STATISTIQUE
SCIENCE DES DONNÉES BIOSTATS
UNIVERSITÉ DE MONTPELLIER

Master 2 Statistiques et Sciences des Données – BioStats

Année universitaire 2024-2025

Comparaison par paire généralisée pour plusieurs critères de jugement classés par ordre de priorité : win ratio vs. Net benefit of treatment vs. win Odds

Auteur : Arthur TENA

Encadrante : Maïlis AMICO

Jury : Ali Gannoun, Nicolas Meyer, Élodie Brunel

IDESP
INSTITUT DESBREST
D'ÉPIDÉMIOLOGIE ET DE SANTÉ PUBLIQUE

Inserm
La science pour la santé
From science to health

Année 2024 – 2025

Abréviations

Ces abréviations seront utilisées tout au long du rapport :

NB(T) : Net Benefit (of Treatment)
tte : time-to-event
GGW : Gehan's Generalized Wilcoxon
WR : win ratio
WO : win Odds
T : individus appartenant au groupe traité
C : individus appartenant au groupe contrôle
HR : Hazard Ratio
Otc : Outcome

Table des figures

Liste des tableaux

REMERCIEMENTS

Tout d'abord je souhaite remercier ma tutrice de stage Mailïs Amico pour son accueil, sa confiance et sa bienveillance. Je te remercie pour ton soutien qui m'a permis de grandir personnellement et professionnellement. Le stage s'est déroulé au sein d'un environnement agréable et joyeux qui m'a permis de m'épanouir.

Je remercie également toute l'équipe de l'IDESP pour son accueil chaleureux, son accompagnement et les échanges enrichissants dont j'ai pu bénéficier.

Table des matières

1 Introduction

1.1 Cadre du stage

Mon stage se déroule à l'**IDESP** (Institut Desbrest d'Épidémiologie et de Santé Publique), une unité de recherche rattachée à la fois à l'**INSERM** (Institut National de la Santé et de la Recherche Médicale) et à l'**Université de Montpellier (UM)**.

L'IDESP a pour vocation de modéliser et d'optimiser la prévention et la prise en charge des maladies chroniques, en s'appuyant sur l'analyse de données massives issues du domaine de la santé et de l'environnement. Ses travaux s'articulent autour de trois axes de recherche principaux :

- **Axe 1** : étude des déterminants environnementaux des maladies chroniques ;
- **Axe 2** : analyse des déterminants influençant les stratégies de prévention et de soins chez les patients atteints de maladies chroniques ;
- **Axe 3** : exploration des interactions entre données multimodales et trajectoires menant aux maladies chroniques.

C'est au sein de ce troisième axe que s'inscrit l'équipe dans laquelle je réalise mon stage.
L'objectif scientifique global de l'IDESP est de mieux comprendre le développement et l'évolution des maladies chroniques, qui concernent plus de 30 % de la population et représentent environ 80 % des dépenses de santé. L'enjeu est d'imaginer des parcours de soin et de prévention personnalisés et efficaces pour des groupes de patients cibles. Les premières modélisations portent notamment sur les maladies respiratoires et allergiques, le diabète, l'obésité, ainsi que les pathologies rénales et vasculaires.
Cette unité de recherche multidisciplinaire, dirigée par Pascal Demoly et Isabella Annesi-Maesano, et codirigée par Nicolas Molinari et Gregory Ninot, regroupe 83 membres. Elle accueille des chercheurs et praticiens issus de l'INSERM, de l'**INRIA** (Institut National de Recherche en Informatique et en Automatique), de l'Université de Montpellier, des CHU de Nîmes et Montpellier, de l'Institut du Cancer de Montpellier ainsi que du Département Universitaire de Médecine Générale de Montpellier.

1.2 Contexte

Dans l'évaluation des traitements en recherche médicale, comparer l'efficacité de deux groupes repose souvent des critères multiples, tels que la survie, les évènements indésirables, ou encore la qualité de vie. Pendant longtemps, les analyses statistiques utilisées dans ce domaine ont principalement mobilisé des méthodes classiques — telles que les tests usuels, les régressions logistiques ou les modèles de Cox — qui sont adaptées pour analyser séparément plusieurs critères de jugement bien définis. Ce cadre d'analyse est généralement accepté et fonctionne bien tant que les critères sont traités individuellement. En revanche, des limites méthodologiques apparaissent lorsque l'on utilise un critère de jugement principal composite. En effet, les approches traditionnelles ne tiennent pas compte de la priorité potentielle entre les composantes du critère composite, ce qui conduit à comparer les sujets sur des bases hétérogènes. Pour remédier à cela, des méthodes d'évaluation des critères composites tenant compte de la hiérarchie ou de la priorité entre les composantes ont été développées afin d'améliorer la pertinence clinique et la robustesse des comparaisons.

