et mathématiques utilisés dans nos travaux ainsi qu'une revue de la littérature des différents modèles de survie, modèles de taux de guérison et modèles de croissance tumorale, des différentes méthodes de généralisation de modèles, d'inclusion des variables indépendantes dans un modèle ainsi que les méthodes numériques utilisées.

Les autres chapitres seront réservés pour notre propre contribution. **Le deuxième chapitre** présente les nouveaux modèles de survie créés dans le cadre de cette thèse. Les fonctions de survie des modèles créées convergent vers une constante qui représente le taux de guérison lorsque le temps tend vers l'infini. Une estimation robuste des paramètres des modèles créés a été réalisé pour avoir la valeur de cette constante et ainsi prédire la probabilité de survie. Les modèles sont ensuite appliqués à des bases de données complètes et censurées décrivant des maladies différentes : cancer de la vessie, cancer du sang, cancer du sein et la maladie neurologique nommée la sclérose latérale amyotrophique.

Le troisième chapitre est consacré à l'analyse de l'effet des variables indépendantes, ou encore les facteurs de risque, sur la survie des patients. Cette analyse prend deux axes différents : le premier axe de modélisation où un modèle statistique de régression linéaire est développé en se basant sur le modèle statistique de survie créé dans le chapitre précédent. Ce modèle est comparé avec des modèles de la littérature. Une application a une base de données récente de patients atteints de la sclérose latérale amyotrophique a été créé et des résultats très intéressants ont été

trouvés. Le deuxième axe porte sur la pandémie de COVID-19 en cours. Une étude inférentielle est présentée pour découvrir le modèle statistique qui décrit le mieux la survie des patients atteints de coronavirus. L'effet de quelques facteurs de risque a été quantifié. Les résultats trouvés sont en totale harmonie avec les résultats de la statistique descriptive réalisés par l'OMS.

Le quatrième chapitre présente une étude de l'effet des paramètres cliniques utilisés par les pneumologues, radiologues et oncologues sur l'estimation de la survie des patients atteints d'un cancer du poumon. Un résumé exhaustif sur les biomarqueurs d'imagerie permettant aux médecins de quantifier la réponse des tumeurs solides aux différents traitements est présenté. Un aperçu sur les paramètres cliniques primaires souvent utilisés en cas d'indisponibilité de l'information sur la survie globale des patients, ce qui est un cas très fréquent. Une étude inférentielle a été effectuée dans le but de tirer des conclusions sur la mesure dans laquelle on pourrait considérer la survie sans progression comme étant un paramètre de substitution fiable à la survie globale. Ce chapitre est basé sur des données primaires récemment collectées du service radiologie de l'Institut Salah Azaiez de Tunis.

Le cinquième chapitre, un aperçu sur les différentes équations différentielles ordinaires utilisées dans la littérature à travers les années qui permettent la modélisation de la croissance tumorale. Les cas d'utilisation et les limites de ces modèles sont mis en évidence. Les capacités biotiques constantes et dynamiques sont distinguées. Un résumé de l'immunité antitumorale simulant Volterra-Lotka prédateurs-proies est présenté.

La conclusion générale contient un récapitulatif sur les contributions présentées dans cette thèse ainsi que quelques commentaires sur les résultats trouvés. Un ensemble de perspectives et idées intéressantes qui pourront faire l'objet de prochaines études pour enrichir notre contribution et améliorer nos résultats.

Chapitre 1

État de l'art

1.1 Introduction

La mort, l'arrêt cardiaque, la rechute de la toxicomanie, la récidive du cancer, la défaillance d'un appareil électronique ... Ce sont des évènements marquants qui s'observent au fil du temps. L'analyse de ces évènements se fait à l'aide de méthodes et modèles statistiques qui se développent jour après jour. Des données sur les durées de vie ou de survie s'accumulent chaque jour pour former la richesse du XXIe siècle. Des informations cruciales peuvent être tirées à partir de ces données. C'est là qu'intervient l'analyse de survie. Outre la prédiction et l'estimation des durées de survie, l'analyse de survie permet de juger sur la significativité des facteurs de risques. Ces informations sont d'une importance capitale pour l'aide à la décision médicale, chirurgicale, industrielle, économique etc. et pour traiter les facteurs de risques de sorte à aboutir à un résultat optimal quel que soit le domaine. Tous ces concepts généraux seront introduits dans ce premier chapitre.

1.2 Analyse de survie

1.2.1 Historique

Le terme "analyse de survie" a été utilisé dans le domaine de la démographie depuis le XVIIe avec la première table de mortalité (voir figure 1.1) réalisée par John Graunt en 1662 [1, 2]. La durée de survie est le temps écoulé avant un certain événement tel que le décès, l'apparition d'une maladie ou la rechute d'une affection. Au cours des siècles, l'analyse de survie a été uniquement liée à l'estimation des durées de vies, à partir des registres de décès, de la longévité, l'effectif etc.

Au XIXe, ces analyses ont été affinées en classant les durées de survie en catégories prédéfinies connues par les variables exogènes comme le sexe, la profession, la nationalité etc. Durant ce même siècle, des premières modélisations concernant de la probabilité de mourir après un certain temps sont apparus. D'où, aussi l'apparition du concept de "la fonction de risque". Au XXe siècle, des analystes des données de survie ont pensé en dehors du cadre de la démographie pour investir à l'actuariat, la fiabilité etc.

Jusqu'à la première moitié du XXe, les contributions en analyse de données de survie n'avaient pas été très nombreuses. La principale publication était celle de Greenwood (1926) [3], qui proposa une formule qui permet d'obtenir l'erreur standard à partir d'une table de survie.

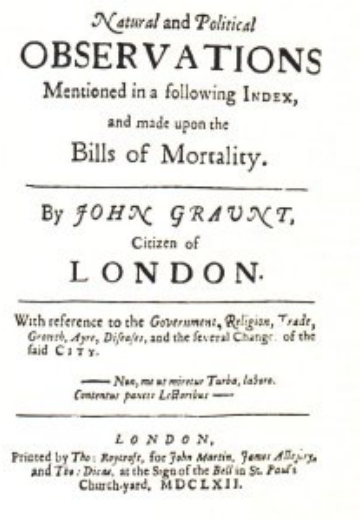
En 1951, Waloddi Weibull a conçu une loi de probabilité continue dans un contexte fiabiliste mais qui a été fréquemment utilisée en analyse de survie par la suite. La contribution de Kaplan et Meier en 1958 à l'estimation non paramétrique des probabilités de survie et des taux de risque, a conduit à des améliorations révolutionnaires dans le domaine d'analyse de survie. L'année 1972 est une date qui a marqué l'histoire de l'analyse de survie : en effet, c'est l'année au cours de laquelle le modèle statistique semi-paramétrique à risques proportionnels de Sir David Cox a vu le jour [4]. Le modèle de Cox a deux composantes. La première représente le risque de base qui décrit la variation du risque dans le temps. La deuxième est une fonction exponentielle d'une combinaison linéaire des variables exogènes indépendantes du temps. Ce modèle est sans doute le plus utilisé dans le domaine de l'analyse de survie. Il sert, essentiellement, à comparer les forces relatives de mortalité de deux vies ou de deux groupes de vies homogènes, à stratifier l'effet des covariables, décrire une dépendance vis-à-vis du temps, prendre en considération une interdépendance éventuelle des durées de vie observées.

Au cours des dernières décennies, plusieurs modèles statistiques ont été développés avec des caractéristiques différentes. Ces modèles et méthodes statistiques pour l'analyse de survie s'appliquent aujourd'hui non seulement aux domaines de fiabilité et actuariat, mais aussi à la sociologie, la criminologie, le marketing et plus intéressant encore, le biomédical.

1.2.2 Distributions de la durée de survie

Soit T une variable aléatoire (v.a), continue, positive ou nulle représentant une durée de survie, alors sa loi de probabilité peut être définie par une des fonctions caractéristiques introduites dans ce qui suit.

Les cinq fonctions ci-dessous sont équivalentes dans le sens où on peut obtenir chacune des fonctions à partir de l'une des autres fonctions.



Natural and *Political*
OBSERVATIONS
Mentioned in a following INDEX,
and made upon the
Bills of Mortality.

By JOHN GRAUNT,
Citizen of
LONDON.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Ayre*, *Diseases*, and the several Changes of the said CITY.

Nec, me ut miretur Turba, laboro,
Contentus paucis Lectoribus

LONDON,
Printed by *Tho: Roycroft*, for *John Martin*, *James Allestry*, and *Tho: Dicas*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXII.

FIGURE 1.1 – La première page de la première édition de la première table de mortalité réalisée par Graunt en 1662.

1.2.2.1 Fonction de survie S

La fonction de survie est donnée par l'équation (1.1) :

$$S(t) = P(T > t), \quad t \geq 0, \tag{1.1}$$

avec t est la variable temps, T est la durée de survie et P est la fonction de probabilité. La fonction de survie signifie, la probabilité de survivre jusqu'à l'instant t.

1.2.2.2 Fonction de répartition F

La fonction de répartition, ou la fonction de densité cumulative (CDF) est donnée par (1.2) :

$$F(t) = P(T \leq t), \quad t \geq 0. \tag{1.2}$$

Elle signifie la probabilité que l'évènement d'intérêt ait lieu avant l'instant t.

1.2.2.3 Fonction densité de probabilité f

La fonction densité de probabilité (PDF) (1.3) est une fonction positive et intégrable d'intégral 1 telle que, $\forall t \geq 0$,

$$f(t) = \frac{dF(t)}{dt}, \tag{1.3}$$

où $F(t)$ est la fonction de répartition définie dans 1.2.2.2. La fonction densité de probabilité signifie la probabilité que l'évènement d'intérêt ait lieu dans un petit intervalle de temps après l'instant t. Informellement, elle peut être vue comme la limite de l'histogramme décrivant les durées de vie observées.

1.2.2.4 Fonction de risque h

Le risque instantané ou taux de hasard est donné par (1.4) en fonction des fonctions définies dans 1.2.2.3 et 1.2.2.1 :

$$h(t) = \frac{f(t)}{\mathrm{S}(t)} = -\frac{d\ln(\mathrm{S}(t))}{dt}. \tag{1.4}$$

Le taux de hasard signifie la probabilité de mourir dans une petite période de temps après l'instant t sous condition que le sujet ait survécu jusqu'à l'instant t.

1.2.2.5 Fonction de risque cumulé H

La fonction de risque cumulé, étant l'intégrale du taux de hasard instantané définit dans 1.2.2.4, a la forme de l'équation (1.5) :

$$\mathrm{H}(t) = \int_0^t h(u)du = -\ln(\mathrm{S}(t)). \tag{1.5}$$

1.2.3 Modèles de survie

1.2.3.1 Modèles de survie non paramétriques

Pour tirer une inférence sur la distribution d'une v.a., basée sur un échantillon de données censurées à droite, l'estimateur de Kaplan-Meier représente une alternative à la fonction de distribution empirique des données. La méthode de Kaplan-Meier, aussi appelée produit-limite est un estimateur non paramétrique [5].

On pose Y_i le nombre d'individus qui n'ont toujours pas subi l'évènement d'intérêt jusqu'à l'instant T_i, et d_i le nombre d'individus qui ont subi l'évènement d'intérêt en T_i. La probabilité conditionnelle qu'un individu qui survit juste avant le temps T_i subisse l'événement d'intérêt au temps T_i est égal à $q_i = \frac{d_i}{\mathrm{Y}_i}$. La proportion de la population dont la durée de vie dépasse le temps t, est définie par la fonction (1.6) :

$$\hat{\mathrm{S}}(t) = \prod_{\mathrm{T}_i \leq t} (1 - q_i) \tag{1.6}$$

La survie estimée est considérée égale à 1 quand t est inférieure à la plus petite valeur de durée de vie observée. Le fait que cet estimateur ne soit pas bien défini lorsque t est supérieur à la plus grande valeur de durée de vie observée est une limitation à tenir en compte. L'estimateur de Greenwood [3] de la variance de l'estimateur de Kaplan-Meier, obtenue en appliquant la méthode delta, est donnée par (1.7) :

$$\widehat{Var}(\hat{\mathrm{S}}(t)) = \hat{\mathrm{S}}(t)^2 \sum_{i:\mathrm{T}_i \leq t} \frac{d_i}{\mathrm{Y}_i(\mathrm{Y}_i - d_i)}. \tag{1.7}$$

1.2.3.2 Modèles de survie semi-paramétriques

Les modèles semi-paramétriques joignent un modèle paramétrique pour certaines parties du modèle et conservent une estimation non paramétrique pour d'autres parties. Soit $x^{\mathrm{T}}(t) = (x_1(t), \ldots, x_k(t))$ un vecteur de covariables, $b^{\mathrm{T}} = (b_0, b_1, \ldots, b_k)$ est un vecteur de coefficients, T est le temps d'un évènement. Il y a plusieurs méthodes semi-paramétriques en analyse de survie :

Modèle de transformation linéaire : Le modèle de transformation linéaire est généralisé comme dans (1.8) :

$$h(\mathrm{T}) = -\mathbf{x}^{\mathrm{T}} b + \epsilon, \tag{1.8}$$

où ϵ est une erreur aléatoire.

Modèle de taux de risque additif : Le taux de risque au temps t, pour chaque individu, est exprimé par la combinaison linéaire (1.9) :

$$h(t|\mathbf{x}(t)) = b_0(t) + \sum_{i=1}^{k} b_i(t)\mathbf{x_i}(t), \tag{1.9}$$

Les k fonctions de régression peuvent être positives ou négatives, mais les valeurs sont contraintes car $h(t|\mathbf{x}(t))$ doit être positive. L'estimation des modèles additifs est généralement effectuée par des méthodes non paramétriques [6].

Modèle de taux de risque multiplicatif : Les modèles de taux de risque multiplicatif ont la forme suivante :

$$h(t, \mathbf{x}, b) = h_0(t)\, a(\mathbf{x}, b), \tag{1.10}$$

où $h_0(t)$ pourrait avoir n'importe quelle forme paramétrique arbitraire ou n'importe quelle fonction non négative de t, et $a(\mathbf{x}, b)$ est une fonction non négative de covariables qui ne dépend pas de t. Ce modèle contient deux composantes, l'une est paramétrique, l'autre est non paramétrique. Le modèle est formé de telle sorte que la fonction $h(t, \mathbf{x}, b)$ soit positive. La fonction $h_0(t)$ est la fonction de risque de base, qui représente le cas où les covariables pour tous les sujets sont égales à zéro. Quand $a(\mathbf{x}, b) = 1$, la fonction de risque est égale à la fonction de risque de base. Le rapport du modèle (1.10), pour deux individus ayant x_1 et x_2 comme covariables, est donné dans (1.11) :

$$\Gamma(t, \mathbf{x_1}, \mathbf{x_2}) = \frac{h(t, \mathbf{x_1}, b)}{h(t, \mathbf{x_2}, b)} = \frac{h_0(t)\, a(\mathbf{x_1}, b)}{h_0(t)\, a(\mathbf{x_2}, b)} = \frac{a(\mathbf{x_1}, b)}{a(\mathbf{x_2}, b)}. \tag{1.11}$$

On voit que le taux de hasard Γ ne dépend que de la fonction $a(\mathbf{x}, b)$. L'estimation est principalement basée sur la forme de la partie paramétrique $a(\mathbf{x}, b)$. Cox (1972) [4], l'un des leaders de

l'analyse de survie, a suggéré que $a(\mathbf{x}, b)$ soit égale à $e^{\mathbf{x}^T b}$. Dans ce cas, le taux de hasard sera donné par (1.12) :

$$\Gamma(t, \mathbf{x_1}, \mathbf{x_2}) = e^{(\mathbf{x_1}-\mathbf{x_2})^T b}. \tag{1.12}$$

La fonction de survie aura donc la forme de l'équation (1.13) :

$$S(t, \mathbf{x}, b) = e^{-H(t,\mathbf{x},b)}, \tag{1.13}$$

où $H(t, \mathbf{x}, b)$ est la fonction de risque cumulé au moment t pour un individu de covariabale $\mathbf{x}$. La fonction de risque cumulé est trouvé de la façon suivante :

$$H(t, \mathbf{x}, b) = \int_0^t h(u, \mathbf{x}, b)\, du = a(\mathbf{x}, b) \int_0^t h_0(u)\, du = a(\mathbf{x}, b) H_0(t),$$

où $H_0(t)$ est la fonction de risque cumulé de base. La fonction survie sera donc donnée par (1.14) :

$$S(t, \mathbf{x}, b) = [e^{-H_0(t)}]^a(\mathbf{x}, b) = [S_0(t)]^a(\mathbf{x}, b). \tag{1.14}$$

avec $S_0(t) = e - H_0(t)$ la fonction de survie de base. Si on applique la suggestion de Cox à cette fonction de survie, on trouve (1.15) :

$$S(t, \mathbf{x}, b) = \{S_0(t)\}^{e^{\mathbf{x}^T b}}. \tag{1.15}$$

On trouve plus d'informations sur le modèle de Cox dans [4]. Lorsqu'on veut comparer deux groupes ou plus de données de survie, on peut avoir recours au modèle à risques proportionnels de Cox. Ce faisant, on n'aura pas besoin de spécifier la fonction de risque de base dans l'analyse de survie de ces groupes. Quand il s'agit d'une évaluation de l'effet des covariable sur les durées de survie, le modèle de Cox est le modèle le plus utilisé.

1.2.3.3 Modèles de survie paramétriques

Les modèles de survie paramétriques ont été largement utilisés pour l'analyse des données de survie. On adapte les modèles de survie paramétriques aux données pour permettre la description de leur comportement au cours du temps. On peut extraire beaucoup d'informations à partir de la fonction de survie et des fonctions équivalentes comme indiqué dans 1.2.2.

loi de Gompertz : Dans cette partie, le modèle de Gompertz va être introduit comme exemple de modèle de survie paramétrique : il s'agit une distribution de probabilité continue. Sa densité de probabilité est donnée par (1.16) :

$$f(t, \alpha, \beta) = \alpha e^{\beta t} e^{-\frac{\alpha}{\beta}(e^{\beta t}-1)}, \tag{1.16}$$

avec $a > 0$, $b > 0$ et $t > 0$. Dans cette paramétrisation, α est un paramètre de forme et β est un paramètre d'échelle. La fonction de survie est donnée par (1.17) :

$$S(t,\alpha,\beta) = e^{-\frac{\beta}{\alpha}(e^{\alpha t}-1)}. \tag{1.17}$$

Il existe plusieurs autres modèles paramétriques qui sont utilisés en analyse de survie. Les courbes de risque instantané des modèles paramétriques peuvent avoir plusieurs formes : constante (loi exponentielle), monotone (loi de Weibull, loi Gamma ...), en baignoire, en forme de cloche, ...

Les modèles paramétriques peuvent être généralisés pour donner naissance à un autre modèle, qui pourrait être plus flexible.

1.2.4 Données de survie

Parmi les caractéristiques des données de survie, le fait que, communément, les durées sont recueillies partiellement. En d'autres termes, l'information n'est souvent pas observée intégralement. Ce cas est intensivement étudié dans la littérature et connu comme le phénomène de censure ou de troncature dans les données. L'incomplétude des données peut être due à plusieurs causes. À titre d'exemple, en collectant des données de durées de survie des patients, ces derniers peuvent être perdus de vue (e.g. à cause d'un déménagement), morts d'une autre cause (e.g. un accident). Parfois, l'étude s'arrête alors que le patient est toujours vivant. Des informations sur la date de début de la maladie peuvent être indisponibles ... Tous ces exemples peuvent être modélisés mathématiquement. On distingue plusieurs types de censure.

1.2.4.1 Censure de type I

On parle de censure de type I quand la durée de survie n'est pas observée au-delà d'une durée maximale, fixée d'avance. Dans le cas de censure fixe, au lieu d'observer $T_1,\ldots,T_n$, on prend le minimum entre la $i^{ème}$ observation et la censure fixe C. L'observation considérant la censure à droite est donc donnée par (1.18) :

$$X_i = min(T_i, C), \qquad i = 1,\ldots,n. \tag{1.18}$$

Quand la durée n'est pas observée avant un temps fixe, on dit que les observations sont censurées à gauche. Au lieu d'observer $T_1,\ldots,T_n$, on prend le maximum entre la $i^{ème}$ observation et la censure fixe C. L'observation considérant la censure à gauche est donc donnée par (1.19) :

$$X_i = max(T_i, C), \qquad i = 1,\ldots,n. \tag{1.19}$$

1.2.4.2 Censure de type II

Les durées de vie sont observées pour n individus jusqu'à ce que l'évènement d'intérêt se produit pour R individus. Au lieu d'observer $T_1, \ldots, T_n$, on observe (1.20) :

$$T_1 \leq T_2 \leq \ldots \leq T_R. \tag{1.20}$$

1.2.4.3 Censure de type III

La censure aléatoire à droite : la durée de vie est dite censurée à droite si le patient n'a pas subi l'évènement d'intérêt à sa dernière observation. Si C_i est une censure aléatoire, l'observation est alors donnée par (1.21) :

$$X_i = min(T_i, C_i), \quad \text{et} \quad \delta_i = 1_{T_i \leq C_i} \quad \text{pour} \quad i = 1, \ldots, n. \tag{1.21}$$

La censure aléatoire à gauche : la durée de vie est dite censurée à gauche si le patient a déjà subi l'évènement d'intérêt avant qu'il soit observé. L'observation est alors donnée par (1.22) :

$$X_i = max(T_i, C_i), \quad \text{et} \quad \delta_i = 1_{T_i \geq C_i} \quad \text{pour} \quad i = 1, \ldots, n. \tag{1.22}$$

La censure par intervalle : quand il s'agit d'une censure à droite et à gauche.

1.3 Méthodologie

1.3.1 Quelques méthodes de généralisation de modèles

1.3.1.1 Marshall-Olkin

Marshall et Olkin (1997) [7] ont proposé une méthode de généralisation de modèles. Leur idée consiste à étendre une distribution de base en y ajoutant un paramètre supplémentaire r de sorte qu'elle soit un cas particulier de la généralisation produite. C'est-à-dire, quand on pose $r = 1$ dans la distribution généralisée, on retrouve la distribution originelle.

Si $S(t)$ désigne la fonction de survie de base, alors la fonction de survie étendue par Marshall-Olkin est donnée par :

$$S_{MO}(t) = \frac{rS(t)}{1-(1-r)S(t)}, \qquad -\infty < t < +\infty, \qquad r > 0. \tag{1.23}$$

La généralisation de Marshall-Olkin s'avère être extrêmement stable géométriquement [8]. Cette famille de distributions a été utilisée dans la littérature pour plusieurs modèles, pour n'en nommer que quelques-uns : Weibull [9] , Weibull bivarié [10], Lindley généralisé [11] et Gamma généralisé [12] . Si $f(t)$ désigne la PDF de la distribution de base, alors la PDF étendue par la famille Marshall-Olkin est donnée par (1.24) :

$$f_{MO}(t) = \frac{rf(t)}{(1-(1-r)S(t))^2}. \tag{1.24}$$

Des études plus concises et une analyse mathématique sur la famille de distributions Marshall-Olkin se trouvent dans [8] .

1.3.1.2 Substitution par la puissance

Supposons qu'une v.a. T suit la loi avec la fonction de répartition $F(t)$ sur $(0,b)$. Alors, pour $p>0$, les fonctions de répartition $G_1(t)$ et $G_2(t)$ dans les équations (1.25) et (1.26) sont des généralisations de $F(t)$.

$$G_1(t) = \frac{F(t^p)}{F(b^p)}, \quad \forall t \in (0,b), \tag{1.25}$$

$$G_2(t) = F(t^p), \quad \forall t \in (0, b^{\frac{1}{p}}). \tag{1.26}$$

Pour chacune des généralisations, on obtient la CDF de base à $p=1$. À titre d'exemple, quand on applique ce type de généralisation à la fonction exponentielle, on obtient la fonction de Weibull. La distribution de Burr de type XII généralise la distribution de Lomax de la même manière [13].

1.3.1.3 Exponentiation

La méthode initiale d'exponentiation consiste à ajouter un paramètre en tant qu'exposant à la CDF. En d'autres termes, si $F(t)$ est la CDF de base, alors la CDF généralisée $G_3(t)$ est donnée par (1.27) :

$$G_3(t) = (F(t))^p \tag{1.27}$$

L'exponentiation a été utilisé pour obtenir une généralisation du modèle de Weibull afin d'analyser des données dont les taux de risque génèrent une forme de baignoire [14]. Cette méthode a aussi été appliquée au modèle de Gumbel pour donner le modèle de Gumbel exponentié [15]. Plus importante encore, la méthode d'exponentiation a été appliquée au modèle de Gompertz par El-Gohary *et al.* en 2013 [16] pour donner le modèle de Gompertz généralisé. Des détails sur les performances du modèle de Gompertz et sa généralisation seront donnés dans les prochains chapitres.

1.3.1.4 Transmutation

Comme un moyen de généralisation des fonctions de répartition, la carte de transmutation de rang quadratique a été élaborée par [17] et appliquée aux distributions uniformes, exponentielle et normale. La transmutation se fait suivant l'équation (1.28) comme suit :

$$G_4(t) = (1+p)F(t) - pF^2(t) \quad \text{avec} \quad |p| \leq 1. \tag{1.28}$$

Cette méthode a été aussi appliqué au modèle de Weibull, Lomax, Rayleigh et plusieurs autres [13]. Le travail de S.E. Smith [18] pourrait être consulté pour plus de détails.

1.3.1.5 Familles-G

La famille-G, aussi appelée odds est une famille de distributions continues, dont le taux de risque pourrait avoir plusieurs formes [19]. Elle comprend, comme cas particulier, la distribution exponentiée de Weibull largement connue. Supposons qu'une v.a. T a une fonction de répartition $F(t)$, alors la fonction de répartition de la distribution généralisée avec cette méthode est donnée par l'équation (1.29) :

$$G_5(t) = \left(1 - e^{-p_1 \frac{F(t)}{1-F(t)}}\right)^{p_2}. \tag{1.29}$$

Cette méthode de généralisation a été appliquée au modèle de Gompertz, Burr, log-logistique, Burr III et autres.

1.3.1.6 Kumaraswamy

Supposons qu'une variable aléatoire T a une fonction de répartition $F(t)$, la loi généralisée par la famille de distributions Kumaraswamy a la fonction de répartition (1.30) suivante :

$$G_6(t) = 1 - \left(1 - F(t)^{p_1}\right)^{p_2}, \tag{1.30}$$

avec $p_1 > 0$, $p_2 > 0$ des paramètres de forme [20]. Les courbes de la fonction de risque de cette famille de distributions pourraient avoir plusieurs formes. Ceci prouve qu'elle peut décrire une grande variété de bases de données.

1.3.2 Comment inclure des covariables à un modèle paramétrique ?

On garde les notations $\mathbf{x}$ pour les vecteurs covariables et b est le vecteur de coefficients de régression. Il existe plusieurs façons qui permettent la prise en considération et même la quantification des covariables (dites aussi variables prédictives ou variables explicatives).

1.3.2.1 Familles paramétriques

On peut inclure les covariables à un modèle paramétrique en remplaçant un ou plusieurs paramètres du modèle par $e^{\mathbf{x}^T b}$. On prend comme exemple l'application de cette méthode pour le modèle exponentiel. La PDF du modèle exponentiel est donnée par (1.31) :

$$f(t,\lambda) = \lambda e^{-\lambda t}, \tag{1.31}$$

avec λ un paramètre de forme. On peut remplacer λ par le risque relatif $e^{\mathbf{x}^T b}$. Ainsi, le modèle est écrit en fonction des covariables.

1.3.2.2 Hasards proportionnels

Le modèle de régression statistique de Cox à hasard proportionnel est couramment utilisé en analyse de données pour étudier le lien entre le temps de survie des patients et une ou plusieurs covariables. Le modèle a déjà été introduit dans (1.2.3.2). On peut inclure les covariables au modèle en utilisant les hasards proportionnels comme cela est décrit dans l'équation (1.32) suivante :

$$h(t) = h_0(t)\, e^{\mathbf{x}^T b}. \tag{1.32}$$

Cette fois, $h_0(t)$ désigne la fonction de risque de base lorsque toutes les covariables sont nulles.

1.3.2.3 Modèles à temps de vie accéléré

Étant une alternative aux modèles à risques proportionnels de Cox, les modèles à temps de vie accéléré incluent aussi les covariables qui peuvent influencer la survie. Le modèle de Cox exige la proportionnalité des taux de hasards des sujets, mais ce n'est pas le cas pour les modèles à temps de vie accéléré. Le modèle à temps de vie accéléré peut être décrit par (1.33) :

$$h(t) = e^{\mathbf{x}^T b}\, h_0(t\, e^{\mathbf{x}^T b}). \tag{1.33}$$

1.3.2.4 Modèles à chances proportionnelles

Le modèle à chances proportionnelles (ou proportional odds en anglais) suppose que chaque variable explicative exerce le même effet sur chaque logit cumulatif. Le logit, étant une fonction mathématique utilisée pour la régression logistique, est exprimé par (1.34) :

$$logit(p) = \ln\left(\frac{p}{1-p}\right), \tag{1.34}$$

le modèle à chances proportionnelles peut avoir la paramétrisation donnée par (1.35) :

$$logit(\mathrm{S}(t)) = logit(\mathrm{S}_0(t))\,\mathrm{e}^{-\mathbf{x}^{\mathrm{T}}b}. \tag{1.35}$$

1.3.3 Modélisation du taux de guérison

Les données de survie sont des durées qui mesurent le temps de suivi depuis un point de départ déterminé jusqu'à l'occurrence d'un événement d'intérêt. Un exemple d'ensemble de données de survie est le temps écoulé entre le premier diagnostic d'une maladie et le décès ou la récurrence de la même maladie chez plusieurs patients. Néanmoins, l'événement d'intérêt n'est pas toujours observé. Cela peut être dû à deux raisons. La première raison est la perte d'informations que l'on appelle censure ou troncature. La deuxième raison est le fait que l'événement d'intérêt ne s'est jamais produit. C'est-à-dire qu'il existe une proportion guérie dans la population étudiée. La proportion guérie au sein de la population est également connue sous le nom de fraction survivante, fraction guérie, survivants à long terme, proportion immunitaire, etc.

Des modèles de taux de guérison ont été introduits et étudiés par de nombreux chercheurs. Une classe générale de modèles de transformation non linéaires a été introduite par Tsodikov en 2002 [21]et en 2003 [22]. La fonction de survie dans cette classe est supposée être donnée par $g(\mathrm{S}(t))$, où $g(.)$ est une fonction de distribution cumulative spécifiée paramétriquement sur $[0,1]$ [23].

Des années plus tard, Koutras et Milienos [24] ont proposé une famille de transformation flexible de modèles de taux de guérison. La proposition est motivée par des aspects biologiques et présente une discussion intéressante sur de nombreux modèles de transformation non linéaires.

Récemment, un modèle de taux de guérison de transformation non linéaire a été proposé par Balakrishnan et Milienos en 2020 [25], fournissant une interprétation réaliste d'une classe spécifique de fonctions de transformation appropriées, pour survivre à la modélisation des fractions. La fraction de guérison est modélisée mathématiquement de différentes manières.

1.3.3.1 Modèles qui nécessitent l'ajout d'un paramètre supplémentaire

Pour modéliser la fraction de guérison on pourrait ajouter un paramètre supplémentaire à une distribution bien fondée. Il existe deux méthodes pour le faire.

Modèles de mélange : La première méthode est connue sous le nom de modèles de mélange. Initiée par Boag en 1949 [26], les modèles de mélange standard sont couramment utilisés avec la fonction de survie ajustée (1.36) :

$$S_1(t) = \lambda + (1-\lambda)S(t), \tag{1.36}$$

où $S(t)$ est la fonction de survie propre et λ est un paramètre dans $]0,1[$ qui représente le taux de guérison ou la proportion de patients immunisés. $S(t)$ tend vers zéro lorsque le temps t tend vers l'infini et peut être remplacé par toute distribution propre telle que la distribution exponentielle, la distribution de Weibull et la distribution de Gompertz, ainsi que leurs versions modifiées et généralisées. Le modèle de mélange $S_1(t)$ converge vers λ lorsque le temps approche de l'infini. Des modèles de mélange ont été proposés pour étudier l'hétérogénéité entre les patients guéris et non guéris.

Modèles sans mélange : La deuxième méthode pour ajouter un paramètre supplémentaire à un modèle bien fondé pour décrire une fraction de guérison est un modèle sans mélange. Ce dernier décrit une asymptote pour le hasard cumulé et par conséquent pour la proportion de guérison, [27]. La fonction de survie du modèle de taux de guérison sans mélange est alors donnée par l'équation (1.37) :

$$S_2(t) = \lambda^{1-S(t)}. \tag{1.37}$$

1.3.3.2 Modèles qui ne nécessitent pas l'ajout d'un paramètre supplémentaire

On peut aussi modéliser la fraction de guérison sans avoir besoin d'ajouter un paramètre supplémentaire. Cette approche est un concept contemporain connu sous le nom de modélisation défectueuse. Elle est utilisée pour décrire les données de survie avec une proportion de survivants. Par nature, la fonction de survie d'une distribution converge vers zéro au fil du temps. L'interprétation médicale de ce fait est que tous les sujets sont sensibles à l'événement d'intérêt. Cette définition ne tient pas compte de l'existence d'une proportion de survivants. Pour prendre en compte cette proportion, Haybittle en 1959 [**?**] a initié l'idée de modifier le domaine des paramètres inconnus d'une distribution bien fondée de telle sorte que la fonction de survie ne converge plus vers

zéro lorsque le temps approche de l'infini. Au lieu de cela, elle converge vers le taux de guérison $\lambda \in]0,1[$. Ainsi, le taux de guérison est modélisé sans avoir à ajouter de paramètre supplémentaire comme les modèles de taux de guérison traditionnels décrits dans ce qui précède 1.3.3.1.

Étant une alternative aux premières approches avec une complexité réduite, plusieurs modèles défectueux ont été publiés récemment. Balka *et al.* [28] ont suggéré plusieurs modifications au modèle Gaussien inverse pour fournir un ajustement amélioré et une estimation des paramètres. L'idée d'utiliser une distribution défectueuse avait prouvé son efficacité et utilité dans la modélisation du taux de guérison. En effet, le concept a été appliqué dans de nombreuses autres distributions bien établies. Pour en nommer quelques-uns, Cancho et Bolfarine [29] ont géré le modèle de Weibull exponentié. Cantor et Shuster (1992) [30] ont appliqué la distribution de Gompertz défectueuse (et l'ont nommée Gompertz modifié) sur les données de survie censurées des patients pédiatriques atteints de leucémie.

Il convient de noter que, récemment, de nouvelles distributions défectueuses ont été générées. Rocha *et al.* ont généralisé les distributions de Gompertz modifié et inverse Gaussienne via la famille Kumaraswamy dans [31] et via la famille Marshall-Olkin [32]. Ils ont également prouvé que l'extension Marshall-Olkin d'une distribution défectueuse est une distribution défectueuse et ont donné 10 cas particuliers de distribution de Weibull étendue défectueuse [33]. Les distributions défectueuses ont plusieurs avantages.

(i) En utilisant un modèle défectueux, il est possible d'estimer un taux de guérison en utilisant directement une distribution naturellement défectueuse au lieu d'estimer la proportion sous forme de modèle de mélange ou sans mélange.

(ii) Il n'est pas nécessaire d'ajouter un paramètre supplémentaire tel que λ dans le modèle de mélange qui ajouterait de la complexité à la distribution, surtout quand la distribution est déjà compliquée, et qui la rendrait tout à fait impossible pour trouver des estimations de paramètres précises.

(iii) Il n'est pas nécessaire de supposer l'existence d'une fraction de guérison dans votre modèle. Une fois que vous avez un modèle défectueux, cela conduira à une fraction de guérison lorsque la procédure d'estimation présente une valeur hors de la plage de paramètres habituelle.

1.3.4 Distribution de Gompertz généralisée et défectueuse

Martinez *et al.* [34] qui, inspiré des travaux de Cantor et Shuster [30] réalisés en 1992, ont changé le domaine du paramètre d'échelle du modèle de Gompertz généralisé proposé par El-Gohary *et al.* [16] en 2013 pour publier la distribution de Gompertz généralisée et défectueuse (MGGD).

Une v.a. T a la distribution de Gompertz généralisée et défectueuse avec les paramètres α, β

et γ, disons MGGD (α, β, γ), si sa fonction de répartition et sa PDF correspondante sont données par les équations (1.38) et (1.39) ci-dessous :

$$F(t) = \left(1 - e^{\frac{\alpha}{\beta}\left(e^{-\beta t} - 1\right)}\right)^{\gamma}, \tag{1.38}$$

et

$$f(t) = \gamma\alpha e^{-\beta t} e^{\frac{\alpha}{\beta}\left(e^{-\beta t} - 1\right)} \left(1 - e^{\frac{\alpha}{\beta}\left(e^{-\beta t} - 1\right)}\right)^{\gamma - 1}, \tag{1.39}$$

où β est un paramètre d'échelle strictement positif et α et γ sont des paramètres de forme strictement positifs. Si T suit la loi MGGD (α, β, γ), alors la fonction de survie de T est donnée par (1.40) :

$$S(t) = 1 - \left(1 - e^{\frac{\alpha}{\beta}\left(e^{-\beta t} - 1\right)}\right)^{\gamma}. \tag{1.40}$$

Puisque MGGD est une distribution défectueuse, $S(t)$ converge vers un paramètre θ en $[0, 1]$ représentant la proportion d'éléments immunitaires dans la population :

$$\theta = \lim_{t \to \infty} S(t) = 1 - \left(1 - e^{-\frac{\alpha}{\beta}}\right)^{\gamma}. \tag{1.41}$$

1.4 Inférences statistiques

En analyse de survie, l'ajustement des modèles paramétriques aux données observées est l'un des soucis majeurs puisque cela permet d'avoir une interprétation naturelle et de calculer les probabilités nécessaires de manière plus adéquate.

1.4.1 Principe de l'estimateur de maximum de vraisemblance

Pour inférer les paramètres d'une distribution d'un échantillon donné, on peut utiliser la méthode de maximum de vraisemblance qui consiste à chercher les valeurs des paramètres qui maximisent la fonction de vraisemblance [35]. En se basant sur les résultats obtenus à partir des échantillons, l'estimateur de maximum de vraisemblance sélectionne le meilleur ensemble de paramètres pour la prétendue distribution des données [36].

La méthode de maximum de vraisemblance possède d'excellentes propriétés asymptotiques pour les échantillons de tailles assez grandes. En plus, avec cette méthode, on peut aisément incorporer la notion de censure. Par conséquent, la méthode de maximum de vraisemblance est largement utilisée pour l'analyse de survie. Pour estimer le paramètre du modèle, deux mesures

doivent être prises : tout d'abord, la fonction de vraisemblance doit être créée à l'aide des hypothèses et de la formulation du modèle. Deuxièmement, les valeurs des paramètres du modèle qui maximisent la fonction de vraisemblance doivent être trouvées. La contribution de l'*observation complète* à la fonction de vraisemblance est décrite avec la fonction de densité de probabilité. Comme alternative, lorsque l'observation est censurée, la fonction de survie est utilisée pour représenter l'observation pour laquelle l'évènement d'intérêt n'est pas survenu. En d'autres termes, si le patient qui n'a pas vécu l'événement d'intérêt au moment de la collecte des données, alors l'indicateur de censure $\delta_i = 1$ (c'est-à-dire que la contribution du patient à la vraisemblance est la fonction densité de probabilité). Sinon, l'indicateur de censure $\delta_i = 0$ (c'est-à-dire que la contribution du patient à la vraisemblance est la fonction de survie $S(t)$). Ainsi, les observations d'un échantillon aléatoire peuvent être divisées en deux sous-ensembles : censurées et non censurées.

On pose (t_i, δ_i) une donnée qui représente l'information sur une durée, une longévité ou une donnée de survie, où i est un entier en $[1, n]$, n est le nombre de patients dans l'étude, δ_i est une valeur binaire indiquant la censure et t_i est la durée observée (i.e. entre l'apparition des premiers symptômes et la date de décès du $i^{ème}$ sujet). Les t_i sont indépendamment et identiquement distribuées suivant une loi de probabilité spécifiée par les fonctions $f(t)$ et $S(t)$ qui représentent, respectivement, la fonction densité de probabilité et la fonction de survie. Si $\Theta = (\theta_1, \ldots, \theta_k)$ est le vecteur de paramètres inconnus du modèle, alors la fonction de vraisemblance est donnée par :

$$L(t_i; \Theta) = \prod_{i=1}^{n} f(t_i; \Theta)^{\delta_i} S(t_i; \Theta)^{1-\delta_i}. \tag{1.42}$$

En appliquant le logarithme népérien à l'équation (1.42), on obtient la fonction
+ log-vraisemblance [1].

Cette expression est valable pour les censures de type I, type II, type III et quand le mécanisme de censure n'est pas informatif. L'estimateur de maximum de vraisemblance est la valeur de Θ qui maximise la fonction (1.42) ou bien, d'une manière équivalente, $l = \ln L(\Theta)$. On dérive la fonction obtenue par rapport à chacun des paramètres du modèle dans Θ. On obtient j équations non linéaires, où j est le nombre de paramètres dans le modèle. Pour obtenir les estimateurs, on résout le système d'équations (1.43) :

$$U(\Theta) = \frac{\partial l(\Theta)}{\partial \theta_j}, \quad \text{pour} \quad j = 1, \ldots, k. \tag{1.43}$$

Généralement, l'estimateur de maximum de vraisemblance n'a pas une expression fermée. Ceci est dû à la complexité des équations qui dépendent évidemment du modèle qui est supposé décrire les données en question, des hypothèses prises en compte sur les paramètres du modèle etc. Une méthode numérique est donc généralement nécessaire pour ce genre de calcul.

Une fois les paramètres estimés, comment pourrait-on juger la qualité du modèle ajusté ?

1.4.2 Méthodes de sélection de modèles

Il existe certaines mesures qui permettent de vérifier la qualité relative des modèles ajustés.

1.4.2.1 Tests statistiques

En statistique inférentielle, on réalise des calculs sur les observations pour pouvoir émettre des conclusions par la suite. Les tests d'hypothèse sont un des moyens pour ce faire. Il s'agit d'une approche qui consiste à rejeter ou accepter une hypothèse nulle en fonction d'un échantillon. L'information sur l'échantillon à tester est résumé par la statistique de test S. Celui-ci est choisie de façon à connaître la loi sous l'hypothèse nulle qu'on note H_0 et qui est considérée vraie a priori. L'hypothèse alternative, notée H_1, est choisie par rapport à H_0 si ce dernier n'est pas considéré comme crédible. Le choix de H_0 et de H_1 est généralement imposé par le test statistique utilisé.

La construction mathématique d'un test se fait grâce au lemme 1.4.1 de Neyman-Pearson [37] et nous donne la forme de la région de rejet .

Lemme 1.4.1. *Soit un test de taille* α, *pour tester* $H_0 : \theta = \theta_0$ *contre* $H_1 : \theta = \theta_1$ *pour un échantillon* $\mathbf{T} = (\mathbf{T_1}, \ldots, \mathbf{T_n})$.

Le test rejette H_0 *en faveur de* H_1 *lorsque* $\dfrac{L(\mathbf{T}, \theta_0)}{L(\mathbf{T}, \theta_1)} \leq k_\alpha$ *où* k_α *est donné par (1.44)*

$$P\left(\frac{L(\boldsymbol{T}, \theta_0)}{L(\boldsymbol{T}, \theta_1)} \leq k_\alpha \middle| H_0\right) = \alpha \tag{1.44}$$

Définitions 1.4.1. *On appelle* ***région de rejet*** *le sous-ensemble* I *de* ***R*** *tel que l'hypothèse nulle est rejetée si la statistique observée appartient à* I.

Pour un ***test bilatéral****, on rejette* H_0 *si la statistique observée est trop grande ou trop petite. La région de rejet a donc la forme :* $]-\infty, a] \cup [b, +\infty[$.

Pour un ***test unilatéral à droite****, on rejette* H_0 *uniquement si la statistique observée est trop grande. La région de rejet a donc la forme :* $[a, +\infty[$.

Pour un ***test unilatéral à gauche****, on rejette* H_0 *uniquement si la statistique observée est trop petite. La région de rejet a donc la forme :* $]-\infty, b]$

La procédure de décision entre les hypothèses pour un test statistique unilatéral se déroule comme suit :

— On met en place l'hypothèse nulle H_0 et l'hypothèse alternative H_1.

— On calcule une statistique qui correspond à une mesure de la distance entre l'échantillon et la loi statistique pour le cas de l'adéquation. Plus la distance mesurée est grande, moins H_0

est probable. Cette variable de décision est basée sur une statistique qui est calculée à partir des données. Par exemple, le test statistique unilatéral correspond à rejeter l'hypothèse nulle si elle dépasse une certaine limite fixée.

— Tout en supposant que H_0 est vraie, on calcule la probabilité d'obtenir une valeur du test statistique supérieure ou égale à la valeur de la statistique obtenue avec l'échantillon. Il s'agit en fait de la valeur-p.

— On émet des conclusions en fonction de la taille α du test.Généralement, $\alpha = 5\%$ est acceptable ; ceci signifie que l'hypothèse nulle n'est vraie que dans 5% des cas au maximum. Au-dessous de ce risque seuil, on peut rejeter H_0. Cependant, le seuil à choisir dépend de la certitude acceptée et de la vraisemblance des autres choix.

— si la valeur-p $< \alpha$, alors H_0 est rejetée, sinon H_0 pourrait être acceptée.

Pour ne pas tomber dans le risque de deuxième espèce (voir remarque 1 ci-dessous), on peut utiliser la probabilité β_2, si elle est petite (inférieure à 5% par exemple), on accepte l'hypothèse nulle, sinon, on ne peut rien affirmer. β_2 représente le risque de ne pas rejeter H_0 quand on devrait le faire. Sa valeur est liée à l'application, et son évaluation peut être difficile, voire impossible. C'est la raison pour laquelle, on utilise principalement la valeur de α comme critère de décision. H_0 est rarement acceptée. Dans le cas où l'hypothèse alternative n'est pas acceptée, on affirme que le test n'est pas concluant.

