**SOUTENANCE**

1.Introduction

Bonjour à toutes et à tous,

Je m’appelle Merlin Simsen Baratault, étudiant en BUT3 Sciences des données parcours EMS, en alternance au CEREDIH à l’hôpital Necker.

J’ai effectué cette alternance dans le cadre de ma formation, pendant la deuxième et troisième année. J’ai été sous la supervision de Mickaël Alligon et Nizar Malhaoui au CEREDIH, et de François Montestruc en tant que tuteur universitaire.

L’objectif de cette soutenance est de vous présenter premièrement le contexte dans lequel j’ai fait cette alternance, mes principales missions, les compétences que j’ai pu développer, et en quoi cette expérience a été déterminante pour mon projet professionnel.

2.Présentation du CEREDIH et du service Data

Le CEREDIH, ou Centre de Référence des Déficits Immunitaires Héréditaires, à été fondé en 2005 dans le cadre du Plan national maladies rares. Il gère le registre national des déficits immunitaires héréditaires, qui compte aujourd’hui plus de 10 000 patients, ce qui en fait la base la plus importante en nombre de patients sur ces maladies rares.

Son rôle est multiple :

* Améliorer la connaissance épidémiologique
* Soutenir la recherche médicale
* Et contribuer directement ou indirectement à la prise en charge et la qualité de vie des patients

Le service Data où j’ai travaillé, joue un rôle clé car il assure la qualité, la structuration, et l’analyse des données, contrairement ou groupe des ARC (attachés de recherche clinique), qui elles collectent les données des patients dans les hôpitaux, nous au service data on fait le nettoyage, la gestion, les contrôles qualité, les analyses statistiques, et la production de rapports.

Concrètement, j’ai occupé un poste de datamanager et « biostatisticien », en soutien direct du statisticien principal, et en lien direct avec les ARC et les médecins.

3.Aperçu global de mes missions

Pendant ces deux années, j’ai eu trois grandes missions, ou trois grands projets.

Le premier était la refonte complète du code de Data Management, le deuxième a été la modernisation du rapport d’activité, et le troisième à été la rédaction d’un article méthodologique sur le modèle de Cox.

Je vais commencer par détailler les deux premiers projets, plus techniques, puis je mettrai l’accent sur la troisième, qui m’a permis d’approfondir mes compétences méthodologiques et scientifique.

4.Mission 1 – Refonte du Data Management

Ma première mission a été de réécrire entièrement le script de Data Management, qui prépare la base avant toute analyse.

Le script initial faisait près de 9 000 lignes. Et il fallait environ deux heures pour le faire compiler, et il s’interrompait souvent en cas d’erreur dans les données, de nouvelle variable ou nouvelle modalité. Cela posait plusieurs problèmes : une difficulté à réagir rapidement à une demande, blocage en cas d’erreur, qui obligeait à tout relancer, et une pèrte de temps et manque de fiabilité.

Les objectifs de ces projets étaient donc de réécrire ce code, afin de gagner en rapidité de compilation, d’améliorer la robustesse face aux erreurs, et de le structurer afin de le rendre plus lisible et maintenable.

Pour faire cela j’ai commencé par lire et analyser le code existant, et d’identifier les parties les plus coûteuse en termes de temps de compilation. Nous avons finalement trouvé que ce qui consommait le plus de temps était les boucles for, qui pour la plupart des créations de nouvelles variables, parcourais la base ligne par ligne et prenait donc beaucoup de temps à répétition. On a donc décidé de remplacer toutes ces boucles for, par des fonctions vectorisées comme apply, mapply, ou lapply.

Par exemple pour créer une variable de date, le script utilisait une boucle qui parcourait la base ligne par ligne, et une boucle qui la parcourais data par data, juste pour créer une seule variable. J’ai donc réécrit ce passage en une seule ligne avec la fonction mapply, ce qui a grandement diminué le temps de compilation.