Ce n'est qu'à partir de 2010, avec l'émergence d'articles pionniers, que des méthodes spécifiques à la comparaison par paires ont vu le jour, ouvrant la voie à une nouvelle façon de concevoir l'analyse comparative dans les essais cliniques. Le changement fondamental apporté par ces approches est la possibilité de hiérarchiser les critères de jugement : on ne se contente plus d'évaluer le premier critère survenu, mais on hiérarchise les évènements selon leur gravité ou leur pertinence médicale. Trois méthodes, le **win Ratio**, le **Net Benefit** et le **win Odds** se sont alors distinguées pour répondre à ces enjeux.

1.3 Plan du rapport

Ce travail a pour objectif de comparer ces trois méthodes, en les décrivant d'abord théoriquement dans la section 2, puis en les testant via des simulations selon différents scénario dans la partie 3, avant

de les appliquer à un jeu de données réel dans la section 4. L'intérêt est double : d'une part mieux comprendre les forces et limites de chaque approche, et d'autre part, offrir un éclairage pratique sur leur utilisation dans un contexte médical, encore trop peu exploré jusqu'ici.

2 Description des méthodes

Un essai clinique est une recherche effectuée sur des volontaires pour évaluer les effets d'un traitement ou d'une intervention médicale.

Lors d'un essai clinique, plusieurs phases se succèdent, chacune ayant un objectif spécifique visant à évaluer la sécurité, l'efficacité et les effets secondaires d'un traitement ou d'une intervention.

- **Phase 0** : Étude préclinique exploratoire, visant à observer le comportement du médicament à très faible dose chez les animaux ou les cellules.
- **Phase I** : Évaluation de la sécurité et du dosage. Cette phase vise principalement à identifier les effets secondaires, la tolérance et la dose maximale tolérée.
- **Phase II** : Étude de l'efficacité et de la posologie. Elle permet d'affiner la dose optimale et d'évaluer les premiers signaux d'efficacité sur une période de quelques mois à un an.
- **Phase III** : Confirmation de l'efficacité et surveillance des effets secondaires sur une population plus large. Cette phase est souvent déterminante pour l'autorisation de mise sur le marché et peut durer plusieurs années.
- **Phase IV** : Surveillance post-commercialisation, destinée à suivre les effets à long terme, détecter d'éventuels effets indésirables rares et évaluer l'utilisation du traitement en conditions réelles.

Au cours de ce travail, plusieurs notions ont nécessité un approfondissement. Tout d'abord, il était important de comprendre la définition d'un **critère de jugement** que nous verrons dans la partie 2.1. Ensuite, dans la partie 2.2, nous allons étudier la **comparaison par paire**, ce que c'est et quelles sont ses utilités. Nous regarderons ensuite trois méthodes de calcul pour définir si un nouveau traitement est meilleur qu'un traitement de contrôle, le **net benefit of treatment** (NBT) dans la partie 2.2.1, le **win ratio** dans la partie 2.2.2 et enfin une approche un peu différente du win ratio, les **win odds** dans la partie 2.2.3.
Nous illustrerons nos méthodes à l'aide d'un exemple d'accident cardiovasculaire.

2.1 Critère de jugement

Le critère de jugement (en anglais *outcomes* ou *outcome*) est expliqué par McCoy (2018), lors de l'évaluation des résultats d'une étude clinique. Lors d'un essai clinique, les **critères de jugement** sont des indicateurs permettant d'évaluer l'efficacité et la sécurité d'un traitement. On distingue généralement un **critère de jugement principal** (*primary outcomes*), qui représente l'événement clinique le plus important pour l'étude, et des **critères de jugement secondaires**(*secondary outcome*), qui sont des évènements pertinents permettant d'augmenter la justesse statistique et de mieux refléter l'impact global du traitement. Il est toute fois possible d'avoir un critère de jugement principal comprenant plusieurs variables, on parlera alors d'un **critère de jugement composé**.
Par exemple, dans une étude évaluant l'efficacité d'un traitement sur la prévention des événements cardiovasculaires majeurs, le critère de jugement principal pourrait être la mortalité cardiovasculaire, c'est-à-dire le décès dû à une cause d'origine cardiaque. Cependant, comme la mortalité cardiovasculaire seule peut être un événement relativement rare sur la durée de l'étude, on peut également définir un critère de jugement principal composé, pouvant inclure plusieurs événements cliniquement significatifs, tels que l'hospitalisation pour infarctus du myocarde (IDM) ou pour un accident vasculaire cérébral (AVC) en plus de la mortalité cardiovasculaire.
Ainsi, le choix entre un critère de jugement principal unique et un critère composé dépend des objectifs de l'étude, de la fréquence des événements et de leur importance clinique.
Avant de présenter les méthodes qui nous intéressent, il est important de comprendre comment les statisticiens faisaient avant leur introduction. L'analyse se focalisait uniquement sur le critère principal avec un test d'hypothèse, lorsque plusieurs critères étaient cliniquement important, on les combinait pour en faire des critères de jugement composite mais seul le premier évènement survenu était utilisé sans prendre en compte sa priorité.
Les méthodes que nous allons présenter quant à elles conservent l'intérêt d'utiliser les critères de jugement composés tout en donnant des priorités à ces critères.