Remarque 1. *Il y a deux risques à considérer lors d'un test statistique :*

1. *Le risque de première espèce : et ce lorsqu'on rejette* H_0 *alors qu'elle est vraie et on note* β_1 *la probabilité de tomber dans cette erreur.*
2. *Le risque de deuxième espèce : et ce lorsqu'on retient* H_0 *alors qu'elle est fausse et on note* β_2 *la probabilité de tomber dans cette erreur.*

Le test statistique n'est pas suffisamment précis pour la prise de décision. On utilise alors la valeur-p pour avoir une vision plus fine que sa simple comparaison avec α. En fait, plus la valeur-p est petite, plus l'hypothèse nulle est loin d'être acceptée. Il existe plusieurs tests statistiques classiques.On introduit dans ce qui suit quelques exemples :

Le test de Kolmogorov-Smirnov : Le test de Kolmogorov-Smirnov [38, 39] permet de tester l'adéquation entre des observations et une distribution de probabilité en calculant l'écart entre la CDF théorique et celle observée. Ce test est particulièrement pratique pour les variables aléatoires continues. Il s'agit d'un test non paramétrique qui permet de prendre la décision par rapport à une hypothèse nulle. Si on considère que T est une v.a. dont F est la CDF telle que les hypothèses de tests sont : H_0 : T suit la loi F contre H_1 : T suit une autre loi. Pour un échantillon $(\mathbf{T_1}, \ldots, \mathbf{T_n})$, la fonction de répartition empirique associée à T est donnée par :

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{]-\infty,t]}(T_i). \tag{1.45}$$

L'équation (1.45) représente la proportion des observations dont la valeur est inférieure ou égale à t. L'écart entre les valeurs observées et celles théoriques dérivées du modèle dont la CDF est F est alors donné par :

$$E_n = \sup_{t \in \mathbf{R}} |F_n(t) - F(t)|. \tag{1.46}$$

La v.a. (1.46) représente la fonction discriminante du test Kolmogorov-Smirnov ou sa variable de décision.

Le test du rapport de vraisemblance : Le test du rapport de vraisemblance permet de tester un modèle paramétrique contraint contre un modèle non contraint [40]. Si θ est le vecteur des paramètres estimés par la méthode de maximum de vraisemblance, l'hypothèse du test de maximum de vraisemblance peut être exprimée par : $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \notin \Theta_0$. Si on note $\hat{\theta}$ l'estimateur de maximum de vraisemblance et $\hat{\theta}_0$ l'estimateur de maximum de vraisemblance sous l'hypothèse nulle H_0, alors la statistique du test du rapport de vraisemblance est donnée par (1.47) :

$$\Gamma = -2(l(\theta_0) - l(\hat{\theta})). \tag{1.47}$$

Le test de Wald : Le test de Wald est un test paramétrique qui permet de tester la vraie valeur du paramètre basé sur l'estimation de l'échantillon [40]. Les hypothèses du test sont : $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. L'estimateur de maximum de vraisemblance $\hat{\theta}$ de θ est comparé à la valeur θ_0. La statistique de Wald est donnée par (1.48) :

$$W = \frac{(\hat{\theta} - \theta_0)^2}{Var(\hat{\theta})}. \tag{1.48}$$

Le test de Cramér-von Mises : Le test de Cramé-von Mises est un test non paramétrique qui utilise la distance quadratique entre la fonction de répartition d'une distribution et la fonction de répartition empirique pour évaluer la qualité d'ajustement [40]. Étant donnée un échantillon, l'hypothèse du test peut être exprimée par : $H_0 : F = F_n$ contre $H_1 : F \neq F_n$. La statistique de Cramér-von Mises est donnée par (1.49) :

$$C_n^2 = n \int_{-\infty}^{+\infty} (F_n - F)^2 dF. \tag{1.49}$$

On rejette l'hypothèse nulle pour des grandes valeurs de C_n^2.

Autres tests : La liste n'étant pas limitative, on peut citer [40] :

- Le test de Student qui permet de comparer, pour un échantillon distribué suivant la loi normale, une valeur désirée à la moyenne observée.
- Le test de Fisher-Snédécor, connu par le test de Fisher, qui permet de comparer deux variances observées.

- Le test d'analyse de la variance, connu par ANOVA qui permet de comparer un plan expérimental prédéterminé aux moyennes observées. Ce test est basé sur une décomposition de la variance en deux parties : variance intergroupe (explicable) et variance résiduelle globale intragroupe supposée distribuée suivant la loi normale.
- Le test du χ^2 de Pearson, connu par le test de χ^2, qui permet de comparer deux distributions observées.

La majorité des tests statistiques introduits sont basés sur la fonction de répartition empirique. Il existe d'autres méthodes qui sont plutôt basées sur le ré-échantillonnage bootstrap [41], une méthode de Monte Carlo basée sur les données qui s'est révélée mathématiquement valable dans un très large éventail de situations. Pour estimer la distribution des statistiques de qualité d'ajustement pour un modèle où les paramètres ont été estimés à partir des données, le bootstrap peut être utilisé de deux manières différentes : le bootstrap paramétrique et le bootstrap non paramétrique. Le bootstrap paramétrique peut être familier à de nombreux praticiens en tant que technique bien établie de création de données simulées à partir du modèle paramétrique par les méthodes de Monte Carlo. Le bootstrap non paramétrique, en revanche, est une réalisation particulière de Monte Carlo de l'EDF observé en utilisant une procédure de «sélection aléatoire avec remplacement». La procédure de bootstrap de sélection de modèle consiste à minimiser les estimations bootstrap de l'erreur de prédiction. Bien que les estimations bootstrap aient de bonnes propriétés, cette procédure de sélection bootstrap est peut être incohérente dans certains cas [42].

1.4.2.2 Critères d'information

Outre que les tests d'ajustement mentionnés ci-dessus, il existe des critères d'information qui permette d'évaluer la qualité relative des modèles ajustés. Ici, on introduit cinq critères d'information couramment utilisés pour comparer les modèles ajustés : le critère d'information Akaike (AIC), la version corrigée du critère d'information Akaike (AICc), le critère d'information Akaike cohérent (CAIC), le critère d'information bayésien (BIC), le critère d'information d'Hannan-Quinn (HQIC).

Ces critères d'information ne sont pas une mesure de qualité en eux-mêmes mais ils sont un outil relativement fiable qui permet de comparer [43]. Ils sont définis comme suit (1.50) :

$$\begin{aligned}
\text{AIC} &= 2j - 2\ln(\text{L}) \\
\text{BIC} &= j\ln(n) - 2\ln(\text{L}) \\
\text{CAIC} &= -2\ln(\text{L}) + j(\ln(\text{L}) + 1) \\
\text{AIC}c &= \text{AIC} + \tfrac{2(j+1)(j+2)}{n-j-2} \\
\text{HQIC} &= -2\text{L} + 2j\ln(\ln(\text{L}))
\end{aligned} \tag{1.50}$$

où j est la taille du vecteur des paramètres Θ ou bien le nombre de paramètres dans le modèle, n est la taille de l'échantillon donné et L est la valeur de la fonction de maximum de vraisemblance fixée en utilisant l'échantillon des données utilisé et les paramètres du vecteur Θ.

1.4.3 Intervalles de confiance asymptotiques

Dériver une distribution exacte des estimateurs du maximum de vraisemblance est tout à fait impossible car Θ n'a pas de forme fermée. C'est pourquoi il est important d'étudier les intervalles de confiance asymptotiques des paramètres des modèles [40]. Afin de quantifier l'information qu'une v.a. T possède sur les paramètres inconnus de la distribution qui modélise T, on appelle $\text{I}(\hat{\Theta})$ la matrice d'information de Fisher en Θ. On obtient $\text{I}(\hat{\Theta})$ par le calcul des dérivées partielles secondes du logarithme de la fonction de vraisemblance par rapport à tous les paramètres inconnus du modèle. La matrice d'information de Fisher observée est donnée par :

$$\text{I}(\hat{\Theta}) = \begin{pmatrix}
\frac{\partial^2 l}{\partial\theta_1^2} & \frac{\partial^2 l}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 l}{\partial\theta_1\partial\theta_k} \\
\frac{\partial^2 l}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 l}{\partial\theta_2^2} & \cdots & \frac{\partial^2 l}{\partial\theta_2\partial\theta_k} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 l}{\partial\theta_k\partial\theta_1} & \frac{\partial^2 l}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial^2 l}{\partial\theta_k^2}
\end{pmatrix}$$

La matrice de variance-covariance, qui est l'inverse de la matrice d'information de Fisher, peut alors être approximée.

$$\text{V}(\hat{\Theta}) = \text{I}^{-1}(\hat{\Theta}) = \begin{pmatrix}
Var(\hat{\theta}_1) & Cov(\hat{\theta}_1,\hat{\theta}_2) & \ldots & Cov(\hat{\theta}_1,\hat{\theta}_k) \\
Cov(\hat{\theta}_2,\hat{\theta}_1) & Var(\hat{\theta}_2) & \ldots & Cov(\hat{\theta}_2,\hat{\theta}_k) \\
\vdots & \vdots & \ddots & \vdots \\
Cov(\hat{\theta}_k,\hat{\theta}_1) & Cov(\hat{\theta}_k,\hat{\theta}_2) & \ldots & Var(\hat{\theta}_k)
\end{pmatrix}$$

Les intervalles de confiance asymptotiques des paramètres d'une distribution peuvent être approximés avec une probabilité de $100(1-\delta)\%$. On obtient les intervalles suivants : $\theta_k \pm Z_{\frac{\delta}{2}}\sqrt{Var(\hat{\theta}_k)}$

où $Z_{\frac{\delta}{2}}$ est le quantile de la distribution normale standard.

1.5 Conclusion

Dans ce premier chapitre nous avons présenté les notions de l'analyse de survie sur lesquelles on s'est basé pour créer nos propres modèles de survie dans ce qui suit. Un aperçu a été donnée sur les différentes fonctions caractéristiques qui décrivent les durées de survie. Les différents types de modèles et de données de survie sont résumés d'une manière exhaustive. Les méthodologies connues dans la littérature pour la généralisation des modèles statistiques et les méthodes de l'inférence statistique sont aussi récapitulées. Tout cela sera utilisé dans le chapitre suivant dans le but de proposer une nouvelle distribution de survie permettant une meilleur description que les distributions existantes.

Chapitre 2

Distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin

2.1 Introduction

La fonction de survie d'une distribution converge intrinsèquement vers zéro à l'infini. L'interprétation médicale de cette définition est que tous les sujets malades sont susceptibles à l'évènement d'intérêt qui est généralement la mort ou bien la récurrence de la maladie après en avoir guéri. Mais réellement, ce n'est pas toujours le cas. Il y a quand même un taux de guérison qui ne doit pas être négligé. Cela a donné l'idée d'entrer des modifications sur les domaines de définitions de quelques paramètres inconnus des modèles. Ces changements agiraient sur la distribution de sorte que la fonction de survie ne tend plus vers zéro, mais vers une quantité entre $[0,1]$ connue par le taux de guérison θ. Dans ce chapitre, nous introduirons une généralisation à trois niveaux de la distribution de Gompertz pour la modélisation du taux de guérison. Cette distribution est appelée la distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin (MO-GDGD) L'un des apports majeurs de cette nouvelle distribution est que sa courbe de taux de hasard peut être : décroissante, croissante, constante ou même en forme d'une baignoire. En outre, cette distribution prend en considération les deux cas de présence et d'absence du taux de guérison.

2.2 Méthodologie

2.2.1 Traçabilité du modèle MO-GDGD

En se basant sur la loi de mortalité, Benjamin Gompertz a établi, en 1825, un des modèles mathématiques classiques [44]. Ce modèle a été conçu pour décrire les tables actuarielles et prédire la mortalité humaine. Mais mieux encore, le modèle de Gompertz a été très utile dans d'autres domaines de recherches. Particulièrement, les statisticiens ont employé cette distribution pour modéliser le taux de guérison chez les patients atteints de maladies cancéreuses. Ceci a été possible puisque des modifications mineures ont été faites sur le modèle de telle sorte qu'il soit défectueux. Cantor et Shuster [30] ont attiré l'attention sur concept des distributions défectueuses en introduisant le modèle de Gompertz modifié. En 1998, Gieser *et al.* [45] ont ajouté des informations sur les covariables à la distribution de Gompertz défectueuse et les ont utilisées comme modèle de régression pour ajuster les données sur le cancer pédiatrique.

El-Gohary *et al.* [16] ont proposé une exponentiation à la distribution de Gompertz en 2013. Ils l'ont appelé la distribution de Gompertz généralisée (GGD). La distribution de Gompertz modifiée et généralisée a été par la suite proposée par Martinez *et al.* en 2017 [34] et par Borges [46] dans la même période. Il s'agissait d'une modification dans le domaine du paramètre d'échelle β du modèle GGD de $]0,+\infty[$ à $]-\infty,0[$. Cette modification a permis la description des données de survie qui comprennent une proportion d'éléments exemptés. Les travaux de Borges [46] et Martinez *et al.* [34] comprenaient des informations sur les covariables du modèle de Gompertz modifié et généralisé, mais de deux manières différentes.

2.2.2 Formulation du modèle

Cette contribution résout un certain nombre de limitations scientifiques déclarées dans la littérature. Nous proposons une généralisation à trois niveaux de la distribution de Gompertz. Chaque niveau suggère une solution à, au moins, une de ces limitations :

Niveau 1 **: Exponentiation**
Limitation : La courbe du taux de hasard de la distribution de Gompertz ne peut être que constante ou décroissante.
Solution proposée : En projetant l'idée de El-Gohary *et al.* [16], nous avons ajouté un paramètre exposant, que nous appellerons par la suite γ, à la fonction de répartition de la distribution de Gompertz.

Niveau 2 : Prise en compte du taux de guérison

Limitation : La distribution propre de Gompertz suggère que toutes les populations de patients sont susceptibles à l'évènement d'intérêt, qui est généralement la mort du patient ou bien la récurrence de la maladie après le traitement sans considérer une sous population de survivants.
Solution proposée : Le deuxième niveau de généralisation consiste à la modification du modèle de telle sorte que la fonction de survie ne tend plus vers zéro, mais vers une valeur entre $[0,1]$ qui représente le taux de guérison θ. La modification proposée est d'élargir le domaine du paramètre d'échelle β de R_+^* à R^*. Dans le cas où β est un réel négatif, on obtient la distribution de Gompertz généralisée propre. Dans le cas où β est un réel positif, on obtient la distribution de Gompertz généralisée impropre ou défectueuse.

Niveau 3 : L'extension Marshall-Olkin
Limitation : Il faut des modèles plus flexibles pour mieux décrire les phénomènes réels.
Solution proposée : Ajouter un paramètre de plus à un modèle bien-fondé est une manière ancestrale d'ajouter de la flexibilité au modèle. Le troisième niveau de généralisation consiste à appliquer la famille de distributions Marshall-Olkin expliquée dans le chapitre précédant 1.3.1.1.

2.2.3 Description du modèle

La fonction de survie de la distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin (MO-GDGD) est obtenue en insérant la fonction de survie (1.40) dans l'équation (1.23). Le nouveau modèle de durée de vie est alors donné par (2.1) :

$$\mathrm{S}_{\mathrm{MO}}\left(t;r,\alpha,\beta,\gamma\right)=\frac{r\left(1-\left(1-\mathrm{e}^{\frac{\alpha}{\beta}\left(\mathrm{e}^{-\beta t}-1\right)}\right)^{\gamma}\right)}{1-(1-r)\left(1-\left(1-\mathrm{e}^{\frac{\alpha}{\beta}\left(\mathrm{e}^{-\beta t}-1\right)}\right)^{\gamma}\right)}, \tag{2.1}$$

où $t>0$, $r,\alpha,\gamma>0$ et $\beta\in\mathrm{R}^*$.
La figure 2.1 montre les scénarios possibles de la fonction densité de probabilité. Les figures 2.2 affiche la fonction de survie de la distribution proposée pour quelques valeurs de r, α, β et γ.

Théorème 2.2.1. *Si r est le paramètre d'inclinaison de Marshall-Olkin, alors le taux de guérison θ_{MO} d'une distribution généralisée par la famille Marshall-Olkin est basé sur le taux de guérison θ de la distribution de base de telle sorte que*

$$\theta_{\mathrm{MO}}=\frac{r\theta}{1-(1-r)\theta}.$$

Démonstration. En se basant sur le fait que la généralisation par Marshall-Olkin d'une distribution défectueuse est aussi une distribution défectueuse prouvée par Rocha *et al.* [33], nous pouvons admettre que le taux de guérison de l'extension par Marshall-Olkin est bien considéré. Main-

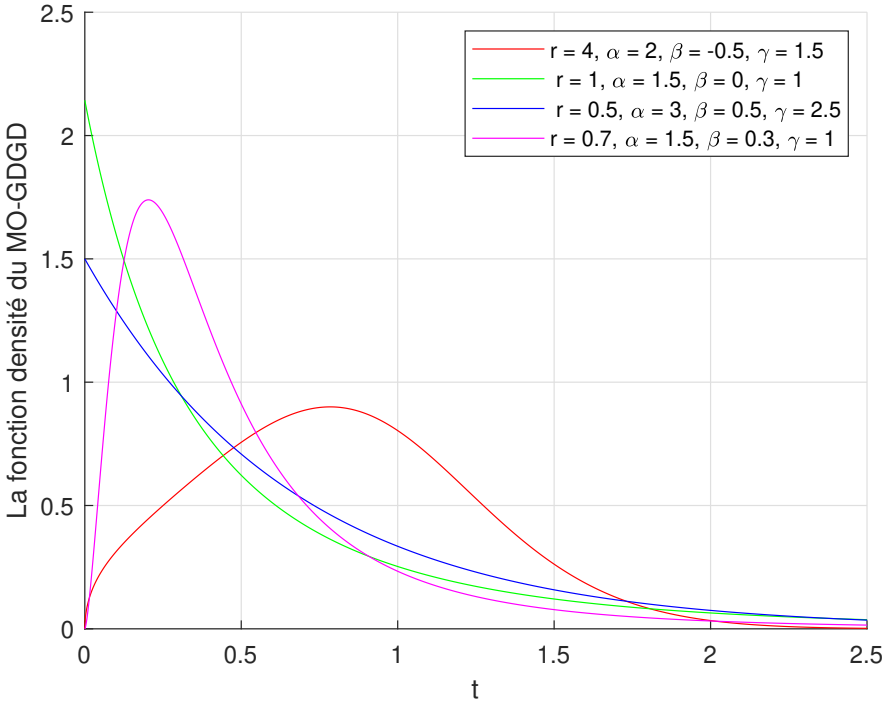


FIGURE 2.1 – La fonction de densité de MO-GDGD pour quelques valeurs de $(r, \alpha, \beta, \gamma)$.

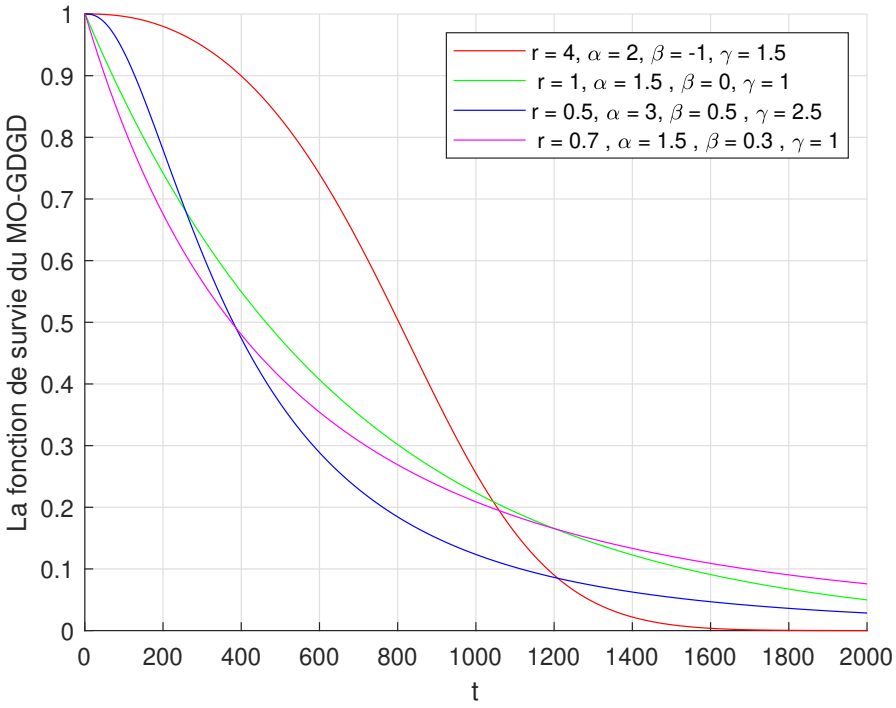


FIGURE 2.2 – La fonction de survie de MO-GDGD pour quelques valeurs de $(r, \alpha, \beta, \gamma)$.

tenant, si on note $S(t)$ la fonction de survie de la distribution défectueuse D et θ son taux de guérison de façon que $\lim_{t\to\infty} S(t) = \theta$, alors $\forall\ \theta \in [0,1]$, le taux de guérison θ_{MO} de l'extension de D par Marshall-Olkin est donnée par (2.2) :

$$\theta_{MO} = \lim_{t\to\infty} S_{MO}(t) = \lim_{t\to\infty} \frac{rS(t)}{1-(1-r)S(t)} = \frac{r\theta}{1-(1-r)\theta}, \tag{2.2}$$

où $S_{MO}(t)$ signifie la fonction de survie de la distribution D. En considérant que $0 \le \theta \le 1$ et $r > 0$, on peut montrer que $\theta_{MO} \in [0,1]$. □

Corollaire 2.2.2. *Tenant compte de (1.41), le taux de guérison de la distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin pour des valeurs de* $\beta > 0$ *est donnée par (2.3) :*

$$\theta_{MO} = \frac{r\left(1-\left(1-e^{-\frac{\alpha}{\beta}}\right)^{\gamma}\right)}{1-(1-r)\left(1-\left(1-e^{-\frac{\alpha}{\beta}}\right)^{\gamma}\right)}. \tag{2.3}$$

Le paramètre d'inclinaison r est ajouté pour l'analyse des données complexe et la maîtrise du mécanisme d'activation de risques latents [47].

La PDF du modèle proposé et la fonction de répartition correspondante sont données par :

$$f_{MO}(t;r,\alpha,\beta,\gamma) = \frac{r\gamma\alpha e^{-\beta t} e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)}\left(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)}\right)^{\gamma-1}}{\left(1-(1-r)(1-(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)})^{\gamma})\right)^2} \tag{2.4}$$

et

$$F_{MO}(t;r,\alpha,\beta,\gamma) = \frac{\left(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)}\right)^{\gamma}}{(1-r)\left(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)}\right)^{\gamma}+r}. \tag{2.5}$$

La fonction du taux de hasard est donc (2.6) :

$$H_{MO}(t;r,\alpha,\beta,\gamma) = \frac{f_{MO}(t;r,\alpha,\beta,\gamma)}{1-F_{MO}(t;r,\alpha,\beta,\gamma)}. \tag{2.6}$$

La fonction du taux de hasard indique le changement du taux de défaillance. Cette fonction présente un intérêt intrinsèque pour la population de patients ayant survécu pour une période bien déterminée et cherche une estimation pronostique. Le taux de défaillance de MO-GDGD peut être croissant, décroissant, en forme d'une baignoire, parabolique ou constant selon les valeurs des paramètres de forme. Les différentes formes de la fonction du taux de hasard sont présentées dans la figure 2.3.

La figure 2.4 montre l'effet du paramètre d'inclinaison ou de forme de Marshall-Olkin sur la fonction de risque ; si $r \ge 1$ alors la courbe est décalée vers le bas ($\forall t \ge 0, H_{MO}(t) \le H(t)$) et si $0 < r \le 1$ alors la courbe est décalée vers le haut ($\forall t \ge 0, H_{MO}(t) \ge H(t)$). Après un certain temps, les courbes superposent. Quant à la fonction densité de probabilité, les courbes présentées dans la figure 2.1 montrent qu'elles peuvent être soit décroissantes ou bien unimodales.

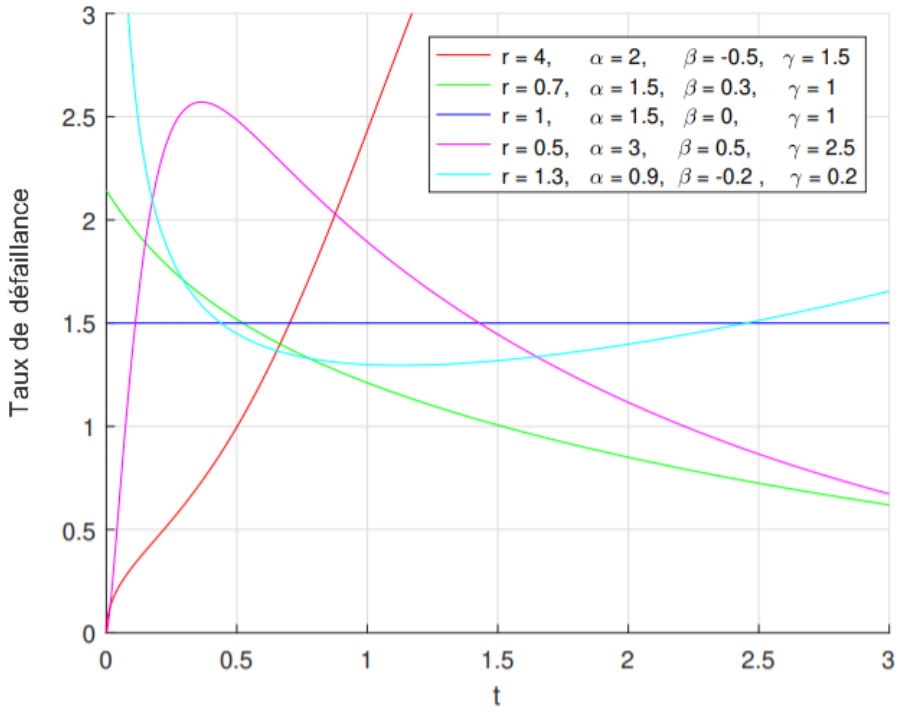


FIGURE 2.3 – Les différentes formes de la fonction du taux de hasard selon les différentes valeurs de $(r, \alpha, \beta, \gamma)$.

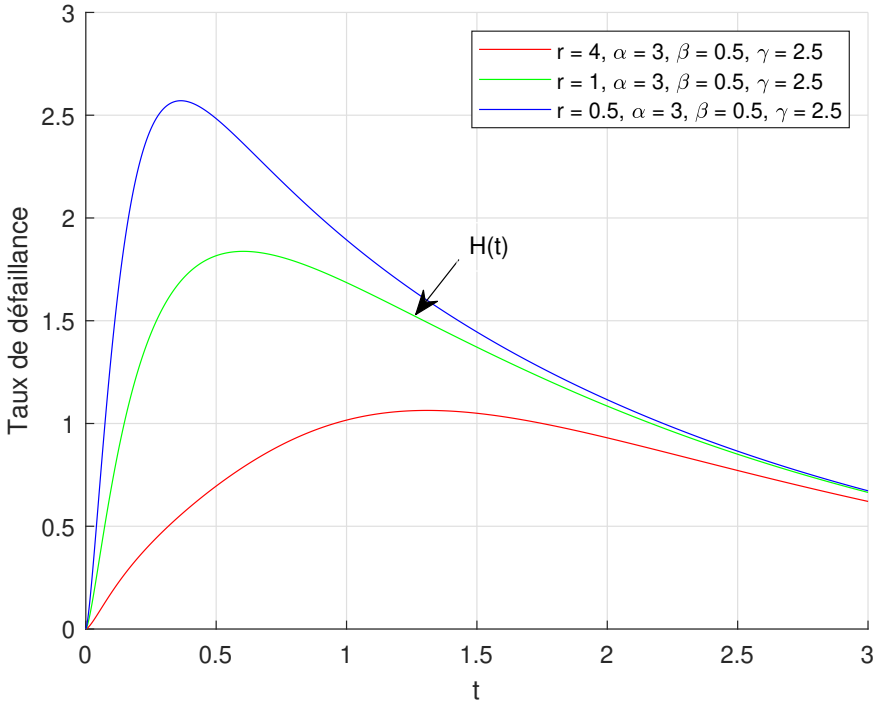


FIGURE 2.4 – L'effet du paramètre de forme de Marshall-Olkin sur la courbe du taux de hasard.

Les propriétés du modèle MO-GDGD proposé sont exposées en Annexe **A** page I.

2.2.4 Cas particuliers

Les sous modèles de la distribution MO-GDGD sont résumés dans le tableau 2.1.

TABLEAU 2.1 – Les cas particuliers du modèle de Gompertz défectueux et généralisé selon Marshall-Olkin.

r	α	β	γ	**Sous-modèle**	**Référence**
1	>0	0	1	Exponentiel	[48]
>0	>0	0	1	Exponentiel généralisé selon MO	[7]
1	>0	<0	1	Distribution de Gompertz (GD)	[44]
1	>0	>0	1	GD modifié (MGD)	[30]
1	>0	<0	>0	GD généraliséd (GGD)	[16]
1	>0	>0	>0	GD modifiéd et généraliséd (MGGD)	[46]
>0	>0	<0	>0	GGD généraliséd selon MO	[11]
>0	>0	>0	1	GD généraliséd selon MO (MO-GD)	[32]
1	>0	$\in \mathbf{R^*}$	>0	GD généraliséd et défectueuse (GDGD)	**Proposé** [49]
>0	>0	$\in \mathbf{R^*}$	>0	MO-GDGD	**Proposé**[49]

Remarque 2. *Le modèle proposé génère, entre autres, un modèle défectueux comme un cas particulier. Ce modèle est une généralisation à deux niveaux du modèle de Gompertz. Dans ce sous modèle, la généralisation selon Marshall-Olkin n'est pas appliquée. On l'appelle le modèle de Gompertz généralisé et défectueux (GDGD).*

2.2.5 Inférence statistique

L'estimateur de maximum de vraisemblance est une méthode paramétrique qui sert à inférer les paramètres inconnus d'une distribution [50]. Soit l'échantillon aléatoire $t_1, t_2, \ldots, t_n$ composé de n observations indépendantes de la distribution $\mathrm{MO-GDGD}(t; r, \alpha, \beta, \gamma)$ dont la fonction de répartition est donnée dans l'équation (2.5). On pose $\Theta = (r, \alpha, \beta, \gamma)^{\mathrm{T}}$ le vecteur de paramètres du modèle MO-GDGD. L'estimation de ces paramètres par la méthode de maximum de vraisemblance est obtenue comme suit :

Si $\mathrm{L}(t_i;\Theta)$ est la fonction de vraisemblance de Θ pour une réalisation $(t_1, \ldots, t_n)$ d'un échantillon, le logarithme népérien de la vraisemblance est :

$$\begin{aligned} l &= \ln \mathrm{L}(t_i;\Theta) = n\ln(r) + n\ln(\alpha) + n\ln(\gamma) - \beta \sum_{i=1}^{n} t_i + \frac{\alpha}{\beta} \sum_{i=1}^{n} \left(\mathrm{e}^{-\beta t_i} - 1\right) \\ &\quad + (\gamma - 1) \sum_{i=1}^{n} \ln(\mathrm{B}(t_i;\Theta)) - 2 \sum_{i=1}^{n} \ln(\mathrm{C}(t_i;\Theta)), \end{aligned} \tag{2.7}$$

où les fonctions non linéaires $B(t;\Theta)$ et $C(t;\Theta)$ sont :

$$\begin{aligned} B(t;\Theta) &= 1-A(t;\Theta), \\ C(t;\Theta) &= 1-(1-r)\left(1-(B(t;\Theta))^{\gamma}\right), \\ &\text{où} \\ A(t;\Theta) &= e^{\frac{\alpha}{\beta}\left(e^{-\beta t}-1\right)}. \end{aligned} \tag{2.8}$$

On calcule les dérivées partielles premières respectives du logarithme de la fonction de vraisemblance (2.7) par rapport à r, α , β et γ. Ensuite, on pose que les équations résultantes sont égales à zéro. On obtient donc un système de quatre équations non linéaires à quatre paramètres inconnus. Les dérivées partielles du logarithme de la fonction de vraisemblance sont données par :

$$\begin{aligned} \frac{\partial l}{\partial \alpha}(t;\Theta) &= \frac{n}{\alpha}-\frac{n}{\beta}+\frac{1}{\beta}\sum_{i=1}^{n} e^{-\beta t_i}+\frac{(1-\gamma)}{\beta}\sum_{i=1}^{n}\frac{\left(e^{-\beta t_i}-1\right)A(t_i;\Theta)}{B(t_i;\Theta)} \\ &\quad +2\frac{\gamma(1-r)}{\beta}\sum_{i=1}^{n}\left(\left(e^{-\beta t_i}-1\right)\left((B(t_i;\Theta))^{\gamma}\right)\frac{A(t_i;\Theta)}{C(t_i;\Theta)}\right), \\ \frac{\partial l}{\partial \beta}(t;\Theta) &= -\sum_{i=1}^{n} t_i-\frac{\alpha}{\beta^2}\left(-n+\sum_{i=1}^{n} e^{-\beta t_i}\right)-\frac{\alpha}{\beta}\sum_{i=1}^{n} t_i e^{-\beta t_i}-\left(\gamma-1\right) \\ &\quad \sum_{i=1}^{n}\frac{A(t_i;\Theta)E(t_i;\Theta)}{B(t_i;\Theta)}-2\gamma(1-r)\sum_{i=1}^{n}\frac{(B(t_i;\Theta))^{\gamma-1}E(t_i;\Theta)A(t_i;\Theta)}{C(t_i;\Theta)}, \\ \frac{\partial l}{\partial \gamma}(t;\Theta) &= \frac{n}{\gamma}+\sum_{i=1}^{n}\ln(B(t_i;\Theta))-2(1-r)\sum_{i=1}^{n}\frac{(B(t_i;\Theta))^{\gamma}\ln(B(t_i;\Theta))}{C(t_i;\Theta)}, \\ \frac{\partial l}{\partial r}(t;\Theta) &= \frac{n}{r}-2\sum_{i=1}^{n}\frac{1-(B(t_i;\Theta))^{\gamma}}{C(t_i;\Theta)}, \end{aligned} \tag{2.9}$$

où les fonctions non linéaires $D(t;\Theta)$, $E(t;\Theta)$ et $G(t;\Theta)$ sont :

$$\begin{aligned} D(t;\Theta) &= \frac{1-(B(t;\Theta))^{\gamma}}{C(t;\Theta)}, \\ E(t;\Theta) &= \frac{\alpha}{\beta^2}(e^{-\beta t}-1)+\frac{\alpha}{\beta}t e^{-\beta t}, \\ &\text{et} \\ G(t;\Theta) &= \frac{2}{\beta}E(t;\Theta)+\frac{\alpha}{\beta}t^2 e^{-\beta t}. \end{aligned} \tag{2.10}$$

Évidemment, les équations de vraisemblance (2.9) n'ont pas de solutions explicites et les estimations doivent donc être calculées numériquement.

Les limites de confiance de (1-δ) de r, α, β et γ sont calculées comme suit :
$\left(\hat{r}\pm Z_{\frac{\delta}{2}}\sqrt{Var(\hat{r})}\right)$, $\left(\hat{\alpha}\pm Z_{\frac{\delta}{2}}\sqrt{Var(\hat{\alpha})}\right)$, $\left(\hat{\beta}\pm Z_{\frac{\delta}{2}}\sqrt{Var(\hat{\beta})}\right)$ et $\left(\hat{\gamma}\pm Z_{\frac{\delta}{2}}\sqrt{Var(\hat{\gamma})}\right)$, où $Z_{\frac{\delta}{2}}$ suit la loi normale centrée réduite et la variance des paramètres estimés est la diagonale de la matrice de variance covariance asymptotique observée obtenue après l'inversion de la matrice d'information de Fisher.

2.2.6 Simulation

2.2.6.1 Algorithme de génération de données simulées

Compte tenu du coût toujours croissant des essais cliniques, de la charge de travail supplémentaire des cliniciens, des difficultés engendrées et des risques encourus par les patients au cours du processus clinique, dont les caractéristiques obéissent à une loi de probabilité connue, nous avons eu recours à des techniques théoriques et numériques telles que la simulation. Il est important de noter que les techniques de simulation permettent de régénérer avec précision des données statistiques à partir du phénomène étudié. Dans la littérature, il existe plusieurs techniques permettant la simulation des données. Parmi ces techniques, on peut citer à titre d'exemple : la méthode de Monte Carlo, la méthode de rejet, la méthode de Box-Muller ainsi que la méthode d'inversion de la CDF que nous illustrons par la suite [51].

La génération de données permet d'évaluer la performance des estimations du maximum de vraisemblance par rapport à la taille de l'échantillon et de montrer, entre autres, que les asymptotes habituelles des estimateurs du maximum de vraisemblance sont toujours valables pour les distributions défectueuses 1.3.3.2. Dans tous les chapitres, les études de simulation sont basées sur la méthode d'inversion de la CDF ayant cette configuration. La description de la génération de données est décrite ci-dessous. Supposons que l'instant d'occurrence d'un événement d'intérêt a une fonction de répartition $\mathrm{F}(t)$. On voudrait simuler un échantillon aléatoire de taille n contenant des temps réels, des temps censurés et une fraction de guérison. Les étapes suivantes sont appliquées :

1. Vérifier que la fonction de répartition $\mathrm{F}(t)$ de la distribution est bien une fonction bijective admettant une fonction réciproque.
2. Générer la fonction inverse $\mathrm{F}^{-1}(t)$ de la CDF.
3. Choisir les valeurs initiales des paramètres inconnus de la distribution et les remplacer dans la réciproque de la fonction de répartition qui sera notée T_i pour désigner les durées de vie simulées en fonction de la variable aléatoire u qui suit la loi uniforme sur $[0,1]$.
4. Générer la v.a. u de façon à ce qu'elle soit uniformément répartie sur l'intervalle $[0,1]$.
5. Répéter les étapes 3 et 4 n fois, pour obtenir une suite d'échantillon de données de durées de vie.
6. Créer le fichier dans lequel les données sont stockées.

La sortie de cet algorithme est un fichier de taille n qui contient les $\mathrm{T}_1,\ldots,\mathrm{T}_i,\ldots,\mathrm{T}_n$ qui représentent les durées de survie prévues pour les sujets dont les caractéristiques peuvent être décrites par la distribution étudiée. Pour le cas de données censurées à droite, une étape supplémentaire doit être ajoutée. Et ce pour fixer le niveau de censure et la valeur après laquelle l'information est considérée comme censurée.

2.2.6.2 Évaluation de la précision des estimateurs

Dans les simulations, on choisit toujours le nombre de simulations S effectué ainsi que la taille de l'échantillon. Dans chaque taille d'échantillon, on pourrait calculer le biais, l'erreur quadratique moyenne, la probabilité de couverture et les longueurs de couverture pour chaque paramètre... Et ce pour évaluer la précision de l'estimation effectuée après l'étape de la simulation des données. Si $\hat{\theta}$ est la moyenne des S paramètres estimées pour chaque simulation, alors les équations (2.11) sont utilisées :

$$\begin{aligned} \mathrm{B}iais(\hat{\theta}) &= \hat{\theta} - \theta, \\ \mathrm{V}ar(\hat{\theta}) &= \frac{1}{S}\sum_{i=1}^{S}(\hat{\theta}_i - \theta)^2, \\ \mathrm{MSE}(\hat{\theta}) &= \mathrm{V}ar(\hat{\theta}) + \left(\mathrm{B}iais(\hat{\theta})\right)^2. \end{aligned} \tag{2.11}$$

La probabilité de couverture est la fréquence à laquelle la valeur réelle du paramètre reste dans la région de confiance, pour chaque simulation. Si l'on considère les résultats des S simulations effectuées et on fixe un niveau de signification égal à 0.95, un test pour vérifier l'équivalence des proportions peut être effectué. Alors, si on a les limites n_1 et n_2 tel que la probabilité que le paramètre étudié soit entre n_1 et n_2 est égale à 0.95. Autrement dit, étant donné un certain nombre de simulations S, nous pouvons nous attendre à ce que la probabilité de couverture reste entre n_1 et n_2 environ 95% du temps. La longueur de couverture est la différence entre les limites de confiance supérieures et inférieures.

2.2.6.3 Résultats de la simulation

En se basant sur la distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin, nous avons élaboré une étude numérique pour simuler les données de durées de vie tenant compte des résultats théoriques présentés ci-dessus (2.5), [52, 53].

La méthode d'inversion a été appliquée pour générer des échantillons aléatoires de la distribution MO-GDGD en utilisant l'équation (A.1) qui décrit la fonction quantile.

Les résultats des estimations en utilisant les données simulées sont présentés dans les tableaux 2.2, 2.3 et 2.4.

TABLEAU 2.2 – MLE, MSE correspondante et le biais pour $\beta > 0$ et r >1.

$\Theta = (r, \alpha, \beta, \gamma)$	**n**		**MLE**	**MSE**	**Biais**
(5,3.5,0.5,3)	30	$\hat{r}$	4.9115	0.0078	0.0885
		$\hat{\alpha}$	3.3913	0.0118	0.1087
		$\hat{\beta}$	0.4444	0.0031	0.0556
		$\hat{\gamma}$	3.0359	0.0013	0.0359
		$\hat{\theta}_{MO}$	0.0072	3.9690×10^{-5}	0.0063
	100	$\hat{r}$	4.9756	0.0006	0.0244
		$\hat{\alpha}$	3.7194	0.0481	0.2194
		$\hat{\beta}$	0.4850	0.0002	0.0150
		$\hat{\gamma}$	2.9439	0.0031	0.0561
		$\hat{\theta}_{MO}$	0.0068	4.5131×10^{-5}	0.0067
	150	$\hat{r}$	4.9309	0.0048	0.0691
		$\hat{\alpha}$	3.4168	0.0069	0.0832
		$\hat{\beta}$	0.4381	0.0038	0.0619
		$\hat{\gamma}$	3.0566	0.0032	0.0566
		$\hat{\theta}_{MO}$	0.0061	0.0001	0.0074
	500	$\hat{r}$	5.0774	0.0060	0.0774
		$\hat{\alpha}$	3.5996	0.0099	0.0996
		$\hat{\beta}$	0.4471	0.0028	0.0529
		$\hat{\gamma}$	3.0186	0.0003	0.0186
		$\hat{\theta}_{MO}$	0.0049	0.0001	0.0086

Des tailles différentes d'échantillon sont utilisés avec différents choix des valeurs de paramètres. Les valeurs des vrais paramètres ont été choisi de telle sorte que les deux cas de présence et d'absence de la fraction survivante sont considérés. En fait, dans le tableau 2.3, on choisit une valeur de β négative pour dire qu'on est dans le cas d'absence d'une proportion de guéris. Par contre, dans les tableaux 2.2 et 2.4, on choisit une valeur de β positive pour indiquer la présence d'une proportion de guéris. Quant au paramètre de forme r de Marshall-Olkin, les trois cas possibles $r > 1$, $r = 1$ et $r < 1$ sont considérés dans les tableaux 2.2, 2.3 et 2.4.

Dans chacun des trois scénarios présentés, les biais ainsi que les erreurs quadratiques moyennes (MSE) pour chacun des paramètres et pour les différentes tailles d'échantillon sont calculés.

À partir de cette étude, nous pouvons conclure que, pour les trois cas, à mesure que la taille de l'échantillon croit, les valeurs des MSE et biais tendent vers zéro.

TABLEAU 2.3 – MLE, MSE correspondante et le biais pour $\beta < 0$ et r=1.

$\Theta = (r, \alpha, \beta, \gamma)$	**n**		**MLE**	**MSE**	**Biais**
(1,4,-0.5,1)	30	$\hat{r}$	0.7389	0.0682	0.2611
		$\hat{\alpha}$	4.1820	0.0331	0.1820
		$\hat{\beta}$	-0.6255	0.0158	0.1255
		$\hat{\gamma}$	0.8562	0.0207	0.1438
	100	$\hat{r}$	0.9527	0.0022	0.0473
		$\hat{\alpha}$	3.9592	0.0017	0.0408
		$\hat{\beta}$	-0.6067	0.0114	0.1067
		$\hat{\gamma}$	0.9447	0.0031	0.0553
	150	$\hat{r}$	0.9959	1.6810×10^{-5}	0.0041
		$\hat{\alpha}$	4.1603	0.0257	0.1603
		$\hat{\beta}$	-0.5648	0.0042	0.0648
		$\hat{\gamma}$	1.0115	0.0001	0.0115
	500	$\hat{r}$	0.9297	0.0049	0.0703
		$\hat{\alpha}$	4.0892	0.0080	0.0892
		$\hat{\beta}$	-0.6079	0.0116	0.1079
		$\hat{\gamma}$	0.9837	0.0003	0.0163

TABLEAU 2.4 – MLE, MSE correspondante et le biais pour $\beta > 0$ et r<1.

$\Theta = (r, \alpha, \beta, \gamma)$	**n**		**MLE**	**MSE**	**Biais**
(0.8,0.5,0.1,1.2)	30	$\hat{r}$	0.7190	00066	0.0810
		$\hat{\alpha}$	0.4967	1.0890×10^{-5}	0.0033
		$\hat{\beta}$	0.0743	0.0007	0.0257
		$\hat{\gamma}$	1.1885	0.0001	0.115
		$\hat{\theta}_{MO}$	0.0011	2.8886×10^{-5}	0.0054
	100	$\hat{r}$	0.7685	9.9225×10^{-4}	0.0315
		$\hat{\alpha}$	0.4884	1.3456×10^{-4}	0.0116
		$\hat{\beta}$	0.0847	2.3409×10^{-4}	0.0153
		$\hat{\gamma}$	1.1919	6.5610×10^{-5}	0.0081
		$\hat{\theta}_{MO}$	0.0019	2.0926×10^{-5}	0.0046
	150	$\hat{r}$	0.9150	0.0132	0.1150
		$\hat{\alpha}$	0.5117	0.0001	0.0117
		$\hat{\beta}$	0.0743	0.0007	0.0257
		$\hat{\gamma}$	1.2023	5.2900×10^{-6}	0.0023
		$\hat{\theta}_{MO}$	0.0011	2.8886×10^{-5}	0.0054
	500	$\hat{r}$	0.8606	0.0037	0.0606
		$\hat{\alpha}$	0.4896	0.0001	0.0104
		$\hat{\beta}$	0.0820	0.0003	0.0180
		$\hat{\gamma}$	1.1832	0.0003	0.0168
		$\hat{\theta}_{MO}$	0.0026	1.5012×10^{-5}	0.0039

2.3 Application et discussion

Dans le but de prouver l'applicabilité du modèle proposé en pratique, cette partie est consacré à l'analyse des données réelles. L'estimation des paramètres inconnus du modèle MO-GDGD est présentée. Des tests de conformité sont établis pour tester la supériorité du modèle proposé par rapport à d'autres en termes d'ajustement ou de fitting des bases des données actuelles. Six sous modèles de la distribution MO-GDGD sont adaptés aux bases de données sélectionnées et sont testés comme des hypothèses nulles.