Après avoir réécrit l’intégralité du code de cette manière, le temps total de compilation est passé d’environ 2 heures, à environ 15 minutes. On a également fait attention à écrire ce code pour qu’il soit le plus robuste possible, et gère au maximum des problèmes comme des nouvelles variables ou modalités. Ce projet a été parfaitement réussi car on peut maintenant se permettre de réaliser plusieurs exports plus fréquemment et à la demande.

Sur le plan technique j’ai renforcé ma maîtrise de R et appris à optimiser du code complexe. Sur le plan méthodologique, j’ai dû analyses un processus long, identifier les points faibles, puis proposer des solutions concrètes. Et sur le plan professionnel, j’ai appris à travailler en autonomie, tout en collaborant avec le statisticien principal.

5.Mission 2 – Rapport d’activité

Ma deuxième mission a été la modernisation du rapport d’activité du CEREDIH. Ce « rapport d’activité c’est un document généré automatiquement à l’aide d’un code R Markdown, qui présente sous la forme d’une présentation PowerPoint, l’état de la base du CEREDIH. En présentant des analyses descriptives pour une grande partie des données de la base.

Ce rapport généré avec R Markdown avait été réaliser il y a quelques années, il donnait des figures et tableaux descriptifs fiables, mais parfois peu lisible, comme on peut voir par exemple des couleurs et échelles parfois peu adaptées, des graphiques figés, peu utilisable en réunion ou congrès.

L’objectif de ce projet était donc de moderniser la présentation, et après ma proposition, on a fait le choix de rendre ce rapport interactif.

J’ai donc recréé tout le rapport, toujours en R Markdown, mais cette fois-ci en refaisant tous les graphiques en utilisant une librairie permettant de créer des graphiques interactif et personnalisable. Au final le rapport est généré sous trois formats différents, un format HTML interactif pour les présentations, mais aussi en PDF et en PowerPoint pour des usages plus classiques.

Au final ce rapport fait environ 70 pages, et le fait d’avoir ajouté de l’interactivité à tout de suite plu, et il a déjà été utilisé à plusieurs reprises lors de présentations, réunions ou congrès, ce qui montre sa valeur pour le CEREDIH.

Sur le plan technique, j’ai renforcé ma maitrise de R Markdown, et découvert la datavisualisation interactive. Sur le plan professionnel, j’ai appris à produire des outils qui ne sont pas seulement statistiques, mais aussi visuels et pédagogique, adaptés à différents publics (médecins, ARC, statisticien, …).

6.Mission 3 – Article méthodologique sur le modèle de Cox

Mon troisième projet, plus scientifique cette fois-ci, a été de travailler sur un article méthodologique consacré au modèle de Cox.

Ce projet fait suite à un précédent travail du CEREDIH sur Kaplan-Meier. L’idée est de fournir aux cliniciens un guide clair pour utiliser correctement le modèle de Cox, en évitant les biais méthodologiques fréquents.

On a donc choisi une approche pédagogique, basée sur la simulation de données.

L’intérêt de ne pas utiliser une base de données toute faites, mais de simuler nos propres données de A à Z, nous permet premièrement d’avoir un contrôle complet des données et de notre environnement, mais également de pouvoir calculer pour chaque cas, une vérité théorique se basant purement sur les données.

On a donc simulé les temps de survie à partir d’une loi, et deux covariables ont été simulés, une binaire avec un effet réel sur la survie, et une continue sans effet réel, ainsi que leur coefficient d’effet sur la survie. Puis au fur et à mesure, on a ajouté progressivement les différents biais, comme la troncature, la censure informative, risques compétitifs, événements récurrents, covariables dépendantes du temps, ou interaction entre les covariables.

Pour chacun de ces biais, nous avons comparé un modèle naïf qui ignore le biais, et un modèle corrigé, adapté au problème. En présentant à la fois les Hazard Ratios avec leurs intervalles de confiances, mais aussi les courbes de survie, comparées à la vérité simulée.