2.2 Comparaison par paire

La comparaison par paire, comme l'explique Marc Buyse (2010) est une méthode de comparaison de deux échantillons, un échantillon T de n patients essayant un nouveau traitement et un échantillon de contrôle C de m patients prenant un traitement de contrôle (ancien traitement ou placebo).
Cette comparaison requiert de constituer des paires d'individus, composé d'un individu du groupe T et d'un individu du groupe C, et de comparer les deux individus de chaque paire. Une paire est dite "**favorable**" si le résultat de l'individu du groupe T est meilleur [1] que celui du groupe C. Elle est dite "**défavorable**" si le résultat de T est moins bon que celui de C. "**Neutre**" si le résultat de T ne change pas de celui de C, et enfin "**non-informatif**" si on ne peut pas déterminer quel groupe a le meilleur résultat.

Explication dans le cas avec un seul outcome

Plusieurs types de variables peuvent être étudiés, des variables binaires, des variables continues ou des variable "tte" (time-to-event). On peut alors voir que pour les variables binaire les quatre cas de figure et le résultat de la paire associé sur les tableaux suivant. Pour les variables continue et tte, un seuil d'acceptation τ est définit. Pour les variables continues trois cas de figure sont étudiés, si la différence de la paire est supérieur à τ alors le résultat est favorable, si la différence est inférieur à $-\tau$ alors le résultat est défavorable et la paire sera désignée neutre si la valeur absolue de la différence est inférieure ou égal à τ. Enfin pour les variables tte, plus de cas de figure sont étudiées car ces variables combinent les variables binaires (valant 1 si l'observation est complète, donc non censurée et 0 si l'observation est censurée) et les variables continues (temps observée). Ainsi douze cas de figure sont à prendre en compte suivant s'il y a observation complète ou non et comment se comporte la différence de la paire par rapport à τ. Tout ceci est mieux détaillé dans les tableaux 1, 2 et 3.
Pour les variables binaire, une paire est dite favorable si la valeur de l'individu du groupe T vaut 1 et celle du groupe C vaut 0, au contraire, une paire sera dite défavorable si la valeur de l'individu du groupe T vaut 0 et celle de l'individu du groupe C vaut 1. Enfin, une paire sera dite neutre si la valeur des deux individus est égale.

X_i dénote le résultat du i-ème individu $i \in \{1, ..., n\}$ du groupe T et Y_j dénote le résultat du j-ème individus $j \in \{1, ..., m\}$ du groupe C. Pour les variables continues, on prendra un seuil τ dépendant de l'étude et de ce que l'on veut.

Comparaison par paire	Paire
$X_i = 1$, $Y_j = 0$	favorable
$X_i = 1$, $Y_j = 1$	neutre
$X_i = 0$, $Y_j = 0$	neutre
$X_i = 0$, $Y_j = 1$	défavorable

TABLE 1 – Comparaison par paire pour des variables binaires

Comparaison par paire	Paire
$X_i - Y_j > \tau$	favorable
$\lvert X_i - Y_j \rvert \leq \tau$	neutre
$X_i - Y_j < -\tau$	défavorable

TABLE 2 – Comparaison par paire pour des variables continues

GPC signifie comparaison par paire généralisée et englobe les 3 méthodes statistiques.
Dans le cadre de variable tte, des censures à droite peuvent arriver, c'est pourquoi les variables ϵ et η dénotent la non-censure de X et Y respectivement c'est à dire les observations non-censurées. Et on note X_i' (resp. Y_j') les observations censurées du groupe T (resp. C).

1. un meilleur résultat dépend de ce que l'on cherche. "Meilleur" peut être plus grand ou plus petit suivant l'étude.

Censure de X et Y	Comparaison par paire	Paire
$\epsilon_i = 1$, $\eta_j = 1$	$X_i - Y_j > \tau$	favorable
	$\lvert X_i - Y_j \rvert \leq \tau$	neutre
	$X_i - Y_j < -\tau$	défavorable
$\epsilon_i = 0$, $\eta_j = 1$	$X'_i - Y_j > \tau$	favorable
	$\lvert X'_i - Y_j \rvert \leq \tau$	non-informative
	$X'_i - Y_j < -\tau$	non-informative
$\epsilon_i = 1$, $\eta_j = 0$	$X_i - Y'_j > \tau$	non-informative
	$\lvert X_i - Y'_j \rvert \leq \tau$	non-informative
	$X_i - Y'_j < -\tau$	défavorable
$\epsilon_i = 0$, $\eta_j = 0$	$X'_i - Y'_j > \tau$	non-informative
	$\lvert X'_i - Y'_j \rvert \leq \tau$	non-informative
	$X'_i - Y'_j < -\tau$	non-informative