(i) $H_0 : r = 1\ ,\ \gamma = 1, \beta < 0$, les données suivent la distribution $GD(\alpha, \beta)$,

(ii) $H_0 : \gamma = 1, \beta < 0$, les données suivent la distribution $MO - GD(r, \alpha, \beta)$,

(iii) $H_0 : r = 1\ ,\ \gamma = 1, \beta > 0$, les données suivent la distribution $MGD(\alpha, \beta)$,

(iv) $H_0 : r = 1, \beta < 0$, les données suivent la distribution $GGD(\alpha, \beta, \gamma)$,

(v) $H_0 : \beta < 0$, les données suivent la distribution $MO - GGD(r, \alpha, \beta, \gamma)$,

(vi) $H_0 : r = 1, \beta > 0$, les données suivent la distribution $MGGD(\alpha, \beta)$

contre la distribution alternative H_1 où les données suivent la distribution $MO-GDGD(r, \alpha, \beta, \gamma)$.

Le test paramétrique du rapport de vraisemblance Γ_{H_0} est appliqué pour tester si les hypothèses nulles sont rejetées en faveur de la distribution proposée. Les critères de sélection des modèles tels que le critère d'information d'Akaike (AIC), le critère d'information cohérent d'Akaike (CAIC) et le critère d'information bayésien (BIC) pour chacune des bases de données sont également présentés ci-après.

Des tests non paramétriques tels que le test de Kolmogorov-Smirnov (KS) et le test de Cramér-von Mises (CvM) avec leurs valeurs-p correspondantes sont également appliqués pour MO-GDGD et ses sous modèles.

À partir de l'équation (2.3), la fraction de survie à long terme est estimée à (2.12) :

$$\hat{\theta}_{MO} = \frac{\hat{r}\left(1 - \left(1 - e^{-\frac{\hat{\alpha}}{\hat{\beta}}}\right)^{\hat{\gamma}}\right)}{1 - (1 - \hat{r})\left(1 - \left(1 - e^{-\frac{\hat{\alpha}}{\hat{\beta}}}\right)^{\hat{\gamma}}\right)}. \tag{2.12}$$

2.3.1 Application au cancer de la vessie

2.3.1.1 Base de données du cancer de la vessie

Le cancer de la vessie est une pathologie provenant des tissus de la vessie dans laquelle les cellules se développent anormalement et se divisent de manière incontrôlable. La base présente des durées de rémission de 128 patients souffrant de cancer de la vessie. Cette base a été rapportée dans l'article de Lee & Wang [54]. La base de données apparaît dans le tableau D.1 en Annexe **D** et ses propriétés statistiques sont présentées dans le tableau 2.5.

Minimum	**Quartile inférieur**	**Médiane**	**Quartile supérieur**	**Maximum**
0.0800	3.3600	6.6050	12	79.050

TABLEAU 2.5 – Le résumé à cinq chiffres de la base de données liées au cancer de la vessie.

2.3.1.2 Analyse inférentielle

Le tableau 2.6 donne les estimations par la méthode de maximum de vraisemblance des paramètres inconnus du modèles MO-GDGD ainsi que certains de ses sous modèles. Les valeurs-p et les résultats du test Γ_{H_0} pour les données du cancer de la vessie sont aussi déclarés dans le même tableau.

TABLEAU 2.6 – Estimation des paramètres inconnus du modèle MO-GDGD et quelques sous modèles, les valeurs-p et le rapport de vraisemblance Γ_{H_0} associés pour les données du cancer de la vessie.

Modèle	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	**Valeur-p**	Γ_{H_0}
$GD(\alpha,\beta)$	–	0.0875	-0.0062	–	< 0.0001	11.6220
$MO-GD(r,\alpha,\beta)$	0.0205	0.0028	-0.0414	–	< 0.0001	19.2090
$MGD(\alpha,\beta)$	–	0.1449	0.0046	–	< 0.0001	16.3670
$GGD(\alpha,\beta,\gamma)$	–	0.0721	-0.0004	0.8652	< 0.0001	20.6498
$MO-GGD(r,\alpha,\beta,\gamma)$	0.0397	0.0180	-0.0014	1.4503	< 0.0001	16.8596
$MGGD(\alpha,\beta,\gamma)$	–	0.1647	0.0027	1.5383	0.0015	10.0648
$MO-GDGD(r,\alpha,\beta,\gamma)$	0.3256	0.0814	0.0019	1.5381	–	–

Les intervalles de confiance à 95% des estimations des paramètres inconnus (r,α,β,γ) pour les données du cancer de la vessie sont générés dans le tableau 2.7.

TABLEAU 2.7 – Les intervalles de confiance à 95% des paramètres du modèle MO-GDGD pour les données du cancer de la vessie.

Paramètre	**Intervalle de confiance à** 95%
$\hat{r}$	(0.2297, 0.4215)
$\hat{\alpha}$	(0.0674, 0.0954)
$\hat{\beta}$	(-0.0107, 0.0145)
$\hat{\gamma}$	(1.3392, 1.7370)

A noter que pour β, une valeur négative peut être tolérée tant qu'elle est proche de zéro. Les erreurs standards de $\hat{r}$, $\hat{\alpha}$, $\hat{\beta}$ et $\hat{\gamma}$ pour les données du cancer de la vessie sont, respectivement, 0.0490, 0.0072, 0.0064 et 0.1015. Compte tenu de la période d'étude, le taux estimé des patients guéris est donné par 0.0011. Les valeurs de la fonction de vraisemblance sous H_0 et les valeurs des critères d'information AIC, BIC et CAIC pour la sélection des modèles en plus des résultats des tests KS et CvM pour les données du cancer de la vessie sont affichés dans les tableaux 2.8 et 2.9 respectivement.

TABLEAU 2.8 – Les valeurs du logarithme de la fonction de vraisemblance l sous H_0 et les critères d'information AIC, BIC et CAIC pour les données du cancer de la vessie.

Modèle	$-l$	**AIC**	**BIC**	**CAIC**
GD(α,β)	417.5862	839.1724	844.8765	846.8765
MO – GD(r,α,β)	421.3797	848.7594	857.3155	860.3155
MGD(α,β)	419.9587	843.9174	849.6215	851.6215
GGD(α,β,γ)	422.1001	850.2002	858.7563	861.7563
MO – GGD(r,α,β,γ)	415.0506	838.1012	849.5093	853.5093
MGGD(α,β,γ)	416.8076	839.6152	848.1713	851.1713
MO – GDGD(r,α,β,γ)	**411.7752**	**831.5504**	**842.9585**	**846.9585**

TABLEAU 2.9 – Résultats des tests de qualité d'ajustement pour les données du cancer de la vessie.

Modèle	KS	Valeur-p	CvM	Valeur-p
GD(α,β)	0.1154	0.0611	0.4764	0.0456
MO – GD(r,α,β)	0.1146	0.0640	0.5236	0.0346
MGD(α,β)	0.1434	0.0093	0.5087	0.0378
GGD(α,β,γ)	0.1477	0.0067	0.6369	0.0180
MO – GGD(r,α,β,γ)	0.0657	0.6149	0.1523	0.3874
MGGD(α,β,γ)	0.0501	0.8881	0.0629	0.7965
MO – GDGD(r,α,β,γ)	**0.0418**	**0.9721**	**0.0437**	**0.9124**

2.3.2 Application au cancer du sang

2.3.2.1 Base de données du cancer du sang

Le cancer du sang est une maladie qui prend naissance dans les tissus hématopoïétiques, par exemple la moelle osseuse où le sang est produit, ou dans les cellules du système immunitaire, ce qui affecte la production et la fonction des cellules sanguines. Il existe trois principaux types de cancer du sang : la leucémie, le lymphome et le myélome. Les durées de vie de 43 patients atteints d'un cancer du sang, obtenues par Abouammoh *et al.* [55], sont utilisées dans ce travail. Il convient de noter que le type de cancer du sang dans les présentes données n'est pas spécifié. L'ensemble de données, présenté dans le tableau D.2 (voir Annexe **D**), est collecté dans un hôpital du ministère de la Santé d'Arabie Saoudite. Les propriétés extraites de cette base sont présentées dans le tableau 2.10.

Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum
115	807	1251	1599	1965

TABLEAU 2.10 – Le résumé à cinq chiffres de la base de données du cancer du sang.

2.3.2.2 Analyse inférentielle

Les estimateurs du maximum de vraisemblance des quatre paramètres du modèle proposé ainsi que certains de ses cas particuliers et leurs valeurs-p associées et les valeurs du test du rapport de vraisemblance pour les données sur le cancer du sang sont présentés dans le tableau 2.11. La valeur de la fonction log-vraisemblance l sous H_0 et les valeurs des critères d'information AIC,

BIC et CAIC sont également affichées dans le tableau 2.12. L'estimation de la fraction de guérison pour le MO-GDGD est de 17,56%.

TABLEAU 2.11 – Estimateurs du maximum de vraisemblance des paramètres inconnus de MO-GDGD et de certains de ses cas particuliers, les valeurs-p associées et les valeurs de test du rapport de vraisemblance Γ_{H_0} pour les données sur le cancer du sang.

Modèle	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	**Valeur-p**	Γ_{H_0}
GD(α,β)	–	0.003	-0.0001	–	< 0.0001	30.2674
MGD(α,β)	–	0.008	0.0021	–	< 0.0001	120.6924
GGD(α,β,γ)	–	0.000125	-0.00003	0.03	< 0.0001	119.5032
MGGD(α,β,γ)	–	0.0055	0.0017	1.2374	< 0.0001	40.6888
MO – GDGD(r,α,β,γ)	12.6053	0.0244	0.0025	290.3021	–	–

TABLEAU 2.12 – La valeur de la fonction log-vraisemblance l sous H_0 et les critères d'information AIC, BIC et CAIC pour les données sur le cancer du sang.

Modèle	$-l$	**AIC**	**BIC**	**CAIC**
GD(α,β)	409.7320	823.4640	826.9864	828.9864
MGD(α,β)	545.9445	913.8890	917.4114	919.4114
GGD(α,β,γ)	454.3499	914.6998	919.9834	922.9834
MGGD(α,β,γ)	414.9427	835.8854	841.1690	844.1690
MO – GDGD(r,α,β,γ)	394.5983	797.1966	804.2414	808.2414

2.3.3 Application au cancer du sein

2.3.3.1 Base de données du cancer du sein

Le cancer du sein est le cancer le plus diagnostiqué chez les femmes à travers le monde. Les cellules cancéreuses, dans le cas d'un cancer du sein, peuvent rester dans le sein comme elles peuvent se propager dans le corps des patients à travers les vaisseaux sanguins. La base de données sur le cancer du sein a été collectée par King *et al.* [56] sur un modèle animal. L'ensemble de données se compose des durées de vie des tumeurs mammaires de 30 rats nourris avec un régime non saturé. Les durées de vie des tumeurs mammaires sont présentées dans le tableau D.3 (voir Annexe **D**) et leurs propriétés statistiques sont affichées dans le tableau 2.13.

Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum
60	68	92.5	112	178

TABLEAU 2.13 – Le résumé à cinq chiffres de la base de données du cancer du sein.

2.3.3.2 Analyse inférentielle

Les estimateurs du maximum de vraisemblance des paramètres inconnus de MO-GDGD ainsi que certains de ses cas particuliers et leurs valeurs-p associées et les valeurs de test Γ_{H_0} pour les données sur les tumeurs mammaires sont présentés dans le tableau 2.14. Les valeurs AIC, BIC et CAIC sont également calculées pour estimer la qualité relative du modèle statistique proposé pour les données sur les tumeurs mammaires. Les résultats sont présentés dans le tableau 2.15. L'estimation de la fraction de guérison pour le MO-GDGD est de 1.1%.

TABLEAU 2.14 – Les estimateurs du maximum de vraisemblance des paramètres inconnus de MO-GDGD et de certains de ses cas particuliers, les valeurs-p associées et les valeurs de test du rapport de vraisemblance Γ_{H_0} pour les données sur les tumeurs mammaires.

Modèle	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	**Valeur-p**	Γ_{H_0}
GD(α,β)	–	0.0075	-0.0053	–	< 0.0001	31.7124
MGD(α,β)	–	0.0147	0.0065	–	< 0.0001	65.2434
GGD(α,β,γ)	–	0.0105	-0.0101	1.9182	< 0.0001	16.8596
MGGD(α,β,γ)	–	0.1059	0.0085	257.5688	0.0037	8.4114
MO – GDGD(r,α,β,γ)	1.2051	0.0968	0.0097	198.8172	–	–

TABLEAU 2.15 – La valeur de la fonction log-vraisemblance l sous H_0 et les critères d'information AIC, BIC et CAIC pour les données sur les tumeurs mammaires.

Modèle	$-l$	**AIC**	**BIC**	**CAIC**
GD(α,β)	160.3068	324.6136	327.4160	329.4160
MGD(α,β)	177.0723	358.1446	360.9470	362.9470
GGD(α,β,γ)	152.8804	311.7608	315.9644	318.9644
MGGD(α,β,γ)	148.6563	303.3126	307.5162	310.5162
MO – GDGD(r,α,β,γ)	144.4506	296.9012	302.5060	306.5060

2.3.4 Application à la sclérose latérale amyotrophique

2.3.4.1 Base de données de la sclérose latérale amyotrophique

La sclérose latérale amyotrophique (ALS) est une maladie neurodégénérative qui touche principalement les cellules nerveuses responsables du contrôle des mouvements musculaires volontaires. La base de données PRO-ACT, disponible sur :
`https://nctu.partners.org/ProACT/Data/Index`, a été exploitées. Le consortium PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials) a été formé en 2011 et mis à jour en 2015. Un sous-échantillon de données complètes a été exploité dans cette partie.

Les sous-échantillons utilisés dans cette partie consistent à des durées de survie de 200 patients souffrant de sclérose latérale amyotrophique (ALS). Les statistiques descriptives de la base de données sont résumées dans le tableau 2.16.

Minimum	**Quartile inférieur**	**Médiane**	**Quartile supérieur**	**Maximum**
7	162	289.5	401	1271

TABLEAU 2.16 – Le résumé à cinq chiffres de la base de données PRO-ACT.

2.3.4.2 Analyse inférentielle

Les paramètres inconnus du modèle MO-GDGD ainsi que ses sous modèles ont été estimés par la méthode de maximum de vraisemblance. Les valeurs-p et les rapports de vraisemblances associés pour le sous ensemble ont été présentés dans le tableau 2.17.

TABLEAU 2.17 – Résultats de l'estimation des paramètres des modèles MO-GDGD et ses cas particuliers, les valeurs-p et le rapport de vraisemblance Γ_{H_0} associés pour les données ALS.

Modèle	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	**Valeur-p**	Γ_{H_0}
$GD(\alpha,\beta)$	–	0.0020	-0.0020	–	< 0.0001	25.3548
$MO-GD(r,\alpha,\beta)$	1.5570	0.0026	-0.0017	–	< 0.0001	19.4944
$MGD(\alpha,\beta)$	–	0.0031	0.0001	–	< 0.0001	71.2062
$GGD(\alpha,\beta,\gamma)$	–	0.0025	-0.0017	1.3721	0.0003	13.0662
$MO-GGD(r,\alpha,\beta,\gamma)$	3.3301	0.0064	-0.0002	1.4891	0.1552	2.0200
$MGGD(\alpha,\beta,\gamma)$	–	0.0042	0.0001	1.7019	< 0.0001	28.6312
$MO-GDGD(r,\alpha,\beta,\gamma)$	9.1180	0.0091	0.0004	1.1216	–	–

Les intervalles de confiance à 95% des estimations des paramètres inconnus $(r, \alpha, \beta, \gamma)$ pour les données de ALS sont générés dans le tableau 2.18.

TABLEAU 2.18 – Les intervalles de confiance de 95% des paramètres du modèle MO-GDGD pour les données de ALS.

Paramètre	**Intervalle de confiance à** 95%
$\hat{r}$	(6.9125, 11.3235)
$\hat{\alpha}$	(0.0084, 0.0098)
$\hat{\beta}$	(0.0001, 0.0007)
$\hat{\gamma}$	(0.8859, 1.3573)

Les erreurs standards de $\hat{r}$, $\hat{\alpha}$, $\hat{\beta}$ et $\hat{\gamma}$ sont, respectivement, 1.1252, 0.0004, 0.0002 et 0.1203. Compte tenu de la période d'étude, le taux estimé par MO-GDGD des patients guéris est donné par 0.0012.

Les valeurs de la fonction de vraisemblance l sous H_0 et les valeurs des critères d'information AIC, BIC et CAIC pour la sélection des modèles en plus des résultats des tests KS et CvM pour les données ALS sont aussi affichés dans les tableaux 2.19 et 2.20 respectivement.

TABLEAU 2.19 – Les valeurs du logarithme de la fonction de vraisemblance l sous H_0 et les critères d'information AIC, BIC et CAIC pour les données ALS.

Modèle	$-l$	**AIC**	**BIC**	**CAIC**
$GD(\alpha, \beta)$	1323.9586	2651.9172	2658.5138	2660.5138
$MO-GD(r, \alpha, \beta)$	1321.0284	2648.0568	2657.9518	2660.9518
$MGD(\alpha, \beta)$	1346.8843	2697.7686	2704.3652	2706.3652
$GGD(\alpha, \beta, \gamma)$	1317.8143	2641.6286	2651.5236	2654.5236
$MO-GGD(r, \alpha, \beta, \gamma)$	1312.2912	2632.5824	2645.7757	2649.7757
$MGGD(\alpha, \beta, \gamma)$	1325.5968	2657.1936	2667.0886	2670.0886
$MO-GDGD(r, \alpha, \beta, \gamma)$	**1311.2812**	**2630.5624**	**2643.7557**	**2647.7557**

2.3.5 Discussion

Le test Γ_{H_0} et les valeurs-p correspondantes pour les bases de données utilisées sont présentés dans les tableaux 2.6, 2.14 , 2.11 et 2.17.

À partir des résultats du test Γ_{H_0}, on remarque que le modèle MO-GDGD est le modèle le plus efficace en terme d'ajustement des données. Les faibles valeurs-p montrent que les quatre

TABLEAU 2.20 – Résultats des tests de qualité d'ajustement pour les données ALS.

Modèle	KS	Valeur-p	CvM	Valeur-p
GD(α,β)	0.0845	0.1085	0.5003	0.0397
MO – GD(r,α,β)	0.0664	0.3259	0.2838	0.1502
MGD(α,β)	0.1589	<0.0001	1.9824	0.0010
GGD(α,β,γ)	0.0657	0.3388	0.1768	0.3210
MO – GGD(r,α,β,γ)	0.0425	0.8471	0.0630	0.7955
MGGD(α,β,γ)	0.0968	0.0440	0.4978	0.0403
MO – GDGD(r,α,β,γ)	**0.0406**	**0.8833**	**0.0498**	**0.8774**

hypothèses nulles sont rejetées en faveur du modèle MO-GDGD à un niveau de signification statistique établie à 5%.

Les résultats des critères AIC, BIC et CAIC présentés dans les tableaux 2.8 , 2.12, 2.15 et 2.19, et les résultats des tests d'ajustement présentés dans les tableaux 2.20 et 2.9, confirment que MO-GDGD permet un meilleur ajustement aux données.

Pour les données du cancer de la vessie, les valeurs AIC des modèles GD et MGGD sont très proches. Ces deux distributions, ont non seulement des valeurs AIC inférieurs à celle de MGD mais aussi ils surpassent nettement la distribution GGD. Cependant, la performance de ces quatre distributions en termes d'ajustement aux données n'est pas aussi bien que le modèle MO-GDGD. À partir des résultats générés en utilisant le critère d'information BIC, la distribution qui décrit le mieux les données du cancer de la vessie est la distribution MO-GDGD proposée, suivie par le modèle GD. MO-GDGD et GD ont également des résultats proches en terme du critère CAIC. D'après les critères de sélection de modèles, nous pouvons conclure que les modèles GD et MO-GDGD sont tous les deux de bons candidats pour modéliser les données actuelles. Cependant, les tests d'ajustement de données sont importants pour décider équitablement du modèle qui correspond le mieux aux données. Pour prendre cette décision cruciale, nous avons eu recours aux tests de Kolmogorov-Smirnov et Cramér-von Mises. Ces deux tests, ainsi que leurs valeurs-p correspondantes montrent clairement que le modèle proposé correspond mieux aux données que GD ce qui implique des valeurs KS et CvM les plus basses et des valeurs-p les plus élevées.

Il est clair à partir des résultats que les données ALS sont en adéquation avec le modèle proposé. Les mesures AIC, BIC et CAIC montrent que MO-GDGD offre de loin le meilleur ajustement. Aussi, le modèle MO-GGD qui donne des résultats très proches de ceux du modèle MO-GDGD. Ceci est tout à fait logique, compte tenu de la très faible valeur du paramètre β et du fait que pour une petite valeur de β MO-GGD converge vers MO-GDGD. Vient ensuite GGD puis MO-GD puis GD puis MGGD. Les valeurs les plus élevées de (beaucoup plus grandes que les précédentes) AIC, BIC et CAIC sont celles de la distribution MGD. Ceci est également confirmé par les tests de qua-

lité d'ajustement KS et CvM. La grande différence entre les valeurs des critères d'information de MO-GDGD par rapport à ses sous modèles montre que les performances du modèle alternatif MO-GDGD sont meilleures que les modèles nuls GD, MGD, GGD et MGGD.

L'adéquation du modèle ajusté MO-GDGD est également prouvée pour les données mammaires. Les valeurs AIC, BIC et CAIC de MO-GDGD surclassent tous ses sous modèles dans puisqu'ils ont les valeurs les plus basses. Les autres sous modèles qui ont des valeurs basses de AIC, BIC et CAIC sont, par ordre, MGGD puis GGD puis GD. Les valeurs les plus élevées d'AIC, BIC et CAIC sont pour le MGD (beaucoup plus grandes que les précédentes). Tous les critères d'information utilisés pour comparer le modèle proposé avec ses sous modèles prouvent clairement que MO-GDGD offre le meilleur ajustement.

Il est bien visible dans les résultats trouvés que les données sur le cancer du sang sont en adéquation avec le modèle proposé. Les mesures AIC, BIC et CAIC montrent que MO-GDGD offre de loin le meilleur ajustement. Vient ensuite GD puis MGGD. Les grandes valeurs (beaucoup plus grandes que les précédentes) AIC, BIC et CAIC sont pour MGD et GGD. La grande différence entre les valeurs des critères d'information de MO-GDGD par rapport à ses sous modèles montre que les performances du modèle alternatif MO-GDGD sont meilleures que les modèles nuls GD, MGD, GGD et MGGD.

Le modèle proposé prend en compte tous les cas possibles de valeurs de fraction de guérison, qui peuvent être en $[0,1]$. Le MO-GDGD a décrit le premier et le dernier ensemble de données où la fraction de guérison est presque nulle. Le modèle décrit également le troisième ensemble de données qui a une très petite fraction de guérison qui est numériquement égale à 0,0110. En outre, il décrit la situation dans laquelle la proportion de survivants est assez significative, comme le deuxième ensemble de données dont la fraction de guérison est égale à 0,1756. Ces résultats confirment le fait que le modèle proposé est suffisamment flexible pour considérer une proportion survivante nulle, petite et significative.

2.4 MO-GDGD en présence de censure

2.4.1 Inférence statistique

Cette partie est consacrée pour l'inférence des paramètres de la distribution MO-GDGD pour des échantillons censurés donnés. La méthode de maximum de vraisemblance est utilisée.

La formulation de la fonction de vraisemblance est fondamentalement basée sur la PDF vue qu'elle est par définition égale à la distribution de la probabilité jointe d'un ensemble donné d'observations. Pourtant, pour le cas traité dans cette partie, les observations sont censurées, c'est-à-dire incomplètes [57]. Par conséquent, les patients dont les observations sont censurées ne contri-

buent pas à la vraisemblance avec leur fonction densité de probabilité $f_{MO}(t,\Theta)$. La fonction de survie $S_{MO}(t,\Theta)$ est plutôt utilisée pour représenter les patients qui n'ont pas vécu l'évènement d'intérêt. Particulièrement, les contributions des patients à la vraisemblance est $f_{MO}(t,\Theta)$ si le patient est mort (i.e. l'heure du décès est disponible) et $S_{MO}(t,\Theta)$ s'il est vivant (i.e. non sensible à l'évènement d'intérêt).

Soit (t_i, δ_i) un ensemble de données de survie, où i est un entier dans $[1,n]$, n est le nombre de sujets dans l'étude, t_i la durée de vie du patient i observée indépendamment, et δ_i l'indicateur de censure qui est égale à 1 si le patient a vécu l'évènement d'intérêt (qui est la mort ou bien la récurrence de la maladie) et 0 si l'information est censurée (patient toujours en vie, a quitté l'étude, a été mort d'une cause qui n'a pas de rapport avec l'étude ...). La durée de vie du patient numéro i est censurée à un temps T_i. Si les T_i sont égales, i.e. le temps de censure est fixé, alors $T_i = T\ \forall i = 1,...,n$. La fonction de vraisemblance a alors la forme de l'équation (2.13) :

$$L(t_i;\Theta) \quad = \quad \prod_{i=1}^{n} f_{MO}(t;\Theta)^{\delta_i} S_{MO}(t;\Theta)^{1-\delta_i}. \tag{2.13}$$

En appliquant le logarithme à l'équation (2.13), on obtient (2.14) :

$$\begin{aligned} l \quad = \quad & \ln L(t_i;\Theta) = \ln(r)\sum_{i=1}^{n}\delta_i + \ln(\alpha)\sum_{i=1}^{n}\delta_i + \ln(\gamma)\sum_{i=1}^{n}\delta_i - \beta\sum_{i=1}^{n}\delta_i t_i + \frac{\alpha}{\beta}\sum_{i=1}^{n}\delta_i(e^{-\beta t_i}-1) \\ & +(\gamma-1)\sum_{i=1}^{n}\delta_i \ln(B(t_i;\Theta)) - 2\sum_{i=1}^{n}\delta_i \ln(C(t_i;\Theta)) + \ln(r)(n-\sum_{i=1}^{n}\delta_i) \\ & +\sum_{i=1}^{n}(1-\delta_i)\ln(D(t_i)), \end{aligned} \tag{2.14}$$

où $B(t;\Theta)$ et $C(t;\Theta)$ ont été définis auparavant dans les équations (2.8) et $D(t;\Theta)$ dans (2.10).

Les dérivées partielles de la fonction logarithme de la fonction de vraisemblance par rapport à r, α, β et γ sont données par :

$$
\begin{aligned}
\frac{\partial l}{\partial r}(t;\Theta) &= \frac{n}{r} - \sum_{i=1}^{n}(1+\delta_i)\mathrm{D}(t_i;\Theta), \\
\frac{\partial l}{\partial \alpha}(t;\Theta) &= \frac{1}{\alpha}\sum_{i=1}^{n}\delta_i + \frac{1}{\beta}\sum_{i=1}^{n}\delta_i(e^{-\beta t_i}-1)[1+(1-\gamma)\tfrac{\mathrm{A}(t_i;\Theta)}{\mathrm{B}(t_i;\Theta)}] \\
&\quad + \tfrac{\gamma(1-r)}{\beta}\sum_{i=1}^{n}(1+\delta_i)(e^{-\beta t_i}-1)\,(\mathrm{B}(t_i;\Theta))^{\gamma-1}\,\tfrac{\mathrm{A}(t_i;\Theta)}{\mathrm{C}(t_i;\Theta)} \\
&\quad + \tfrac{\gamma}{\beta}\sum_{i=1}^{n}(1-\delta_i)(e^{-\beta t_i}-1)\,(\mathrm{B}(t_i;\Theta))^{\gamma-1}\,\tfrac{\mathrm{A}(t_i;\Theta)}{1-(\mathrm{B}(t_i;\Theta))^{\gamma}}, \\
\frac{\partial l}{\partial \beta}(t;\Theta) &= -\sum_{i=1}^{n}\delta_i t_i - \sum_{i=1}^{n}\delta_i \mathrm{E}(t_i;\Theta) + (\gamma-1)\sum_{i=1}^{n}\delta_i\tfrac{\mathrm{A}(t_i;\Theta)\mathrm{E}(t_i;\Theta)}{\mathrm{B}(t_i;\Theta)} \\
&\quad -\gamma(1-r)\sum_{i=1}^{n}(1+\delta_i)\tfrac{(\mathrm{B}(t_i;\Theta))^{\gamma-1}\mathrm{E}(t_i;\Theta)\mathrm{A}(t_i;\Theta)}{\mathrm{C}(t_i;\Theta)} \\
&\quad -\gamma\sum_{i=1}^{n}(1-\delta_i)\tfrac{(\mathrm{B}(t_i;\Theta))^{\gamma-1}\mathrm{E}(t_i;\Theta)\mathrm{A}(t_i;\Theta)}{1-(\mathrm{B}(t_i;\Theta))^{\gamma}}, \\
\frac{\partial l}{\partial \gamma}(t;\Theta) &= \frac{1}{\gamma}\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\delta_i\ln(\mathrm{B}(t_i;\Theta)) - (1-r)\sum_{i=1}^{n}(1+\delta_i)\tfrac{(\mathrm{B}(t_i;\Theta))^{\gamma}\ln(\mathrm{B}(t_i;\Theta))}{\mathrm{C}(t_i;\Theta)} \\
&\quad -\sum_{i=1}^{n}(1-\delta_i)\tfrac{(\mathrm{B}(t_i;\Theta))^{\gamma}\ln(\mathrm{B}(t_i;\Theta))}{1-(\mathrm{B}(t_i;\Theta))^{\gamma}},
\end{aligned}
\tag{2.15}
$$

où les fonctions non linéaires $\mathrm{A}(t;\Theta)$,$\mathrm{E}(t;\Theta)$ et $\mathrm{G}(t;\Theta)$ sont définis auparavant dans les équations (2.8) et (2.10).

Pour trouver les estimations qui maximisent la fonction de vraisemblance, on pose que les équations (2.15) sont égales à zéro et on résout numériquement le système d'équations. La faiblesse majeure de cette méthode est sa sensibilité aux valeurs de départ. Néanmoins, le choix des valeurs de départ appropriées peut contribuer à l'obtention de l'approximation globale optimale.

2.4.2 Simulation

2.4.2.1 Génération des données et résultats

Une étude de simulation exhaustive a été menée pour l'évaluation de l'exactitude des estimateurs du maximum de vraisemblance. Pour générer des échantillons simulés à partir du modèle MO-GDGD, la méthode d'inversion a été utilisée. En se basant sur le fait que la fonction de répartition est une fonction bijective, on peut dériver des données simulées de l'équation (2.16) :

$$
t = -\tfrac{1}{\beta}\ln(\tfrac{\beta}{\alpha}\ln(1-(\tfrac{ur}{1+u(r-1)})^{1/\gamma})+1), \tag{2.16}
$$

où u est une v.a. qui suit une distribution uniforme standards dans $[0,1]$.

Nous élargissons l'étude de simulation à différentes tailles n d'échantillons simulés 30, 100, 200, 500 et 1000. Les résultats de la simulation apparaissent en Annexe **B** sur la page V.

2.4.2.2 Corrélation entre niveaux de censure et taux de guérison

Pour visualiser la relation mutuelle entre le pourcentage d'observations incomplètes et la proportion survivante, nous simulons un échantillon aléatoire de taille 100, dérivé de la CDF de la distribution MO-GDGD. L'estimation des paramètres ainsi que l'estimation du taux de guérison pour chaque niveau de censure (CL) sont présentés dans le tableau 2.21. À partir des résultats, on peut conclure que le niveau de censure et le taux de guérison sont hautement corrélés.

TABLEAU 2.21 – Une étude de la corrélation entre le taux de guérison estimé et le niveau de censure.

CL	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\theta}_{MO}$
0 %	2.1315	0.1866	0.0371	0.4877	0.0068
5 %	2.1434	0.1739	0.0530	0.4903	0.0391
10 %	1.8170	0.1635	0.0631	0.4921	0.0663
15 %	1.7069	0.1592	0.0668	0.4910	0.0767
20 %	1.8132	0.1526	0.0676	0.4903	0.0917
25 %	1.8275	0.1447	0.0698	0.4912	0.1109
30 %	1.8724	0.1338	0.0710	0.4855	0.1349
35 %	1.7899	0.1217	0.0660	0.5146	0.1423
40 %	2.0064	0.1253	0.0673	0.5430	0.1616
45 %	1.7172	0.1304	0.0823	0.5566	0.1896
50 %	1.8404	0.1035	0.0993	0.5011	0.3094
60 %	1.5782	0.0827	0.1390	0.5091	0.4432
70 %	1.6582	0.0839	0.1934	0.5190	0.5440
80 %	2.3781	0.0875	0.2289	0.5243	0.6622
90 %	4.627	0.0816	0.2532	0.5345	0.8222

2.4.3 Application et discussion

Pour illustrer ce qui a précédé, nous consacrons cette partie à l'analyse de données réelles. Afin d'étudier la flexibilité de la distribution MO-GDGD pour décrire différentes tailles de proportions survivantes et différents niveaux de censure, les données médicales ont été soigneusement sélectionnées. Les résultats théoriques sont illustrés à l'aide de deux ensembles de données de survie médicales de deux maladies différentes : le cancer pédiatrique et la sclérose latérale amyotrophique (ALS).

Pour l'analyse des données,

1. les paramètres du modèle MO-GDGD, ainsi que ses sous modèles défectueux, sont estimés en utilisant l'approche du maximum de vraisemblance (voir les tableaux 2.22 et 2.24),
2. l'efficacité du MO-GDGD pour s'adapter aux bases de données actuelles est évaluée à l'aide du rapport de vraisemblance paramétrique (le test statistique Γ_{H_0}) et de sa valeur-p associée (voir les tableaux 2.22 et 2.24),
3. les profils de la fonction log-vraisemblance de chacun des paramètres du modèle sont affichés,
4. les critères d'information (AIC) [58] et Akaike corrigé (AICc) sont utilisés pour prouver l'applicabilité et la supériorité du MO-GDGD par rapport à ses sous modèles défectueux (voir les tableaux 2.23 et 2.24),
5. les courbes de Kaplan-Meier de la fonction de survie, ainsi que son estimation paramétrique et ses bornes inférieure et supérieure, sont présentées.

2.4.3.1 Application au cancer pédiatrique

Base de données du cancer pédiatrique : Il s'agit d'un ensemble de durées de rémission en jours d'un essai randomisé dans le cancer pédiatrique. L'ensemble est composé de 41 observations suivant le même type de traitement [30]. Seules 5 observations sur 41 sont complètes. Les observations complètes sont mises en évidence en gras dans le tableau D.4 (voir Annexe **D**).

Analyse inférentielle : L'estimation de Kaplan-Meier de la fonction de survie, ainsi que la courbe de survie paramétrique, sont affichées dans la figure C.4. Les observations censurées, les limites inférieure et supérieure sont également affichées sur la même figure. La fonction de survie estimée converge vers 0.8724, ce qui représente la proportion guérie comme indiquée sur la figure. La fraction de guérison obtenue correspond à la valeur de la fraction de guérison calculée à l'aide de l'équation (2.3) après avoir remplacé les paramètres $(r, \alpha, \beta, \gamma)$ par leurs estimations $(\hat{r}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})$. La superposition de la courbe de Kaplan-Meier et de la courbe de la fonction de survie estimée indique que le processus d'estimation donne des résultats satisfaisants.

Il est important de noter que, pour cette base sur le cancer pédiatrique, le processus d'estimation donne une valeur du paramètre d'inclinaison de Marshall-Olkin r proche de 1. Ce qui fait que si on pose $r = 1$ dès le début (i.e. on utilise la distribution GDGD au lieu de MO-GDGD, on peut avoir des résultats très proches. Cette remarque est confirmée dans les tableaux 2.22 et 2.23 où les critères d'information sont proches et le rapport de vraisemblance a une petite valeur. Il est aussi à noter que le niveau élevé de censure a donné lieu à un taux important de guérison.

Les fonctions log-vraisemblance des paramètres des modèles r, α, β et γ sont affichées dans la figure C.4 présentée en Annexe en page VII.

Dans le tableau 2.22, le test Γ_{H_0} et sa valeur-p correspondante pour les données sur le cancer

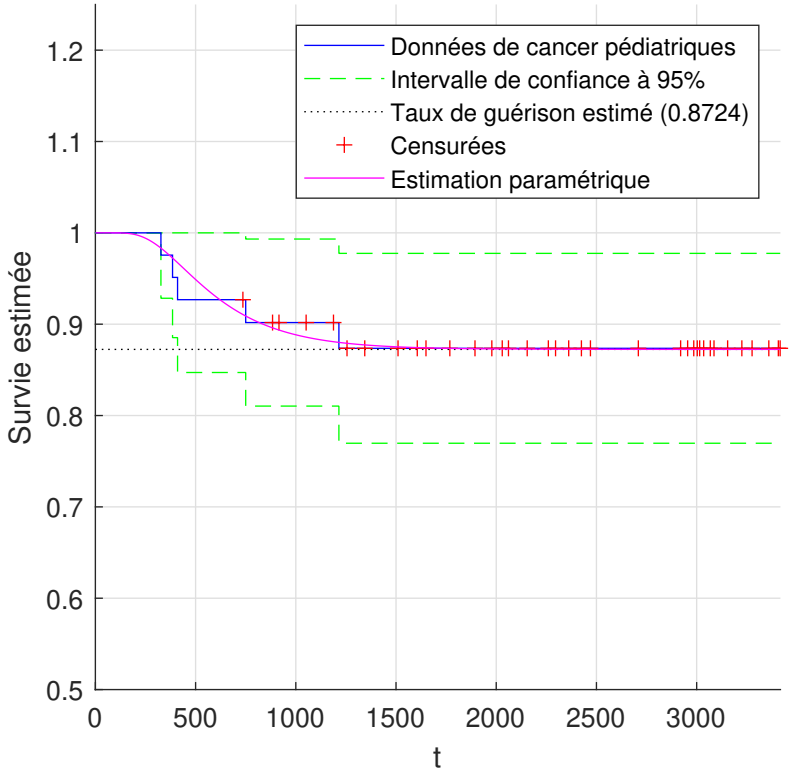


FIGURE 2.5 – Kaplan-Meier et l'estimation paramétrique de la fonction de survie pour les données du cancer pédiatrique.

TABLEAU 2.22 – Les estimations des paramètres inconnus des modèles MO-GDGD et certains de ses sous modèles ainsi que les valeurs-p associées et les valeurs des rapports de vraisemblance Γ_{H_0} pour les données du cancer pédiatriques.

Modèle	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	**Valeur-p**	Γ_{H_0}
$MGD(\alpha,\beta)$	–	0.0002	0.0012	–	0.0446	4.0342
$MO-MGD(r,\alpha,\beta)$	55.2482	0.0035	0.0016	–	0.0724	3.2272
$GDGD(\alpha,\beta,\gamma)$	–	0.0007	0.0018	1.7296	0.1294	2.2994
$MO-GDGD(r,\alpha,\beta,\gamma)$	1.0031	0.0067	0.0035	12.8901	–	–

pédiatrique prouvent l'efficacité du modèle MO-GDGD à s'ajuster aux données actuelles. À partir des valeurs-p, nous rejetons les trois hypothèses nulles en faveur du MO-GDGD à un niveau de signification de 5%. Ceci est également prouvé par les critères AIC et AICc dans la tableau 2.23.

2.4.3.2 Application à la sclérose latérale amyotrophique

Introduction à la base de données : La source de la base de données utilisée dans cette partie a été déjà introduite dans la section 2.3.4.1. Le sous-échantillon exploité se compose de 166 ob-

TABLEAU 2.23 – Les hypothèses nulles H_0, les valeurs de la fonction log-vraisemblance l sous l'hypothèse H_0 et les critères d'information AIC et AICc pour les données du cancer pédiatriques.

Modèle	H_0	l	**AIC**	**AICc**
MGD(α,β)	$r=\gamma=1,\beta>0$	-52.0495	108.0990	105.3917
MO$-$MGD(r,α,β)	$\gamma=1,\beta>0$	-51.6460	109.2920	105.8774
GDGD(α,β,γ)	$r=1,\beta>0$	-51.1821	108.3642	104.9496
MO$-$GDGD(r,α,β,γ)	–	-50.0324	108.0648	104.0404

servations choisies de façon aléatoire de la base PRO-ACT. 50% des observations sont complètes et les 50% restantes sont censurées.

Analyse inférentielle : La figure 2.6 affiche la courbe Kaplan-Meier de la fonction de survie, l'estimation paramétrique de la courbe de survie, les observations censurées, les limites supérieures et inférieures, et le taux de guérison estimé avec l'équation (1.41).

Cependant, cette fois, nous ne tendons pas le temps à l'infini comme nous l'avons fait dans la première application. Nous fixons plutôt le temps sur une valeur maximale. Ensuite, nous calculons la valeur de la fonction de survie à cette valeur fixée (qui est de trois ans dans ce cas). Nous obtenons ainsi l'estimation sur trois ans de la fraction de guérison qui vaut 0.0281.

La superposition de la courbe de Kaplan-Meier des données ALS et de la courbe de la fonction de survie estimée est affichée dans la figure 2.6. Les profils de la fonction log-vraisemblance par rapport à chaque paramètre du modèle MO-GDGD appliqué à la base de données du cancer pédiatrique sont affichés en Annexe **C** en page VII.

TABLEAU 2.24 – Les estimations par la méthode de maximum de vraisemblance des paramètres inconnus du modèle MO-GDGD et certains de ses ces particuliers, les valeurs-p associées et les valeurs du rapport de vraisemblance Γ_{H_0} pour les données ALS.

Modèle	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	**Valeur-p**	Γ_{H_0}
MGD(α,β)	–	0.001	0.0025	–	$<10^{-4}$	147.2118
MO$-$MGD(r,α,β)	8.6772	0.0071	0.0007	–	0.0344	4.4734
GDGD(α,β,γ)	–	0.0011	-0.0024	1.0257	$<10^{-4}$	13.7564
MO$-$GDGD(r,α,β,γ)	1.0141	0.0034	-0.0004	2.2498	–	–

Pour l'ensemble de données actuelles, le test du rapport de vraisemblance Γ_{H_0} et sa valeur-p correspondante montrent que le modèle MO-GDGD surpasse nettement tous ses sous modèles défectueux. La valeur Γ_{H_0} de MGD est très élevée par rapport à MO-MGD et GDGD (voir table 2.24).

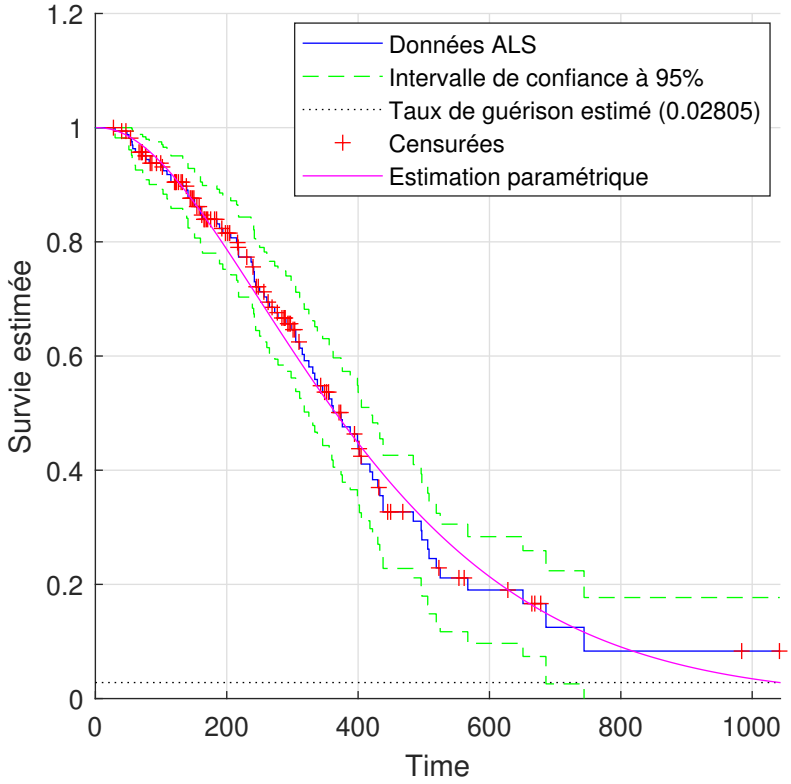


FIGURE 2.6 – Kaplan-Meier et l'estimation paramétrique de la fonction de survie pour les données ALS.

TABLEAU 2.25 – L'hypothèse nulle H_0, les valeurs de la fonction de vraisemblance l sur H_0 et les critères d'information AIC et AICc pout les données ALS.

Modèle	H_0	l	**AIC**	**AICc**
$MGD(\alpha,\beta)$	$r=\gamma=1,\beta>0$	-662.1733	1328.3466	1325.4189
$MO-MGD(r,\alpha,\beta)$	$\gamma=1,\beta>0$	-590.8041	1187.6082	1183.7528
$GDGD(\alpha,\beta,\gamma)$	$r=1,\beta>0$	-595.4456	1196.8912	1193.0358
$MO-GDGD(r,\alpha,\beta,\gamma)$	–	-588.5674	1185.1348	1180.3785

Les valeurs de tous les tests de qualité d'ajustement appliqués prouvent que la MGD est le modèle le moins efficace pour ajuster les données actuelles. En terme d'AIC et d'AICc, la meilleure distribution pour ajuster les données PRO-ACT est MO-GDGD suivi de MO-MGD (voir table 2.24). Quant aux modèles MO-MGD et GDGD, ils ont tous deux des valeurs de critères d'information plus faibles et surpassent considérablement le MGD. Néanmoins, aucun des sous modèles considérés n'est aussi efficace que le MO-GDGD. L'adéquation du modèle ajusté MO-GDGD est ensuite prouvée pour les données ALS.

2.5 Conclusion

Dans ce chapitre, une nouvelle généralisation de la distribution de Gompertz est proposée. Contrairement aux modèles mixtes, cette distribution modélise la possibilité que les données contiennent des survivants à long terme, et cela, sans ajouter de paramètre supplémentaire. L'accent est mis sur la théorie des distributions défectueuses. Une étude exhaustive des performances et propriétés de la distribution proposée est menée. Des cas particuliers du modèle sont générés et une étude de simulation a été établie. Des estimateurs du maximum de vraisemblance des paramètres inconnus sont obtenus. Soulignant l'une des valeurs ajoutées de cette contribution, la distribution proposée est très flexible pour s'adapter à des données réelles complètes et censurées. Et si on lui ajoute des covariables ?

Chapitre 3

Analyse de l'effet des variables explicatives sur la survie

3.1 Introduction

La majorité des données de survie sont affectées par des variables explicatives. Ce chapitre porte sur deux axes essentiels. Le premier axe concerne une étude de modélisation où nous développons une famille de modèles de régression pour l'analyse des données de survie. Nous prenons en compte l'avantage du concept des distributions défectueuses. Les modèles proposés sont basés sur la distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin présentée dans le chapitre précédent et appliqués aux patients traités pour la sclérose latérale amyotrophique. Le deuxième axe concerne une investigation sur le meilleur modèle qui décrit la survie des patients atteints de coronavirus. Une étude de l'effet des facteurs de risque est aussi menée.