Par exemple, pour la partie ou l’on simule une interaction entre les covariables, on a un modèle naïf (qui ne prend pas en compte l’interaction) qui donne des courbes totalement fausses, qui s’écartent de la réalité. Et en fasse le modèle corrigé, prenant en compte l’interaction entre les covariables, qui là nous donnes des courbes qui colle bien mieux aux courbes théoriques.

Donc pour chaque biais, on montre que les ignorer peuvent avoir un réel effet néfaste sur les prédictions, mais on ne se contente pas de ça, on montre et explique pas à pas comment le résoudre, pour avoir un modèle et des résultats fiable.

L’article n’est pas encore publié mais il est en cours de finalisation de notre côté, j’ai réalisé la majorité du code, avec mon Mickaël qui m’a accompagné. Et j’ai également rédigé une première version de l’article en anglais, qui est en train d’être peaufiner par Mickaël et Pr Olivier Bouaziz, un chercheur en mathématique appliquées qui nous accompagne sur ce projet.

Pour moi, c’est une expérience unique qui m’a permis d’avoir un aperçu de ce côté du travail en statistique avec la simulation de données, le modèle de Cox, et la rédaction d’un article scientifique, tout dans le but de fournir un outil réellement pratique pour les cliniciens et pour améliorer au moins un petit peu, leurs modèles de Cox.

Sur le plan technique j’ai appris à simuler des données complexes, à travailler avec le modèle de Cox, et à produire des résultats reproductibles. Su le plan méthodologique, j’ai appris comment correctement construire un modèle de Cox. Et Sur le plan professionnel j’ai acquis une réelle expérience de rédaction scientifique en anglais, et de communication interdisciplinaire.

7.Portfolio – Compétences du BUT SD

Mes différentes missions m’ont permis de développer et de mettre en œuvre les 4 grandes compétences du BUT.

La première, « Traiter des données à des fins décisionnelles », avec principalement le projet de refonte du Data Management avec lequel j’ai appris à manipuler des bases massives et complexes, à automatiser des traitements et à assurer la traçabilité et la reproductibilité. Mais également avec les différentes analyses ponctuelles demandé par des cliniciens.

La deuxième, « Analyser statistiquement des donnée », avec le projet sur le modèle de Cox, j’ai non seulement appliqué des modèles, mais aussi réfléchi à leurs hypothèses, simulé des biais, et évalué leur validité.

La troisième, « Valoriser une production dans un contexte professionnel », avec le projet du rapport d’activité, où j’ai transformé des résultats complexes, en figures clairs et parlantes, utilisables tant avec des statisticiens qu’avec des médecins.

Et pour la quatrième, « Modéliser les données dans un cadre statistique », avec les modèles de Cox avancées que j’ai pu faire lors du projet, où j’ai mis en pratique des notions théoriques parfois abstraites, et je les ai transformées en applications concrètes, le tout pour un public large.

En parallèle, j’ai également développé des compétences transversales comme l’autoformation, en me formant seul sur certains biais de survie, ou en apprenant à réaliser des figures interactives. Mais aussi la communication interdisciplinaire, en expliquant des résultats statistiques à des médecins ou des ARC pas forcément spécialistes en stat. Et également la gestion de projets, en organisant des missions de longue durée, avec des étapes claires, des livrables, et des réunions.

Ces apprentissages rejoignent donc je pense, directement les objectifs du BUT.

8.Bilan et perspéctives

Pour résumé, cette alternance m’a permis de développer des compétences techniques solides, que ce soit en Rn data management, survie, dataviz. Mais également des compétences professionnelles comme la rigueur et la vulgarisation, le travail en équipe et dans un domaine complexe.

Et m’a également permis de confirmer mon choix d’orientation, car je souhaite poursuivre dans un domaine ou mon travail sur les données, aura un impact direct ou indirect sur la qualité de vie des gens, et non juste dans le but de faire plus d’argent pour une entreprise.

Je vous remercie pour votre attention, et je suis disponible pour vos questions si vous en avez.