Table 3 – GPC pour des variables tte

L'idée globale est donc de comparer chaque paire d'individu afin de donner un score qui sera utilisé pour déterminer l'utilité du nouveau traitement.
Nous allons alors créer un score $p_{ijk} \in \{-1, 0, 1\}$ qui nous servira par la suite où,

$$p_{ijk} = \begin{cases} 1 & \text{si la paire est favorable} \\ -1 & \text{si la paire est défavorable} \\ 0 & \text{si la paire est neutre} \end{cases}$$

Remarque : Les variables peuvent également nécessiter une stratification, par exemple en fonction de l'âge, du sexe, du stade de la maladie ou d'un risque que le patient encoure. Cette stratification sera indiquée par l'indice k ($k \in \{1, \ldots, K\}$), définissant ainsi des sous-groupes de taille n_k et m_k pour les groupes T et C, respectivement. Elle sera utile pour comparer des individus semblable selon ses caractéristiques.

Dans l'exemple d'accident cardiovasculaire, la comparaison par paire va comparer 2 à 2 chaque patient de chaque groupe en attribuant ce score suivant si le critère de jugement principal est arrivé. Chaque patient étant comparé avec des patients similaire afin de ne pas fausser l'étude.

Dans le cas particulier où $\tau = 0$, la comparaison par paire est équivalente à un test de Wilcoxon pour les variables continues et à un test de Gehan (GGW, Gehan (1965)) qui est un test de Wilcoxon généralisé aux variables tte.

Extension dans le cas de plusieurs critères

La comparaison par paire peut être généralisée de plusieurs façons : en intégrant plusieurs seuils, en tenant compte d'observations répétées dans le temps pour un même critère de jugement, ou encore en considérant des critères de jugement composés. Chaque extension présente un intérêt particulier. L'utilisation de seuils successifs permet, par exemple, de définir différents stades d'une maladie. L'intégration d'observations répétées offre un suivi des patients et facilite la prédiction des critères de jugement. Enfin, prendre en compte plusieurs critères de jugement est pertinent lorsque l'analyse ne peut se limiter à un seul indicateur. C'est cette dernière approche que nous explorerons dans la suite.

On notera alors l ($l \in \{1, ..., L\}$) le l-ième critère de jugement classé par priorité où 1 est la plus grosse priorité et L la plus basse. On aura alors une mesure $u_{ijk}(l)$ comme suit

$$u_{ijk}(l) = \begin{cases} 1 & \text{si le } l\text{-ième critère est non-informatif} \\ 0 & \text{sinon} \end{cases}$$

et un score associé :

$$p_{ijk}(l) = \begin{cases} 1 & \text{si le } l\text{-ième critère est favorable} \\ -1 & \text{si le } l\text{-ième critère est défavorable} \\ 0 & \text{si le } l\text{-ième critère est neutre} \end{cases} \quad ; \quad \text{et} \quad u_{ijk}(h) = 1 \quad \forall h < l \tag{1}$$

Cette deuxième partie $u_{ijk}(h) = 1 \quad \forall h < l$ joue un rôle clé dans la priorisation des critères de jugement. En d'autres termes, cela signifie que si tous les critères plus prioritaires que l (celui utilisé pour calculer $p_{ijk}(l)$) sont considérés comme non-informatifs, alors on passe au critère suivant. Les conditions précises définissant un critère comme non-informatif sont détaillées dans les tableaux 1, 2 et 3.
Ce score sera utile pour calculer un proportion en faveur du traitement évoqué dans la partie suivante sur l'estimation de la statistique et les tests.

Reprenons notre exemple d'accident cardiovasculaire avec, comme différents critère composé la mortalité cardiovasculaire ayant la priorité maximale, l'hospitalisation pour infarctus du myocarde ayant la seconde priorité et l'AVC avec la dernière priorité. Nous allons devoir comparer chaque patient d'une même strate 2 à 2 afin de voir si les 3 critères, classés par ordre de priorité, sont en faveur du nouveau traitement ou non. Pour ce faire, nous regarderons d'abord si le critère prioritaire (le décès) est non-informatif ou non, s'il ne l'est pas alors on calculera le score p de ce critère, alors que s'il est non-informatif on passera au critère suivant (infarctus du myocarde par exemple) et on calculera le score p correspondant etc.

Les 3 méthodes peuvent être utilisées pour comparer deux groupes sur base d'un critère composé et qui utilise la comparaison par paire.