3.2 Application à la sclérose latérale amyotrophique

Dans ce qui suit, on introduit les modèles de régression proposés ainsi que l'étude inférentielle utilisant la méthode du maximum de vraisemblance en présence de covariables et d'un phénomène de censure. Des variables explicatives sont incorporées dans le modèle suivant l'hypothèse des risques proportionnels afin d'évaluer l'effet des facteurs de risque sur la survie globale. Des méthodes paramétriques, semi-paramétriques et non paramétriques sont appliquées pour l'analyse de survie des patients traités pour la sclérose latérale amyotrophique. Des résultats intéressants sur l'effet de l'utilisation du riluzole et d'autres traitements sur la survie des patients sont obtenus [59].

3.2.1 Méthodologie

3.2.1.1 Modèles de régression propre MOEGG

Pour examiner l'effet des variables indépendantes d'intérêt sur la variable dépendante, nous procédons à une approche de régression. Soit $x_i = (1, x_{i1}, \ldots, x_{ik})^{\mathrm{T}}$ un vecteur de covariables d'un ensemble de données associé à la $i^{ème}$ variable de réponse t_i, où k est le nombre de variables explicatives considérées, et $i = 1, \ldots, \mathrm{N}$, où N est la taille de l'échantillon. En se basant sur la distribution MOEGG, les covariables sont incorporées via les hypothèses des risques proportionnels (3.1) :

$$h_{\mathrm{MO}_{reg}}(t_i; \Theta, b, x_i) = h_{\mathrm{MO}}(t_i; \Theta)\, \mathrm{e}^{x_i^{\mathrm{T}} b}, \tag{3.1}$$

où $b = (b_0, b_1, \ldots, b_k)^{\mathrm{T}}$ est un vecteur de coefficients de régression à estimer et h_{MO} est la fonction du taux de risque du modèle MOEGG précédemment donnée dans l'équation (2.6). La fonction de survie du modèle de régression MOEGG est donnée par (3.2) :

$$\begin{aligned} \mathrm{S}_{\mathrm{MO}_{reg}}(t_i; \Theta, b, x_i) &= \left(\mathrm{S}_{\mathrm{MO}}(t_i; r, \alpha, \beta, \gamma)\right)^{\mathrm{e}^{x_i^{\mathrm{T}} b}} \\ &= \left(\frac{r(1-(1-\mathrm{e}^{\frac{\alpha}{\beta}(\mathrm{e}^{-\beta t}-1)})^{\gamma})}{1-(1-r)(1-(1-\mathrm{e}^{\frac{\alpha}{\beta}(\mathrm{e}^{-\beta t}-1)})^{\gamma})} \right)^{\mathrm{e}^{x_i^{\mathrm{T}} b}}. \end{aligned} \tag{3.2}$$

La fonction densité de probabilité du modèle de régression MOEGG est donnée par (3.3) :

$$f_{\mathrm{MO}_{reg}}(t_i; \Theta, b, x_i) = \mathrm{e}^{x_i^{\mathrm{T}} b} [\mathrm{S}_{\mathrm{MO}}(t_i; \Theta)]^{\mathrm{e}^{x_i^{\mathrm{T}} b}-1} f_{\mathrm{MO}}(t_i; \Theta), \tag{3.3}$$

Où S_{MO} et f_{MO} sont, respectivement, la fonction de survie et la fonction densité de probabilité du modèle MOEGG données précédemment dans les équations (2.1) et (2.4).

3.2.1.2 Modèle de régression défectueux MOEGG

Les informations sur les covariables peuvent être ajoutées au modèle défectueux pour évaluer l'effet des variables explicatives sur le taux de guérison ou de survie. Le premier article à inclure des informations sur les covariables à l'idée de modélisation défectueuse est Gieser *et al.* [45]. De là, certains modèles ont été développés dans le même contexte tels que [60, 61, 62, 63, 64]

Nous proposons dans ce chapitre une modification de la distribution MOEGG. Cette modification consiste à considérer que le paramètre d'échelle β de la fonction de survie (3.2) peut être une valeur réelle positive et négative non nulle (i.e. $\beta \in \mathrm{R}^*$), le modèle de régression MOEGG proposé est alors un modèle partiellement défectueux permettant de prendre en compte la fraction de guérison ou de survie. Notez qu'un modèle partiellement défectueux prend en considération les données avec (quand $\beta > 0$) ou sans ($\beta < 0$) une proportion survivante (ou un taux guérison). Cela améliorerait la flexibilité du modèle pour décrire les données de durée de vie remarquablement mieux que les distributions propres.

Il s'ensuit que si MOEGG est utilisée comme étant un modèle défectueux (i.e. si la procédure d'estimation présente une valeur positive du paramètre β), alors le taux de guérison $\theta_{MO_{reg}}$ qui appartient à $[0,1]$ est donné par (3.4) :

$$\theta_{MO_{reg}} = \lim_{t\to\infty} S_{MO_{reg}}(t) = \left(\frac{r\theta}{1-(1-r)\theta}\right)^{e^{x_i^T b}}, \tag{3.4}$$

où $\theta = 1-(1-e^{-\frac{\alpha}{\beta}})^{\gamma}$ dans $[0,1]$. Notons que θ représente la proportion d'éléments survivant dans la population obtenue en calculant la limite, lorsque t tend vers l'infini, de la fonction de survie de la distribution de Gompertz généralisée modifiée proposée par [34]. Il convient de mentionner que si $e^{x_i^T b} < 1$ alors le taux de guérison s'élève et si $e^{x_i^T b} > 1$ alors le taux diminue, [65]. En présence de données de durée de vie avec une proportion de survivants, et dans le cas particulier où $r = 1$, on obtient le modèle de régression de Gompertz généralisé proposé par Borges [46].

3.2.1.3 Inférence statistique

Estimation du maximum de vraisemblance : Les variables explicatives sont incluses dans le modèle proposé à travers l'hypothèse des modèles à risques proportionnels. Nous supposerons que tous les vecteurs de covariables ont la même taille que le vecteur de durée de vie observé t_i. On pose N la taille des données et x_j un vecteur de covariables $\in \mathrm{R}^{\mathrm{N}}$ pour $j = 1,\ldots,k$, où k est le nombre de covariables considérées.

La durée de vie observée t_i peut être soumise à une censure à droite. Soit δ_i l'indicateur de censure : si l'observation est censurée alors $\delta_i = 0$, et si elle est complètement observée alors $\delta_i = 1$.

Il existe d'autres moyens d'incorporer des covariables dans le modèle, telles que les familles paramétriques, les modèles à temps de vie accéléré et les modèles de chances proportionnelles. Néanmoins, la sensibilité de la méthode d'incorporation des covariables ne sera pas abordée ici. La procédure inférentielle est basée sur l'approche du maximum de vraisemblance. Considérons l'échantillon aléatoire $t_1, t_2, \ldots, t_{\mathrm{N}}$ qui se compose de N observations indépendantes et identiquement distribuées suivant le modèle MOEGG ayant comme fonction de survie l'équation (3.2). Les estimations du maximum de vraisemblance (MLE) des paramètres $\Theta_1 = (\Theta, b)$ du modèle peuvent être obtenus à partir de la fonction de vraisemblance notée par $\mathrm{L}(t_i;\Theta_1;x_i)$ dans (3.5) :

$$\mathrm{L}(t_i;\Theta_1;x_i) = \prod_{i=1}^{\mathrm{N}} f_{MO_{reg}}(t_i;\Theta_1;x_i)^{\delta_i} S_{MO_{reg}}(t_i;\Theta_1;x_i)^{1-\delta_i}. \tag{3.5}$$

En utilisant l'expression de $S_{MO_{reg}}$ et $f_{MO_{reg}}$ données dans les équations (3.2) et (3.3), la fonc-

tion log-vraisemblance l peut être exprimée par (3.6) :

$$l = \ln \mathrm{L}(t_i;\Theta_1;x_i) = \sum_{i=1}^{\mathrm{N}} \delta_i x_i^{\mathrm{T}} b + \delta_i \ln(f_{\mathrm{MO}}(t_i;\Theta_1;x_i)) + \sum_{i=1}^{\mathrm{N}} (\mathrm{e}^{x_i^{\mathrm{T}} b} - \delta_i) \ln(\mathrm{S}_{\mathrm{MO}}(t_i;\Theta_1;x_i)). \tag{3.6}$$

La fonction log-vraisemblance est ensuite différencié par rapport aux paramètres fondamentaux Θ du modèle ainsi que les coefficients de régression b. On fixe ensuite les dérivées à zéro et on les résout en utilisant une procédure numérique itérative appelée l'algorithme de Newton-Raphson dans le but de trouver les estimations du maximum de vraisemblance.

Les limites de confiance asymptotiques : Il est impossible de dériver une distribution exacte des estimateurs du maximum de vraisemblance vu que les paramètres du modèle de régression MOEGG Θ_1 n'ont pas de forme fermée. C'est la raison pour laquelle il est important d'étudier les limites de confiance asymptotiques (CI) des paramètres du modèle. Pour quantifier l'information qu'une variable aléatoire T_i pourrait posséder sur les paramètres inconnus Θ_1 du modèle de régression MOEGG qui modélise les observations t_i, soit $\mathrm{I}(\hat{\Theta}_1$ la matrice d'information de Fisher à Θ_1.

On obtient $\mathrm{I}(\hat{\Theta}_1)$ en calculant la dérivée partielle seconde de la fonction log-vraisemblance (3.6) par rapport à chacun des paramètres du modèle. $\mathrm{I}(\hat{\Theta}_1)$ est une matrice carrée de taille $p = 4 + k$, où k est le nombre de variables explicatives considérées et 4 est le nombre de paramètres fondamentaux du modèle MOEGG.

Comme la matrice d'information de Fisher est symétrique, le nombre de dérivées partielles à calculer est égal à $n_d = \sum_{i=0}^{p-1} (p-i)$.

La matrice de variance-covariance asymptotique observée $\mathrm{V}(\hat{\Theta}_1)$ est égale à $\mathrm{I}^{-1}(\hat{\Theta}_1)$. En outre, les deux limites de confiance de $(1-\delta)100\%$ de Θ_1 sont calculées comme suit :

$\left(\hat{\Theta}_{1_i} \pm \mathrm{Z}_{\frac{\delta}{2}} \sqrt{Var(\hat{\Theta}_{1_i})}\right)$, où $i = 1,\ldots,p$ et $\mathrm{Z}_{\frac{\delta}{2}}$ est le quantile d'ordre $\frac{\delta}{2}$ de la distribution normale standard.

3.2.2 Simulation

Cette partie est réservée pour une étude de simulation des données de durées de vie et évaluer la justesse des estimateurs. La méthode inverse a été utilisée en se basant sur la fonction de répartition pour générer des échantillons aléatoires simulés à partir du modèle défectueux de régression MOEGG.

La simulation est étendue à différentes tailles d'échantillons $n = 30, 100$ et 200 avec $\Theta_1 =$

$(2, 0.7, 1.5, 4, 1, 1.5, -1, -2, 3)$ ce qui est un taux de guérison qui vaut 0.7970 . Les moyennes des estimations du maximum de vraisemblance (AMLE), les erreurs quadratiques moyennes (MSE), et les biais des estimations du maximum de vraisemblance pour les données simulées sont présentées dans le tableau 3.1. Plus la taille de l'échantillon augmente, plus les valeurs de MSE et biais approchent de zéro. L'étude de simulation montre que le modèle proposé fournit une bonne performance surtout avec les données qui ont des taux de guérison élevés.

3.2.3 Base de données PRO-ACT

À titre d'illustration, le modèle de régression MOEGG est appliquée à une base de données sur une maladie neurologique appelée Amyotrophic Lateral Sclerosis (ALS). La base de données a été introduite dans 2.3.4.1. Cette base de 8600 patients atteints de ALS a été sous-échantillonnée en fonction de la disponibilité de toutes les informations nécessaires sur chaque sujet, ce qui nous conduit à travailler avec 1715 patients. L'ensemble des données utilisé est partitionné longitudinalement selon les catégories démographiques et de traitement. En d'autres termes, des informations sur l'âge (≤ 50 ans ou > 50 ans), le sexe (masculin ou féminin), le traitement utilisé (actif ou placebo) et l'utilisation du riluzole (oui ou non) sont connues pour tous les patients. Le riluzole est un médicament ayant une activité antagoniste de la glutamine, qui retarde le début de la dépendance à la ventilation non invasive chez certaines personnes. Il a été démontré que le riluzole prolonge la survie globale des patients ALS de deux à trois mois avec une probabilité que 9% des patients survivent pendant une année supplémentaire [66].

Le résumé en cinq nombres des durées de vie observées en jours pour le sous-échantillon de PRO-ACT est présenté dans le tableau 3.11.

Une première étude pivot sur les caractéristiques des patients avec des données PRO-ACT est résumée dans la figure 3.1. Le nombre de sujets pour chaque stratification ainsi que son pourcentage par rapport au total et le niveau de censure sont représentés.

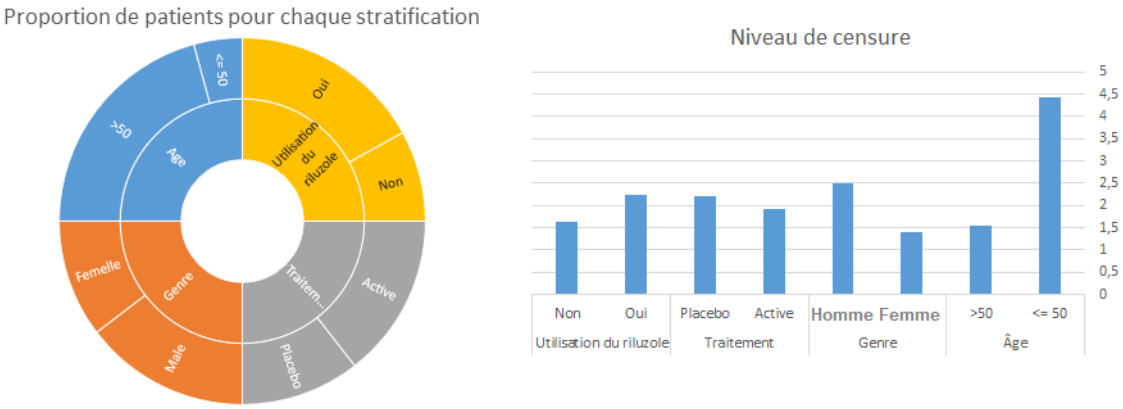


FIGURE 3.1 – Résumé sur les données des patients et des informations sur la censure.

Pour visualiser l'effet de chaque covariable sur la survie globale des patients ALS, la survie

médiane pour chaque stratification des données est présentée dans la figure 3.2. Cinq conditions sont considérées : sans tenir compte du traitement utilisé, les patients qui ont utilisé le riluzole, les patients qui n'ont pas utilisé le riluzole, les patients qui ont utilisé un traitement actif et les patients qui ont utilisé un traitement placebo. Les résultats présentés dans la figure 3.2 sont ensuite discutés dans la section 3.2.4.

3.2.4 Application du modèle de régression MOEGG

3.2.4.1 Hypothèses inférentielles

Pour le sous-échantillon ALS utilisé, nous considérons quatre covariables (c-à-d. K = 4) et supposons que tous les vecteurs de covariables ont la même taille que le vecteur de durée de vie observé (N = 1715). Les variables catégorielles et affectations suivantes sont prises en compte pour chaque patient :

- t_i est la durée de vie du patient observé (en jours).
- x_{i1} est l'âge du patient, réparti en deux sous-groupes : inférieur ou égal à 50 ans ($x_{i1} = 0$) et supérieur à 50 ans ($x_{i1} = 1$).
- x_{i2} est le sexe du patient, classé comme homme ($x_{i2} = 0$) versus femme ($x_{i2} = 1$).
- x_{i3} est le traitement utilisé par le patient, classé comme placebo ($x_{i3} = 0$) et actif ($x_{i3} = 1$).
- x_{i4} est pour l'utilisation du riluzole, non ($x_{i4} = 0$) et oui ($x_{i4} = 1$).

Dans un premier temps, nous modélisons indépendamment l'effet de chaque covariable. Ensuite, nous incluons tous les facteurs de risque en considérant un modèle complet avec toutes les covariables mentionnées précédemment.

3.2.4.2 Effet sur chaque covariable indépendamment des autres

Nous formulons dans cette partie quatre modèles différents pour décrire indépendamment l'effet de chaque facteur de risque. Considérant que $x_i^{\mathrm{T}} b = b_{0j} + b_{1j} x_{ij}$ pour $j = 1, \ldots, 4$, on ajuste le modèle 1 (c-à-d $j = 1$) en ne prenant en compte que l'effet d'âge (x_{i1}), modèle 2 ($j = 2$) ne considérant que l'effet du sexe du patient (x_{i2}), modèle 3 ($j = 3$) avec seulement l'effet de le traitement utilisé (x_{i3}), modèle 4 ($j = 4$) avec uniquement l'effet de l'utilisation du riluzole (x_{i4}).

Les estimations du maximum de vraisemblance pour les modèles de régression ajustés (1-4) pour l'ensemble de données ALS ainsi que les intervalles de confiance à 95%, et les erreurs standard pour chaque estimation sont présentées dans le tableau 3.2. Le taux de hasard des coefficients de régression et les intervalles de confiance correspondants sont affichés dans le tableau 3.3.

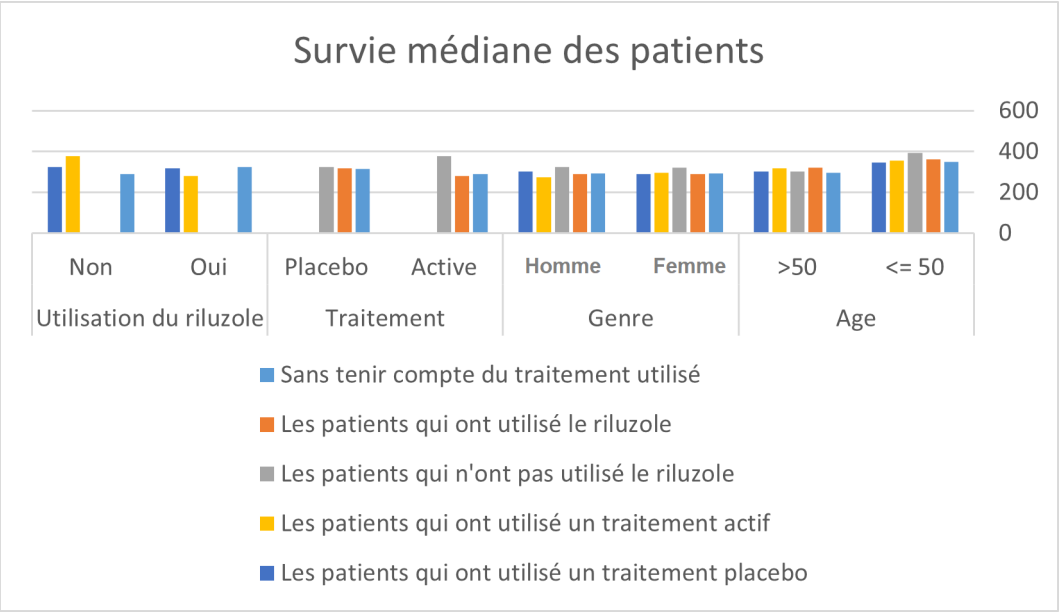


FIGURE 3.2 – Survie médiane pour chaque stratification.

À titre d'illustration, nous avons également ajusté (a) des modèles de régression paramétrique MOEGG univariés et (b) des estimateurs de Kaplan-Meier, en tant que technique non paramétrique de modélisation des fonctions de survie, pour chaque strate. Dans la figure 3.3, on trace la fonction de survie estimée dérivée des modèles de régression univariés MOEGG, y compris comme covariable : l'âge dans la sous-figure en haut à gauche, le sexe dans la sous-figure en haut à droite, le type de traitement dans la sous-figure en bas à gauche et l'utilisation du riluzole dans la sous-figure en bas à droite. Les estimateurs de Kaplan-Meier ajustés pour chacune des strates considérées sont également affichés dans la figure 3.3.

Beaucoup d'informations sont cachées derrière les résultats obtenus dans la figure 3.2, le tableau 3.2, le tableau 3.3 et la figure 3.3.

Dans le modèle de régression ajusté 1, on note que l'âge du sujet a un effet significatif sur sa survie globale puisque l'intervalle de confiance à 95% du coefficient de régression $\hat{b}_{11}$ ne contient pas de valeur nulle. De plus, l'intervalle de confiance à 95% du taux de risque ne contient pas 1.

Ceci est également confirmé, premièrement, par les courbes de survie de la figure 3.3 (courbes en haut à gauche), où les courbes en bleu qui représentent l'estimation de la survie des patients âgés de 50 ans ou moins sont décalées au-dessus les courbes en rouge qui représentent la survie des patients de plus de 50 ans. Deuxièmement, à partir de la figure 3.2, il est évident que la survie médiane des patients âgés de 50 ans ou moins est significativement plus élevée que leur complémentaire dans toutes les stratifications mentionnées précédemment. Mieux encore, avoir moins de 50 ans peut prolonger la survie globale jusqu'à 92 jours. On peut conclure que l'âge influence remarquablement la survie des patients et que le risque augmente avec l'âge.

Le modèle 2, montre que l'estimation de survie est presque la même que le patient soit de

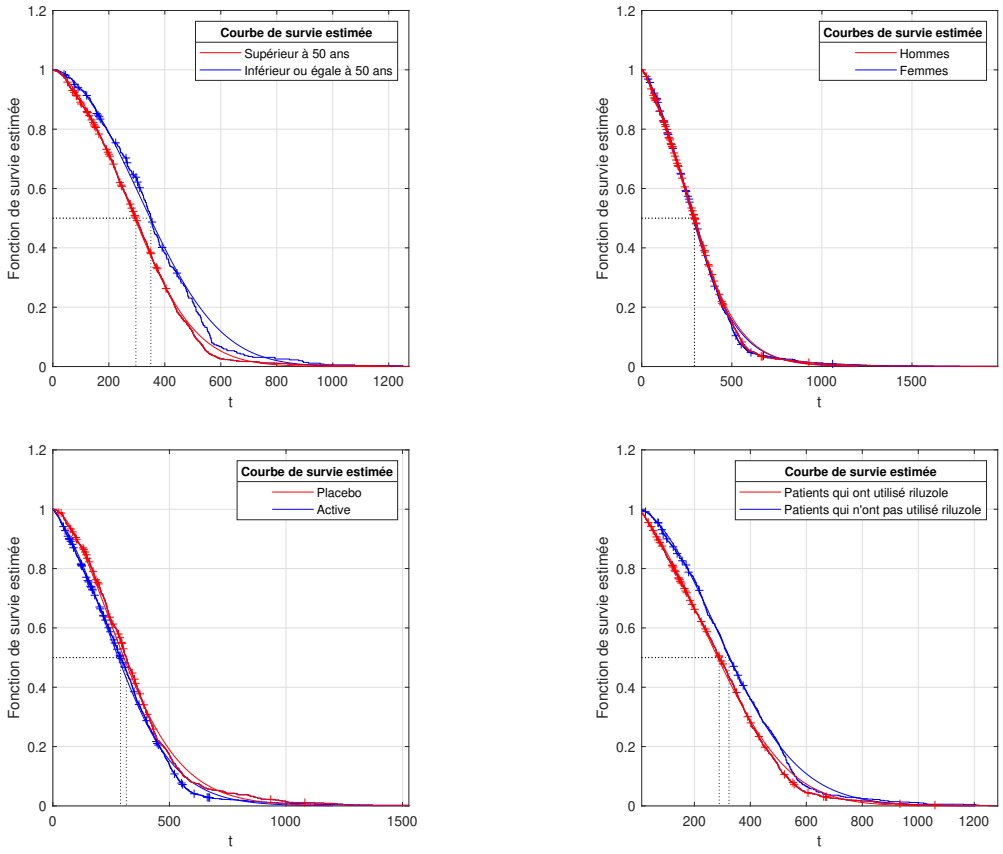


FIGURE 3.3 – Modèles MOEGG univariés et estimateurs de Kaplan-Meier ajustés pour chaque strate.

sexe masculin ou féminin, puisque la valeur du coefficient de régression $\hat{b}_{12}$ est proche de zéro et son intervalle de confiance à 95% comprend valeur zéro.

L'intervalle de confiance du HR contient la valeur 1. Ce fait est également un indicateur que le sexe du patient n'a pas une importance dans ce contexte. Ceci est également justifié par les courbes de survie de la figure 3.3 (courbes en haut à droite), où les courbes en rouge, représentant la population masculine, et les courbes en bleu, représentant la population féminine, sont superposées. La figure 3.2 confirme également ce résultat. Les colonnes femmes et hommes de la figure 3.2 présentent des valeurs médianes de survie qui sont très proches dans chacun des sous-groupes considérés.

Par conséquent, nous concluons que le sexe du patient ALS n'a pas d'incidence sur la durée de survie.

Le modèle 3 suggère que, que le patient ait utilisé un traitement actif ou un traitement placebo, l'effet sur l'estimation de la survie n'est pas significatif puisque, premièrement, le coefficient de régression $\hat{b}_{13}$ est faible et son intervalle de confiance à 95% inclut une valeur nulle et, deuxièmement, l'intervalle de confiance à 95% du HR ne contient pas la valeur 1.

Sur la figure 3.3, (courbes en bas à gauche), l'estimation de la survie de Kaplan-Meier des patients ayant utilisé un traitement placebo (courbe en rouge) n'est que légèrement supérieure à l'estimation de la survie de Kaplan-Meier des patients ayant utilisé un traitement actif (courbe en bleu). En revanche, la figure 3.2 donne plus de détails sur l'effet du traitement en fonction des stratifications considérées.

Quant au modèle 4, notre étude a révélé que le coefficient de régression $\hat{b}_{14}$ est significatif et que son intervalle de confiance à 95% n'inclut pas de valeur nulle.

La figure 3.3 montre que les courbes d'estimation de survie (en bleu) des patients qui n'ont pas utilisé de riluzole sont décalées au-dessus des courbes de survie (en rouge) des patients qui l'ont fait. Cela signifie que la survie médiane de ceux qui ont consommé du riluzole est inférieure à celle de ceux qui ne l'ont pas fait, ce qui est également présenté dans la figure 3.2. Cependant, il a été signalé que la consommation de riluzole peut prolonger la survie de deux à trois mois [66, 67].

Cela signifie que les résultats de cette étude ne sont pas en harmonie avec la littérature. Nous mentionnons dans la figure 3.1 que, pour l'échantillon en cours d'étude, 67.93% des patients sont des utilisateurs de riluzole. Il convient de noter que, dans l'ensemble de données PRO-ACT, les patients sont classés, selon un enregistrement binaire : «oui» pour ceux qui ont utilisé riluzole et «non» pour ceux qui n'ont pas utilisé riluzole. Cela implique que même si l'utilisation du riluzole est incohérente ou incomplète, le patient est classé comme utilisateur du riluzole [68]. La consommation de médicaments pendant une durée limitée ou inexacte ne modifie probablement pas la progression de la maladie. De plus, Fournier et Glass [69] ont déclaré que la population ALS dans la base de données PRO-ACT pouvait être des patients avec un meilleur état de santé que ceux de la population ALS réelle. D'où, ils sont moins dépendants du médicament utilisé. Ce qui explique les résultats obtenus.

Il est intéressant aussi de noter que pour le sous-ensemble de données en cours d'étude, la majorité des modèles proposés ne donnent pas de distribution défectueuse, sauf pour le modèle 1 où la procédure d'estimation a produit une valeur positive du paramètre $\hat{\beta}$. Comme mentionné ci-dessus, les limites de confiance à 95% sont calculées à l'aide de la matrice d'information de Fisher. Pour les modèles $1-4$, la taille de la matrice de Fisher est de (6×6) et le nombre de dérivées secondes calculées est $n_d = 21$.

3.2.4.3 Considération de tous les facteurs de risque

Nous formulons également un modèle qui décrit les effets de toutes les covariables considérées simultanément [70]. Fondamentalement, nous incorporons $x_i^{\mathrm{T}} b = b_{05} + \sum_{j=1}^{4} b_{j5} x_{ij}$ dans le modèle à risques proportionnels. Ensuite, nous trouvons la fonction de densité de probabilité et la fonction de survie correspondantes pour calculer la fonction log-vraisemblance comme dans l'équation (3.6). Nous procédons ensuite de la manière traditionnelle expliquée dans la section

1.4.1 pour trouver les estimations du maximum de vraisemblance.

Les limites de confiance à 95% sont calculées à l'aide de la matrice d'information de Fisher. Pour le modèle 5, la taille de la matrice de Fisher est (9×9) et le nombre de dérivées secondes calculées est $n_d = 45$.

Le tableau 3.4 présente les estimations du maximum de vraisemblance, les intervalles de confiance et l'erreur standard pour chaque estimation. Le taux de hasard et l'intervalle de confiance correspondant des coefficients de régression sont affichés dans le tableau 3.5.

L'estimation des paramètres du modèle 5 donne une valeur positive du paramètre β. Cela signifie que le modèle de régression MOEGG 5 est défectueux, ce qui implique l'existence d'une fraction de survie dans la population. Le taux de survie est calculé à l'aide de l'équation (3.4). En utilisant les estimations obtenues et les vecteurs de covariables, la valeur moyenne du taux de survie sur toutes les observations est donnée par 0.0003. Cette valeur est bien comprise dans l'intervalle $[0,1]$. D'après la valeur du taux de survie obtenue, il convient de noter que la proportion de sujets dans la base de données PRO-ACT qui sont survécu ALS est estimée faible.

Parmi les covariables considérées dans le modèle 5, la covariable ayant l'effet le plus significatif est l'âge du patient puisque le coefficient de régression correspondant $\hat{b}_{15}$ a une valeur de MLE relativement élevée. En tenant compte de tous les facteurs de risque dans un seul modèle, certains faits sont confirmés, comme par exemple le fait que le sexe du patient a peu d'importance dans la prédiction de la survie.

Certains autres faits sont mis en évidence, comme l'importance de l'effet du traitement sur la survie des patients. Même si le modèle 3, qui modélise uniquement l'effet du traitement, n'a pas montré l'importance de l'effet placebo/actif sur la survie des patients, le modèle 5 affirme ce fait comme l'intervalle de confiance n'inclut pas la valeur zéro. Ces faits prouvent que l'ajustement d'un modèle multivarié est plus significatif que l'ajustement de différents modèles chacun décrivant l'effet d'une seul covariable. Donc, pour analyser l'effet des facteurs de risque, on peut se contenter d'un modèle multivarié.

3.2.4.4 Considération de l'effet d'interactions entre facteurs de risque

On parle de l'interaction lorsque les variables explicatives affectent les changements de la variable de sortie en fonction de la (des) valeur(s) d'une ou plusieurs autres variables explicatives. Dans un modèle de régression, l'effet d'interaction est exprimé comme le produit de deux ou plusieurs variables explicatives [34]. Dans cette partie, nous limitons notre étude à l'interaction bidirectionnelle entre les variables explicatives.

Nous formulons le modèle 6 pour voir si l'effet de l'utilisation du riluzole sur la vie des patients change ou pas à des âges, des sexes et des catégories de traitement différents. Pour ce

faire, nous suggérons d'incorporer $x_i^{\mathrm{T}} b = b_{06} + \sum_{j=1}^{3} b_{j6} x_{ij} x_{i4}$ dans le modèle de régression à risques proportionnels. Les résultats de l'estimation du maximum de vraisemblance, les intervalles de confiance à 95% et les erreurs standard du modèle 6 tenant compte des effets d'interaction entre l'utilisation du riluzole et d'autres variables indépendantes sont présentés dans le tableau 3.4.

Le modèle 6 suggère que l'effet de l'utilisation du riluzole sur la survie des patients change avec le changement de l'âge des patients. L'intervalle de confiance à 95% du coefficient de régression $\hat{b}_{16}$ ne comprend pas de valeur nulle. Cette observation indique que l'effet d'interaction est significatif. Ce fait est également confirmé dans la figure 3.4. Les deux premières sous-figures en haut représentent la fonction de survie empirique et la fonction de survie estimée en fonction de l'âge des patients ayant utilisé le riluzole (courbes en haut à gauche) et de ceux qui n'ont pas utilisé le riluzole (courbes en haut à droite). Les sous-figures ne sont pas similaires. La différence entre les courbes en bleu représentant les patients dont l'âge est inférieur ou égale à 50 ans et les courbes en rouge représentant les patients dont l'âge est supérieur à 50 ans est plus importante pour les patients qui n'ont pas utilisé de riluzole que pour ceux qui l'ont utilisé.

Cependant, les petites valeurs des coefficients de régression $\hat{b}_{26}$ dans le tableau 3.6 indiquent qu'il n'y a pas d'interaction importante entre le sexe, le traitement et l'utilisation du riluzole. Ceci n'est pas seulement affirmé par les courbes de deuxième colonne de la figure 3.4, mais également par les lignes en pointillées indiquant la survie médiane. Nous remarquons également dans les courbes du bas de la figure 3.4 que l'estimation non paramétrique par les courbes de Kaplan-Meier suggère que l'effet de l'utilisation du riluzole sur la survie des patients n'est pas si différent pour les utilisateurs de traitement actifs par rapport aux utilisateurs de placebo. Cependant, le tableau 3.2 fournit quelques détails sur les interactions traitement-riluzole. La survie médiane des populations utilisateurs du traitement actif et utilisateurs du traitement placebo varie en fonction de la stratification associée. Le tableau indique que, pour les patients ayant utilisé du riluzole, la survie médiane des patients traités par placebo est supérieure à celle de ceux qui ont utilisé un traitement actif. Par contre, pour les patients qui n'ont pas utilisé de riluzole, la survie médiane de ceux qui sont traités par placebo est inférieure à celle de ceux qui sont traités par un traitement actif. On peut conclure à l'importance d'envisager l'utilisation du riluzole pour juger l'effet des traitements actifs et placebo.

Nous formulons le modèle 7 pour voir si l'effet du type de traitement sur la vie du patient qui l'a utilisé dépend ou pas de son âge et son genre. Soit $x_i^{\mathrm{T}} b = b_{07} + b_{17} x_{i1} x_{i3} + b_{27} x_{i2} x_{i3}$. Nous incorporons $x_i^{\mathrm{T}} b$ dans le modèle à risques proportionnels proposé. Les résultats de l'inférence statistique sont disponibles dans le tableau 3.7.

Le modèle 7 suggère que l'effet du placebo ou des traitements actifs utilisés par le patient sur sa survie est quelque peu différent selon les catégories d'âge. Le modèle 7 suggère également que l'effet du traitement utilisé sur la survie des patients est négligeable selon les catégories de sexe. Ces phénomènes peuvent également être observés sur la figure 3.5. La matrice d'information

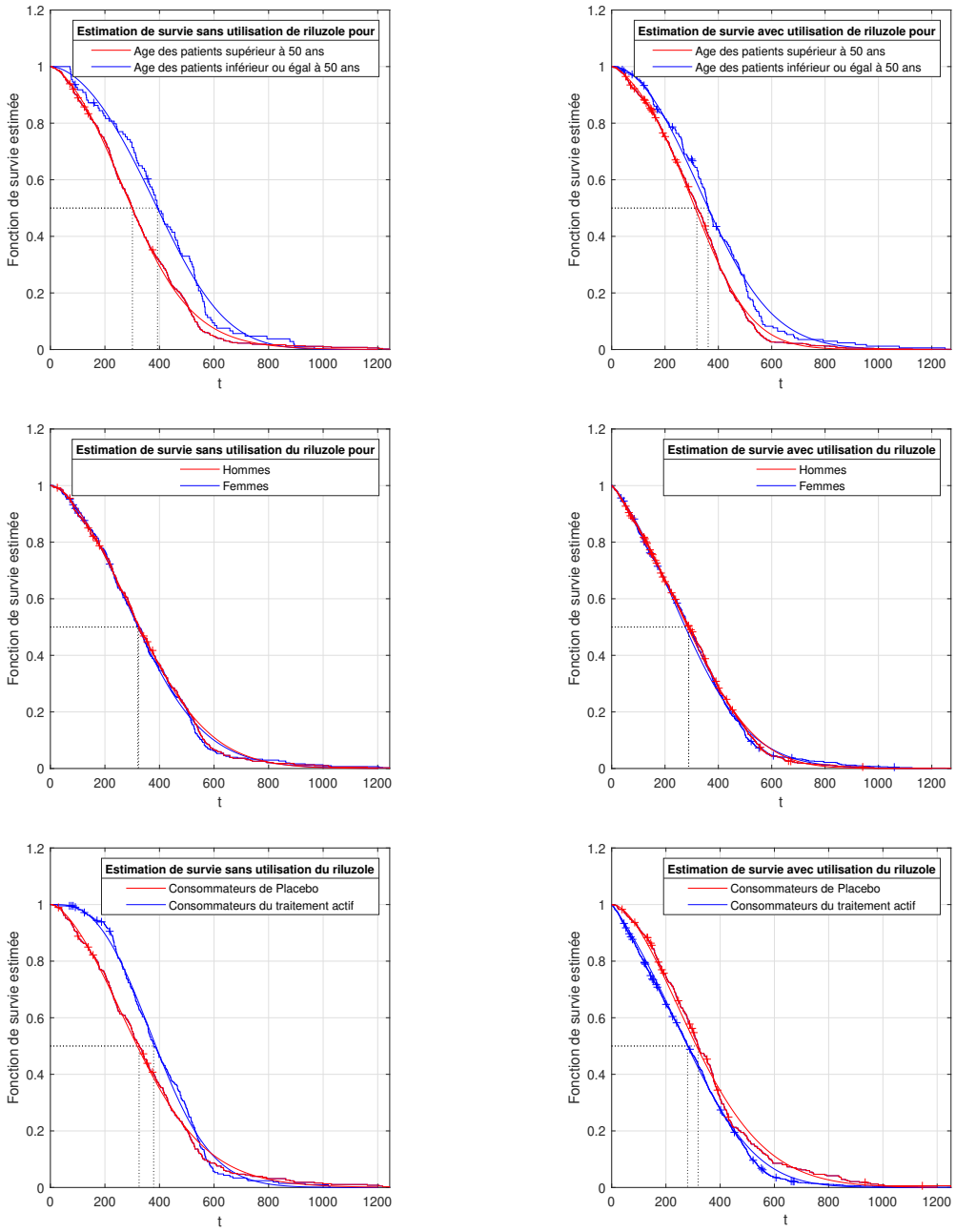


FIGURE 3.4 – Kaplan-Meier et les courbes paramétriques de survie stratifié par âge, sexe et type de traitement selon l'utilisation du riluzole pour les données ALS.

de Fisher utilisée pour calculer les intervalles de confiance pour les modèles 6 et 7 est de tailles respectivement (8×8) et (7×7). Les nombres de dérivées secondes calculées pour les modèles 6 et 7 sont alors respectivement $n_d = 36$ et $n_d = 28$.

3.2.5 Modèle de Cox à risque proportionnel

Le modèle à risques proportionnels de Cox [4] introduit dans 1.2.3.2 est un modèle semi-paramétrique populaire pour l'analyse des données de survie. Le modèle cox a été utilisé ici pour

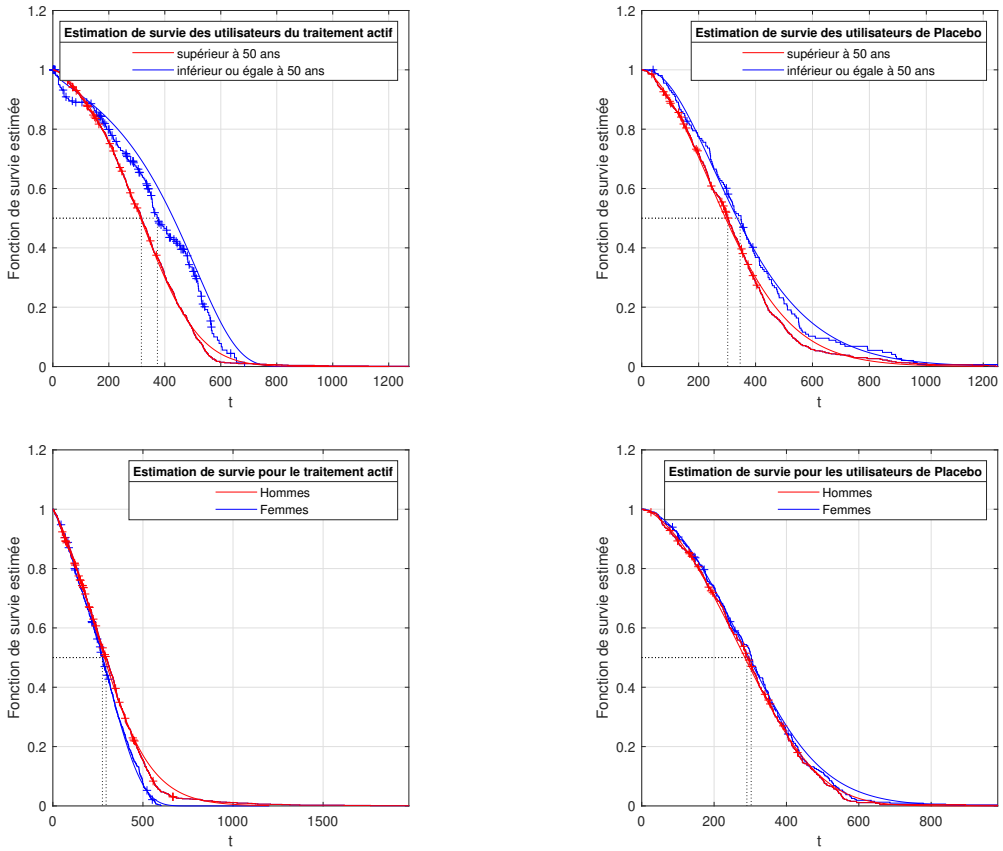


FIGURE 3.5 – Kaplan-Meier et les courbes de survie paramétriques stratifiées par âge et sexe selon le type de traitement pour les données ALS.

évaluer l'effet de l'âge, du sexe, du traitement actif et placebo et de l'utilisation du riluzole sur la survie globale des patients atteints de ALS.

Les modèles univariés ont confirmé les résultats obtenus par le modèle de régression MOEGG (modèles 1 à 4). Les résultats du modèle de Cox univarié sont affichés dans le tableau 3.8. Le modèle de régression multivariée de Cox a validé les résultats obtenus par le modèle 5 proposé où tous les facteurs de risque sont pris en compte. Les résultats du modèle de Cox multivarié sont présentés dans le tableau 3.9.

La régression avec le modèle de Cox est largement utilisée par les praticiens. Cependant, une version paramétrique du modèle de régression de Cox peut conduire à une estimation plus précise des probabilités de survie et, par conséquent, aider les praticiens à mieux comprendre le phénomène en cours d'étude.

La méthode de ré échantillonnage bootstrap a été utilisée pour la sélection du modèle. Les performances des modèles prédictifs de régression de Cox et du MOEGG sont évaluées en estimant les erreurs de prédiction (PE) à l'aide de la méthode .632+ Bootstrap [71].

D'après les résultats rapportés dans le tableau 3.10, le modèle proposé surpasse le modèle

candidat en termes de capacité prédictive. Les performances des modèles de régression de survie ont également été testées à l'aide de l'indice de concordance (C-index) [72].

Le modèle proposé fournit une valeur d'indice de concordance plus élevée que Cox et s'avère donc être le modèle le plus approprié. L'erreur standard du C-index ainsi que les résultats de la comparaison sont reportés dans le tableau 3.10.

TABLEAU 3.1 – Moyennes des estimations du maximum de vraisemblance (AMLE), erreur quadratique moyenne (MSE) et biais des estimations du maximum de vraisemblance pour les données simulées.

n		AMLE	MSE	Biais
30				
	$\hat{r}$	1.6855	0.0460	0.3145
	$\hat{\alpha}$	0.5625	0.0189	0.1375
	$\hat{\beta}$	1.3369	0.0266	0.1631
	$\hat{\gamma}$	4.3454	0.1193	0.3454
	$\hat{b}_0$	1.2838	0.0806	0.2838
	$\hat{b}_1$	1.6594	0.0254	0.1594
	$\hat{b}_2$	-0.8492	0.0227	0.1508
	$\hat{b}_3$	-1.9377	0.0039	0.0623
	$\hat{b}_4$	3.1550	0.0240	0.1550
	θ_{MO}	0.8866	0.0080	0.0896
100				
	$\hat{r}$	1.7792	0.0487	0.2208
	$\hat{\alpha}$	0.8185	0.0140	0.1185
	$\hat{\beta}$	1.5960	0.0092	0.0960
	$\hat{\gamma}$	3.8842	0.0134	0.1158
	$\hat{b}_0$	1.2271	0.0516	0.2271
	$\hat{b}_1$	1.5855	0.0073	0.0855
	$\hat{b}_2$	-0.9303	0.0049	0.0696
	$\hat{b}_3$	-1.9420	0.0034	0.0580
	$\hat{b}_4$	3.1517	0.0230	0.1517
	θ_{MO}	0.8095	0.0002	0.0125
200				
	$\hat{r}$	2.0219	0.0005	0.0219
	$\hat{\alpha}$	0.6681	0.0010	0.0319
	$\hat{\beta}$	1.5480	0.0023	0.0480
	$\hat{\gamma}$	3.9267	0.0054	0.0733
	$\hat{b}_0$	1.0123	0.0002	0.0123
	$\hat{b}_1$	1.57695	0.0059	0.0770
	$\hat{b}_2$	-0.9694	0.0009	0.0306
	$\hat{b}_3$	-1.9892	0.0001	0.0108
	$\hat{b}_4$	3.0075	0.0001	0.0075
	θ_{MO}	0.7911	3.4810×10^{-5}	0.0059

TABLEAU 3.2 – Les estimations du maximum de vraisemblance, l'intervalle de confiance à 95%, et les erreurs standards des modèles de régressions ajustés indépendamment aux données ALS en considérant une seule covariable à la fois.

Paramètre	**MLE**	**SE**	**Intervalle de confiance à** 95%
Modèle 1			
$\hat{r}$	34.4326	1.1621	$(32.1549, 36.7103)$
$\hat{\alpha}$	1.0548×10^{-2}	1.5834×10^{-5}	$(1.0517 \times 10^{-2}, 1.0579 \times 10^{-2})$
$\hat{\beta}$	0.1064×10^{-2}	7.9525×10^{-6}	$(0.1048 \times 10^{-2}, 0.1080 \times 10^{-2})$
$\hat{\gamma}$	1.3446	0.0424	$(1.2615, 1.4277)$
$\hat{b}_{01}$	0.5939	0.1321	$(0.3350, 0.8528)$
$\hat{b}_{11}$	0.2714	0.0270	$(0.2185, 0.3243)$
Modèle 2			
$\hat{r}$	1.2550	0.0395	$(1.1775, 1.3325)$
$\hat{\alpha}$	0.1930×10^{-2}	1.0299×10^{-6}	$(0.1928 \times 10^{-2}, 0.1932 \times 10^{-2})$
$\hat{\beta}$	-0.0912×10^{-2}	7.7946×10^{-6}	$(-0.0927 \times 10^{-2}, -0.0897 \times 10^{-2})$
$\hat{\gamma}$	2.0259	0.0032	$(1.9615, 2.0903)$
$\hat{b}_{02}$	1.1168	0.1411	$(0.8403, 1.3933)$
$\hat{b}_{12}$	-0.00432	0.0380	$(-0.1178, 0.0314)$
Modèle 3			
$\hat{r}$	1.2546	0.0395	$(1.1773, 1.3319)$
$\hat{\alpha}$	0.1928×10^{-2}	1.0284×10^{-6}	$(0.1926 \times 10^{-2}, 0.1930 \times 10^{-2})$
$\hat{\beta}$	-0.0913×10^{-2}	7.8084×10^{-6}	$(-0.0928 \times 10^{-2}, -0.0898 \times 10^{-2})$
$\hat{\gamma}$	2.0254	0.0328	$(1.9610, 2.0898)$
$\hat{b}_{03}$	1.0882	0.1436	$(0.8068, 1.3696)$
$\hat{b}_{13}$	0.0190	0.0322	$(-0.0441, 0.0821)$
Modèle 4			
$\hat{r}$	0.0657	0.0026	$(0.0605, 0.0071)$
$\hat{\alpha}$	5.4019×10^{-7}	9.9919×10^{-12}	$(5.4017 \times 10^{-7}, 5.4020 \times 10^{-7})$
$\hat{\beta}$	-0.0100	0.0001	$(-0.0101, -0.0099)$
$\hat{\gamma}$	0.6026	0.0051	$(0.5926, 0.6126)$
$\hat{b}_{04}$	0.6059	0.0862	$(0.4370, 0.7748)$
$\hat{b}_{14}$	0.2542	0.0304	$(0.1947, 0.3137)$

TABLEAU 3.3 – Le taux de risque et son intervalle de confiance à 95% des coefficients de régression des modèles de régression ajustés indépendamment.

Modèle	**Coefficient**	**HR**	**Intervalle de confiance à** 95%
Modèle 1	$\hat{b}_{01}$	1.8110	(1.3979, 2.3462)
	$\hat{b}_{11}$	1.3118	(1.2442, 1.3831)
Modèle 2	$\hat{b}_{02}$	3.0551	(2.3169, 4.0283)
	$\hat{b}_{12}$	0.9957	(0.9242, 1.0727)
Modèle 3	$\hat{b}_{03}$	2.9689	(2.2406, 3.9340)
	$\hat{b}_{13}$	1.0192	(0.9549, 1.0877)
Modèle 4	$\hat{b}_{04}$	1.8329	(1.1137, 3.0166)
	$\hat{b}_{14}$	1.2894	(1.2148, 1.3686)

TABLEAU 3.4 – Les estimateurs du maximum de vraisemblance, l'intervalle de confiance à 95%, et les erreurs standards du modèle de régression ajusté considérant tous les facteurs de risque à la fois.

Paramètre	**MLE**	**SE**	**Intervalle de confiance à** 95%
$\hat{r}$	34.0426	1.1124	(31.8622, 36.2230)
$\hat{\alpha}$	1.0511×10^{-2}	1.5489×10^{-2}	$(1.0481 \times 10^{-2}, 1.0541 \times 10^{-2})$
$\hat{\beta}$	0.1041×10^{-2}	7.6546×10^{-6}	$(0.1026 \times 10^{-2}, 0.1056 \times 10^{-2})$
$\hat{\gamma}$	1.3443	0.0412	(1.2635, 1.4251)
$\hat{b}_{05}$	0.5350	0.1621	(0.2173, 0.8527)
$\hat{b}_{15}$	0.2823	0.0262	(0.2309, 0.3337)
$\hat{b}_{25}$	-0.0632	0.0371	$(-0.1359, 0.0094)$
$\hat{b}_{35}$	-0.0641	0.0314	$(-0.1257, -0.0025)$
$\hat{b}_{45}$	0.1525	0.0290	(0.0957, 0.2093)

TABLEAU 3.5 – Le taux de hasard et l'intervalle de confiance à 95% des coefficients de régression du modèle ajusté considérant tous les facteurs de risque.

Coefficient	**HR**	**Intervalle de confiance à** 95%
$\hat{b}_{05}$	1.7074	(1.2427, 2.3460)
$\hat{b}_{15}$	1.3262	(1.2598, 1.3961)
$\hat{b}_{25}$	0.9388	(0.8729, 1.0096)
$\hat{b}_{35}$	0.9379	(0.8819, 0.9974)
$\hat{b}_{45}$	1.1647	(1.1004, 1.2329)

TABLEAU 3.6 – Estimation du maximum de vraisemblance, intervalles de confiance à 95% et erreurs standard du modèle de régression ajusté tenant compte de l'effet d'interaction entre le riluzole et d'autres facteurs de risque.

Paramètre	**MLE**	**SE**	**Intervalle de confiance à** 95%
$\hat{r}$	32.7630	1.0819	$(30.6424, 34.8836)$
$\hat{\alpha}$	0.0104	1.4805×10^{-5}	$(1.0371 \times 10^{-2}, 1.0429 \times 10^{-2})$
$\hat{\beta}$	0.0010	7.1561×10^{-6}	$(0.0986 \times 10^{-2}, 0.1014 \times 10^{-2})$
$\hat{\gamma}$	1.3486	0.0417	$(1.2669, 1.4303)$
$\hat{b}_{06}$	0.6978	0.2533	$(0.2014, 1.1942)$
$\hat{b}_{16}$	0.2463	0.0320	$(0.1835, 0.3091)$
$\hat{b}_{26}$	-0.0720	0.0456	$(-0.1614, 0.0174)$
$\hat{b}_{36}$	-0.0464	0.0349	$(-0.1148, 0.0220)$

TABLEAU 3.7 – Estimation du maximum de vraisemblance, intervalles de confiance à 95% et erreurs standard du modèle de régression ajusté tenant compte des interactions entre le traitement vs l'âge et le sexe.

Paramètre	**MLE**	**SE**	**Intervalle de confiance à** 95%
$\hat{r}$	33.3561	1.1134	$(31.1738, 35.5384)$
$\hat{\alpha}$	0.0105	1.6052×10^{-5}	$(1.0469 \times 10^{-2}, 1.0531 \times 10^{-2})$
$\hat{\beta}$	0.0011	8.4680×10^{-6}	$(0.1083 \times 10^{-2}, 0.1117 \times 10^{-2})$
$\hat{\gamma}$	1.3496	0.0421	$(1.2671, 1.4321)$
$\hat{b}_{07}$	0.7743	0.1786	$(0.4242, 1.1244)$
$\hat{b}_{17}$	0.0774	0.0351	$(0.0086, 0.1462)$
$\hat{b}_{27}$	-0.0064	0.0491	$(-0.1025, 0.0897)$

TABLEAU 3.8 – Résultats du modèle de régression à risques proportionnels univarié de Cox.

	Coefficient	**SE**	**Valeur-p**	**Wald**	**HR**	**CI du HR**
Age	0.2788	0.0644	1.4770×10^{-5}	4.3321	1.3216	$(1.1649, 1.4992)$
Sexe	-0.0093	0.0491	0.8498	-0.1893	0.9907	$(0.8998, 1.0908)$
Traitement	-0.0409	0.0495	0.04094	-0.8250	0.9600	$(0.8711, 1.0578)$
Riluzole	0.1356	0.0519	0.0090	2.6115	1.1452	$(1.0344, 1.2679)$

TABLEAU 3.9 – Résultats du modèle de régression à risques proportionnels multivariés de Cox.

	Coefficient	**SE**	**Valeur-p**	**Wald**	**HR**	**CI du HR**
Age	0.2856	0.0651	1.1542×10^{-5}	4.3861	1.3306	(1.1711, 1.5117)
Sexe	−0.0455	0.0497	0.3601	−0.9152	0.9956	(0.8669, 1.0533)
Traitement	−0.0981	0.0525	0.0617	−1.8682	0.9066	(0.8179, 1.0048)
Riluzole	0.1642	0.0551	0.0029	2.9810	1.1784	(1.0578, 1.3127)

TABLEAU 3.10 – Comparaison entre le modèle de Cox à risques proportionnels et le modèle proposé.

	PE	**C-index**	**SE**
Cox	0.3436	0.5547	0.0082
MOEGG	0.2733	0.5618	0.0080

TABLEAU 3.11 – Le résumé en cinq nombre du sous-échantillon de PRO-ACT utilisé.

Minimum	**Premier quartile**	**Médiane**	**Troisième quartile**	**Maximum**
4	211	330	442	1271

3.3 Application à COVID-19

La maladie de coronavirus 2019 (COVID-19) a causé une morbidité et une mortalité considérables dans le monde depuis décembre 2019. Cependant, les informations sur la modélisation du taux de guérison et l'ajustement des modèles de survie chez les patients atteints de COVID-19 sont limitées.

Dans ce chapitre, nous avons réalisé une étude comparative entre quelques distributions bien fondées. Nous avons adapté des modèles défectueux aux données de survie de la COVID-19. Par la suite, nous avons utilisé ces modèles pour estimer le taux de guérison des sujets. L'inférence statistique a été effectuée en utilisant l'estimation du maximum de vraisemblance.

Des données de survie des patients atteints de la COVID-19 ont été analysées par des méthodes paramétriques, semi-paramétriques et non paramétriques. Le taux de guérison de cette population a été estimé en fonction des caractéristiques démographiques.

3.3.1 Base de données de COVID-19

Les coronavirus sont une famille de virus causant des maladies comme le rhume, le syndrome respiratoire du Moyen-Orient (MERS-Cov) et le syndrome respiratoire aigu sévère (SARS-Cov). Dernièrement, un nouveau coronavirus a été identifié comme étant la cause d'une épidémie de pneumonie virale originaire de la ville de Wuhan, province du Hubei en Chine depuis décembre 2019. L'organisation mondiale de la santé (OMS) a nommé la nouvelle maladie Coronavirus Disease 2019 (COVID-19). Le 11 mars 2020, l'OMS déclare une pandémie [73]. Affectant plus d'un million de personnes dans le monde, le COVID-19 est maintenant devenu un événement de santé publique mondiale qui est très préoccupant.

3.3.2 Quelle est la distribution qui décrit le mieux la survie?

3.3.2.1 Introduction aux modèles de durées de vie

Différentes distributions ont été proposées par les statisticiens au fil des ans. Ici, nous avons sélectionné 12 distributions bien fondées. Nous proposons de déterminer la distribution qui représente le mieux les données de survie de COVID-19. Les noms des distributions, la fonction de densité de probabilité correspondante (PDF) ainsi que la plage et le type des paramètres sont donnés dans le tableau 3.12.

Il est important de noter que la procédure d'estimation de la distribution de Gompertz (GD)

TABLEAU 3.12 – Des données sur quelques distributions.

Distribution	PDF	Paramètres
Weibull	$f(t)=\frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}e^{-(\frac{t}{\beta})^{\alpha}}$	$\alpha>0$: forme, $\beta>0$: échelle
Exponentiel	$f(t)=\alpha e^{-\alpha t}$	$\alpha>0$: inverse de l'échelle
Burr	$f(t)=\frac{\alpha\beta}{\gamma}(\frac{t}{\gamma})^{\alpha-1}(1+(\frac{t}{\gamma})^{\alpha})^{-\beta-1}$	$\alpha>0$: échelle, $\beta,\gamma>0$: forme
Gamma	$f(t)=\frac{\beta^{\alpha}t^{\alpha-1}e^{-\beta t}}{(\alpha-1)!}$	$\alpha>0$: forme, $\beta>0$: inverse échelle
Valeurs extrêmes	$f(t)=\frac{1}{\beta}e^{\frac{t-\alpha}{\beta}}e^{-e^{\frac{t-\alpha}{\beta}}}$	$\alpha>0$: position, $\beta>0$: échelle
Log-logistique	$f(t)=\frac{(\frac{\beta}{\alpha})(\frac{t}{\alpha})^{\beta-1}}{(1+(\frac{t}{\alpha})^{\beta})^2}$	$\beta>0$: forme, $\alpha>0$: échelle
Logistique	$f(t)=\frac{e^{-(\frac{t-\alpha}{\beta})}}{\beta(1+e^{-\frac{t-\alpha}{\beta}})^2}$	$\alpha>0$: position, $\beta>0$:échelle
Log-normale	$f(t)=\frac{1}{t}\frac{1}{\beta\sqrt{2\pi}}e^{-\frac{1}{2\beta^2}(lnt-\alpha)^2}$	$\alpha\in R$: position, $\beta>0$: échelle
Nakagami	$f(t)=\frac{2\alpha^{\alpha}}{(\alpha-1)!\beta^{\alpha}}t^{2\alpha-1}e^{-\frac{\alpha}{\beta}t^2}$	$\alpha\geq\frac{1}{2}$: forme, $\beta>0$: échelle
Normale	$f(t)=\frac{1}{\beta\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{1}{\beta}(t-\alpha))^2}$	$\alpha\in R$: position, $\beta>0$: échelle
MGD	$f(t)=\alpha e^{\beta t}e^{-\frac{\alpha}{\beta}(e^{\beta t}-1)}$	$\alpha>0$: forme, $\beta<0$: échelle
MGGD	$f(t)=\gamma\alpha e^{\beta t}e^{-\frac{\alpha}{\beta}(e^{\beta t}-1)}(1-e^{-\frac{\alpha}{\beta}(e^{\beta t}-1)})^{\gamma-1}$	$\alpha,\gamma>0$: forme, $\beta<0$: échelle

[44] et de la distribution de Gompertz généralisée (GGD) [16] a donné des valeurs négatives du paramètre d'échelle β (voir les détails dans la section 2.2.2). Par conséquent, leurs versions défectueuses Modified Gompertz Distribution (MGD) [30] et Modified Generalized Gompertz Distribution (MGGD) [34], [46] respectivement, sont introduites dans le tableau 3.12 à la place des distributions GD et GGD.

Sachant que les survies $S(t)$ de MGD et MGGD sont respectivement données par (1.17) et (1.40), le taux de guérison de MGD est donné par (3.7) :

$$\theta=\lim_{t\to\infty}S(t)=e^{\frac{\alpha}{\beta}}, \tag{3.7}$$

et le taux de guérison de MGGD est donné par (3.8) :

$$\theta=\lim_{t\to\infty}S(t)=1-(1-e^{\frac{\alpha}{\beta}})^{\gamma}. \tag{3.8}$$

3.3.2.2 Résultats de l'inférence statistique

Il est à noter que nous considérons ici que le temps de censure est fixé au dernier jour de collecte des données. Il s'agit donc d'une censure fixe à droite. Les résultats de l'estimation du maximum de vraisemblance pour chacun des modèles sélectionnés, les erreurs standard correspondantes et les intervalles de confiance à 95% sont présentés dans le tableau 3.13. Il convient de

mentionner que, puisque le paramètre de forme α n'autorise que des valeurs positives, les valeurs négatives de la limite inférieure de l'intervalle de confiance à 95% sont remplacées par zéro.

TABLEAU 3.13 – Estimations du maximum de vraisemblance (MLE), l'erreur standard (SE) correspondante, la borne inférieure et la borne supérieure de l'IC à 95% pour chaque distribution.

Distribution	**Paramètre**	**MLE**	**SE**	**Intervalle de confiance à** 95%
Weibull	$\hat{\alpha}$	1388.4400	659.159	(547.55, 3520.7)
	$\hat{\beta}$	0.7448	0.0886	(0.5899, 0.9406)
Exponentiel	$\hat{\alpha}$	540.2380	68.0636	(428.2145, 703.0437)
Burr	$\hat{\alpha}$	6.4544	3.1913	(2.4490, 17.0107)
	$\hat{\beta}$	1.3652	0.2844	(0.9074, 2.0538)
	$\hat{\gamma}$	0.0271	0.0115	(0.0117, 0.0627)
Gamma	$\hat{\alpha}$	0.7388	0.0914	(0.5797, 0.9416)
	$\hat{\beta}$	1667.3700	909.2830	$(572.5798, 4.8554 \times 10^3)$
Valeurs extrêmes	$\hat{\alpha}$	112.1063	8.8073	(94.8442, 129.3684)
	$\hat{\beta}$	26.9675	3.0084	(21.6711, 33.5582)
Log-logistique	$\hat{\alpha}$	7.1059	0.4627	(6.1989, 8.0129)
	$\hat{\beta}$	1.3134	0.1553	(1.0417, 1.6562)
Logistique	$\hat{\alpha}$	108.4490	8.5430	(91.7048, 125.1929)
	$\hat{\beta}$	26.1034	2.8974	(20.9997, 32.4474)
Log-normale	$\hat{\alpha}$	8.1139	0.5688	(6.9990, 9.2289)
	$\hat{\beta}$	2.9829	0.3255	(2.4085, 3.6942)
Nakagami	$\hat{\alpha}$	0.6345	0.0437	(0.2881, 0.4612)
	$\hat{\beta}$	1.2604 10^6	1.1216 10^6	$(2.2032 \times 10^5, 7.2111 \times 10^5)$
Normale	$\hat{\alpha}$	119.753	9.9627	(100.2268, 139.2799)
	$\hat{\beta}$	53.9710	5.7397	(43.8163, 66.4790)
MGD	$\hat{\alpha}$	**0.0038**	**0.0007**	(**0.0024**, **0.0053**)
	$\hat{\beta}$	**-0.0469**	**0.0121**	(**-0.0706**, **-0.0232**)
MGGD	$\hat{\alpha}$	0.0117	0.0068	(0.0000, 0.0250)
	$\hat{\beta}$	-0.0693	0.0173	$(-0.0353, 0.1032)$
	$\hat{\gamma}$	1.4128	0.2782	(0.8676, 1.9580)

Le tableau 3.14 donne la valeur du logarithme de la fonction de vraisemblance en fonction des paramètres estimés ainsi que les valeurs des critères d'information. Le modèle qui décrit le mieux les données est le modèle avec les valeurs les plus basses des critères d'information. Les valeurs AIC, BIC, CAIC, AICc et HQIC montrent que MGD a nettement surpassé les autres modèles.

La figure 3.6 présente le tracé de la courbe d'estimation non paramétrique de Kaplan-Meier

TABLEAU 3.14 – Valeur de log-vraisemblance L et critères d'information pour chaque distribution.

Distribution	L	AIC	BIC	CAIC	AICc	HQIC
Weibull	-455.9490	915.8980	925.8748	927.8748	912.9091	919.6750
Exponentiel	-459.3970	920.7940	925.7824	926.7824	918.7977	922.6825
Burr	-451.3390	908.6780	923.6432	926.6432	904.7001	914.3435
Gamma	-456.1220	916.2440	926.2208	928.2208	913.2551	920.0210
Valeurs extrêmes	-503.7201	1011.4400	1021.4168	1023.4168	1008.4511	1015.2170
Log-logistique	-455.6590	915.3180	925.2948	927.2948	912.3291	919.0950
Logistique	-502.963	1009.9260	1019.9028	1021.9028	1006.9371	1013.7030
Log-normale	-453.5080	911.0160	920.9928	922.9928	908.0271	914.7930
Nakagami	-456.24	916.4568	926.4568	928.4568	913.4911	920.2570
Normale	-497.584	999.1680	1009.1448	1011.1448	996.1791	1002.9450
MGD	**-450.2200**	**902.4400**	**914.4168**	**916.4168**	**901.4511**	**908.2170**
MGGD	-460.7900	927.5800	942.5452	945.5452	923.6021	933.2455

ainsi que les courbes de survie de tous les modèles propres ajustés. Plus le modèle paramétrique est proche de la courbe de Kaplan-Meier, meilleur est l'ajustement.

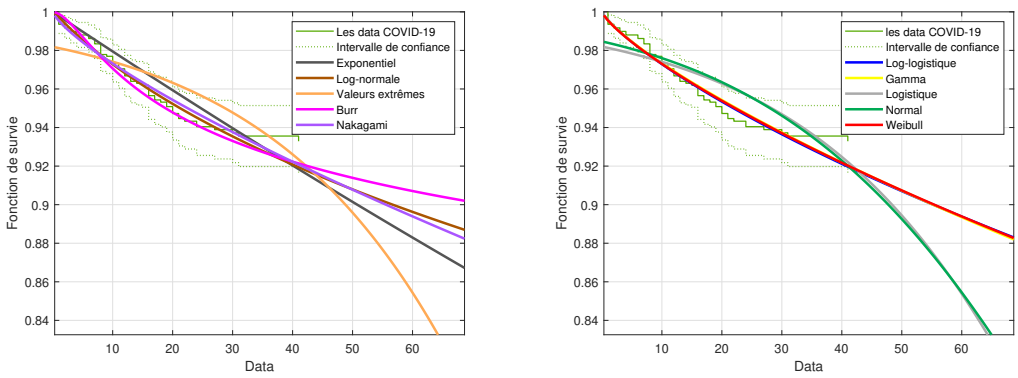


FIGURE 3.6 – Kaplan-Meier et estimation paramétrique de la fonction de survie de quelques distributions pour les données des patients atteints de COVID-19.

La figure 3.7 illustre le Kaplan-Meier pour l'estimation de la fonction de survie à partir de l'ensemble de données COVID-19 ainsi que les courbes de survie des distributions défectueuses MGD et MGGD respectivement. Pour ce type de modèle, cette étape permet d'estimer le taux de guérison dans la population étudiée. En fait, c'est la valeur vers laquelle converge la fonction de survie estimée.

Le taux de guérison estimé par la distribution de Gompertz modifiée est donné par la valeur 0.9222. Le taux de guérison estimé par la distribution Gompertz généralisée et modifiée est donné par la valeur 0.9280. Ces estimations sont en harmonie avec la littérature récente [74]. Pourtant,

d'autres facteurs peuvent également influencer le taux de guérison.

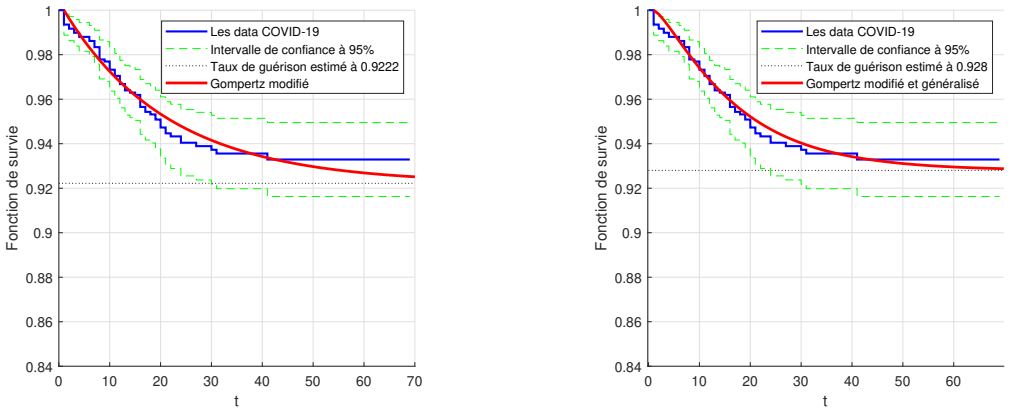


FIGURE 3.7 – Kaplan-Meier, l'intervalle de confiance à 95% et l'estimation paramétrique de la fonction de survie de MGD en (a) et de MGGD en (b) pour les données COVID-19 avec le taux de guérison estimé.

3.3.3 Analyse de sensibilité des variables explicatives

Pour visualiser l'effet de certaines covariables démographiques sur la survie globale des patients atteints de COVID-19, des approches d'estimation paramétrique, semi-paramétrique et non paramétrique ont été menées. La distribution modifiée de Gompertz (MGD) qui présentait le modèle le mieux ajusté est utilisée pour estimer le taux de guérison pour chaque stratification démographique : âge (≤ 60, > 60) et sexe.

3.3.3.1 Approche semi-paramétrique

La régression à risques proportionnels de Cox a confirmé que les facteurs de risque indépendants, l'âge et le sexe, ont une influence significative sur la survie des patients atteints de COVID-19. Les analyses de régression de Cox univariée et multivariée, présentées respectivement dans les tableaux 3.15 et 3.16, montrent que l'âge au-dessus de 60 ans et le sexe étaient associés à la survie globale des patients atteints de coronavirus.

Les valeurs des coefficients de régression négatives pour la variable sexe ($-0,8662$ et $-0,9113$) indiquent que le taux de survie des femmes est plus élevé que celui des hommes.

Les valeurs des coefficients de régression positifs (2.0656 et 2.0857) pour la variable âge indiquent que plus le patient est âgé, plus le risque de décès est élevé (et donc plus le taux de guérison est faible).

Les valeurs statistiques Wald (z) sont le rapport du coefficient de régression à son erreur

standard $z = coefficient/\text{SE}(coefficient)$. Les valeurs statistiques de Wald sont significativement différentes de zéro. Nous pouvons conclure que les deux variables âge et sexe ont des coefficients qui sont statistiquement très significatifs [52]. Une autre caractéristique à noter dans l'analyse de régression est la valeur du taux de risque calculée comme HR = $e^{coefficient}$. Le fait que les bornes de confiance à 95% du HR dans les deux variables ne contiennent pas zéro, est également une preuve de l'importance de l'effet des covariables étudiées sur la survie des patients COVID-19.

Pour conclure, le modèle de régression à risques proportionnels de Cox a montré un risque de décès important chez les hommes et encore plus important chez les patients âgés.

TABLEAU 3.15 – Résultats de l'étude univarié du modèle de Cox.

Covariable	**Coefficient**	**SE**	**Valeur-p**	**z**	**HR**	**CI à** 95%
Âge	2.0656	0.3004	$6.1198\ 10^{-12}$	6.8768	7.8902	(4.3793, 14.2159)
Sexe	-0.8662	0.3069	0.0048	-2.8229	0.4205	(0.2305, 0.7674)

TABLEAU 3.16 – Résultats de l'étude multivarié du modèle de régression de Cox.

Covariable	**Coefficient**	**SE**	**Valeur-p**	**z**	**HR**	**CI à** 95%
Âge	2.0857	0.3006	$3.9775\ 10^{-12}$	6.9380	8.0498	(4.4658, 14.5102)
Sexe	-0.9113	0.3070	0.0030	-2.9684	0.4020	(0.2202, 0.7338)

3.3.3.2 Approche paramétrique et approche non paramétrique

Dans la figure 3.8, les courbes de Kaplan–Meier non paramétriques et les courbes de la fonction de survie décrivent la survie globale en fonction du sexe et de l'âge des patients atteints de coronavirus. Les courbes paramétriques sont basées sur la distribution de Gompertz modifiée, avec des paramètres estimés par l'approche du maximum de vraisemblance.

Comme le montrent les figures, les courbes de survie ont atteint un plateau. Il est alors recommandé d'utiliser des modèles de taux de guérison car ils conduisent à des résultats plus précis [75].

Le plateau de la courbe de Kaplan-Meier est plus faible pour la population masculine et la population âgée des données COVID-19 et, par conséquent, le taux de guérison estimé associé est plus faible.

L'estimation du maximum de vraisemblance des paramètres MGD qui ont permis de réaliser les courbes de survie estimées pour chaque stratification considérée (voir figure 3.8) sont

données dans le tableau 3.17. Les erreurs standard correspondantes et les intervalles de confiance à 95% sont également indiqués dans le même tableau. Le taux de guérison Θ_{est} pour chaque sous-population est estimé en se basant sur la distribution de Gompertz modifiée, après l'utilisation de l'équation (3.7). Le taux de guérison empirique Θ_{emp} est naïvement calculé comme la proportion de patients qui sont encore en vie, ou bien le niveau de censure (à droite) dans la sous-population, qui est également le plateau dans les courbes de survie. Les taux de guérison estimés et empiriques ont des valeurs proches.

Kaplan-Meier et l'analyse paramétrique de survie ont démontré que les patientes et, plus précisément, celles qui sont âgées de moins de 60 ans ont une chance de survie significativement plus élevée.

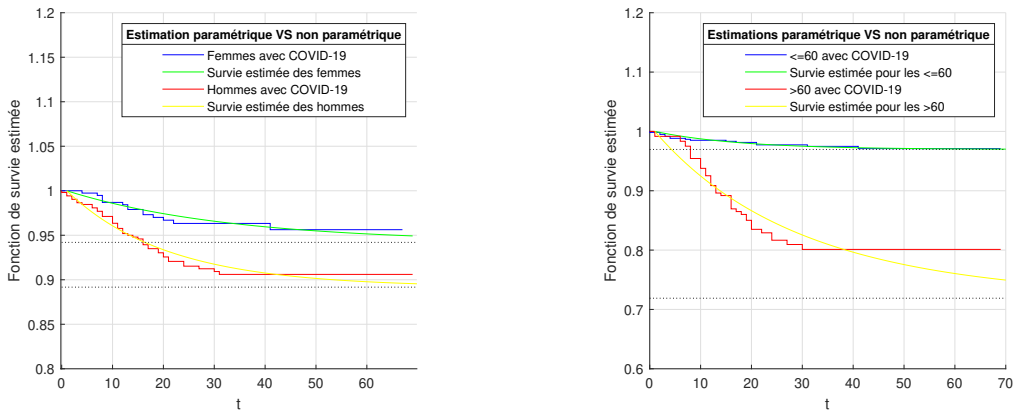


FIGURE 3.8 – Estimation non paramétrique par Kaplan-Maier et les courbes issues de la fonction paramétrique de survie globale selon le sexe dans (a) et l'âge dans (b).

TABLEAU 3.17 – MLE des paramètres de MGD, SE correspondante et le CI à 95%, le taux de guérison empirique et le taux de guérison estimé pour chaque stratification considérée.

	Age		**Sexe**	
	≤ 60	***>60***	***Femme***	***Homme***
$\hat{\alpha}$	0.0018	0.0099	0.0018	0.0055
$SE_{\hat{\alpha}}$	0.0007	0.0024	0.0008	0.0013
95% $CI_{\hat{\alpha}}$	$(0.0004, 0.0031)$	$(0.0053, 0.0143)$	$(0.0003, 0.0034)$	$(0.0030, 0.0080)$
$\hat{\beta}$	-0.0583	-0.0300	-0.0301	-0.0480
$SE_{\hat{\beta}}$	0.0253	0.0136	0.0229	0.0141
95% $CI_{\hat{\beta}}$	$(-0.1079, -0.0088)$	$(-0.0568, -0.0033)$	$(-0.0751, 0.0148)$	$(-0.0755, -0.0204)$
Θ_{emp}	0.9751	0.8216	0.9634	0.9154
Θ_{est}	0.9696	0.7189	0.9420	0.8917

3.4 Conclusion

Dans ce chapitre, une nouvelle famille de modèles de régression pour les données de survie a été présenté. Le modèle prend en compte la présence et l'absence d'une fraction guérie parmi les patients vu que pour certaines valeurs du paramètre d'échelle, le modèle proposé peut être défectueux. Une étude approfondie a été menée pour évaluer la performance de modèles à ajuster les données de survie ainsi que l'effet des facteurs de risque du modèle sur la survie globale des patients. L'effet de chaque facteur de risque a été modélisé indépendamment, simultanément et considérant ou non l'effet d'interactions entre eux. Le modèle proposé a été utilisé pour déterminer laquelle des covariables, parmi celles considérées, a le plus grand effet sur la survie médiane et les courbes des fonctions de survie empiriques et estimées. Les résultats sont validés à l'aide du modèle semi-paramétrique à risques proportionnels de Cox. L'application de données réelles prouve la flexibilité du modèle proposé pour décrire des données actuelles. Avoir des informations sur la survie globale des patients a permis d'avoir tous ces résultats. Mais si l'on n'avait pas ces informations ? Y a-t-il des alternatives à la survie globale des malades ? Si oui, à quel point peut on dire que ces alternatives sont fiables ?

Chapitre 4

Analyse de l'effet des paramètres cliniques : Application à une base de données de cancer du poumon récemment collectée

4.1 Introduction

Récemment, une attention particulière a été accordée à l'évaluation des paramètres primaires qui substituent la survie globale (OS) dans l'évaluation de l'effet d'une intervention dans divers cancers. Des paramètres de substitution valides pour la survie globale permettront de faire des études avec des tailles d'échantillons réduits, des durées du suivi réduites et des coûts des essais beaucoup plus réduits. Par exemple, l'agence américaine des produits alimentaires et médicamenteux des États-Unis [76] a approuvé de nouveaux médicaments dans plusieurs cancers en utilisant la survie sans progression (PFS) plutôt que la survie globale (OS) comme critère d'approbation.

Dans ce chapitre, nous introduisons une base de données nouvellement collectée de l'Institut Salah Azaiez (ISA) de Tunis. Cette base de données sera exploitée dans le but de mettre en évidence l'effet de la substitution de la survie globale par la survie sans progression dans l'étude inférentielle des données des patients atteints d'un cancer du poumon. En d'autres termes, on veut étudier la fiabilité de l'information extraite quand on travaille avec les données de survie sans progression au lieu de la survie globale.

4.2 Contexte médical

4.2.1 Le cancer du poumon

4.2.1.1 Introduction à la maladie

Le cancer du poumon est la principale cause de décès par cancer chez les hommes et les femmes du monde entier depuis plusieurs décennies. En fait, selon le classement de l'organisation mondiale de la santé (OMS), le cancer du poumon est le cancer le plus fréquemment diagnostiqué dans le monde (1,8 million, 13,0 % du total) et représente plus de décès chaque année que le cancer du sein, de la prostate et du côlon réunie [77].

Le cancer du poumon commence dans les cellules qui forment le tissu pulmonaire. Les cellules cancéreuses sont des cellules anormales dans le sens où elles se divisent plus rapidement que les cellules habituelles et leur accumulation entraîne une croissance cancéreuse, dites tumeurs. Un facteur important qui démontre la dangerosité du cancer du poumon est sa capacité à envahir les tissus ou les organes voisins et à traverser les vaisseaux lymphatiques ou les vaisseaux sanguins, ce que l'on appelle les métastases médicales.

4.2.1.2 Histoire

Les statiques présentées ci-dessus ne l'ont pas toujours été. Il y a environ 150 ans, le cancer du poumon était une maladie rare. En 1878, les tumeurs pulmonaires malignes ne représentaient que 1% de tous les cancers (vus à l'autopsie à l'Institut de pathologie de l'Université de Dresde en Allemagne). En 1918, ce pourcentage est passé à près de 10% et en 1927 à plus de 14% [78]. Lorsque Marie Curie et Pierre Curie ont découvert les radiations à la fin du 19e siècle et que Wilhelm Conrad Roentgen a découvert les rayons-X, les physiciens ont profité de ces découvertes pour sonder le corps humain [78]. Ainsi, des approches de traitement non chirurgical du cancer sont apparues. Par suite, les collaborations entre les chirurgiens et les radiologues hospitaliers ont commencé. En 1968, une quantité, relativement grande, de données sur le cancer a commencé à être compilée à l'aide d'ordinateurs. De grands efforts ont été consacrés au cours des 50 dernières années.

4.2.1.3 Facteurs de risque du cancer du poumon

Le tabac est la cause la plus fréquente de cancer du poumon. En fait, 90% des patients atteints d'un cancer du poumon sont actuellement d'anciens fumeurs [77]. De nombreux autres facteurs de risque sont à l'origine du cancer du poumon, comme l'exposition à la fumée secondaire,

les antécédents familiaux de cancer du poumon et une exposition antérieure à la radiothérapie, etc. [78]

4.2.1.4 Types de cancer du poumon

Il existe deux principaux types de cancer du poumon. Ces deux types se propagent de diverses manières et sont traités différemment : le cancer du poumon non à petites cellules (NSCLC) est le type le plus courant. Il représente environ 85% de tous les cancers du poumon. Ses sous-types les plus courants sont l'adénocarcinome et le carcinome épidermoïde. Ce type de cancer peut être traité par chirurgie, radiothérapie, chimiothérapie ou une combinaison de ceux-ci. Le traitement dépend du stade de la tumeur. Le cancer du poumon à petites cellules est beaucoup moins fréquent que le premier. Il représente en fait 15% de tous les cancers du poumon. Cependant, il est plus agressif que le NSCLC et tend à se propager plus tôt que le premier. Ce type de cancer du poumon peut être traité par chirurgie ou en utilisant une combinaison de chimiothérapie et de radiothérapie [77].

4.2.1.5 Dépistage du cancer du poumon

Le dépistage du cancer est conçu pour détecter le cancer chez les patients présentant des facteurs de risque mais sans symptômes. Il est destiné à détecter le cancer à ses premiers stades, au moment où il est le plus traitable. Il y a plusieurs années, une étude appelée National Lung Screening Trial ou NLST a démontré que le dépistage des patients par tomodensitométrie pouvait réduire le risque de décès par cancer du poumon de 20% chez les personnes à haut risque. Nous déterminons les personnes à haut risque comme étant des fumeurs actuels ou anciens, âgés de 55 à 74 ans et ayant des antécédents de tabagisme de plus de 30 paquets-années.

4.2.1.6 Diagnostic du cancer du poumon

Les patients présentant des symptômes de cancer du poumon et/ou des résultats suspects à la radiographie ou à la tomodensitométrie peuvent nécessiter d'autres tests de diagnostic. Ces tests peuvent être effectués pour obtenir des tissus pour le diagnostic communément appelé biopsie. Certains patients atteints d'un cancer du poumon à un stade précoce probable peuvent être référés pour une chirurgie sans biopsie diagnostique.

4.2.1.7 Stadification du cancer du poumon

Le stade du cancer fait référence à la taille et à l'étendue de la propagation de la tumeur. Il est important pour déterminer les options de traitement et le pronostic. Le stade est déterminé à l'aide d'une combinaison de tests non invasifs (radiographie) et parfois de tests invasifs (endoscopie et/ou biopsie).

4.2.1.8 Modalités d'imagerie couramment effectuée

L'utilisation des modalités d'imagerie est d'une grande importance dans le suivi de l'évolution de la maladie en fonction du temps et des enregistrements des durées de survie. L'imagerie représente 90% des arguments en faveur du choix thérapeutique, et indique l'évaluation thérapeutique de la tumeur. Lorsque les radiations et les rayons-X ont été découverts à la fin du 19e siècle, les physiciens ont profité de ces découvertes pour sonder le corps humain et des approches de traitement non chirurgical du cancer sont apparues.

Nous introduisons, dans ce qui suit, les modalités d'imagerie ou les tests qui sont couramment pratiqués pour déterminer le stade du cancer du poumon.

La tomodensitométrie thoracique ou abdominale : La tomodensitométrie est l'une des meilleures méthodes actuellement disponibles et reproductibles pour mesurer les lésions cibles sélectionnées pour l'évaluation de la réponse. La tomodensitométrie conventionnelle doit être réalisée avec des coupes de $10mm$ ou moins d'épaisseur de tranche contiguës. La tomodensitométrie en spirale doit être réalisée à l'aide d'un algorithme de reconstruction contigu de $5mm$. Elle peut être appliquée aux tumeurs thoraciques, abdominales et du bassin. L'apparence des deux types de cancer du poumon peut paraître similaire à la radiographie pulmonaire ou à la tomodensitométrie. Des spécimens peuvent être examinés au microscope afin de faire la distinction entre les deux types de cancer du poumon.

En général, les cliniciens commencent le diagnostic du cancer du poumon par une radiographie thoracique pour vérifier si la tumeur existe ou non. Cependant, si la tumeur mesure moins de $1cm$ de diamètre ou si elle est cachée derrière un organe, par exemple un os ou un ganglion lymphatique, le cancer peut ne pas paraître à la radio. D'autre part, si la tumeur est visible par la radiographie pulmonaire, son estimation de sa taille peut être effectuée par tomodensitométrie (TDM). La tomodensitométrie thoraco-abdominale fournit des images détaillées de la tumeur pulmonaire et de l'anatomie. Du coût, la TDM est importante pour la mise en scène et la planification du traitement. La tomodensitométrie thoracique est utilisée pour la stadification du cancer et la tomodensitométrie abdominale est utilisée pour la localisation des métastases.

La tomographie par émission de positons (TEP) : L'utilisation du sucre radioactif est mise en œuvre car les cellules cancéreuses utilisent le sucre rapidement. Important pour identifier la propagation aux ganglions lymphatiques ou à d'autres organes.

L'imagerie par résonance magnétique (IRM) : L'IRM est également l'une des meilleures méthodes actuellement disponibles et reproductibles pour mesurer les lésions cibles sélectionnées pour l'évaluation de la réponse. La tomodensitométrie et l'IRM conventionnelles doivent être effectuées avec des coupes contiguë inférieures ou égales à $10mm$ d'épaisseur. Une IRM du cerveau peut être nécessaire pour déterminer si la tumeur s'est propagée au cerveau.

4.2.2 Réponse objective

L'évaluation des changements tumoraux est cruciale pour la prise de décision clinique des thérapies anticancéreuses. Le rétrécissement de la tumeur indique la réponse positive au traitement utilisé. Cependant, la croissance tumorale indique la progression de la maladie. Le rétrécissement et la croissance sont des paramètres utiles dans les essais cliniques. À cette fin, de nombreuses définitions et hypothèses des critères de réponse tumorale ont été explorées au fil des ans. Afin de savoir si l'état de santé des patients atteints d'un cancer s'améliore «répond au traitement», reste le même «est stable» ou s'aggrave «progresse» pendant le traitement, différents ensembles de règles publiées ont été définis.

Suivant les critères de l'OMS qui sont à l'origine de tous les autres biomarqueurs, la réponse objective d'une maladie mesurable peut être confirmée 4 semaines après le début du traitement et dépend du pourcentage d'évolution du produit des diamètres les plus longs. Quatre types de réponses ont été identifiés :

Réponse complète (CR) : La disparition de toutes les maladies connues et de tous les signes de cancer en réponse au traitement. La réponse complète ne signifie pas toujours que le cancer a été guéri et est également appelé rémission complète et c'est bien la réponse positive que le traitement vise à donner.

La réponse partielle (PR) est une diminution d'au moins 50% par rapport à la taille initiale de la tumeur avant le traitement.

La progression tumorale (PD) est un cancer qui se développe, se propage ou s'aggrave. C'est, en fait, quand il s'agit d'une augmentation de plus de 25% d'une ou plusieurs lésions ou de l'apparition de nouvelles lésions.

La stabilité tumorale (SD) est lorsque la tumeur n'a pas insuffisamment diminué pour définir une réponse partielle et/ou n'a pas augmenté suffisamment pour définir une progression tumorale. Précisément, la stabilité tumorale est une diminution inférieure à 50% ou une augmen-

tation inférieure à 25% par rapport au test précédent.

La réponse objective des tumeurs au traitement a été définie il y a plus de 50 ans. Puis au fil des années, des critères morphologiques, appelés biomarqueurs d'imagerie quantitative, ont été développés, tels que :

- Les critères de l'OMS
- Les critères d'évaluation de la réponse dans les tumeurs solides version 1.0 (RECIST 1.0)
- Les critères d'évaluation de la réponse dans les tumeurs solides version 1.1 (RECIST 1.1), dont nous discuterons plus en détail dans la section suivante [79].

4.2.3 Biomarqueurs d'imagerie quantitative

Dans cette partie, nous donnons un aperçu des principaux critères anatomiques existants utilisés au fil du temps pour l'évaluation de la réponse dans les tumeurs solides, en particulier les tumeurs pulmonaires.

Les critères de l'OMS ont été publiés en 1981 [80] et continuait à être utilisés même plus d'une décennie après l'introduction des critères RECIST 1.0, publiés en 2000 [81], et leur version révisée RECIST 1.1, publiée en 2009 [82]. Ces ensembles de critères ont été développés pour estimer la réponse aux agents chimiothérapeutiques cytotoxiques et pour observer le changement en taille de la tumeur au cours du traitement. Les critères de l'OMS sont basés sur une mesure bidimensionnelle comme le montre la figure 4.1. La somme des produits des diamètres les plus longs (DL) avec sa plus grande perpendiculaire dans la lésion cible, ou en d'autres termes la surface du rectangle qui entoure la cible.

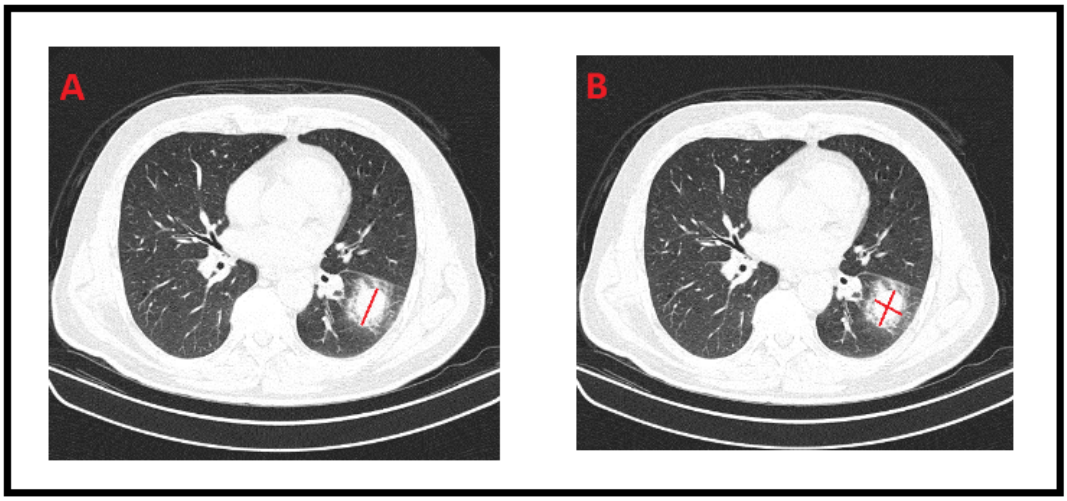


FIGURE 4.1 – La différence entre : A : RECIST une mesure unidimensionnelle et B : OMS une mesure bidimensionnelle.

Les critères de l'OMS ont été progressivement abandonnés en raison de son manque de normalisation dans le temps. Les critères d'évaluation de la réponse dans les tumeurs solides (RECIST) sont basés sur la mesure du diamètre le plus long des lésions, comme le montre la figure 4.1 ; unidimensionnel plutôt que des mesures bidimensionnelles pour évaluer la charge tumorale. Il introduit la notion de taille minimale mesurable, et le nombre maximal de lésions qui seront pris en compte (jusqu'à 10, un maximum de 5 par organe), d'où la notion de lésion cible et non cible. Les images tomodensitométriques présentées à la figure 4.1 sont fournis avec l'autorisation du service de radiologie de l'Institut Salah Azaiez en Tunisie. La version 1.1 du guide RECIST révisé (RECIST 1.1) a été présentée par le groupe de travail RECIST, en partie sur la base d'investigations utilisant une base de données comprenant plus de 6500 patients avec environ 18000 lésions cibles. Les principaux changements dans RECIST 1.1 comprenaient la mesure des ganglions lymphatiques, le nombre maximal de lésions cibles et la définition de la progression de la maladie. Le nombre maximum de lésions cibles a été abaissé de dix à cinq au total et de cinq à deux par organe. La progression de la maladie a été clarifiée. RECIST 1.1 a montré un accord presque parfait avec RECIST 1.0 dans l'évaluation de la réponse tumorale des patients atteints d'un cancer du poumon non à petites cellules.