2.2.1 Net Benefit of treatment

Le net benefit of treatment (NBT), d'après Buyse (2010), s'appuie sur 2 notions, la *proportion en faveur du traitement* (noté Δ) et la *proportion en faveur du traitement cumulé* (noté δ).
Dans le NBT, nous comparons toutes les paires d'individus, nous avons donc $n.m$ paires qui peuvent être classées par strate, toujours dans le but de comparer des individus identique.
La **proportion en faveur du traitement** Δ est la différence entre le nombre de paire favorable et le nombre de paire défavorable divisé par le nombre de paire totale.

$$\Delta = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{k=1}^{m_k} p_{ijk}}{\sum_{k=1}^{K} n_k.m_k} \tag{2}$$

Avec $n = \sum_{k=1}^{K} n_k$, $m = \sum_{k=1}^{K} m_k$ et $\sum_{k=1}^{K} n_k.m_k \leq n.m$.
Δ variant entre -1 et 1, où une valeur de -1 signifie que l'ancien traitement est uniformément meilleur que le nouveau, et une valeur de 1 signifie que le nouveau traitement est uniformément meilleur que l'ancien.
Ce qui est différent de la **proportion en faveur du traitement cumulé** dépendant elle du critère de jugement l,

$$\delta(l) = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{k=1}^{m_k} p_{ijk}(l)}{\sum_{k=1}^{K} n_k.m_k} \tag{3}$$

On a alors que $\Delta(l) = \sum_{i \leq l} \delta(i)$ et $\Delta = \Delta(L)$

Afin de vérifier s'il existe bien une différence significative entre les groupes, des tests d'hypothèses sont effectués. L'hypothèse nulle signifie qu'il n'y a pas de différence entre les groupes c'est à dire $H_0 : \Delta = 0$ et l'hypothèse alternative correspond au fait qu'il existe une différence entre les groupes $H_1 : \Delta \neq 0$.

Les tests de randomisation seront réalisés sur la valeur de Δ sous l'hypothèse nulle.
Nous pouvons alors créer, à l'aide d'un nombre de test de randomisation suffisant (S), des intervalles de confiance à $(1-\alpha)\%$ basé sur Δ_{obs}, la valeur de Δ observée avant le test de permutation. Cet intervalle de confiance est calculé comme suit :

$$[\Delta_{obs} + \Delta_{\alpha/2}, \Delta_{obs} + \Delta_{1-\alpha/2}]$$

où $\Delta_{\alpha/2}$ est la valeur de Δ_i ($i \in \{1, \ldots, S)$ qui laisse au plus $\alpha/2$ des valeurs de Δ_i à sa gauche et $\Delta_{1-\alpha/2}$ la valeur qui laisse au plus $\alpha/2$ des valeurs de Δ_i à sa droite.
Les tests sur la proportion en faveur du traitement peuvent être effectués de la même façon pour la proportion en faveur du traitement cumulé dans le cas où nous avons plusieurs critères de jugements.

Le test de randomisation est ici important car il permet de se placer sous l'hypothèse H_0 en randomisant le groupe des individus, les valeurs des individus reste inchangé mais leur groupe, lui, est randomisé. Il permet d'évaluer si les différences observées dans les données peuvent être attribuées à un effet réel ou si elles sont dû au hasard. Pour ce faire, on calcul la statistique observées Δ_{obs} pour les données réelles, puis on permute les données en réaffectant aléatoirement les étiquettes du groupe un grand nombre de fois (ici noté S). On calcul la statistique d'intérêt à chaque permutation pour obtenir une distribution de la statistique sous l'hypothèse nulle (vecteur Δ_i) et nous comparons la statistique observée avec la distribution des statistiques permutées pour obtenir la probabilité d'obtenir une différence aussi grande ou plus grande que celle observée sous l'hypothèse nulle.

Relation entre les tests de Wilcoxon, de Gehan et de Fisher avec la GPC

Le NBT peut être vu comme un cas particulier des tests de Wilcoxon et de GGW dans le cas où $\tau = 0$. C'est ce que nous allons regarder. Alors que lorsque la variable est binaire, le NBT est un cas particulier du test de Fisher à la condition qu'il y ait le même nombre d'individus dans les deux groupes T et C.

Dans le cadre de variable continue, nous utiliserons des tests de rang de Wilcoxon ou des des tests de Mann-Whitney.
Pour rappel, le test de rang de Wilcoxon est construit comme suit. Par soucis de simplicité, nous n'aborderons pas le cas où les variables sont stratifiées. Soit $S_1 < S_2 < ... < S_n$ les rangs des observations du groupe T, la statistique de Wilcoxon est alors

$$W_S = \sum_{i=1}^{n} S_i$$

La statistique de Mann-Witney s'écrit

$$W_{MW} = W_S - \frac{n.(n+1)}{2}$$

Dans le cadre de variable tte, Gehan (1965) a proposé un généralisation de la statistique de Wilcoxon en incluant la possibilité d'avoir des évènements censurés. Gehan propose alors de séparer les observations de X et de Y suivant si elles sont censurées ou non. On aura alors :

$$\begin{aligned} x'_1, ..., x'_{r_1}, &\quad r_1 \text{ censurés} \\ x_{r_1+1}, ..., x_n, &\quad n - r_1 \text{ échec} \\ y'_1, ..., y'_{r_2}, &\quad r_2 \text{ censurés} \\ y_{r_2+1}, ..., y_m, &\quad m - r_2 \text{ échec} \end{aligned}$$

où les valeurs de x (resp. y) sont les temps jusqu'à l'échec des personnes ayant pris le traitement T (resp. C), x' (resp. y') est le temps jusqu'à la censure des personnes ayant pris le traitement T (resp. C).