4.2.3.1 Critères RECIST 1.1

Actuellement, la métrique standard, par laquelle la progression de la maladie est mesurée, est l'ensemble des lignes directrices RECIST (version 1.1). Au départ, il existe deux types de lésions tumorales : les lésions mesurables et les lésions non-mesurables. Une lésion mesurable peut être mesurée avec précision dans une ou plusieurs dimensions. Le plus long diamètre d'une lésion mesurable dépasse $20mm$. Une lésion est dite non mesurable lorsqu'elle ne peut pas être mesurée avec précision. Son plus long diamètre est inférieur à $20mm$.

La réponse objective ne concerne que les patients avec des maladies mesurables dès le départ. Le nombre de lésions mesurables doit être inférieur à deux lésions par organe pathologique et à cinq lésions au total. Ces lésions mesurables sont considérées comme des lésions cibles et doivent être enregistrées au départ. Une lésion cible est une lésion mesurable dont le diamètre le plus long est supérieur à $10mm$ si elle est évaluée cliniquement ou par scanner, et supérieur à $20mm$ si elle est radiographiquement évaluée. Si la lésion cible est un ganglion lymphatique, les mesures à axe court doivent être utilisées et enregistrées $\geq 5mm$. Toutes les lésions qui ne sont pas considérées comme cibles (ainsi que les lésions non mesurables et les ganglions lymphatiques pathologiques) doivent également être enregistrées au départ. Les ganglions lymphatiques sont inférieurs à $15mm$, ils sont considérés comme des lésions non cible. Seule la présence ou l'absence d'une telle lésion est notée. Ou rarement lorsque la progression est évidente, ce type de réponse doit également être noté. Contrairement aux biomarqueurs d'imagerie quantitative mentionnés précédemment, les critères de quantification de la réponse des lésions cibles, selon RECIST 1.1, peuvent être :

Réponse complète (CR) : où toutes les lésions cibles disparaissent.

Réponse partielle (RP) : où il y a une diminution de plus de 30% de la somme du diamètre le plus long des lésions cibles, en prenant comme référence la somme de base du diamètre le plus long.

Stabilité tumorale (SD) : où les lésions n'ont pas suffisamment rétréci (pour que les praticiens considèrent la réponse comme partielle) ni n'ont augmenté suffisamment (pour que les praticiens considèrent la maladie comme progressive). Ici, nous prenons comme référence la plus petite somme des plus longs diamètres.

Progression tumorale (PD) : où il y a plus de 20% d'augmentation de la somme du plus long diamètre des lésions cibles [82]. Ici, nous prenons comme référence la plus petite somme des plus longs diamètres ou si une nouvelle lésion apparaît.

La figure 4.2 ci-dessous illustre la quantification de la réponse des lésions cibles.

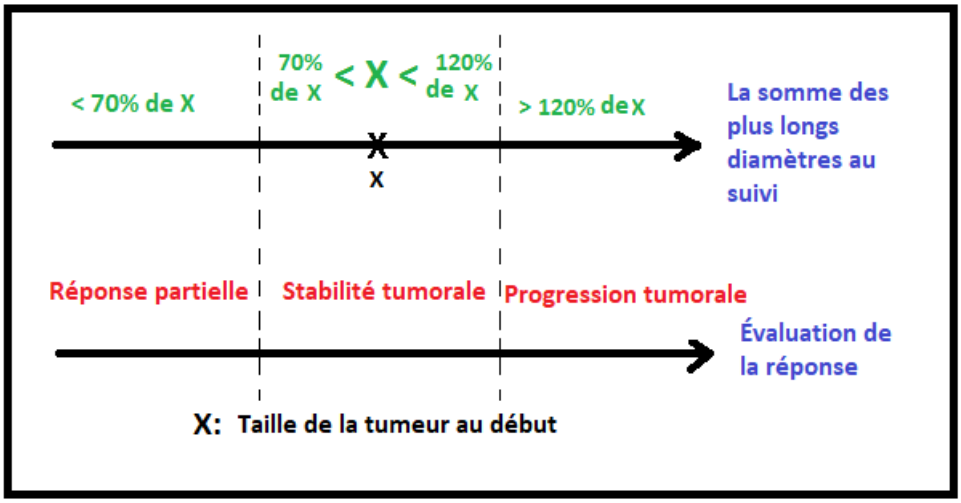


FIGURE 4.2 – Schéma représentatif de la quantification de la réponse des lésions cibles selon RECIST 1.1

Les critères de quantification de la réponse des lésions non cibles peuvent être : *Réponse complète (RC) :* où toutes les lésions non cibles disparaissent. *Stabilité tumorale ou réponse incomplète (SD) :* lorsqu'une ou plusieurs lésions non cibles persistent. *Progression tumorale (PD) :* où une ou plusieurs nouvelles lésions apparaissent ou une lésion non cible déjà existante se développe visiblement.

4.2.3.2 Étude critique des différents critères

Pour se mettre dans le contexte, la corrélation entre les différents ensembles de critères est présentée et les avantages et les inconvénients de chacun sont mis en évidence. En outre, l'utilité

de ces critères pour la communauté médicale et mathématique est également étudiée.

Dans les tableaux 4.1, 4.2 et 4.3, un résumé exhaustif des avantages et des limites de la méthode de l'OMS, RECIST 1.0 et RECIST 1.1 est présenté. Même si RECIST 1.1 a ses lacunes, nous pensons qu'il s'agit de l'ensemble de critères les plus fiables par rapport aux méthodes actuellement disponibles.

TABLEAU 4.1 – Étude critique des critères de l'OMS

Avantages	**Inconvénients**
Relativement facile et rapide à faire.	La planification de telles études demande parfois trop de temps.
Il est possible de déterminer la quantité d'effet du traitement sur la tumeur.	La taille minimale des lésions et le nombre de lésions par organe à mesurer diffèrent d'un groupe de recherche à l'autre.
Le calcul de la distance entre deux points est généralement disponible.	La progression tumorale est définie comme la variation considérable d'une seule lésion par certains groupes de chercheurs et définie comme la variation considérable de la charge tumorale globale par d'autres.
Même si les critères de l'OMS sont anciens, ils sont toujours applicables et sont efficaces dans de nombreuses évaluations de la réponse clinique.	L'intégration de mesures tridimensionnelles dans l'évaluation de la réponse est devenue déroutante depuis l'arrivée de nouvelles modalités d'imagerie comme la tomodensitométrie et l'imagerie par résonance magnétique IRM.
Il est possible de comparer les médicaments en fonction de leur efficacité.	Les mesures de volume ne sont pas incluses dans les critères de l'OMS à cause de la limitation des techniques d'imagerie ainsi que de la restriction des méthodes de mesure disponibles.
Meilleure détection du changement de tumeur que RECIST.	Résultats instables lors de l'acquisition de l'angle de coupe et de l'évaluation de la cavitation tumorale.
Les critères de l'OMS peuvent montrer une progression de la tumeur plus rapidement que RECIST.	L'intégration du changement de taille des lésions mesurables dans l'évaluation de la réponse varie selon les groupes de recherche.

TABLEAU 4.2 – Étude critique des critères RECIST

Avantages	**Inconvénients**
RECIST 1.0 est une mesure standardisée de la réponse tumorale qui pas besoin d'appareils coûteux en calcul.	RECIST 1.0 peut être trompeur dans certains cas et son utilité est controversé si on a affaire à de nouvelles modalités d'imagerie comme tomodensitométrie multi-détecteurs (MDCT).
L'application des critères RECIST est plus simple et plus pratique que les critères de l'OMS ainsi que son calcul.	Le nombre de lésions cibles à traiter est assez grand.
Il n'y a pas de divergence majeure avec les critères de l'OMS (en particulier pour les patients atteints de cancer du poumon) ce qui facilite l'application des critères d'évaluation de la réponse aux cliniciens.	Les critères RECIST unidimensionnels ne permettent pas de prédire la survie globale aussi précisément que les mesures volumétriques
C'est un point de terminaison antérieur à la survie. C'est donc vraiment un substitut.	L'étape de la confirmation peut perdre beaucoup de temps.

TABLEAU 4.3 – Étude critique des critères RECIST 1.1

Avantages	**Inconvénients**
Son avantage majeur par rapport aux méthodes de réponse antérieures était de simplifier le nombre de lésions à mesurer et les types de mesures qu'elle prenait : une mesure au lieu de deux mesures et deux lésions dans chaque organe au lieu de dix. En bref, RECIST 1.1 a été publié pour simplifier, optimiser et normaliser les critères d'origine.	RECIST 1.1 peut générer une réponse mixte (c'est-à-dire que certaines lésions deviennent plus grosses et d'autres deviennent plus petites et parfois il y a une nouvelle lésion) en particulier lorsque des traitements immunitaires sont utilisés. Cependant, si vous attendez suffisamment longtemps, la tumeur peut disparaître et donc la réponse peut même être complète.
Puisque RECIST 1.1 évalue un maximum de 5 tumeurs (contre 10 dans RECIST), elles se traduisent par un taux de réponse complète plus élevé que les critères RECIST d'origine (au moins dans les ganglions lymphatiques).	L'une des nombreuses faiblesses de RECIST est qu'il ne concerne que les tumeurs solides. En fait, il n'y a pas de critères pour les tumeurs non solides, en particulier que, dans certains cancers, les tumeurs peuvent passer de solides à non solides.
RECIST 1.1 effectue relativement la meilleure prédiction de la survie globale.	RECIST ne traite pas le cas où la croissance tumorale est non sphérique ou asymétrique.

4.2.4 Paramètres primaires

À la suite d'une intervention thérapeutique, des mesures des résultats doivent être prises. Ces mesures sont considérées comme des paramètres cliniques. Il s'agit d'analyse distincte faisant référence à des caractéristiques de la maladie observée dans un essai de recherche clinique ou une étude (e.g. un symptôme, la survenue d'une maladie ...). Les paramètres cliniques reflètent l'effet de l'intervention.

Le critère d'évaluation principal d'un essai clinique est désigné ici par paramètre primaire. À la fin d'une étude, le paramètre primaire permet de voir si un traitement donné a fonctionné. Il est décidé avant le début de l'étude. Les paramètres secondaires sont des mesures supplémentaires, également pré-spécifiés, qui sont moins importants que les paramètres primaires et pour lesquels l'étude peut ne pas être alimentée.

4.2.4.1 Survie globale OS

La survie globale (ou OS pour "Overall Survival" en anglais) est le temps écoulé entre le début de la randomisation et le décès par n'importe quelle cause. La survie globale mesure le bénéfice clinique pour un patient suite à une intervention thérapeutique. Les patients vivants ou dont l'information n'est pas entièrement observée sont considérés comme censurés. Cliniquement, la survie globale offre le gain le plus grand [76]. Elle peut être mesurée facilement, sans ambiguïté ni subjectivité. La survie globale est jugée médicalement significative et non sensible au moment de l'évaluation [83]. Une limitation majeure de ce critère d'évaluation est le fait qu'il nécessite un grand nombre de sujets ainsi qu'une durée de suivie assez longue. Ces conditions retardent potentiellement le résultat clinique (approbation d'un nouveau médicament ou agent...). De plus, l'OS n'est pas applicable en tant que paramètre primaire dans tous les types de cancer. L'existence de signes ultérieures de traitement est une autre limitation de l'OS. D'ailleurs, l'interprétation de la survie globale prise comme critère primaire pourrait être difficile. Une alternative à la survie globale est donc nécessaire comme critère d'évaluation [84].

4.2.4.2 Survie sans progression PFS

La survie sans progression (ou PFS pour 'Progression Free Survival' en anglais) est souvent utilisée comme alternative à la survie globale. La PFS est utile pour analyser les résultats des traitements des maladies en stades avancés. Contrairement à la survie globale, l'évènement pour la PFS est que la maladie progresse (s'aggrave), ou que le sujet décède de n'importe quelle autre cause. La PFS est pratiquement le paramètre primaire à adopter pour les situations métastatiques [83].

4.2.4.3 Survie sans maladie DFS

La survie sans maladie (ou DFS pour 'Disease Free Survival' en anglais) est le critère tout à fait pertinent pour statuer sur l'effet d'un nouvel anticancéreux [83]. L'évènement d'intérêt pour le DFS est la rechute plutôt que la mort. Les sujets qui rechutent survivent encore mais ils ne sont plus indemnes de maladie.

4.2.4.4 Temps jusqu'à progression TTP

Le délai de progression ou le temps jusqu'à progression (ou TTP pour 'Time To Progression' en anglais) est un paramètre primaire qui permet de statuer sur l'effet thérapeutique. Le TTP est similaire à la PFS mais, contrairement à la PFS, le TTP ignore les patients qui meurent avant que la maladie progresse. Des définitions séparées sont données et la PFS est préférable dans la majorité des cas [83].

4.2.4.5 État de l'art de la relation entre les différents paramètres primaires

En se référant aux travaux de [85, 86, 87, 88, 83], on constate que le niveau de preuve disponible soutenant une relation entre la survie globale et ses alternatives (PFS, TTP ...) varie considérablement selon le type de cancer et n'est pas toujours cohérent, même au sein d'un type de cancer spécifique.

4.3 Analyse de l'effet des paramètres cliniques

4.3.1 Construction des données de survie à partir des images médicales

La base de données utilisée dans ce chapitre se compose d'images tomodensitométriques de patients diagnostiqués pour le cancer du poumon. Il s'agit de 600 TDM de patients avec des caractéristiques différentes suivis et traités dans l'Institut Salah Azaiez de Tunisie. Pour chaque TDM, nous disposons :

- des images de type DICOM (une abréviation pour Digital Imaging and Communications in Medicine) qui est un outil standard pour la gestion des données médicales basées sur l'image. Jusqu'à 1300 coupes sagittales, frontales et transverses sont retraitées pour donner des images volumiques. Ces scanners sont lisibles à l'aide d'un logiciel appelé *Radiant*.
- des comptes rendus rédigés par les radiologues de l'Institut Salah Azaiez contenant : les renseignements cliniques, la technique utilisée pour faire le TDM et les résultats du scanner

à l'étage thoracique, à l'étage abdomino-pelvien ... L'étude médiastinale, l'étude pleuro-parenchymateuse (et/ou pariétale) et l'étude de la fenêtre osseuse sont aussi disponibles dans les comptes rendus. L'étude du scanner cérébral est aussi parfois disponible. Les médecins radiologues concluent sur l'état du patient à partir de ces résultats. Dans certains comptes rendus et quand il s'agit d'une tumeur solide dans les poumons, les critères RECIST (introduit dans 4.2.3.1) sont adoptés pour évaluer la réponse du malade à un traitement (souvent la chimiothérapie et/ou radiothérapie et/ou chirurgie).

En utilisant ces informations, nous étions capables d'établir une base qui fournit, pour chaque patient : l'âge, le sexe, le traitement, la présence d'une masse et la présence de nodules, des informations sur la censure à gauche, à droite ou par intervalle, la réponse objective et la durée de survie en jours. On considère que l'évènement d'intérêt est la progression ou l'aggravation ou la rechute de l'état du patient. Donc lorsque l'application des critères RECIST résulte une réponse objective autre que la progression de la maladie, la durée de survie est considérée comme censurée à droite. Les durées de survie globale sont les périodes calculées depuis le début du traitement ou bien le premier diagnostic jusqu'au dernier diagnostic ou le décès du patient. Alors que les durées de survie-sans-progression sont les périodes calculées depuis le début du traitement ou le premier TDM disponible jusqu'au premier TDM indiquant une progression tumorale selon les critères RECIST 1.1 soit par l'augmentation de la somme des diamètres par plus que 20% des lésions cibles, soit par l'apparition de nouvelles lésions. Dans les cas où les critères RECIST 1.1 indiquent une stabilité tumorale, la durée de survie globale est égale à la durée de survie-sans-progression. De même pour le cas où les critères RECIST 1.1 donnent une réponse complète ou une réponse partielle. À partir des dates disponibles dans cette base, des intervalles de survie sont aussi construits. On s'en servira pour faire l'étude de la censure par intervalle. L'analyse de ces données de survie est présentée dans ce qui suit.

4.3.2 Méthodologie et inférence statistique

Le but de cette partie est d'examiner les preuves disponibles liant la PFS à la OS particulièrement dans le cancer du poumon. Quand on ne connait pas la survie globale d'un patient (parce qu'il est encore vivant ou bien parce que le patient est perdu de vue), dans quelle mesure la survie sans progression peut-elle être un paramètre de substitution fiable et robuste ? Cette question est abordée en partant d'une perspective mathématique.

L'ajustement d'une distribution à une base de données décrivant les mesures répétées d'un phénomène variable a pour but la prédiction et la prévision de la fréquence de la survenue de l'ampleur du phénomène dans un intervalle de temps bien déterminé.

Un grand nombre de distributions a été développé à travers l'histoire. Parmi ces distributions, il y en a qui peuvent être ajustées plus étroitement aux observations que d'autres. Les caractéristiques du phénomène et les spécificités de la distribution jouent un rôle très important en ce

sujet. La distribution qui donne l'ajustement le plus étroit et serré de la courbe de Kaplan-Meier est considérée comme la meilleure distribution pour la description de la base et pour l'obtention de bonnes prédictions. La sélection de la loi la plus adéquate aux données est donc nécessaire.

Peu de travaux ont été réalisés pour ajuster des modèles statistiques à des durées de survie sans progression. En supposant que le TTP et le OS sont distribués d'une manière exponentielle, une distribution qui étudie la relation entre le PFS et la OS a été développé par Fleischer *et al.* [89]. Un modèle multi-états, randomisation-progression-décès, a aussi été élaboré par [90] pour analyser leur dépendance.

Pour la série de données de cancer du poumon décrite ci-dessus 4.3.1, nous avons ajusté plusieurs distributions bien fondées. Ces distributions sont introduites dans 3.3.2.1.

Les 10 distributions qui ont convergé et qui ont donné les meilleurs résultats sont : la loi exponentielle, la loi log-logistique, la loi gamma, la loi nakagami, la loi weibull, la loi log-normale, la loi inverse-Gaussienne,la loi logistique, la loi de Student (t Location-Scale),la loi d'extremum. D'autres distributions ont été testé mais qui ont, soit divergé, soit donné de mauvais résultats. Parmi ces distributions, on cite : la loi de Birnbaum-Saunders, la loi de Burr, la loi de Gompertz, la loi de Gompertz modifiée (DG) et des généralisations de la loi de Gompertz.

La survie sans progression est parfois adoptée comme un paramètre de substitution valide lors de l'établissement du bénéfice clinique d'un traitement quand les données matures sur la survie globale ne sont pas disponibles [91]. Dans le même contexte, une certaine quantification du rapport entre la survie globale et la survie sans progression peut être bénéfique dans le sens où on pourrait peupler le modèle "économique" pour qu'il soit réellement un marqueur de substitution valide à la modélisation directe de la survie globale à partir des données d'essai.

4.3.3 Résultats de l'étude de la survie sans progression (PFS)

L'ajustement des 10 distributions aux données de cancer du poumon prenant en compte les durées de survie sans progression est affiché dans les figures 4.3- 4.6. Une étude non paramétrique utilisant l'estimateur de Kaplan-Meier ainsi que les intervalles de confiance sont aussi disponibles dans les mêmes figures.

L'estimation des paramètres inconnus des 10 modèles qui ont convergés est présentée dans le tableau 4.4 ainsi que les bornes supérieures et inférieures de l'intervalle de confiance de chaque paramètre.

Des tests de conformité sont établis pour tester la supériorité de quelques modèles par rapport à d'autres en termes d'ajustement aux durées de survie sans progression des données de cancer du poumon. Les critères de sélection et de jugement sur les modèles tels que le critère

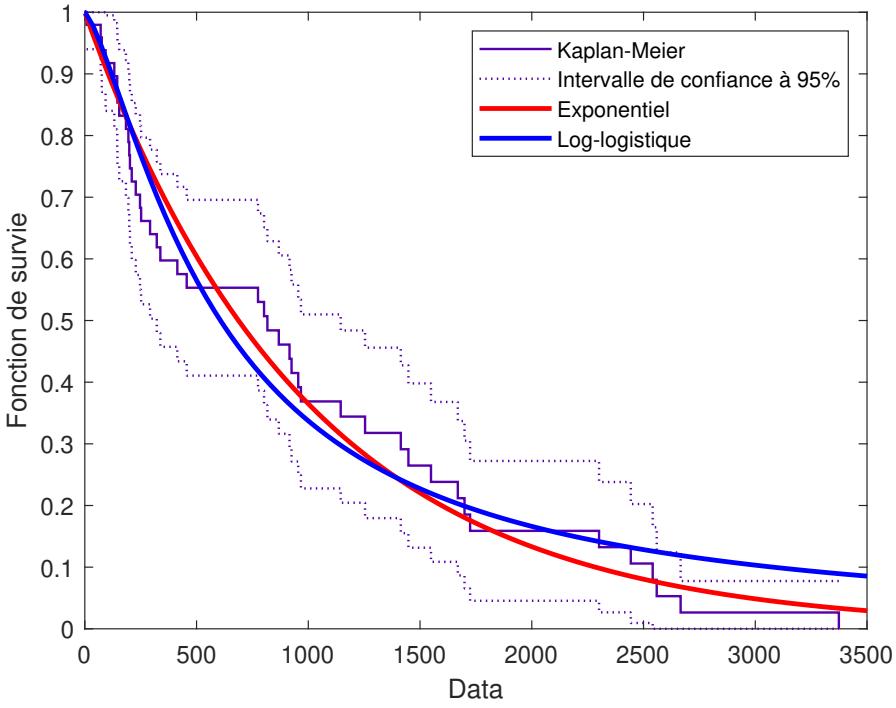


FIGURE 4.3 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution exponentielle et la distribution log-logistique pour les durées de survie sans progression de patients atteints d'un cancer du poumon.

d'information d'Akaike (AIC), le critère d'information bayésien (BIC), le critère d'information cohérent d'Akaike (CAIC), le critère d'information d'Akaike corrigé (AICc) et le critère d'information de Hannan–Quinn (HQIC) sont calculés. Les valeurs de la fonction de vraisemblance et les valeurs des critères d'information sont présentées dans le tableau 4.5.

Les statistiques descriptives des données de cancer du poumon tenant compte des durées de survie sans progression sont résumées dans le tableau 4.6.

4.3.4 Résultats de l'étude de la survie globale

L'ajustement des 10 distributions aux données de cancer du poumon prenant en compte les durées de survie globale est affiché dans les figures 4.7- 4.10. Une étude non paramétrique utilisant l'estimateur de Kaplan-Meier ainsi que les intervalles de confiance sont aussi disponibles dans les mêmes figures.

Le tableau 4.7 présente l'estimation des paramètres inconnus des 10 distributions ainsi que les bornes supérieures et inférieures de l'intervalle de confiance de chaque paramètre.

Des tests de conformité sont établis pour tester la supériorité de quelques modèles par rapport à d'autres en termes d'ajustement aux durées de survie globale des données de cancer

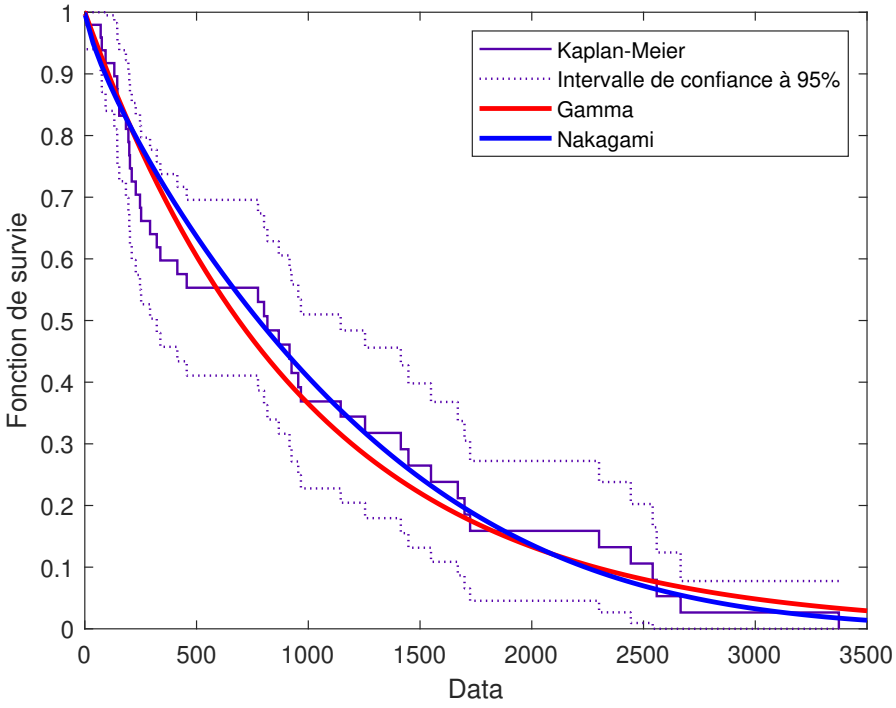


FIGURE 4.4 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution Gamma et la distribution de Nakagami pour les durées de survie sans progression de patients atteints d'un cancer du poumon.

du poumon. Les critères de sélection et de jugement sur les modèles tels que le critère d'information d'Akaike (AIC), le critère d'information bayésien (BIC), le critère d'information cohérent d'Akaike (CAIC), le critère d'information d'Akaike corrigé (AICc) et le critère d'information de Hannan–Quinn (HQIC) sont calculés. Les valeurs de la fonction de vraisemblance et les valeurs des critères d'information sont présentées dans le tableau 4.8.

Les statistiques descriptives des données de cancer du poumon tenant compte des durées de survie globale sont résumées dans le tableau 4.9.

4.3.5 Discussion

En plus de l'objectif habituel de l'ajustement des modèles statistiques aux données de cancer du poumon, on vise également l'estimation des corrélations entre différents points cliniques déterminants.

Si on applique le théorème annoncé dans [89] qui dit que la corrélation entre la PFS et la OS est le rapport entre la durée médiane de la survie sans progression et la durée médiane de la survie globale, on peut l'estimer tel que :

$$Corr(\text{PFS}, \text{OS}) = \frac{med_{\text{PFS}}}{med_{\text{OS}}} = 0.79$$

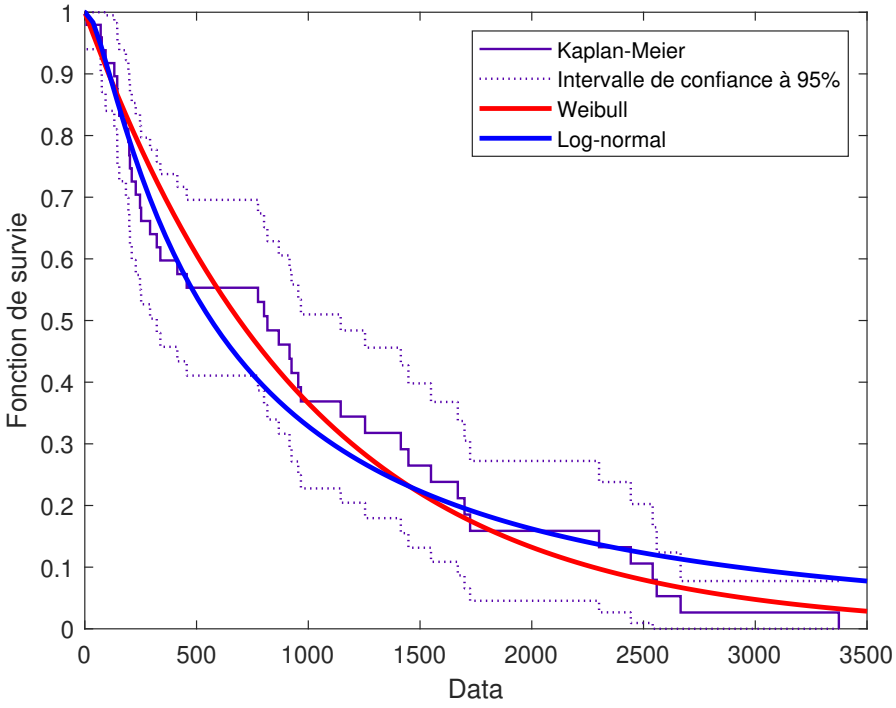


FIGURE 4.5 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution de Weibull et la distribution log-normale pour les durées de survie sans progression de patients atteints d'un cancer du poumon.

En se basant sur ce résultat et sur la figure 4.11 qui affiche les courbes de Kaplan-Meier des durées de survie globale et des durées de survie sans progression, on peut dire que la PFS est fortement corrélée à la OS.

Suivant le cadre de [89], la corrélation entre le PFS et le OS peut aussi être estimée analytiquement à l'aide des estimations par la méthode de maximum de vraisemblance des paramètres des modèles statistiques. Plus de détails pour le modèle de Weibull et le modèle exponentiel sont disponibles dans [90].

Les résultats de l'estimation paramétrique affichés dans les tableaux 4.4 et 4.7 sont bien proches en termes de valeurs. Pour toutes les distributions testées, les valeurs des paramètres estimées, autant pour les durées de survie sans progression que pour les durées de survie globale, ont dans les mêmes ordres de grandeur et leurs intervalles de confiance se chevauchent. Pour les deux paramètres cliniques considérés, les distributions de Weibull, Nakagami, exponentielle et Gamma ont les plus petites valeurs de la fonction log-vraisemblance ce qui pourrait donner un aperçu sur les résultats qu'on aurait avec les critères d'information évalués. Les distributions qui ont donné les plus grandes valeurs de la fonction log-vraisemblance sont : la distribution d'extremum, logistique, t Location Scale et inverse-Gaussienne. Ces résultats sont renforcés par les courbes de survie estimée affichées dans les figures 4.3- 4.6 et 4.7-4.10.

Les valeurs des critères d'information présentées dans les tableaux 4.5 et 4.8 confirment les

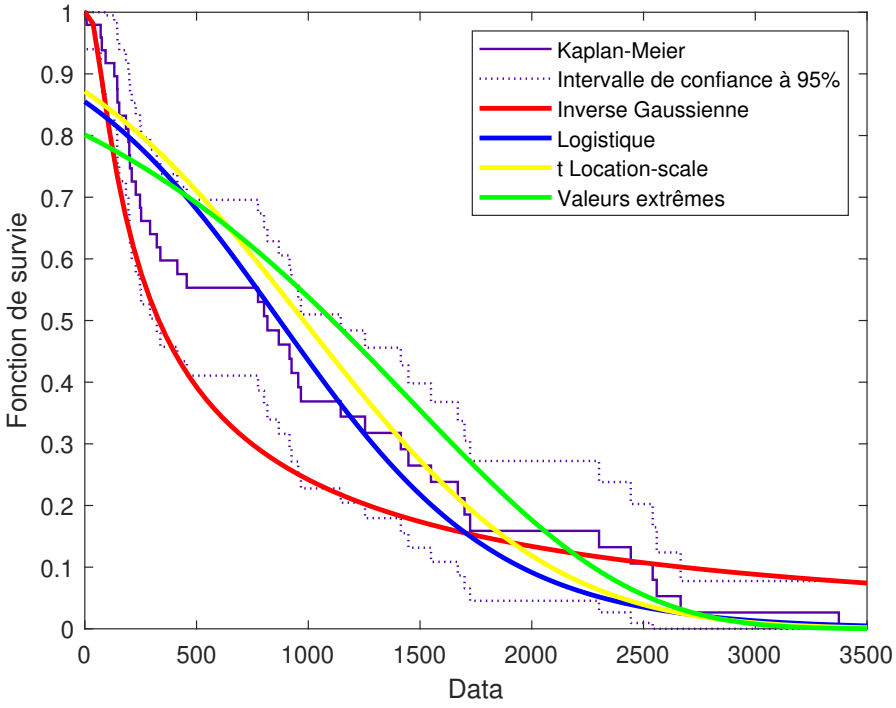


FIGURE 4.6 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution Inverse-Gaussienne, la distribution logistique, la distribution de Student et la distribution d'extremum pour les durées de survie sans progression de patients atteints d'un cancer du poumon.

résultats mentionnés ci-dessus.

Bien que les résultats des critères d'information donnent des valeurs très proches pour les quatre premières distributions, qui sont la distribution Exponentielle, Gamma, Nakagami et Weibull, la loi exponentielle et la distribution exponentielle a les plus petites valeurs des critères d'information AIC, BIC, CAIC, AICc et HQIC dans les deux cas de figure. En se basant sur ces valeurs, on admet que la loi exponentielle est la meilleure loi à décrire les durées de survie sans progression et les durées de survie globale.

Avoir la même distribution à décrire les deux bases de données nous permet de dire qu'au moins, à cette étape, les deux bases de données sont comparables. Il convient d'ajouter à cela qu'avoir des estimations des paramètres très proches, et des intervalles de confiance fortement chevauchés nous permet de dire que le comportement des deux bases de données se ressemble et que les variations des courbes de survie au cours du temps sont très proches. La figure 4.12 illustre les estimations du paramètre de forme de la distribution exponentielle sur une ligne numérique pour les deux cas de figure et met en évidence le chevauchement des intervalles de confiance.

Pour résumer, la distribution qui décrit les durées de survie sans progression est la même distribution qui décrit les durées de survie globale. L'estimation des paramètres dans les deux cas sont très proches. Les deux bases de données sont, dans une certaine mesure, similairement distribuées.

TABLEAU 4.4 – Estimations du maximum de vraisemblance (MLE), la borne inférieure et la borne supérieure de l'IC à 95% des paramètres de quelques distributions pour les durées de survie sans progression de patients atteints d'un cancer du poumon.

Distribution	**Paramètre**	**MLE**	**Intervalle de confiance à** 95%
Exponentielle	$\hat{\alpha}$	**992.047**	(**751.394**, **1370.79**)
Gamma	$\hat{\alpha}$	1.0019	(0.6990, 1.4359)
	$\hat{\beta}$	990.064	(609.968, 1607.01)
Nakagami	$\hat{\alpha}$	0.3828	(0.2763, 0.5303)
	$\hat{\beta}$	1.7602×10^6	$(1.0746 \times 10^6, 2.8833 \times 10^6)$
Weibull	$\hat{\alpha}$	994.602	(733.226, 1349.15)
	$\hat{\beta}$	1.0085	(0.7966, 1.2769)
Log-normale	$\hat{\alpha}$	6.3384	(5.9684, 6.7084)
	$\hat{\beta}$	1.2808	(1.0372, 1.5814)
Log-logistique	$\hat{\alpha}$	6.4085	(6.0327, 6.7843)
	$\hat{\beta}$	0.7391	(0.5812, 0.9400)
Inverse-Gaussienne	$\hat{\alpha}$	1120.54	(297.62, 1943.46)
	$\hat{\beta}$	218.124	(130.412, 305.836)
Logistique	$\hat{\alpha}$	871.672	(620.616, 1122.73)
	$\hat{\beta}$	491.495	(383.948, 629.168)
t Location-Scale	$\hat{\alpha}$	978.777	(721.964, 1235.59)
	$\hat{\beta}$	863.884	(697.926, 1069.3)
	$\hat{\gamma}$	3.27944×10^6	$(2.3726 \times 10^6, 4.5327 \times 10^6)$
Valeurs extrêmes	$\hat{\alpha}$	1464.84	(1162.39, 1767.28)
	$\hat{\beta}$	971.407	(793.119, 1189.77)

De ce point de vue, on peut dire que la survie sans progression n'est pas le meilleur point clinique déterminant tant qu'on peut collecter la survie globale des patients. Mais, dans le cas où il est impossible d'avoir les durées de survie globale, ce qui est un cas fréquent, on pourrait avoir recours aux durées de survie sans progression. L'étude statistique et les prédictions sont ainsi fiables.

D'une façon générale, même s'il y a des preuves solides et cohérentes soutenant une corrélation entre la PFS et la OS, on ne sait pas comment cela devrait être converti en une relation quantifiée. Par conséquent, toute étude qui émet une hypothèse forte concernant la relation entre la PFS et la OS doit être traitée avec prudence et étayée par une explication transparente de la façon dont la relation est quantifiée dans les modèles utilisés. L'étude serait plus performante si elle est accompagnée d'une analyse de sensibilité explorant l'incertitude associée. Cela permettrait aux décideurs de juger de la pertinence de l'étude en se basant sur les preuves disponibles dans le

TABLEAU 4.5 – La valeur de la fonction log-vraisemblance (L) et les critères d'information pour les durées de survie sans progression de patients atteints d'un cancer du poumon pour chacune des distributions.

Distribution	L	**AIC**	**BIC**	**CAIC**	**AICc**	**HQIC**
Exponentielle	**-339.6901**	**681.3802**	**683.2720**	**684.2720**	**679.4618**	**682.0980**
Gamma	-339.6901	683.3802	687.1638	689.1638	680.6251	684.8157
Nakagami	-339.571	683.1420	686.9256	688.9256	680.3869	684.5775
Weibull	-339.688	683.3760	687.1596	689.1596	680.6209	684.8115
Log-normale	-342.406	688.8120	692.5956	694.5956	686.0569	690.2475
Log-logistique	-342.636	689.2720	693.0556	695.0556	686.5169	690.7075
Inverse-Gaussienne	-353.175	710.3500	714.1336	716.1336	707.5949	711.7855
Logistique	-355.331	714.6620	718.4456	720.4456	711.9069	716.0975
t Location-Scale	-354.867	715.7340	721.4095	724.4095	712.2238	717.8873
Valeurs extrêmes	-363.3	730.6	734.3836	736.3836	727.8449	732.0355

Minimum	**Quartile inférieur**	**Médiane**	**Quartile supérieur**	**Maximum**
8	197	611	1334	3374

TABLEAU 4.6 – Le résumé à cinq chiffres de la base de données de cancer du poumon prenant en compte les durées de survie sans progression.

domaine spécifique de la maladie étudiée.

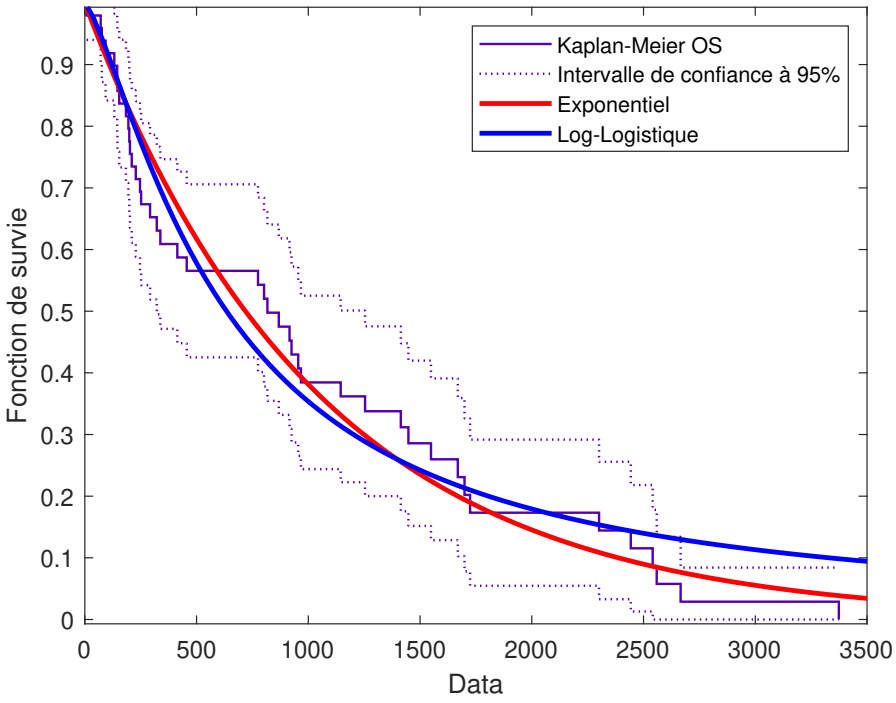


FIGURE 4.7 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution exponentielle et la distribution log-logistique pour les durées de survie globale de patients atteints d'un cancer du poumon.

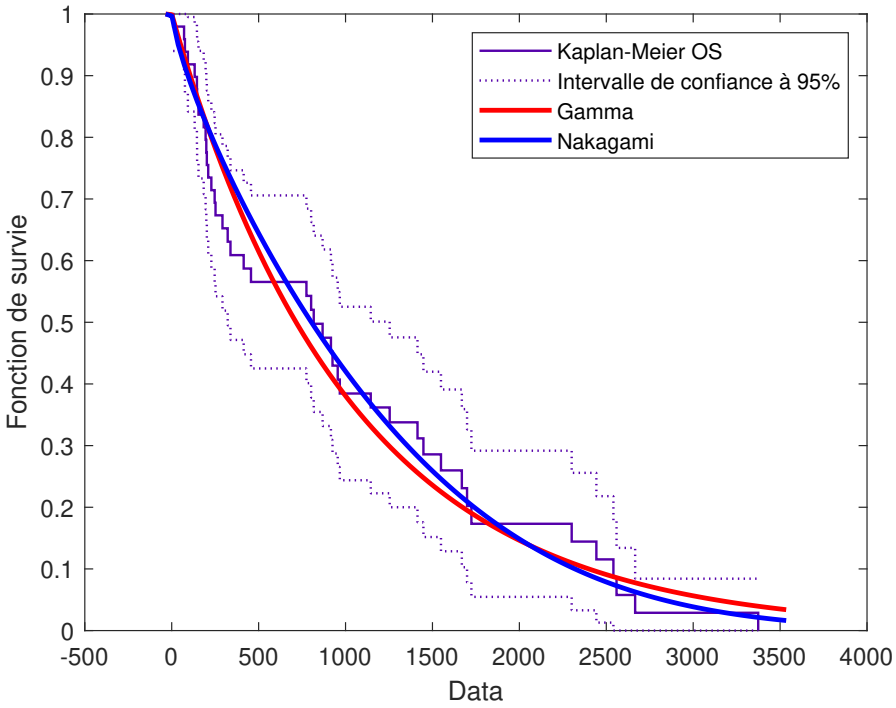


FIGURE 4.8 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution Gamma et la distribution de Nakagami pour les durées de survie globale de patients atteints d'un cancer du poumon.

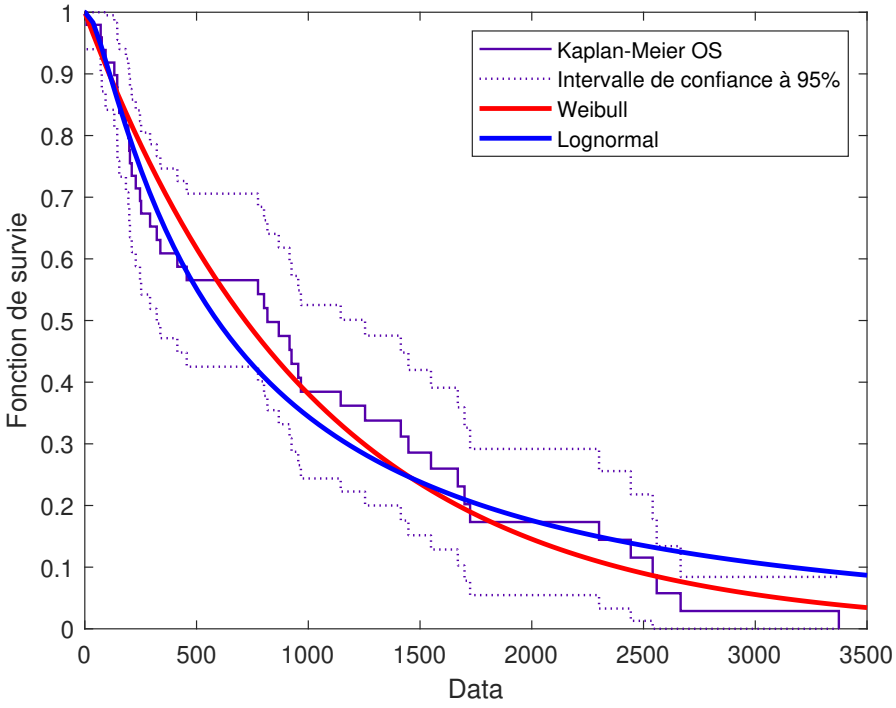


FIGURE 4.9 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution de Weibull et la distribution log-normale pour les durées de survie globale de patients atteints d'un cancer du poumon.

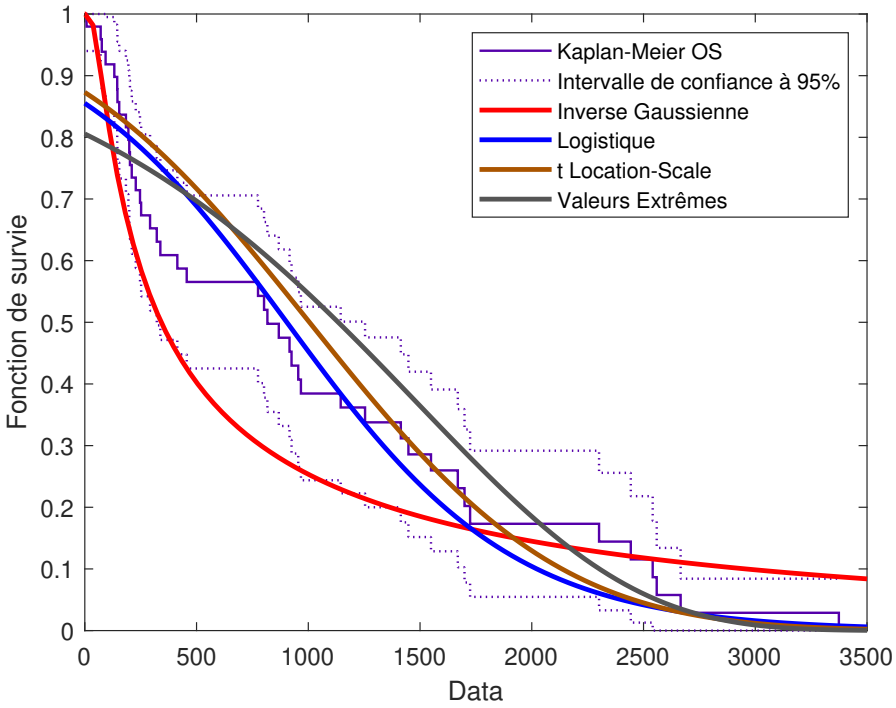


FIGURE 4.10 – L'estimateur de Kaplan-Meier et l'ajustement de la distribution Inverse-Gaussienne, la distribution logistique, la distribution de Student et la distribution d'extremum pour les durées de survie globale de patients atteints d'un cancer du poumon.

TABLEAU 4.7 – Estimations du maximum de vraisemblance (MLE), la borne inférieure et la borne supérieure de l'IC à 95% des paramètres de quelques distributions pour les durées de survie globale de patients atteints d'un cancer du poumon.