Gehan définit alors sa statistique comme suit :

$$U_{ij} = \begin{cases} 1 & \text{si } x_i > y_j \text{ ou } x'_i \geq y_j \\ -1 & \text{si } x_i < y_j \text{ ou } x_i \leq y'_j \\ 0 & \text{si } x_i = y_j \text{ ou } (x'_i, y'_j) \text{ ou } x'_i < y_j \text{ ou } y'_j < x_i \end{cases} \tag{4}$$

On peut alors calculer la statistique W, $W = \sum_{i,j} U_{i,j}$. On peut montrer que Δ est en fait la U-statistique pour le test de Wilcoxon,

$$U = \Delta = \frac{1}{n.m} \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}$$

où

Pour des variables continues :

$$U_{ij} = \begin{cases} 1 & \text{si } X_i > Y_j \\ -1 & \text{si } X_i < Y_j \\ 0 & \text{sinon} \end{cases}$$

Pour des variables tte :

$$U_{ij} = \begin{cases} 1 & \text{si } x_i > y_j \text{ ou } x'_i \geq y_j \\ -1 & \text{si } x_i < y_j \text{ ou } x_i \leq y'_j \\ 0 & \text{si } x_i = y_j \text{ ou } (x'_i, y'_j) \text{ ou } x'_i < y_j \text{ ou } y'_j < x_i \end{cases}$$

2.2.2 win Ratio

Explication

Pocock et al. (2012) propose une approche de l'analyse des critères de jugement basés sur leurs priorité en définissant le **win Ratio**.

Lors d'une approche win Ratio, nous allons utiliser une méthode comparable au NBT, il faut donc prendre en compte que deux patients n'auront pas les mêmes risques suivant leur profil (âge, sexe, diabète etc) afin de s'assurer que la différence soit due aux traitements et non pas aux profils.

Le *win Ratio* propose deux approches : une approche **appariée** et une approche **non-appariée**. L'approche appariée compare un nombre fixe de paires d'individus comme étant le minimum entre les n individus du groupe T et les m individus du groupe C, chaque paire étant formée d'un individu du premier groupe et d'un individu du second groupe selon leur risque et seront alors comparés. L'approche non-appariée compare toutes les combinaisons possibles de paires entre les deux groupes, c'est-à-dire que chaque individu du premier groupe est comparé à tous les individus du second groupe.

Le win Ratio sera aussi calculé à l'aide des $p_{ijk}(l)$ mais d'une façon différente du NBT, cette fois-ci, trois quantités sont calculées, le nombre de win N_w le nombre de lose N_l et le nombre d'égalité N_t.

$$\begin{cases} N_w = Card(\{p_{ijk}(l) = 1\}) \\ N_l = Card(\{p_{ijk}(l) = -1\}) \\ N_t = Card(\{p_{ijk}(l) = 0\}) \end{cases}$$

Où $p_{ijk}(l)$ est définit par l'équation (1)

Le win Ratio (WR) sera alors :

$$WR = \frac{N_w}{N_l}$$

Le WR varie sur $\mathbb{R}_+^*$ où une valeur de 1 signifie qu'il n'existe pas de différence entre les groupes.

Afin de calculer l'intervalle de confiance du WR, il faudra calculer la proportion de win p_w, valant $p_w = \frac{N_w}{N_w + N_l}$ on pourra alors établir des intervalles de confiance à 95% de cette proportion :

$$\left[p_w \pm 1.96 \left(\frac{p_w(1 - p_w)}{N_w + N_l} \right)^{1/2} \right] = [p_L, p_U]$$

où p_L est la borne inférieure de l'intervalle de p_w et p_U est la borne supérieure.

Comme le win ratio vaut $\frac{p_w}{1-p_w}$, on établit l'intervalle à 95% du win ratio comme étant :

$$\left[\frac{p_L}{(1 - p_L)}, \frac{p_U}{(1 - p_U)} \right]$$

De même, lorsque nous voulons faire un test d'hypothèse afin d'évaluer s'il existe une différence entre les groupes, on peut calculer le z-score, $z = (p_w - 0.5)/[p_w(1-p_w)/(N_w+N_l)]^{1/2}$ suivant une distribution normale standardisée. Nous pouvons alors trouver la p-valeur dans la table de la loi normale-centrée réduite.