Distribution	**Paramètre**	**MLE**	**Intervalle de confiance à** 95%
Exponentielle	$\hat{\alpha}$	**1037**	(**785.443**, **1432.9**)
Gamma	$\hat{\alpha}$	0.9853	(0.6867, 1.4139)
	$\hat{\beta}$	1053.53	(644.674, 1721.69)
Nakagami	$\hat{\alpha}$	0.3810	(0.2748, 0.5281)
	$\hat{\beta}$	1.8882×10^6	$(1.1457 \times 10^6, 3.1118 \times 10^6)$
Weibull	$\hat{\alpha}$	1036.62	(763.141, 1408.12)
	$\hat{\beta}$	0.9986	(0.7861, 1.2687)
Log-normale	$\hat{\alpha}$	6.3836	(6.0089, 6.7584)
	$\hat{\beta}$	1.3062	(1.0558, 1.6159)
Log-logistique	$\hat{\alpha}$	6.4523	(6.0732, 6.8316)
	$\hat{\beta}$	0.7549	(0.5928, 0.9613)
Inverse-Gaussienne	$\hat{\alpha}$	1254.26	(235.561, 2272.96)
	$\hat{\beta}$	217.926	(130.143, 305.708)
Logistique	$\hat{\alpha}$	904.053	(645.632, 1162.47)
	$\hat{\beta}$	509.051	(397.87, 651.3)
t Location-Scale	$\hat{\alpha}$	1005.25	(754.816, 1255.69)
	$\hat{\beta}$	881.514	(715.086, 1086.68)
	$\hat{\gamma}$	3.5125×10^6	$(2.2144 \times 10^6, 5.5717 \times 10^6)$
Valeurs extêmes	$\hat{\alpha}$	1490.83	(1188.85, 1792.8)
	$\hat{\beta}$	973.394	(792.548, 1195.51)

TABLEAU 4.8 – La valeur de la fonction log-vraisemblance (L) et les critères d'information pour les durées de survie globale de patients atteints d'un cancer du poumon pour chacune des distributions.

Distribution	L	**AIC**	**BIC**	**CAIC**	**AICc**	**HQIC**
Exponentielle	**-341.596**	**685.1920**	**687.0838**	**688.0838**	**683.2736**	**685.9098**
Gamma	-341.593	687.1860	690.9696	692.9696	684.4309	688.6215
Nakagami	-341.364	686.7280	690.5116	692.5116	683.9129	688.1635
Weibull	-341.596	687.1920	690.9756	692.9756	684.4369	688.6275
Log-normale	-344.64	693.2800	697.0636	699.0636	690.5249	694.7155
Log-logistique	-344.44	692.8800	696.6636	698.6636	690.1249	694.3155
Inverse-Gaussienne	-355.261	714.5220	718.3056	720.3056	711.7669	715.9575
Logistique	-357.259	718.5180	722.3016	724.3016	715.7629	719.9535
t Location-Scale	-356.524	719.0480	724.7235	727.7235	715.5378	721.2013
Valeurs extrêmes	-364.486	732.9720	736.7556	738.7556	730.2169	734.4075

Minimum	**Quartile inférieur**	**Médiane**	**Quartile supérieur**	**Maximum**
8	206	775	1431	3374

TABLEAU 4.9 – Le résumé à cinq chiffres de la base de données de cancer du poumon prenant en compte les durées de survie globale.

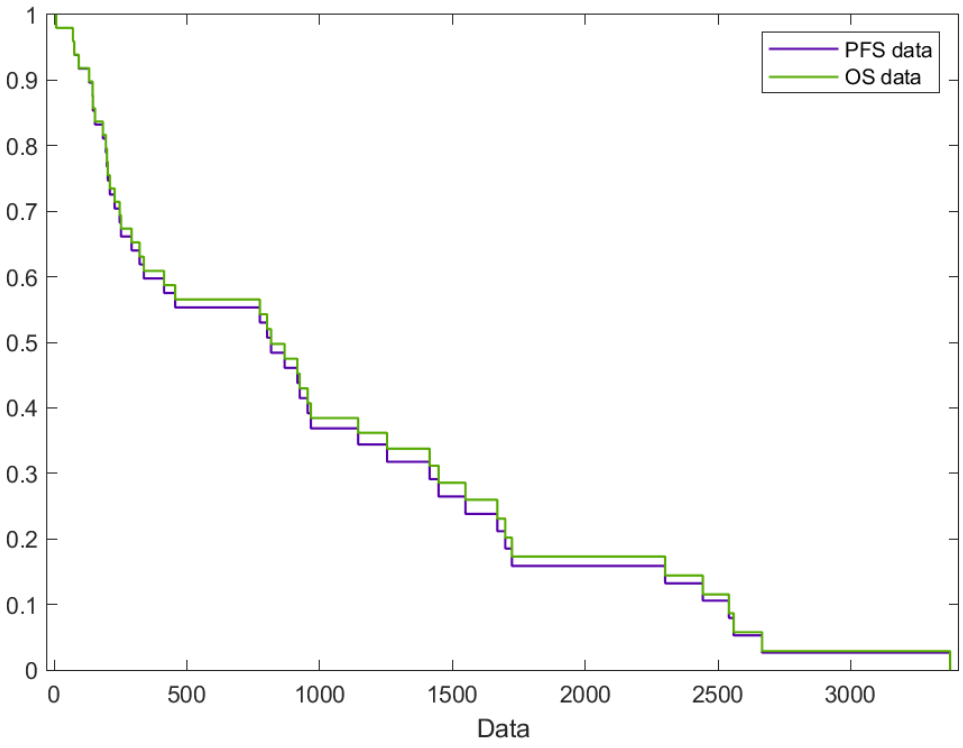


FIGURE 4.11 – Représentation de la différence entre les durées de survie globale et les durées de survie sans progressions dans les données de cancer du poumon en utilisant l'estimateur de Kaplan-Meier

FIGURE 4.12 – Chevauchement des intervalles de confiance du paramètre de forme de la loi exponentielle adaptée, en rouge, aux durées de survie globale et, en vert, aux durées de survie sans progression pour les patients atteints d'un cancer du poumon.

4.4 Conclusion

L'analyse des données de patients atteints d'un cancer du poumon a été présenté dans ce chapitre. Les données ont été organisées de telle sorte que les durées de survie sans progression et les durées de survie globale sont prises en considération. Les critères RECIST 1.1 ont été adoptés pour la prise de décision de la réponse objective des tumeurs pulmonaires solides aux traitements utilisés. Les caractéristiques des deux paramètres cliniques primaires, survie globale et survie sans progression, ont été induites par des techniques d'inférence statistique. 10 distributions ont été ajustées sur les deux populations. L'étude inférentielle est une fenêtre pour voir dans quelle mesure la survie sans progression peut-elle remplacer la survie globale dans l'analyse des données de cancer du poumon. Nous avons trouvé que la survie sans progression est une alternative suffisamment fiable pour être considérée comme point clinique déterminant de l'effet des traitements sur les patients.

Chapitre 5

Modélisation mathématique de la croissance tumorale

5.1 Introduction

Le cancer a fait l'objet de modélisations mathématiques intensives au cours des dernières décennies. Un nombre croissant de modèles mathématiques simulant des systèmes biologiques a été appliqué à la croissance tumorale. Ici, un aperçu de ces modèles, en particulier ceux basés sur des équations différentielles ordinaires, est présenté. De plus, les avantages et les inconvénients des différents modèles, ainsi que la corrélation entre eux, sont mis en évidence. Des cas d'utilisation de modèles de croissance tumorale tenant compte d'une capacité biotique de l'hôte qui varie au cours du temps sont développés. Le potentiel du modèle prédateurs-proies en oncologie est mis en évidence, en supposant que les prédateurs sont des anticorps et les proies sont des antigènes.

5.2 Phénomène biologique

Pour bien comprendre et dériver un modèle mathématique de croissance tumorale, il est indispensable de comprendre le processus biologique. Contrairement aux cellules normales, les cellules cancéreuses se divisent très rapidement et leur accumulation entraîne une croissance cancéreuse, appelée tumeur. Un facteur important qui démontre la malignité du cancer est sa capacité à envahir les tissus ou organes voisins et à voyager à travers les vaisseaux lymphatiques ou les vaisseaux sanguins, ce que l'on appelle en médecine «métastase». La complexité du processus de croissance tumorale est représentée par deux phases principales : avasculaire (sans vaisseaux sanguins) et vasculaire (contenant des vaisseaux qui font circuler des fluides) [92]. Pour proliférer,

les tumeurs ont besoin de nutriments et d'oxygène. En ce qui concerne les tumeurs avasculaires, l'oxygène et les nutriments sont fournis par diffusion à travers les tissus environnants jusqu'à ce que la tumeur atteigne $2-3mm$. Ces suppléments n'atteignent pas le centre de la tumeur. Les cellules cancéreuses ont besoin de leur propre approvisionnement afin de poursuivre la prolifération. C'est la raison pour laquelle de nouveaux vaisseaux sanguins se forment. Le phénomène est appelé «angiogenèse». Cette notion est expliquée à plusieurs reprises dans la littérature [48].

La figure 5.1 montre l'évolution d'une population de cellules cancéreuses. La courbe augmente avec le temps puisque les cellules tumorales se divisent et prolifèrent au fil du temps. La courbe a également un point d'inflexion dans la deuxième phase. Finalement, la courbe converge asymptotiquement vers un volume maximum de la tumeur. Ce rythme est divisé en trois parties : la première phase (1) est l'équilibre entre multiplication et perte de cellules (qui est la phase d'un sujet sain). Le cancer, c'est quand cet équilibre est inaccessible et que les cellules deviennent immortelles. La deuxième phase (2) est lorsque la croissance de la population devient exponentielle avec le développement de nouveaux vaisseaux sanguins, appelée plus tôt angiogenèse, et la connexion avec le système circulaire. La troisième phase (3) est lorsque la tumeur cancéreuse est exposée à un manque d'oxygène et de nutriments (l'apport sanguin est dépassé) et forme un noyau nécrotique. Par conséquent, la croissance tumorale ralentit progressivement jusqu'à atteindre une limite asymptotique.

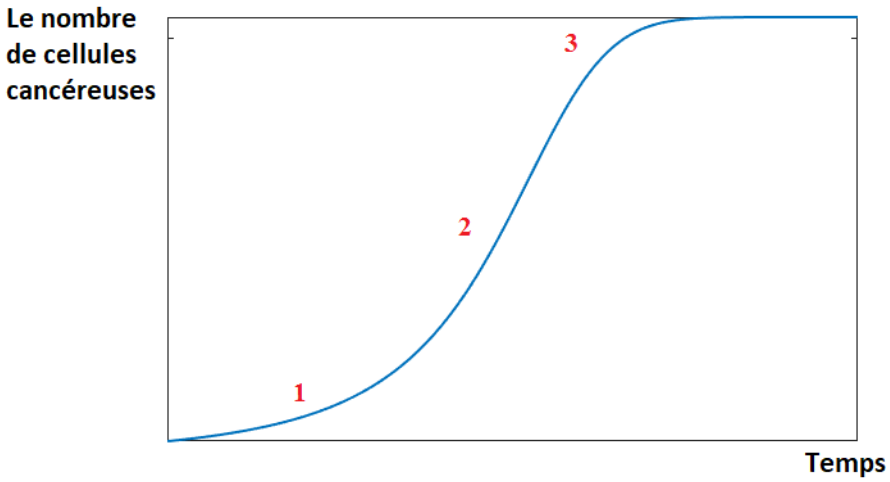


FIGURE 5.1 – Évolution de la population des cellules cancéreuses

En conclusion, cette évolution est plutôt une abstraction mathématique, ce qui conduit au paragraphe suivant qui reprend les différents modèles mathématiques de croissance tumorale.

5.3 Domaines de définition des modèles de croissance tumorale

Le suivi et le traitement du cancer sont aujourd'hui des enjeux majeurs de santé publique. La modélisation mathématique de la croissance tumorale est l'un des thèmes les plus anciens mais toujours actif [93, 48, 94]. Les modèles mathématiques actuels de croissance tumorale peuvent aider dans les interventions guidées et fournir de meilleurs soins aux patients. Les domaines des modèles de croissance tumorale peuvent être : temporels, spatiaux, hybrides, avec système vasculaire, sans système vasculaire, etc. Il existe trois échelles différentes pour lesquelles la croissance tumorale se produit [92] : l'échelle atomique, l'échelle microscopique et l'échelle macroscopique. Pour l'échelle macroscopique, les tumeurs sont considérées comme un organe. Les propriétés sur lesquelles tournent l'échelle macroscopique peuvent être la forme et la morphologie de la tumeur, son invasion, etc. Selon la littérature, les modèles mathématiques à l'échelle macroscopique de la croissance tumorale sont classés comme suit : une catégorie qui considère l'effet de masse (qui est la déformation du tissu environnant) et une catégorie qui ne tient pas compte de l'effet de masse.

Dans la section **??**, nous décrivons les équations différentielles ordinaires dédiées à la modélisation de la croissance tumorale en tenant compte des capacités de charge constantes et dynamiques. La modélisation de la réponse cancéreuse à la chimiothérapie et à l'immunothérapie en utilisant le système prédateurs-proies Volterra-Lotka d'équations différentielles non linéaires de premier ordre sera récapitulée, en supposant que le nombre de cellules tumorales joue le rôle de la proie et que la thérapie joue le rôle du prédateur.

5.4 Modèles de croissance tumorale naturelle

??

5.4.1 Modèles à capacité biotique constante

5.4.1.1 Modèle exponentiel

Le premier modèle auquel on peut légitimement penser pour décrire la croissance tumorale est le modèle exponentiel [48] décrit dans l'équation (5.1) :

$$\begin{cases} \mathrm{P}(t) = \mathrm{P}_0\, \mathrm{e}^{\lambda t} \\ \mathrm{P}(t=0) = \mathrm{P}_0 \end{cases}. \tag{5.1}$$

où P est la prolifération ou la réaction ou la taille de la tumeur au fil du temps, P_0 est la taille de la tumeur lorsque $t = 0$ et $\lambda = \frac{\ln(2)}{T_c}$ où T_c est la durée de vie d'une cellule et λ est le taux de prolifération ou le taux de croissance relatif qui est considéré comme une constante dans ce modèle. Il est à noter que cette fonction satisfait l'équation différentielle ordinaire (5.2) :

$$\frac{dP}{dt} = \lambda P. \tag{5.2}$$

Ce modèle n'est appliqué que dans l'hypothèse suivante :

- Toutes les cellules doivent avoir les mêmes propriétés.
- La mort cellulaire n'est pas prise en compte.
- Il n'y a aucune limitation à la prolifération cellulaire.
- La durée de vie des cellules est considérée comme une fonction constante.

La principale limitation de ce modèle est le fait que le modèle exponentiel est un modèle à taux de croissance constant. Cependant, dans les années 1960, des observations de courbes expérimentales de croissance tumorale sur une longue période ont montré que ce taux de croissance relative n'est pas constant dans le temps. Ainsi, il ne décrit pas la phase terminale de la maladie, en particulier compte tenu du fait que le taux de croissance ralentit à mesure que la taille de la tumeur augmente. Ce phénomène est appelé «nécrose tumorale». Sur la base de cette observation, de nombreux modèles ont été proposés.

5.4.1.2 Modèle logistique de Verhulst

L'idée générale derrière ce modèle est que les cellules ne prolifèrent pas indéfiniment. Plutôt, elles croissent jusqu'à une taille maximale dite «capacité biotique», notée ici K, au-delà de laquelle la population tend à diminuer. En 1838, Verhulst [95] a proposé l'équation différentielle (5.3) comme modèle pour décrire le changement en taille de la population de cellules cancéreuses au fil du temps $\frac{dP}{dt}$:

$$\begin{cases} \frac{dP}{dt} = \theta P\left(1 - \frac{P}{K}\right) \\ P(t=0) = 1\, mm^3 \end{cases}, \tag{5.3}$$

où θ est un coefficient lié à la cinétique de prolifération. Ce modèle a été généralisé en (5.4) avec sa solution explicite et donnée dans l'équation (5.5) comme suit :

$$\begin{cases} \frac{dP}{dt} = \theta P\left(1 - \left(\frac{P}{K}\right)^{\nu}\right) \\ P(t=0) = 1\, mm^3 \end{cases}, \tag{5.4}$$

$$P(t) = \frac{P_0 K}{\left(P_0 + (K^{\nu} - P_0{}^{\nu})\, e^{-\theta \nu t}\right)^{\frac{1}{\nu}}}. \tag{5.5}$$

Le modèle logistique de Verhulst est un modèle de croissance dynamique. Pourtant, il s'agit d'un modèle déterministe qui n'inclut aucune composante stochastique. En conséquence, la fluctuation du niveau de capacité de charge ne peut pas être décrite par ce modèle.

5.4.1.3 Modèle de Gompertz

Plusieurs modèles décrivant la croissance dynamique des cellules tumorales ont été discutés. Parmi ces modèles, il a été démontré que le célèbre modèle de Gompertz tient compte de la décélération de la croissance biologique avec la taille de la population [44]. Ce modèle tire son nom de Benjamin Gompertz, mathématicien autodidacte, qui a introduit la loi de mortalité humaine pour les applications actuarielles en 1825 qui suppose que, à mesure que l'âge augmente, la résistance d'un être humain à la mort diminue (probabilité de survie en fonction de l'âge). Le modèle de Gompertz pour la croissance tumorale s'est avéré particulièrement approprié dans un grand nombre de systèmes expérimentaux, ainsi que dans les données cliniques. Le modèle de Gompertz est défini par l'équation (5.6) :

$$\begin{cases} \frac{d\mathrm{P}}{dt} = \alpha\, \mathrm{e}^{-\beta t}\, \mathrm{P} \\ \mathrm{P}(t=0) = 1\ mm^3 \end{cases}, \tag{5.6}$$

Où α est le taux de prolifération initial (lorsque P = 1 mm^3) et β est le taux de ralentissement en croissance exponentielle de diffusion. La formule analytique (5.7) dérivée de (5.6) est :

$$\mathrm{P}(t) = \mathrm{P}_0\, \mathrm{e}^{\frac{\alpha}{\beta}\left(1-\mathrm{e}^{-\beta t}\right)}. \tag{5.7}$$

Il faut noter que, pour le modèle de Gompertz, le volume tumoral converge vers la capacité biotique comme indiqué dans (5.8) :

$$\lim_{t\to\infty} \mathrm{P}(t) = \mathrm{P}_0\, \mathrm{e}^{\frac{\alpha}{\beta}} = \mathrm{K}. \tag{5.8}$$

L'inconvénient du modèle de Gompertz est le fait que c'est un modèle phénoménologique qui ne décrit que la dynamique de la courbe de croissance et ne repose pas sur une explication biologique du phénomène. Le modèle de Gompertz n'est pas un modèle empirique. En outre, un seul ensemble de paramètres de prolifération est insuffisant pour décrire parfaitement la croissance tumorale. Il y a plusieurs autres variables à prendre en compte telles que le type de cancer, les conditions environnementales, etc.

Après le modèle de Gompertz, un certain nombre de modèles de croissance tumorale basés sur des équations différentielles ordinaires ont été établis. Nous nous concentrons uniquement sur les modèles les plus couramment utilisés en oncologie.

5.4.1.4 Modèle puissance

Des modèles basés sur des considérations métaboliques et des principes énergétiques de base ont également été proposés pour la description de la croissance tumorale comme l'équation proposée par Von Bertalanffy [96] en 1969 donnée par (5.9) :

$$\begin{cases} \frac{d\mathrm{P}}{dt} = a\mathrm{P}^{\sigma} - b\mathrm{P} \\ \mathrm{P}(t=0) = 1\ mm^3 \end{cases} . \tag{5.9}$$

Il est à noter que l'équilibre de la synthèse et de la destruction génère le taux de croissance net. La faible allométrie (c'est-à-dire la désignation conventionnelle en biologie des phénomènes de croissance différentielle des organes) a été prise en considération [97]. La solution analytique de l'équation de Bertalanffy est dans (5.10) :

$$\mathrm{P}(t) = \left(\frac{a}{b} + \left({\mathrm{P}_0}^{1-\sigma} - \frac{a}{b}\right)\mathrm{e}^{-b(1-\sigma)t}\right)^{\frac{1}{1-\sigma}} \tag{5.10}$$

Von Bertalanffy a défini les coefficients a et b comme le coefficient d'anabolisme et le coefficient de catabolisme, respectivement. Dans le cas particulier où $\sigma = \frac{2}{3}$, le modèle est appelé «modèle de Von Bertalanffy à croissance de second type». Dans le cas où le coefficient de catabolisme est nul ($b = 0$), le modèle est appelé «le modèle puissance».

5.4.2 Modèles à capacité biotique dynamique

Indépendamment des détails spécifiques à la tumeur, la croissance tumorale est fortement influencée par l'environnement local. L'organe où la tumeur se développe est considéré comme un modèle déformable. Il est exposé à la pression physique, à l'expansion des membranes basales, à l'augmentation de la masse cellulaire, etc. Cela ajoute de l'importance des zones spatiales dans l'étude des modèles de croissance tumorale.

La diminution du volume de la tumeur peut être provoquée soit en stimulant la mort des cellules cancéreuses, soit en réduisant la taille maximale qu'une tumeur non vascularisée dormante peut atteindre (la capacité biotique). Dans la majorité des modèles de croissance tumorale, la capacité biotique est une valeur fixe [98]. Cependant, dans les contextes pathologiques, la capacité biotique n'est pas constante. Plutôt, elle dépend du temps [97, 99]. En 1999, Hahnfeldt *et al.* [100] ont construit un modèle mathématique qui décrit la croissance tumorale en considérant une capacité biotique variable. Les inhibiteurs de l'angiogenèse empêchent la croissance des cellules cancéreuses et la croissance de nouveaux vaisseaux sanguins. Si l'on néglige l'inhibition de l'angiogenèse, le modèle de capacité biotique dynamique est défini par le système d'équations (5.11) :

$$\begin{cases} \frac{dP}{dt} = \theta P \ln\left(\frac{K}{P}\right) \\ \frac{dK}{dt} = \phi K P^{\frac{2}{3}} \\ P(t=0) = 1\ mm^3,\ K(t=0) = K_0 \end{cases} \tag{5.11}$$

Si on prend en considération l'inhibition de l'angiogenèse, le modèle de capacité biotique dynamique est défini par le système d'équation (5.12) :

$$\begin{cases} \frac{dP}{dt} = \theta P \ln\left(\frac{K}{P}\right) \\ \frac{dK}{dt} = \psi P - \phi K P^{\frac{2}{3}} \\ P(t=0) = 1\ mm^3,\ K(t=0) = K_0 \end{cases} \tag{5.12}$$

Sachant que P et K sont mesurés en mm^3, ψ est une constante positive qui indique le taux de stimulation de l'angiogenèse et ϕ est une constante positive qui indique le taux d'inhibition de l'angiogenèse. Le phénomène décrit par ce modèle est l'accélération initiale de la croissance tumorale et l'effet inhibiteur lorsque la taille de la population augmente.

Après avoir décrit théoriquement la croissance tumorale quantitative et expliqué sa dynamique, Hahnfeldt *et al.* [100] ont introduit la notion de capacité biotique variable sous contrôle angiogénique et l'ont comparée à la capacité biotique statique (ou au niveau de subsistance fixe de la tumeur). Ils ont construit un modèle mathématique qui représente le développement de la tumeur après un traitement anti angiogénique. Ce modèle se caractérise par une paramétrisation minimale.

Le modèle mathématique standard de Hahnfeldt *et al.* [100] a été utilisé seize ans plus tard pour établir un lien entre la dynamique tumorale et l'âge de l'hôte [101]. Ils ont prouvé que la progression du cancer peut être déterminée en utilisant le vieillissement comme un hub réglementaire, c'est-à-dire que la dynamique tumorale dépend fortement de l'âge de l'hôte. Ils ont estimé les quatre paramètres θ, ψ , ϕ et k_0 en utilisant des souris comme hôtes.

5.5 Modèles de croissance tumorale décrivant l'effet thérapeutique

5.5.1 Modèles à capacité biotique dynamique pour l'évaluation de l'effet thérapeutique

Wilkie *et al.* [102] ont prouvé que la progression du cancer ou la réponse au traitement peut être déterminée par la sensibilité des cellules tumorales aux signaux environnementaux à médiation immunitaire. Ceci a été réalisé en utilisant un modèle logistique généralisé modifié qui permet de prédire l'état dormant des tumeurs. Les paramètres de ce modèle ont été estimés par un algorithme de Monte-Carlo par chaîne de Markov à l'aide de mesures expérimentales.

Kareva [103] a proposé une théorie sur la réponse de l'hôte à l'ablation de la tumeur primaire en disant que la progression rapide des métastases après la résection des tumeurs primaires peut être due à la diminution de la quantité de stimulateurs de l'angiogenèse et au soulagement de l'inhibition de la tumeur secondaire. Le modèle mathématique (5.13) a été conçu afin d'évaluer cette hypothèse. Ce modèle de croissance cubique simple prend en compte le taux de croissance λ, la capacité biotique avant et après la formation d'un nouveau système vasculaire (k et $\mathrm{K}(t)$ respectivement) et le seuil de viabilité m de la tumeur reséquée tout en considérant la dynamique de la tumeur :

$$\frac{d\mathrm{P}}{dt} = \lambda \mathrm{P}(t)\left(\frac{\mathrm{P}(t)}{m} - 1\right)\left(1 - \frac{\mathrm{P}(t)}{k + \mathrm{K}(t)}\right). \tag{5.13}$$

Hutchinson *et al.* [104] ont développé un modèle de croissance des tumeurs vasculaires en utilisant des données de tailles longitudinales des seins. Ce modèle prédit la coordination la plus efficace de la chimiothérapie et de la radiothérapie avec la thérapie antiangiogénique. S'inspirant du modèle de capacité biotique dynamique de Hahnfeldt susmentionné, les auteurs ont considéré que la capacité biotique tumorale dépend de la densité et l'architecture vasculaire locale. Contrairement à d'autres travaux qui ne considèrent que la dépendance au temps, la progression tumorale dans ce modèle dépend de la densité vasculaire. Pour les tumeurs non traitées, le modèle mathématique de la croissance tumorale est donné par l'équation (5.14) suivante :

$$\begin{cases} \frac{d\mathrm{P}}{dt} = \theta \mathrm{P}\left(1 - \frac{\mathrm{P}}{\mathrm{K}}\right) \\ \frac{d\mathrm{K}}{dt} = \sigma \mathrm{P}^{\beta} \mathrm{K}^{\mu} \end{cases}, \tag{5.14}$$

Où σ est le taux de croissance de la capacité biotique vasculaire, β et μ des valeurs fixées physiologiquement comme suit : $\beta = \frac{2}{3}$ et $\mu = 1 - \beta = \frac{1}{3}$.

Pour les tumeurs traitées par thérapie antiangiogénique, les auteurs ont incorporé les effets thérapeutiques au modèle (5.14) conçu pour la croissance tumorale non traitée. Le modèle non linéaire à effets mixtes est alors donné par l'équation (5.15) suivante :

$$\begin{cases} \frac{dP}{dt} = \theta P\left(1 - \frac{P}{KN}\right) \\ \frac{dK}{dt} = \sigma P^{\beta} K^{\mu} - \zeta K \\ N(t) \end{cases}, \tag{5.15}$$

avec

$$N(t) = \begin{cases} 1 & \text{for} \quad t \leq t_{n_1} \quad \text{and} \quad t \geq t_{n_2} \\ N_{max} & \text{for} \quad t_{n_1} < t < t_{n_2} \end{cases},$$

où ζ représente le taux de mort des vaisseaux, t_{n_1} et t_{n_2} sont respectivement les temps de début et de fin de croissance tumorale et N est le facteur par lequel K est augmenté.

Il est à noter que la prise en compte de l'environnement tumoral est d'importance majeure (les immunothérapies anti-Pd-1 ou anti-PdL-1, par exemple, interagissent avec l'environnement et pas forcément directement avec les cellules tumorales) [105].

5.5.2 Système prédateur-proie Volterra-Lotka

Le modèle de Volterra-Lotka tire son nom de : Alfred James Lotka, mathématicien, chimiste et démographe, qui a introduit le système d'équations (5.16) en 1910 pour la modélisation de réactions chimiques hypothétiques [106]. Et Vito Volterra, mathématicien et physicien, qui a appliqué le même modèle en 1926 en biologie mathématique, afin de décrire les niveaux oscillatoires des prises de poissons dans le nord de la mer Adriatique [107, 108].

$$\frac{dN}{dt} = aN - bPN \tag{5.16a}$$

$$\frac{dP}{dt} = cNP - dP, \tag{5.16b}$$

Où $N(t)$ est la population de proies, $P(t)$ est la population de prédateurs et a, b, c et d sont des constantes positives.

- aN : représente la croissance de la proie (Malthusienne) en l'absence du prédateur.
- $-bPN$: représente l'effet du prédateur sur la croissance de la population de proies qui diminue en présence du prédateur.

- cNP : représente l'effet de la proie sur la croissance de la population de prédateurs qui augmente en présence de la proie.
- $-d$P : représente la diminution exponentielle de la croissance de la population de prédateurs en l'absence de proie.

Le concept de modèle Volterra-Lotka prédateurs-proies a été largement utilisé dans diverses applications [109, 110, 111].

5.5.3 Simulation de la croissance tumorale avec le modèle prédateur-proie

Les équations prédateurs-proies décrivent un problème d'interaction entre deux populations. Lutte pour l'existence, la population de prédateurs se nourrit de proies pour grandir. Ce concept a été adapté au domaine médical permettant à certains chercheurs de simuler le mécanisme immunitaire et notamment l'immunité antitumorale avec le modèle prédateur-proie depuis 1973 avec George I. Bell [112]. Il est supposé que les prédateurs sont des anticorps et que les proies sont des antigènes, Bell a ouvert la voie à un domaine d'étude très prometteur.

5.5.3.1 Modèle prédateur-proie et immunothérapie

L'immunothérapie est l'une des nombreuses solutions proposées pour le traitement du cancer. Aussi connue sous le nom de thérapie biologique car son objectif principal est de renforcer les défenses naturelles du corps humain pour le préparer à arrêter la croissance des cellules cancéreuses, arrêter les métastases et détruire les cellules cancéreuses en stimulant les cellules effectrices. Même si l'immunothérapie est efficace contre le cancer, elle est inefficace contre les tumeurs ayant atteint des tailles importantes.

Considérant le système immunitaire comme étant une seule population : Étant mathématiquement intéressante, l'immunothérapie et son effet sur la croissance tumorale font l'objet d'investigations actives depuis de nombreuses années maintenant [113, 111, 106, 92, 114, 115]. Considérant le système immunitaire comme une seule population qui est la population de cellules effectrices, Mishkin [114] a adopté une application directe du modèle prédateur-proie. En supposant que la proie $\mathrm{N}(t)$ est le nombre de cellules tumorales et que le prédateur $\mathrm{P}(t)$ est le nombre de cellules effectrices du système immunitaire.

Le même modèle de Volterra-Lotka (5.16) est adopté avec quelques modifications dans la deuxième équation du système. La signification des termes du système a été adaptée à la mise en

œuvre de l'immunothérapie dans (5.16a) du modèle comme suit :

- aN : représente la croissance exponentielle de la tumeur en l'absence de l'effet des cellules effectrices sur la tumeur en supposant que la constante positive a est le taux de croissance continue de la tumeur.
- $-b$PN : représente l'effet des cellules effectrices sur la destruction de la population de cellules tumorales en supposant que la constante positive b représente l'efficacité des cellules effectrices à tuer les tumeurs.

Pour l'équation (5.16b) du système, une légère modification du troisième paramètre c a été apportée pour modéliser le fait qu'au contact des cellules tumorales, les cellules effectrices libèrent des signaux pour stimuler le corps afin de renforcer la production de cellules effectrices. À partir des travaux fondateurs de Kuznetsov *et al.* [111], Mishkin a remplacé c par $c = \frac{e}{(f+\mathrm{N})}$ où, e et f sont des constantes positives. L'équation (5.16b) devient (5.17) :

$$\frac{d\mathrm{P}}{dt} = \frac{e}{f+\mathrm{N}}\mathrm{NP} - d\mathrm{P}. \tag{5.17}$$

La signification des termes de l'équation a été adaptée à l'implémentation de l'immunothérapie dans l'équation (5.17) comme suit :

- $\frac{e}{f+\mathrm{N}}$NP : représente l'effet de la population de cellules tumorales sur la croissance de la population de cellules effectrices qui augmente en présence des cellules tumorales tout en considérant que lorsque le nombre de cellules tumorales approche de l'infini, ce terme se rapprochera de eN.
- $-d$P : représente la diminution exponentielle de la croissance de la population de cellules effectrices en l'absence de la cellule tumorale.

Ce modèle, étant une analogie « directe » entre l'interaction naturelle prédateur-proie et l'interaction naturelle des cellules effectrices tumorales, présente de nombreuses limites. Pour n'en citer que quelques-unes : d'abord, l'immunothérapie est considérée dans ce modèle comme une solution parfaite qui détruit les cellules tumorales puisque la croissance des cellules effectrices et la croissance tumorale sont considérées comme proportionnelles, ce qui n'est clairement pas le cas comme mentionné précédemment. Ainsi, le modèle révèle un manque de réalisme frappant. Mais cela ouvre la voie à des efforts continus pour modéliser l'interaction entre le cancer et l'immunothérapie.

Considérant le système immunitaire comme deux populations : Sarkar et Banerjee [116] ont également proposé un modèle de type prédateur-proie pour la dynamique d'interaction tumeur-système immunitaire. En supposant que les cellules tumorales jouent le rôle de proie et que les cellules immunitaires (Lymphocytes T Cytotoxiques (LTC) et macrophages) jouent le rôle de prédateurs, Sarkar et Banerjee ont considéré que les cellules immunitaires ont deux états : l'état de chasse, où les LTC reçoivent des informations sur la proie, et l'état de repos, où les LTC stockent les

informations reçues et engloutissent les cellules cancéreuses. Le modèle proposé est donné par l'équation (5.18) :

$$\frac{d\mathrm{N}}{dt} = q + r\mathrm{N}\left(1 - \frac{\mathrm{N}}{k_1}\right) - \alpha \mathrm{N}\mathrm{P}_\mathrm{H} \tag{5.18a}$$

$$\frac{d\mathrm{P}_\mathrm{H}}{dt} = \beta \mathrm{P}_\mathrm{H}\mathrm{P}_\mathrm{R} - d_1 \mathrm{P}_\mathrm{H}, \tag{5.18b}$$

$$\frac{d\mathrm{P}_\mathrm{R}}{dt} = s\mathrm{P}_\mathrm{R}\left(1 - \frac{\mathrm{P}_\mathrm{R}}{k_2}\right) - \beta \mathrm{P}_\mathrm{H}\mathrm{P}_\mathrm{R} - d_2 \mathrm{P}_\mathrm{R}, \tag{5.18c}$$

Où $\mathrm{N}(t)$ est le nombre de cellules tumorales, $\mathrm{P}_\mathrm{H}(t)$ est le nombre de cellules de prédateur chassant et $\mathrm{P}_\mathrm{R}(t)$ est le nombre de cellules de prédateur au repos.

- r : une constante positive qui représente le taux de croissance des cellules cancéreuses proies.
- q : une constante fixe positive qui représente le taux de conversion des cellules saines en cellules tumorales.
- k_1 : est la capacité biotique maximale des cellules tumorales proies ($k_1 > k_2$).
- k_2 : est la capacité biotique maximale des cellules de prédateurs au repos.
- α : une constante positive qui représente le taux de prédation/destruction des cellules tumorales par la chasse aux cellules.
- β : une constante positive qui représente le taux de conversion des cellules au repos en cellules de chasse.
- d_1 : une constante positive qui représente le taux de mortalité naturelle des cellules de chasse ou le taux d'apoptose.
- d_2 : une constante positive qui représente le taux de mortalité naturelle des cellules au repos ou le taux d'apoptose.
- s : une constante positive qui représente le taux de croissance des cellules au repos.

Kaur *et al.* [113], inspiré par Sarkar et Banerjee, et El-Gohary [116, 117], a modifié le modèle de Volterra-Lotka prédateur-proie (5.18) mentionné ci-dessus et plus spécifiquement les équations (5.18b) et (5.18c) du modèle (5.19) :

$$\frac{d\mathrm{N}}{dt} = q + r\mathrm{N}\left(1 - \frac{\mathrm{N}}{k_1}\right) - \alpha_1 \mathrm{N}\mathrm{P}_\mathrm{H}, \tag{5.19a}$$

$$\frac{d\mathrm{P}_\mathrm{H}}{dt} = \beta \mathrm{P}_\mathrm{H}\mathrm{P}_\mathrm{R} - d_1 \mathrm{P}_\mathrm{H} - \alpha_2 \mathrm{P}_\mathrm{H}\mathrm{N}, \tag{5.19b}$$

$$\frac{d\mathrm{P}_\mathrm{R}}{dt} = s\mathrm{P}_\mathrm{R}\left(1 - \frac{\mathrm{P}_\mathrm{R}}{k_2}\right) - \beta \mathrm{P}_\mathrm{H}\mathrm{P}_\mathrm{R} - d_2 \mathrm{P}_\mathrm{R} + \frac{\rho \mathrm{P}_\mathrm{R}\mathrm{N}}{\mathrm{T} + \nu}. \tag{5.19c}$$

Des modifications ont été introduites dans le modèle en ajoutant les termes : $-\alpha_2 \mathrm{P}_\mathrm{H}\mathrm{N}$ dans l'équation (5.19b) et $\frac{\rho \mathrm{P}_\mathrm{R}\mathrm{N}}{\mathrm{T}+\nu}$ dans l'équation (5.19c), où :

- α_1 : une constante positive qui représente le taux de prédation/destruction des cellules tumorales par la chasse aux cellules.
- α_2 : une constante positive qui représente le taux de prédation/destruction des cellules de chasse par les cellules tumorales.
- ρ : représente le taux de prolifération des cellules au repos.
- ν : représente la demi-saturation pour le terme de prolifération.

5.5.3.2 Modèle prédateur-proie et chimiothérapie

La chimiothérapie est l'une des nombreuses solutions proposées pour le cancer. La chimiothérapie empêche les cellules de se développer, soit en les tuant, soit en les empêchant de se diviser. Parmi les divers effets secondaires néfastes de la chimiothérapie, il y a le fait qu'elle détruise également les cellules effectrices. À notre connaissance, il n'existe aucun modèle officiel ou validé décrivant l'effet de la chimiothérapie sur la croissance du cancer à l'aide du modèle prédateur-proie de Volterra-Lotka [118]. Pourtant, certaines tentatives ont été développées dans le travail de Mishkin [114]. Comme mentionné précédemment, la chimiothérapie attaque à la fois les cellules tumorales et les cellules effectrices du système immunitaire. Les travaux de Mishkin [114] ont considéré que le prédateur est le produit chimiothérapeutique et que la proie est à la fois des cellules tumorales et des cellules effectrices. Le système à trois équations proposé (en négligeant la décomposition et la croissance logistique de la tumeur) est donné par l'équation (5.20) :

$$\frac{d\mathrm{N}}{dt} = a\mathrm{N} - b\mathrm{CN}, \tag{5.20a}$$

$$\frac{d\mathrm{P}}{dt} = p + c\mathrm{NP} - d\mathrm{P}, \tag{5.20b}$$

$$\frac{d\mathrm{C}}{dt} = v(t) - g\mathrm{C}, \tag{5.20c}$$

Où le nombre de cellules tumorales N et le nombre de cellules effectrices P sont la population de proies tandis que la concentration du produit chimiothérapeutique C est la population de prédateurs.

- p : représente le nombre de cellules effectrices produites par jour.
- $v(t)$: représente la quantité de drogue injectée à un moment donné.
- représente le taux de décroissance de la concentration du produit chimiothérapeutique.

Les auteurs ont remplacé la fonction cNM par $v(t)$ pour rendre le modèle plus réaliste. Sinon, le modèle considérera que le nombre de cellules tumorales influe sur la concentration du produit chimiothérapeutique.

5.6 Conclusion

Dans cette dernière partie nous avons présenté un aperçu des différents modèles mathématiques basés sur des équations différentielles ordinaires pour décrire la croissance tumorale. Les cas d'utilisation et les limites de ces modèles sont mis en évidence. Nous avons fourni les différents modèles mathématiques qui considèrent une capacité biotique constante pour souligner le modèle de Hahnfeldt qui considère une capacité biotique dynamique. Un résumé de l'immunité antitumorale simulant Volterra-Lotka prédateurs-proies est présenté. On peut prévoir l'importance de ces modèles pour décrire avec précision l'évolution tumorale. Nous visons par ce résumé à ouvrir la voie à de futures tentatives de modélisation des cellules cancéreuses.

Conclusion générale et perspective

Cette thèse s'inscrit dans le cadre de la problématique globale portant sur l'analyse des données médicales, en particulier, celles de survie. Le travail ne se limite pas à une maladie bien déterminée. Au contraire, nous avons élargi notre application à des maladies cancéreuses, neurologiques, infectieuses etc. Nous avons pour but d'apporter aux praticiens hospitaliers (en oncologie, neurologie et virologie) une aide à la décision médicale et une vision globale des patients pris en charge. En prenant appui sur des méthodes rigoureuses, les techniques robustes de biostatistique et de modélisation biomathématique, permettent aux médecins de formuler des hypothèses physiopathologiques, de randomisation et d'essais thérapeutiques.

Pour ce faire, nous avons traité des problèmes de modélisation de données médicales en utilisant le concept des distributions défectueuses qui permettent de, non seulement décrire le comportement des données de survie, mais aussi de prédire et quantifier la présence éventuelle d'une fraction survivante. Nous avons travaillé de façon à explorer les distributions bien fondées de la littérature pouvant être utilisée pour la modélisation de données de survie avec un taux de guérison. Ces distributions sont à être utilisées d'une manière compétitive. Des nouvelles distributions très flexibles ont été proposées. Leur surperformance a été démontrée à l'aide des techniques d'inférence statistique et de tests d'ajustement paramétriques et non paramétriques. Les particularités des populations, notamment, lorsque le risque de décès variable au cours du temps est envisagé selon les catégories de patients en cours d'étude, ont été adaptées en utilisant des méthodes de régression linéaire, des méthodes non paramétriques comme Kaplan-Meier, et des méthodes semi-paramétriques comme la régression de Cox à travers des exemples réels. Le phénomène de censure, habituellement rencontré dans l'analyse de survie, a été aussi traité.

Des résultats intéressants sur l'effet de l'utilisation du riluzole et d'autres traitements sur la survie des patients traités pour la sclérose latérale amyotrophique sont obtenus. L'étude inférentielle proposée pour l'analyse des données de patients qui ont souffert de Coronavirus-19 a donné des résultats prometteurs quant à l'ajustement des distributions statistiques à la maladie infectieuse. Une étude bibliographique approfondie sur les réponses objectives permettant l'évaluation des changements tumoraux à partir des images médicales a été présentée. Une étude, reposant sur une base de données primaires récemment collectée de l'Institut Salah Azaiez de Tunis, a été menée. Le but de ces études est double : d'une part, on veut mettre en évidence l'effet de substitution

du paramètre clinique primaire, qui est la survie globale, par une de ses alternatives, qui est la survie sans progression. Et d'autre part, on souhaite examiner cet effet d'une perspective purement mathématique en utilisant la statistique inférentielle. Cette thèse, étant une étude mathématique des développements des maladies, il est primordial d'y faire intégré les modèles mathématiques de croissance tumorale. Un résumé des équations différentielles ordinaires décrivant l'évolution tumorale a été présenté. Ce résumé surligne un déficit important dans la littérature quant à la simulation de la croissance tumorale avec le modèle Prédateur-Proie et la chimiothérapie.

Cela nous conduit aux perspectives envisagées. Tout d'abord, le déficit mentionnée ci-dessus, peut être comblé par la proposition de quelques changements dans le modèle dans le but de mieux décrire l'effet de la chimiothérapie sur les patients atteints d'un cancer. Nous envisageons quelques pistes pour poursuivre le développement de la théorie des distributions défectueuses. On pourrait adopter les techniques d'estimation Bayesienne pour évaluer les paramètres inconnus des modèles ce qui pourrait conduire à une meilleure prédiction du taux de guérison. L'algorithme espérance-maximisation pourrait aussi aboutir à des meilleurs résultats quant à l'estimation des paramètres de maximum de vraisemblance. Aussi comme perspective aux travaux effectués dans le chapitre 4, la quantification de la relation entre la survie sans progression et la survie globale pourrait être révolutionnaire dans le domaine de l'analyse de données.

On pourrait aussi penser à une reparamétrisation qui réduirait le nombre de paramètres des modèles proposés en gardant leur efficacité. Un modèle a été proposé dernièrement appelé la distribution de Gompertz inverse [119]. Il serait intéressant d'appliquer le concept des modèles défectueux à cette nouvelle distribution et de comparer les résultats. Une quantification du gain apporté par le concept des modèles défectueux par rapport aux modèles de mélange pourrait aussi être faite pour voir ce qu'on gagne lorsqu'on réduit un paramètre dans le modèle de mélange. Une étude comparative pourrait aussi être menée entre les différentes méthodes de généralisation résumée dans le premier chapitre. On peut également appliquer les modèles proposés sur des données tronquées et des données censurées à gauche ou par intervalle pour voir s'ils sont suffisamment flexibles pour les décrire. La modélisation multi-états et la loi de probabilité conditionnelle pourraient être appliquées sur les données de cancer du poumon utilisées dans le chapitre 4 pour mieux modéliser la stabilité-progression-décès des patients.

Annexe A

Propriétés du modèle MO-GDGD

Dans cette partie, nous allons exposer quelques propriétés du modèle MO-GDGD introduit à la page 30.