2.2.3 win odds

Le **win Odds** évoqué par Dong et al. (2020) et Brunner et al. (2021) est une adaptation des win ratio incluant la possibilité d'avoir des égalités. De la même façon que pour le win ratio, il existe deux approches pour le win odds, l'approche appariée et l'approche non-appariée. Les deux papiers proposent d'inclure dans le ratio la moitié du nombre d'égalité au numérateur et au dénominateur comme suit :

$$WO = \frac{N_w + 0.5N_t}{N_l + 0.5N_t}$$

Le WO prendra ses valeurs dans $\mathbb{R}_+^*$, comme le WR.

L'hypothèse nulle est la même que le WR, $H_0 : WO = 1 \quad ; \quad H_1 : WO \neq 1$. Autrement dit, si le WO vaut 1 il n'y a pas de différence entre les groupes alors que s'il est inférieur à 1, le groupe C est meilleur que le groupe T et s'il est supérieur à 1 le groupe T est meilleur que le groupe C.

Brunner et al. (2021) nous montre comment définir les intervalles de confiance et les tests d'hypothèse.

2.2.4 Transformation du NBT

Je me suis aussi penché sur un transformation du NBT, compris entre -1 et 1 pour faire un parallèle avec le WO et le WR. Une transformation classique à regarder pour passer d'un objet appartenant à [-1,1] à un objet appartenant à $\mathbb{R}_+$ est :

$$x = \frac{y+1}{y-1}$$

où $y \in [-1, 1]$. Et je me suis rendu compte de ceci, en utilisant la formule du NBT suivante $y = \frac{W-L}{W+L+T}$, on obtient ceci :

$$\begin{aligned} \frac{NB+1}{1-NB} &= \frac{\frac{W-L}{W+L+T}+1}{1-\frac{W-L}{W+L+T}} \\ &= \frac{2W+T}{(W+L+T)}\frac{(W+T+L)}{2L+T} \\ &= \frac{W+0.5T}{L+0.5T}\text{def}{=}WO \end{aligned}$$

Où W, L et T représentent respectivement le nombre de win, de loose et de tie.

La différence entre les deux statistiques est donc la façon dont le statisticien souhaite interpréter le résultat. En effet, bien que la GPC et le win Odds soient liés par une transformation mathématique, la différence principale réside dans leur interprétation. La GPC permet d'estimer une proportion de paires favorables, ce qui est souvent plus intuitif pour interpréter l'effet d'un traitement d'un point de vue clinique. En revanche, le win Odds se lit comme un rapport de probabilités de gagner par rapport à perdre, ce qui le rend plus proche des statistiques classiques comme les hazard ratios. De plus, la GPC permet d'examiner séparément les contributions de chaque critère de jugement à la statistique globale, ce qui peut être particulièrement utile pour identifier les composantes dominantes dans l'effet observé. Ainsi, le choix entre ces deux mesures dépend de l'objectif de l'analyse : mettre en avant une interprétation clinique ou obtenir une mesure synthétique adaptée à des tests statistiques.

Exemples

Dans cette partie, nous allons voir comment sont calculées les méthodes.

	Endpoint 1		Endpoint 2	
id	T	C	T	C
1	0.5	0.1	0.2	0.4
2	0.4	0.4	0.6	0.4
3	0.3	0.3	0.6	0.6
4	0.9	0.8	0.7	0.5
5	0.6	0.6	0.5	0.6
6	0.8	0.8	0.4	0.4
7	0.7	0.8	0.2	0.4

TABLE 4 – Exemple utilisé

Approche appariée :
Nombre de win : 3
Nombre de loose : 2
Nombre de tie : 2
GPC = $\frac{3-2}{3+2+2} = \frac{1}{7} = 0.143$
WR = $\frac{3}{2} = 1.5$
WO = $\frac{3+1}{2+1} = \frac{4}{3} = 1.33$

Approche non-appariée :
Nombre de win : 24
Nombre de loose : 21
Nombre de tie : 3
GPC = $\frac{24-16}{24+21+3} = \frac{8}{47} = 0.17$
WR = $\frac{24}{21} = 1.143$
WO = $\frac{24+1.5}{21+1.5} = 1.133$

On voit comment la fonction marche sur la gestion des égalité pour les individus 2, 3 et 5. Si le premier outcome (ou outcome) est une égalité alors on regarde l'outcome suivant pour le classer en win (id. 2) ou loose (id.5) mais si le dernier outcome est une égalité alors le résultat final sera tie (id.3). Cela fonctionne comme ça aussi pour l'approche non-appariée à la différence près que toutes les combinaison de paire seront effectuées.