A.1 Stabilité

Théorème A.1.1. *L'extension par Marshall-Olkin a la propriété de stabilité; quand on étend une distribution par Marshall-Olkin plus qu'une fois, on n'obtient rien de nouveau.*

Démonstration. Si $S(t)$ est la fonction de survie d'une distribution D, qu'elle soit propre ou défectueuse, $S^*(t)$ est la fonction de survie de l'extension de la distribution D par Marshall-Olkin, et $S^{**}(t)$ est la fonction de survie de la double extension de D par Marshall-Olkin, alors, $\forall -\infty < t < +\infty, \forall r > 0$, l'expression de $S^*(t)$ et $S^{**}(t)$ sont données par :

$$S^*(t) = \frac{rS(t)}{1-(1-r)S(t)} \quad \text{et} \quad S^{**}(t) = \frac{rS^*(t)}{1-(1-r)S^*(t)}.$$

En remplaçant $S^*(t)$ par son expression dans $S^{**}(t)$, on obtient :

$$S^{**}(t) = \frac{r(\frac{rS(t)}{1-(1-r)S(t)})}{1-(1-r)(\frac{rS(t)}{1-(1-r)S(t)})}.$$

À travers un calcul simple, on peut voir que :

$$S^{**}(t) = \frac{r^2S(t)}{1-(1-r^2)S(t)}.$$

En substituant r^2 par le paramètre de forme strictement positif r', on obtient la fonction de survie $S(t)$ de l'extension de la distribution par Marshall-Olkin :

$$S^{**}(t) = \frac{r'S(t)}{1-(1-r')S(t)},$$

ce qui complète la démonstration. □

A.2 Quantiles

Théorème A.2.1. *Les quantiles t_q de MO-GDGD $(t; r, \alpha, \beta, \gamma)$ sont donnés par :*

$$t_q = -\frac{1}{\beta}\ln\left(\frac{\beta}{\alpha}\ln\left(1-\left(\frac{qr}{1+q(r-1)}\right)^{1/\gamma}\right)+1\right), \quad 0<q<1. \tag{A.1}$$

Démonstration. Le quantile d'ordre q d'une variable aléatoire T est la valeur t_q tel que $q = P(T \leq t_q) = F(t_q), \quad t_q > 0$. En utilisant la fonction de répartition de MO-GDGD 2.5 et les abréviations introduites dans 2.8, on peut écrire :

$$q = F(t_q) = \frac{B^{\gamma}(t_q)}{(1-r)B^{\gamma}(t_q)+r}, \quad 0<q<1.$$

sachant que :

$$B(t_q) = \left(\frac{qr}{1-q(1-r)}\right)^{\frac{1}{\gamma}}.$$

En remplaçant $B(t_q)$ par son expression, on a :

$$e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)} = 1-\left(\frac{qr}{1-q(1-r)}\right)^{\frac{1}{\gamma}}.$$

En appliquant le logarithme deux fois des deux côtés, on obtient :

$$e^{-\beta t}-1 = \frac{\beta}{\alpha}\ln\left(1-\left(\frac{qr}{1-q(1-r)}\right)^{\frac{1}{\gamma}}\right).$$

d'où

$$(t_q)_{\mathrm{MO-GDGD}} = -\frac{1}{\beta}\ln\left(\frac{\beta}{\alpha}\ln\left(1-\left(\frac{qr}{1-q(1-r)}\right)^{\frac{1}{\gamma}}\right)\right),$$

ce qui complète la démonstration. □

A.3 Médiane

Corollaire A.3.1. *La médiane A.2 de la distribution MO-GDGD est obtenu en posant $q = \frac{1}{2}$ dans l'équation A.1.*

$$(M\acute{e}diane)_{\mathrm{MO-GDGD}} = -\frac{1}{\beta}\ln\left(\frac{\beta}{\alpha}\ln\left(1-\left(\frac{r}{1+r}\right)^{\frac{1}{\gamma}}\right)\right). \tag{A.2}$$

A.4 Mode

Dans l'application réelle, il est important de connaître la valeur la plus probable de la variable aléatoire. Le mode de la distribution de Gompertz défectueuse et généralisée selon Marshall-Olkin est obtenu en calculant la dérivée par rapport à t de la fonction densité de probabilité donnée dans l'équation 2.4 pour avoir A.3 :

$$f'(t) = -f(t)[\beta + \alpha e^{-\beta t} - \alpha(\gamma - 1)e^{-\beta t}\frac{A(t)}{B(t)} + 2\gamma\alpha(1-r)e^{-\beta t}\frac{A(t)B^{\gamma-1}(t)}{C(t)}]. \tag{A.3}$$

Le mode est la solution de l'équation A.4 par rapport à t. Il n'y a pas de solution analytique à l'équation A.4. On doit donc recourir à une étude numérique en utilisant un package mathématique.

$$B(t)C(t)[\beta + \alpha e^{-\beta t}] - \alpha(\gamma - 1)e^{-\beta t}A(t)C(t) + 2\gamma\alpha(1-r)e^{-\beta t}A(t)B^{\gamma}(t) = 0. \tag{A.4}$$

Remarque 3. *On peut, simplement, dériver les modes des cas particuliers de la distribution MO-GDGD en définissant les paramètres inconnus tels que ceux indiqués dans la colonne* H_0 *du tableau 2.1 et en les replaçant dans l'équation A.4.*

A.5 Moment d'ordre m

Soit T une variable aléatoire dont $f_{MO}(t; r, \alpha, \beta, \gamma)$ est la fonction densité de probabilité, le moment d'ordre m de T est dérivé en utilisant l'expansion suivante comme dans [11]. Pour $|z| < 1$ et $\delta > 0$,

$$(1-z)^{-\delta} = \sum_{j=0}^{\infty}\binom{\delta + j - 1}{j} z^j$$

et pour $x > 0$, l'expansion binomiale de $(1 - e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)})^{\gamma(i+1)-1}$ est donnée par :

$$(1 - e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)})^{\gamma(i+1)-1} = \sum_{j=0}^{\infty}\binom{\gamma(i+1)-1}{j}(-1)^j e^{\frac{j\alpha}{\beta}(e^{-\beta t}-1)}.$$

Si $r > 1$, alors le moment d'ordre m de T est donné par :

$$\mu^{(m)} = \sum_{i,j=0}^{\infty}\binom{\gamma(i+1)-1}{j}\frac{(-1)^j m!\gamma(i+1)\nu_i e^{\frac{-(j+1)\alpha}{\beta}}}{(j+1)\beta^m}\mu_{1,m-1}(\frac{-\alpha}{\beta}(j+1))$$

Si $0 < r < 1$, alors le moment d'ordre m de T est donné par :

$$\mu^{(m)} = \sum_{i,j=0}^{\infty}\binom{\gamma(i+1)-1}{j}\frac{(-1)^j m!\gamma(i+1)\omega_i e^{\frac{-(j+1)\alpha}{\beta}}}{(j+1)\beta^m}\mu_{1,m-1}(\frac{-\alpha}{\beta}(j+1))$$

où

$$\mu_{s,k}(z) = \frac{1}{k!}\int_1^\infty \ln(t)^k t^{-s} \mathrm{e}^{-zt}\, dt$$
$$\omega_i = \frac{r(-1)^i}{i+1} \sum_{j=i}^{\infty} \binom{j}{i} (j+1)(1-r)^j$$
$$\nu_i = \frac{(1-\frac{1}{r})^2}{r}$$

Annexe B

Résultats de la simulation

Le tableau B.1 représente les résultats de l'étude de simulation pour différentes valeurs de paramètres. Les cas traités sont :

- le cas où la présence d'une proportion de survivants est considérée, le modèle est défectueux ($\beta > 0$) et le paramètre de forme de Marshall-Olkin est pris en compte ($r > 1$).
- le cas d'absence d'une proportion de survivants. Dans ce cas le modèle est propre ($S_{MO}(t;\Theta)$ et converge vers zéro lorsque t tend vers l'infini et $\beta < 0$) et le paramètre de forme de Marshall-Olkin est négligé ($r = 1$).
- le cas où la présence d'une proportion de survivants est considérée ($\beta > 0$), et le paramètre de forme de Marshall-Olkin est supposé ($r < 1$).

Pour chacun des cas prédéfinis, l'erreur quadratique moyenne des estimateurs sont calculés dans le but d'évaluer la qualité de l'estimation (voir le tableau B.1). À partir de l'étude de simulation, on peut conclure que, pour les cas présentés, plus l'échantillon est grand, plus les valeurs des MSE et biais tendent vers zéro.

TABLEAU B.1 – Résultats de l'étude de simulation pour différentes valeurs des paramètres.

$\Theta = (r, \alpha, \beta, \gamma)$	n	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\theta}_{MO}$
(5, 3.5, 0.5 ,3)	30	4.9856	3.6816	0.4026	2.9044	0.0015
		(0.0002)	(0.0330)	(0.0095)	(0.0091)	(0.0001)
	100	4.9981	3.5188	0.4130	2.9966	0.0030
		(0.0000)	(0.0004)	(0.0076)	(0.0000)	(0.0001)
	200	4.9963	3.4335	0.4505	3.0241	0.0074
		(0.0000)	(0.0044)	(0.0025)	(0.0006)	(0.0000)
	500	5.0492	3.3733	0.4371	3.0797	0.0069
		(0.0024)	(0.0161)	(0.0040)	(0.0064)	(0.0000)
	1000	5.0936	3.3079	0.4360	3.1102	0.0080
		(0.0088)	(0.0369)	(0.0041)	(0.0121)	(0.0000)
(1, 4 -0.5, 1)	30	0.9330	4.0587	-0.5460	0.8834	–
		(0.0045)	(0.0034)	(0.0021)	(0.0136)	
	100	1.1106	3.9550	-0.4379	1.0222	–
		(0.0122)	(0.0020)	(0.0039)	(0.0005)	
	200	1.0917	3.8854	-0.4008	0.9816	–
		(0.0084)	(0.0131)	(0.0098)	(0.0003)	
	500	1.0700	4.0983	-0.4039	0.9845	–
		(0.0049)	(0.0097)	(0.0092)	(0.0002)	
	1000	1.0353	3.9792	-0.6173	0.9152	–
		(0.0012)	(0.0004)	(0.0138)	(0.0072)	
(0.8, 0.5, 0.1, 1.2)	30	0.8640	0.4593	0.0406	1.2946	$1.3663\ 10^{-5}$
		(0.0041)	(0.0017)	(0.0035)	(0.0089)	(0.0000)
	100	0.8056	0.5098	0.0553	1.2180	$9.7296\ 10^{-5}$
		(0.0000)	(0.0001)	(0.0020)	(0.0003)	(0.0000)
	200	0.8165	0.4743	0.0642	1.2241	$6.1844\ 10^{-4}$
		(0.0003)	(0.0007)	(0.0013)	(0.0006)	(0.0000)
	500	0.8066	0.4925	0.0779	1.2085	0.0018
		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	1000	0.7938	0.5049	0.0747	1.2074	0.0011
		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Annexe C

Profils des fonctions log-vraisemblance par rapport à chaque paramètre du modèle MO-GDGD.

C.1 Application aux données de cancer pédiatrique

Les fonctions log-vraisemblance des paramètres des modèles r, α, β et γ sont affichées dans la figure C.4.

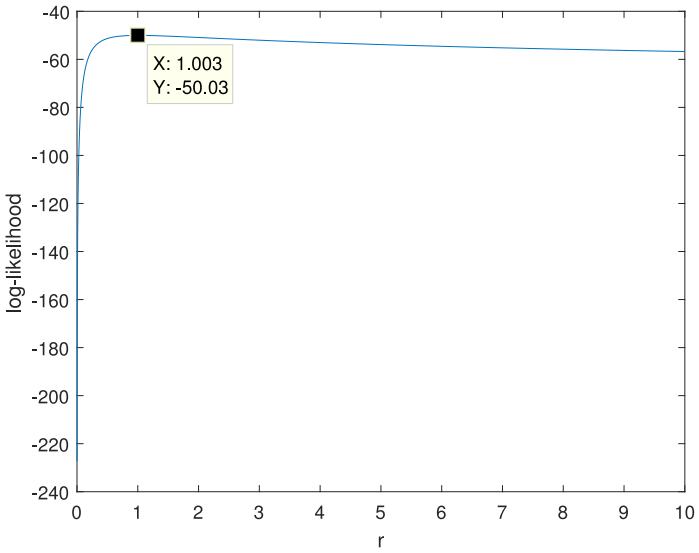


FIGURE C.1 – Le profil de la fonction log-vraisemblance par rapport au paramètre r du modèle MO-GDGD pour les données de cancer pédiatrique.

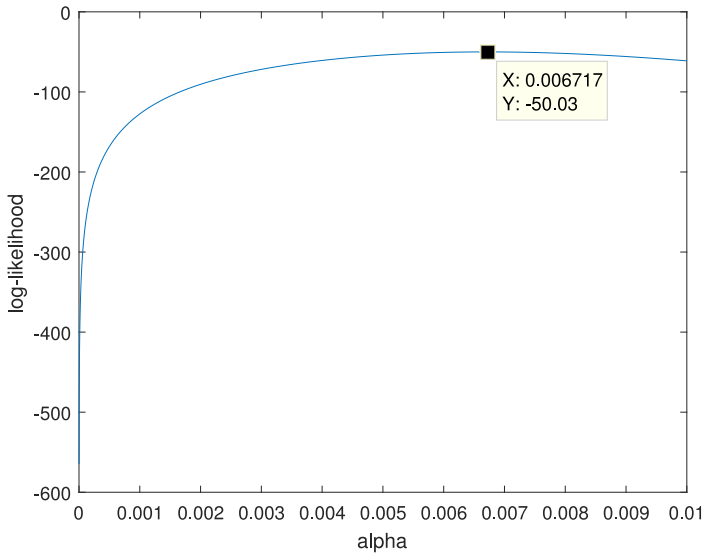


FIGURE C.2 – Le profil de la fonction log-vraisemblance par rapport au paramètre α du modèle MO-GDGD pour les données de cancer pédiatrique.

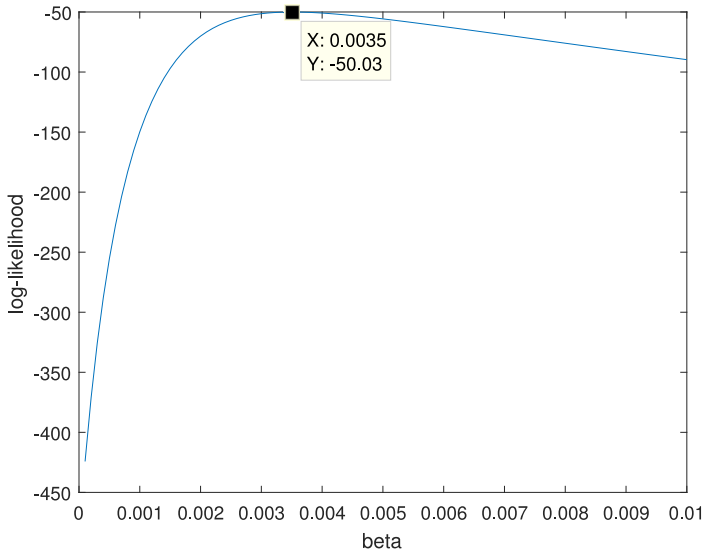


FIGURE C.3 – Le profil de la fonction log-vraisemblance par rapport au paramètre β du modèle MO-GDGD pour les données de cancer pédiatrique.

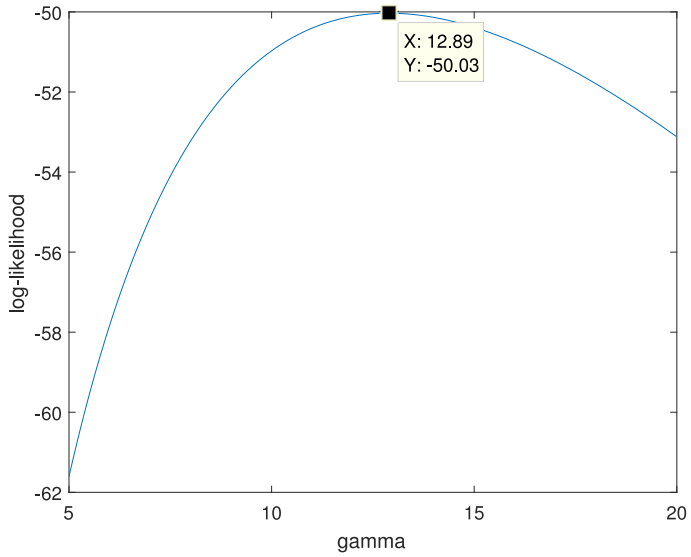


FIGURE C.4 – Le profil de la fonction log-vraisemblance par rapport au paramètre γ du modèle MO-GDGD pour les données de cancer pédiatrique.

C.2 Application aux données PRO-ACT

Les profils des fonctions log-vraisemblance par rapport aux paramètres du modèle MO-GDGD appliqué à la base de données de cancer pédiatrique mentionné en page 54 sont affichés dans la figure C.8 indiquent que le processus d'estimation donne des résultats satisfaisants.

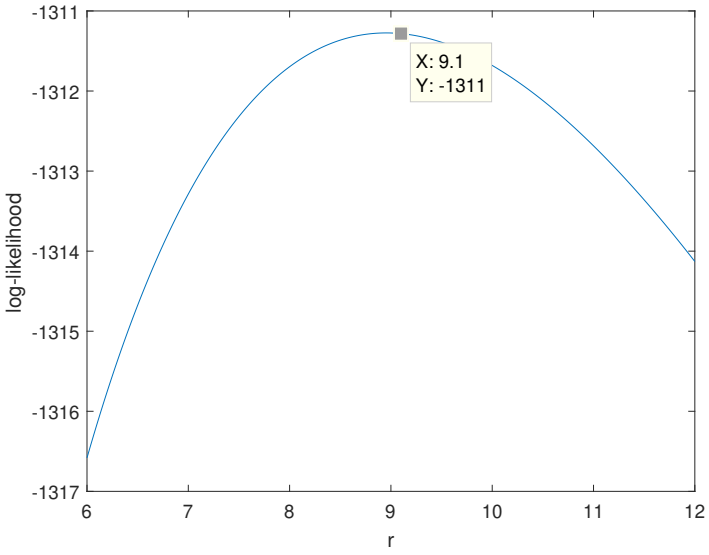


FIGURE C.5 – Le profil de la fonction log-vraisemblance par rapport au paramètre r du modèle MO-GDGD pour les données PRO-ACT.

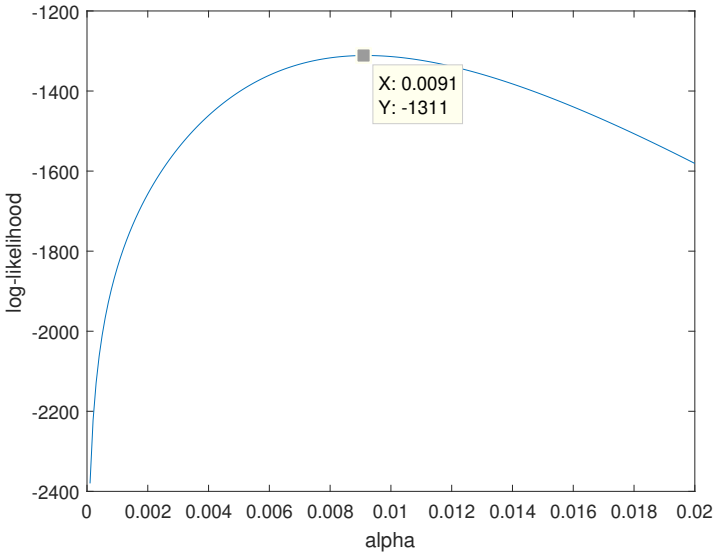


FIGURE C.6 – Le profil de la fonction log-vraisemblance par rapport au paramètre α du modèle MO-GDGD pour les données PRO-ACT.

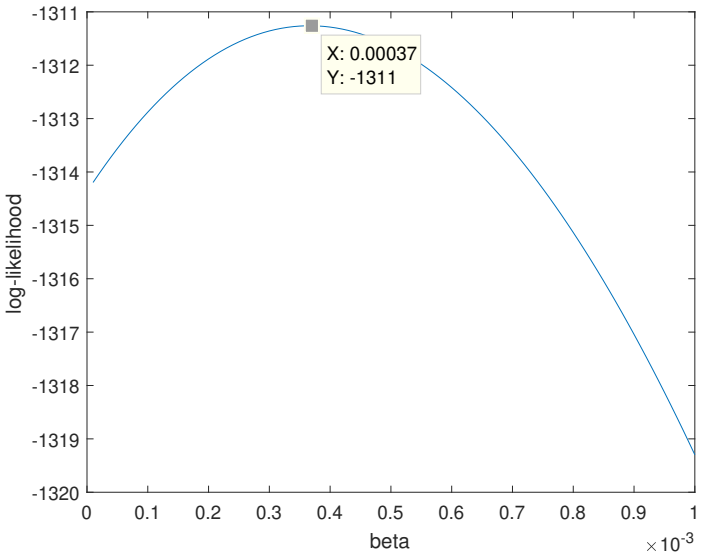


FIGURE C.7 – Le profil de la fonction log-vraisemblance par rapport au paramètre β du modèle MO-GDGD pour les données PRO-ACT.

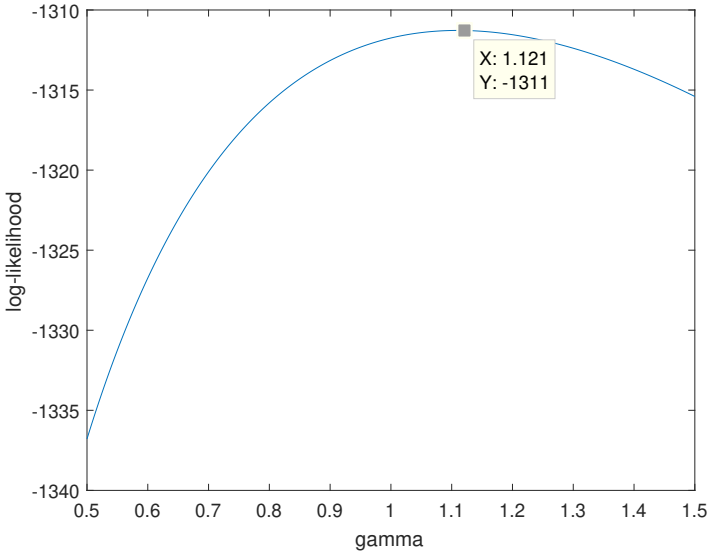


FIGURE C.8 – Le profil de la fonction log-vraisemblance par rapport au paramètre γ du modèle MO-GDGD pour les données PRO-ACT.

Annexe D

Bases de données utilisées

D.1 Cancer de la vessie

TABLEAU D.1 – Base de données de cancer de la vessie.

Durée de rémission de 128 patients atteints d'un cancer de la vessie										
0.08	0.20	0.40	0.50	0.51	0.81	0.90	1.05	1.19	1.26	1.35
1.40	1.46	1.76	2.02	2.02	2.07	2.09	2.23	2.26	2.46	2.54
2.62	2.64	2.69	2.69	2.75	2.83	2.87	3.02	3.25	3.36	3.36
3.48	3.52	3.57	3.64	3.70	3.82	3.88	4.18	4.23	4.26	4.33
4.34	4.40	4.50	4.51	4.87	4.98	5.06	5.09	5.17	5.32	5.32
5.34	5.41	5.41	5.46	5.62	5.71	5.85	6.25	6.54	6.67	6.93
6.94	6.97	7.09	7.26	7.28	7.32	7.39	7.59	7.62	7.63	7.66
7.87	7.93	8.26	8.37	8.53	8.65	8.66	9.02	9.22	9.47	9.74
10.06	10.34	10.66	10.75	11.25	11.64	11.79	11.98	12.02	12.03	12.07
12.63	13.11	13.29	13.31	13.80	14.24	14.76	14.77	14.83	15.96	16.62
17.12	17.14	17.36	18.10	19.13	20.28	21.73	22.69	23.63	25.74	25.82
26.31	32.15	34.26	36.66	43.01	46.12	79.05				

D.2 Cancer du sang

TABLEAU D.2 – Base de données de cancer du sang.

Durées de survie de 43 patients atteints d'un cancer du sang.										
115	181	255	418	441	461	516	739	743	789	807
865	924	983	1025	1062	1063	1165	1191	1222	1222	1251
1277	1290	1357	1369	1408	1455	1478	1519	1578	1578	1599
1603	1605	1696	1735	1799	1815	1852	1899	1925	1965	

D.3 Cancer du sein

TABLEAU D.3 – Base de données de cancer du sein.

Durées de survie des tumeurs mammaires de 30 rats.									
112	68	84	109	153	143	60	70	98	164
63	63	77	91	91	66	70	77	63	66
66	94	101	105	108	112	115	126	161	178

D.4 Cancer pédiatrique

TABLEAU D.4 – Base de données de cancer pédiatrique

Durées de rémission de 41 patients suivant le même traitement.									
327	**385**	**410**	722	**750**	878	890	942	1161	**1215**
1241	1270	1418	1602	1610	1689	1848	1940	2016	2044
2078	2231	2289	2303	2420	2430	2510	2908	2932	2978
2993	3013	3017	3053	3083	3090	3219	3232	3321	3399
3416									

Bibliographie

[1] J.P. Klein and M.L. Moeschberger. Survival analysis : Statistical methods for censored and truncated data. 2nd, 2003. 6, 21

[2] B Benjamin and John Graunt. John graunt's 'observations'. *Journal of the Institute of Actuaries*, 90(1) :1–61, 1964. 6

[3] M. Greenwood. A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33), 1926. 7, 9

[4] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202, 1972. 7, 10, 11, 70

[5] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282) :457–481, 1958. 9

[6] J.P. Klein and M.L. Moeschberger. Statistics for biology and health. *Stat. Biol. Health, New York*, 27238, 1997. 10

[7] A. W. Marshall and I. Olkin. A new method for adding a parameter to a family of distributions with application to the exponential and weibull families. *Biometrika*, 84(3) :641–652, 1997. 13, 34

[8] W. Barreto-Souza, A.J. Lemonte, and G.M. Cordeiro. General results for the marshall and olkin's family of distributions. *Anais da Academia Brasileira de Ciências*, 85(1) :3–21, 2013. 14

[9] M.E. Ghitany, E.K. Al-Hussaini, and R.A. Al-Jarallah. Marshall–olkin extended weibull distribution and its application to censored data. *Journal of Applied Statistics*, 32(10) :1025–1034, 2005. 14

[10] X. Bai, Y. Shi, B. Liu, and Q. Fu. Statistical inference of marshall-olkin bivariate weibull distribution with three shocks based on progressive interval censored data. *Communications in Statistics-Simulation and Computation*, 48(3) :637–654, 2019. 14

[11] Lazhar Benkhelifa. The marshall-olkin extended generalized gompertz distribution. *Journal of Data Science*, 15(2) :239–266, 2017. 14, 34, III

[12] G.D.C. Barriga, G.M. Cordeiro, D. K. Dey, V.G. Cancho, F. Louzada, and A.K. Suzuki. The marshall-olkin generalized gamma distribution. *Communications for Statistical Applications and Methods*, 25(3) :245–261, 2018. 14

[13] Kenneth S Lomax. Business failures : Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49(268) :847–852, 1954. 14, 15

[14] G.S. Mudholkar and D.K. Srivastava. Exponentiated weibull family for analyzing bathtub failure-rate data. *IEEE transactions on reliability*, 42(2) :299–302, 1993. 15

[15] S. Nadarajah. The exponentiated gumbel distribution with climate application. *Environmetrics : The official journal of the International Environmetrics Society*, 17(1) :13–23, 2006. 15

[16] A. El-Gohary, A. Alshamrani, and A.N. Al-Otaibi. The generalized gompertz distribution. *Applied Mathematical Modelling*, 37(1-2) :13–24, 2013. 15, 19, 29, 34, 79

[17] W.T. Shaw and I.R.C. Buckley. The alchemy of probability distributions : beyond gram-charlier expansions, and a skew-kurtotic-normal distribution from a rank transmutation map. *arXiv preprint arXiv :0901.0434*, 2009. 15

[18] S.E. Smith. The hazard-product method of generalization : Application to the gompertz and exponentiated half-normal distributions. *Communications in Statistics-Theory and Methods*, 48(5) :1177–1192, 2019. 15

[19] M.H. Tahir, G.M. Cordeiro, M. Alizadeh, M. Mansoor, M. Zubair, and G.G. Hamedani. The odd generalized exponential family of distributions with applications. *Journal of Statistical Distributions and Applications*, 2(1) :1–28, 2015. 15

[20] P. Kumaraswamy. A generalized probability density function for double-bounded random processes. *Journal of hydrology*, 46(1-2) :79–88, 1980. 15

[21] A. Tsodikov. Semi-parametric models of long-and short-term survival : an application to the analysis of breast cancer survival in utah by age and stage. *Statistics in medicine*, 21(6) : 895–920, 2002. 17

[22] A. Tsodikov. Semiparametric models : a generalized self-consistency approach. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 65(3) :759–774, 2003. 17

[23] A.D. Tsodikov, J.G. Ibrahim, and A.Y. Yakovlev. Estimating cure rates from survival data : an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464) :1063–1078, 2003. 17

[24] M.V. Koutras and F.S. Milienos. A flexible family of transformation cure rate models. *Statistics in medicine*, 36(16) :2559–2575, 2017. 17

[25] N. Balakrishnan and F.S. Milienos. On a class of non-linear transformation cure rate models. *Biometrical Journal*, 62(5) :1208–1222, 2020. 17

[26] J.W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society : Series B (Methodological)*, 11(1) :15–44, 1949. 18

[27] P.C. Lambert, P.W. Dickman, C.L. Weston, and J.R. Thompson. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 59(1) :35–55, 2010. 18

[28] J. Balka, A.F. Desmond, and P.D. McNicholas. Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *Journal of Applied Statistics*, 38(1) :127–144, 2011. 19

[29] V.G. Cancho and H. Bolfarine. Modeling the presence of immunes by using the exponentiated-weibull model. *Journal of Applied Statistics*, 28(6) :659–671, 2001. 19

[30] A.B. Cantor and J.J. Shuster. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, 11(7) :931–937, 1992. 19, 29, 34, 54, 79

[31] R. Rocha, S. Nadarajah, V. Tomazella, F. Louzada, and A. Eudes. New defective models based on the kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*, 26(4) :1737–1755, 2017. 19

[32] R. Rocha, S. Nadarajah, V. Tomazella, and F. Louzada. A new class of defective models based on the marshall–olkin family of distributions for cure rate modeling. *Computational Statistics & Data Analysis*, 107 :48–63, 2017. 19, 34

[33] R. Rocha, S. Nadarajah, V. Tomazella, and F. Louzada. Two new defective distributions based on the marshall–olkin extension. *Lifetime data analysis*, 22(2) :216–240, 2016. 19, 30

[34] E.Z. Martinez and J.A. Achcar. The defective generalized gompertz distribution and its use in the analysis of lifetime data in presence of cure fraction, censored data and covariates. *Electronic Journal of Applied Statistical Analysis*, 10(2) :463–484, 2017. 19, 29, 61, 68, 79

[35] F.Y. Edgeworth. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71(2) :381–397, 1908. 20

[36] R.J. Rossi. *Mathematical statistics : an introduction to likelihood based inference*. John Wiley & Sons, 2018. 20

[37] J. Neyman and E.S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706) :289–337, 1933. 22

[38] A. Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4 :83–91, 1933. 23

[39] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2) :279–281, 1948. 23

[40] A. Pickles. *An introduction to likelihood analysis*. Number 42. Geo Books, 1985. 24, 26

[41] B. Effron. Bootstrap methods : Another look at the jackknife. *Annals of statistics*, 7(1) :1–26, 1979. 25

[42] J. Shao. Bootstrap model selection. *Journal of the American statistical Association*, 91(434) : 655–665, 1996. 25

[43] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998. 25

[44] B. Gompertz. Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, (115) :513–583, 1825. 29, 34, 79, 118

[45] P.W. Gieser, M.N. Chang, P.V. Rao, J.J. Shuster, and J. Pullen. Modelling cure rates using the gompertz model with covariate information. *Statistics in medicine*, 17(8) :831–839, 1998. 29, 60

[46] P. Borges. Em algorithm-based likelihood estimation for a generalized gompertz regression model in presence of survival data with long-term survivors : an application to uterine cervical cancer data. *Journal of Statistical Computation and Simulation*, 87(9) :1712–1722, 2017. 29, 34, 61, 79

[47] F. Cooner, S. Banerjee, and A.M. McBean. Modelling geographically referenced survival data with a cure fraction. *Statistical methods in medical research*, 15(4) :307–324, 2006. 32

[48] V.P. Collins. Observation on growth rates of human tumors. *Am J Roentgenol*, 76 :988–1000, 1956. 34, 115, 116

[49] T. Hamdeni and S. Gasmi. The marshall–olkin generalized defective gompertz distribution for surviving fraction modeling. *Communications in Statistics-Simulation and Computation*, pages 1–14, 2020. 34

[50] A. Nasr, S. Gasmi, and F. Ben Hmida. Parameter estimation of the flexible weibull distribution for type i censored samples. *Journal of Applied Statistics*, 44(14) :2499–2512, 2017. 34

[51] S. Ghnimi and S. Gasmi. Parameter estimations for some modifications of the weibull distribution. *Open Journal of Statistics*, 4(08) :597, 2014. 36

[52] S. Gasmi, C.E. Love, and W. Kahle. A general repair, proportional-hazards, framework to model complex repairable systems. *IEEE Transactions on Reliability*, 52(1) :26–32, 2003. 37, 83

[53] S. Gasmi and M. Berzig. Parameters estimation of the modified weibull distribution based on type-i censored samples. *Applied Mathematical Sciences*, 5(59) :2899–2917, 2011. 37

[54] E.T. Lee and J. Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003. 42

[55] A.M. Abouammoh, S.A. Abdulghani, and I.S. Qamber. On partial orderings and testing of new better than renewal used classes. *Reliability Engineering & System Safety*, 43(1) :37–41, 1994. 44

[56] M.M. King, D.M. Bailey, D.D. Gibson, J.V. Pitha, and P.B. McCay. Incidence and growth of mammary tumors induced by 7, 12-dimethylbenz [a] anthracene as related to the dietary content of fat and antioxidant. *Journal of the National Cancer Institute*, 63(3) :657–663, 1979. 45

[57] T. Hamdeni and S. Gasmi. Parameter estimation of generalized gompertz distribution under type i censoring. Djerba, Tunisia, 2018. Euro-Mediterranean Conference on Mathematical Reliability. 50

[58] W. Nissas and S. Gasmi. A hybrid decision dependent maintenance model of failure rate and virtual age classes using modified weibull intensity. *Communications in Statistics-Simulation and Computation*, pages 1–15, 2019. 54

[59] T. Hamdeni and S. Gasmi. A proportional-hazards model for survival analysis and long-term survivors modeling : application to amyotrophic lateral sclerosis data. *Journal of Applied Statistics*, pages 1–15, 2020. 59

[60] V.F. Calsavara, A.S. Rodrigues, R. Rocha, V. Tomazella, and F. Louzada. Defective regression models for cure rate modeling with interval-censored data. *Biometrical Journal*, 61(4) :841–859, 2019. 60

[61] V.F. Calsavara, A.S. Rodrigues, R. Rocha, F. Louzada, V. Tomazella, A.C.R.L.A. Souza, R.A. Costa, and R.P.V. Francisco. Zero-adjusted defective regression models for modeling lifetime data. *Journal of Applied Statistics*, 46(13) :2434–2459, 2019. 60

[62] R. Rocha, S. Nadarajah, V. Tomazella, F. Louzada, and A. Eudes. New defective models based on the kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*, 26(4) :1737–1755, 2017. 60

[63] P. Borges. Estimating the turning point of the log-logistic hazard function in the presence of long-term survivors with an application for uterine cervical cancer data. *Journal of Applied Statistics*, pages 1–11, 2020. 60

[64] J. Scudilio, V.F. Calsavara, R. Rocha, F. Louzada, V. Tomazella, and A.S. Rodrigues. Defective models induced by gamma frailty term for survival data with cured fraction. *Journal of Applied Statistics*, 46(3) :484–507, 2019. 60

[65] R. Rocha, S. Nadarajah, V. Tomazella, F. Louzada, and A. Eudes. New defective models based on the kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*, 26(4) :1737–1755, 2017. 61

[66] R.G. Miller, J. D. Mitchell, and D.H. Moore. Riluzole for amyotrophic lateral sclerosis (als)/motor neuron disease (mnd). *Cochrane database of systematic reviews*, (3), 2012. 63, 67

[67] M. Hinchcliffe and A. Smith. Riluzole : real-world evidence supports significant extension of median survival times in patients with amyotrophic lateral sclerosis. *Degenerative neurological and neuromuscular disease*, 7 :61, 2017. 67

[68] Z. Huang, H. Zhang, J. Boss, S.A. Goutman, B. Mukherjee, I.D. Dinov, and Y. Guan. Complete hazard ranking to analyze right-censored data : An als survival study. *PLoS computational biology*, 13(12) :e1005887, 2017. 67

[69] C. Fournier and J.D. Glass. Modeling the course of amyotrophic lateral sclerosis. *Nature biotechnology*, 33(1) :45–47, 2015. 67

[70] S.A. Adham and S.G. Walker. A multivariate gompertz-type distribution. *Journal of Applied Statistics*, 28(8) :1051–1065, 2001. 67

[71] B. Efron and R. Tibshirani. Improvements on cross-validation : the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438) :548–560, 1997. 71

[72] F.E. Harrell Jr, K.L. Lee, and D.B. Mark. Multivariable prognostic models : issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4) :361–387, 1996. 72

[73] WHO. Coronavirus disease 2019 (covid-19) situation report-51. 2020. URL https://www.who.int/docs/default-source/coronaviruse/situationreports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57-10. 78

[74] L. Cao, T. Huang, J. Zhang, Q. Qin, S. Liu, H. Xue, Y. Gong, C. Ning, X. Shen, J. Yang, Y. Mi, X. Xiao, and X. Cao. Estimation of instant case fatality rate of covid-19 in wuhan and hubei based on daily case notification data. *medRxiv*, 2020. 81

[75] S. Kim, D. Zeng, Y. Li, and D. Spiegelman. Joint modeling of longitudinal and cure-survival data. *Journal of statistical theory and practice*, 7(2) :324–344, 2013. 83

[76] M. Othus, W. van Putten, B. Lowenberg, S.H. Petersdorf, S. Nand, H. Erba, F. Appelbaum, R. Hills, N. Russell, A. Burnett, and E. Estey. Relationship between event-free survival and

overall survival in acute myeloid leukemia : a report from swog, hovon/sakk, and mrc/ncri. *Haematologica*, 101(7) :e284, 2016. 86, 98

[77] World Health Organization. Cancer. URL https://www.who.int/news-room/fact-sheets/detail/cancer. 87, 88

[78] C.E. DeSantis, C.C. Lin, A.B. Mariotto, R.L. Siegel, K.D. Stein, J.L. Kramer, R. Alteri, A.S. Robbins, and A. Jemal. Cancer treatment and survivorship statistics, 2014. *CA : a cancer journal for clinicians*, 64(4) :252–271, 2014. 87, 88

[79] T. Tirkes, M.A. Hollar, M. Tann, M.D. Kohli, F. Akisik, and K. Sandrasegaran. Response criteria in oncologic imaging : review of traditional and new criteria. *Radiographics*, 33(5) :1323–1341, 2013. 91

[80] A.B. Miller, B.F.A.U. Hoogstraten, M.F.A.U. Staquet, and A. Winkler. Reporting results of cancer treatment. *cancer*, 47(1) :207–214, 1981. 91

[81] P. Therasse, S.G. Arbuck, E.A. Eisenhauer, J. Wanders, R.S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A.T. van Oosterom, M.C. Christian, and S.G. Gwyther. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3) :205–216, 2000. 91

[82] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours : revised recist guideline (version 1.1). *European journal of cancer*, 45(2) :228–247, 2009. 91, 93

[83] R. Pazdur. Endpoints for assessing drug activity in clinical trials. *The oncologist*, 13 :19–21, 2008. 98, 99

[84] R. de Sahb-Berkovitch, M.C. Woronoff-Lemsi, and M. et les participants de la Table Ronde n°7 de Giens 2009 Molimard. Critères et méthodologie d'évaluation au remboursement des anticancéreux. *Therapies*, 65(4) :367–372, 2010. 98

[85] B. Amzal, S. Fu, J. Meng, J. Lister, and H. Karcher. Cabozantinib versus everolimus, nivolumab, axitinib, sorafenib and best supportive care : a network meta-analysis of progression-free survival and overall survival in second line treatment of advanced renal cell carcinoma. *PloS one*, 12(9) :e0184423, 2017. 99

[86] H. Akamatsu, K. Mori, T. Naito, H. Imai, A. Ono, T. Shukuya, T. Taira, H. Kenmotsu, H. Murakami, M. Endo, H. Harada, T. Takahashi, and N. Yamamoto. Progression-free survival at 2 years is a reliable surrogate marker for the 5-year survival rate in patients with locally advanced non-small cell lung cancer treated with chemoradiotherapy. *BMC cancer*, 14(1) :1–5, 2014. 99

[87] E.D. Saad and A. Katz. Progression-free survival and time to progression as primary end points in advanced breast cancer : often used, sometimes loosely defined. *Annals of oncology*, 20(3) :460–464, 2009. 99

[88] L.M. Hess, A. Brnabic, O. Mason, P. Lee, and S. Barker. Relationship between progression-free survival and overall survival in randomized clinical trials of targeted and biologic agents in oncology. *Journal of Cancer*, 10(16) :3717, 2019. 99

[89] F. Fleischer, B. Gaschler-Markefski, and E. Bluhmki. A statistical model for the dependence between progression-free survival and overall survival. *Statistics in Medicine*, 28(21) :2669–2686, 2009. 101, 103, 104

[90] Y. Li and Q. Zhang. A weibull multi-state model for the dependence of progression-free survival and overall survival. *Statistics in medicine*, 34(17) :2497–2513, 2015. 101, 104

[91] K.R. Abrams. Chte2020 sources and synthesis of evidence; update to evidence synthesis methods. 2020. 101

[92] N. Meghdadi, M. Soltani, H. Niroomand-Oscuii, and F. Ghalichi. Image based modeling of tumor growth. *Australasian physical & engineering sciences in medicine*, 39(3) :601–613, 2016. 114, 116, 123

[93] C.M. Newton. Biomathematics in oncology : Modeling of cellular systems. *Annual review of biophysics and bioengineering*, 9(1) :541–579, 1980. 116

[94] A. Rivaz, M. Azizian, and M. Soltani. Various mathematical models of tumor growth with reference to cancer stem cells : a review. *Iranian Journal of Science and Technology, Transactions A : Science*, 43(2) :687–700, 2019. 116

[95] N. Bacaër. Verhulst and the logistic equation (1838). In *A short history of mathematical population dynamics*, pages 35–39. Springer, 2011. 117

[96] A. Talkington and R. Durrett. Estimating tumor growth rates in vivo. *Bulletin of mathematical biology*, 77(10) :1934–1954, 2015. 119

[97] S. Benzekry, C. Lamont, A. Beheshti, A. Tracz, J.M.L. Ebos, L. Hlatky, and P. Hahnfeldt. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput Biol*, 10(8) :e1003800, 2014. 119

[98] R.K. Sachs, L.R. Hlatky, and P. Hahnfeldt. Simple ode models of tumor growth and anti-angiogenic or radiation treatment. *Mathematical and Computer Modelling*, 33(12-13) : 1297–1305, 2001. 119

[99] H. Enderling and M. A.J. Chaplain. Mathematical modeling of tumor growth and treatment. *Current pharmaceutical design*, 20(30) :4934–4940, 2014. 119

[100] P. Hahnfeldt, D. Panigrahy, J. Folkman, and L. Hlatky. Tumor development under angiogenic signaling : a dynamical theory of tumor growth, treatment response, and postvascular dormancy. *Cancer research*, 59(19) :4770–4775, 1999. 119, 120

[101] A. Beheshti, S. Benzekry, J. T. McDonald, L. Ma, M. Peluso, P. Hahnfeldt, and L. Hlatky. Host age is a systemic regulator of gene expression impacting cancer progression. *Cancer research*, 75(6) :1134–1143, 2015. 120

[102] K.P. Wilkie and P. Hahnfeldt. Tumor–immune dynamics regulated in the microenvironment inform the transient nature of immune-induced tumor dormancy. *Cancer research*, 73(12) : 3534–3544, 2013. 121

[103] I. Kareva. Angiogenesis regulators as a possible key to accelerated growth of secondary tumors following primary tumor resection. *arXiv preprint arXiv :1703.09994*, 2017. 121

[104] L.G. Hutchinson, H.J. Mueller, E.A. Gaffney, P.K. Maini, J. Wagg, A. Phipps, C. Boetsch, H.M. Byrne, and B. Ribba. Modeling longitudinal preclinical tumor size data to identify transient dynamics in tumor response to antiangiogenic drugs. *CPT : pharmacometrics & systems pharmacology*, 5(11) :636–645, 2016. 121

[105] V.T. DeVita, T.S. Lawrence, and S.A. Rosenberg. *Cancer : principles & practice of oncology : primer of the molecular biology of cancer*. Lippincott Williams & Wilkins, 2012. 122

[106] A.J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3) :271–274, 2002. 122, 123

[107] V. Volterra. Variations and fluctuations of the number of individuals in animal species living together. *Animal ecology*, pages 409–448, 1926. 122

[108] J.M. Ginoux. The paradox of vito volterra's predator-prey model. *Lettera Matematica*, 5(4) : 305–311, 2017. 122

[109] J.M. Ginoux, R. Naeck, Y.B. Ruhomally, M.Z. Dauhoo, and M. Perc. Chaos in a predator–prey-based mathematical model for illicit drug consumption. *Applied Mathematics and Computation*, 347 :502–513, 2019. 123

[110] J.M. Ginoux, R. Naeck, Y.B. Ruhomally, and M.Z. Dauhoo. Predator-prey model for illicit drug consumption. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–6. IEEE, 2018. 123

[111] V.A. Kuznetsov, I.A. Makalkin, M.A. Taylor, and A.S. Perelson. Nonlinear dynamics of immunogenic tumors : parameter estimation and global bifurcation analysis. *Bulletin of mathematical biology*, 56(2) :295–321, 1994. 123, 124

[112] G.I. Bell. Predator-prey equations simulating an immune response. *Mathematical Biosciences*, 16(3-4) :291–314, 1973. 123

[113] G. Kaur and N. Ahmad. On study of immune response to tumor cells in prey-predator system. *International scholarly research notices*, 2014, 2014. 123, 125

[114] A. Mishkin. *Modeling Cancer Growth Using Lotka-Volterra Predator Prey Model in Conjunction with Bifurcation Analysis*. PhD thesis, Florida Atlantic University, 2013. 123, 126

[115] F.A. Rihan, M. Safan, M.A. Abdeen, and D. Abdel Rahman. Qualitative and computational analysis of a mathematical model for tumor-immune interactions. *Journal of Applied Mathematics*, 2012, 2012. 123

[116] R.R. Sarkar and S. Banerjee. Cancer self remission and tumor stability–a stochastic approach. *Mathematical Biosciences*, 196(1) :65–81, 2005. 124, 125

[117] A. El-Gohary. Chaos and optimal control of steady-states of tumor with drug. *Applied Mathematical Sciences*, 3(16) :779–788, 2009. 125

[118] I. Abdulrashid, H. Ghazzai, X. Han, and Y. Massoud. Optimal control treatment analysis for the predator-prey chemotherapy model. In *2019 31st International Conference on Microelectronics (ICM)*, pages 296–299. IEEE, 2019. 126

[119] M.S. Eliwa, M. El-Morshedy, and M. Ibrahim. Inverse gompertz distribution : properties and different estimation methods with application to complete and censored data. *Annals of data science*, 6(2) :321–339, 2019. 129