2.3 Résumé

Ces méthodes ont en commun l'attribution d'un score permettant d'évaluer si le traitement testé est supérieur, inférieur ou équivalent au traitement de référence. La différence entre elles réside dans la manière dont cette évaluation est effectuée.

3 Simulations de données cliniques

Le code et les résultats se trouvent aussi sur mon github (`https://github.com/Arthur-tena/Comparaison-par-pa`
J'ai utilisé dans cette partie la bibliothèque WINS [2].

La simulation de données joue un rôle fondamental en statistique, notamment lors du développement et de l'évaluation de nouvelles méthodes statistiques. Elle permet d'explorer le comportement de ces méthodes dans un cadre contrôlé, en définissant des scénarios précis (par exemple avec différentes distributions d'échantillons ou des taux de censure variables), ce qui serait difficile à faire uniquement à partir de données réelles, souvent limitées ou partielles.
Dans le cadre de ce stage, j'ai donc rédigé un fichier `Rmarkdown` afin de tester un grand nombre de scénarios, dans le but d'évaluer et de comparer les méthodes. J'ai notamment exploré l'effet du changement de seuils de décision, simulé des groupes avec des distributions identiques dans un premier temps, puis de plus en plus divergentes, afin d'observer l'impact de ces variations sur les statistiques produites.

Pour simuler les données de temps de survie (tte, pour *time-to-event*), j'ai utilisé la méthode de la transformation inverse. Celle-ci consiste à tirer une variable aléatoire $u \sim \mathcal{U}(0,1)$, puis à appliquer à u l'inverse de la fonction de répartition de la loi choisie. Dans ce travail, j'ai alors considéré deux modèles, le modèle de Cox (où le *hazard ratio* est constant) et le modèle AFT (où le *hazard ratio* varie dans le temps) en considérant une distribution de Weibull pour le modèle de Cox et log-normale pour le modèle AFT.
Les données tte sont donc construites à partir de la fonction de survie conditionnelle à une covariable Z [3]. Pour le modèle de Cox considérant une distribution Weibull pour la fonction de survie baseline, cette fonction de survie s'écrit :

$$S(t|Z) = \exp\left(-\lambda t^{\gamma} \exp(\beta Z)\right)$$

où $\lambda > 0$ est le paramètre d'échelle, $\gamma > 0$ est le paramètre de forme de la loi de Weibull, et β est le coefficient associé à la covariable Z.

1. Modèle de Cox (HR constant). Dans ce cadre, on suppose que le *hazard ratio* entre deux individus reste constant dans le temps. On conserve donc une forme proportionnelle des risques.

Partons de :

$$S(t|Z) = \exp\left(-\lambda t^{\gamma} e^{\beta Z}\right)$$

Posons $S(t|Z) = 1 - u$ avec $u \sim \mathcal{U}(0,1)$, on obtient :

$$1 - u = \exp\left(-\lambda t^{\gamma} e^{\beta Z}\right)$$

Prenons le logarithme des deux côtés :

$$\log(1-u) = -\lambda t^{\gamma} e^{\beta Z}$$

Puis, en isolant t :

$$t = \left(\frac{-\log(1-u)}{\lambda e^{\beta Z}}\right)^{1/\gamma}$$

C'est la formule utilisée pour générer les tte sous un modèle de Cox à risques proportionnels.

2. `https://cran.r-project.org/web/packages/WINS/WINS.pdf`
3. Dans mon cas je n'utiliserai uniquement la covariable "patient traité, oui/non"

2. Modèle AFT (HR non constant). Le modèle *Accelerated Failure Time* ne suppose pas des risques proportionnels. Au lieu de modéliser la fonction de risque, il modélise directement le temps de survie via une transformation linéaire du logarithme du temps.

On considère le modèle **Accelerated Failure Time** (AFT) suivant :

$$\log(T) = \beta Z + \sigma\varepsilon$$

Où ε est i.i.d. et dont la distribution n'est pas spécifiée, T est le temps observé, β est un vecteur de paramètres Z est l'ensemble des covariables. Ce qui implique

$$T = \exp(\beta Z + \varepsilon)$$

La fonction de survie du modèle est

$$S(t|z) = S_0\left(\frac{t}{\exp(\beta^t z)}\right)$$

Où $S_0(.)$ est la fonction de survie de base, c'est à dire la probabilité de survie en fonction du temps pour un individu de référence (avec des covariables égales à zéro).

On peut noter aussi que la fonction des hazard ratio pour un modèle AFT est la suivante :

$$\lambda(t|z) = \lambda_0\left(\frac{t}{\exp(\beta^t z)}\right)(\exp(\beta^t z))^{-1}$$

Cela montre que, contrairement au modèle de Cox, le hazard ratio n'est pas constant dans le modèle AFT. en effet, on ne peut pas factoriser proprement le temps t à cause du terme $\frac{t}{\exp(\beta^t z)}$, ce qui rend la proportionnalité des risques inapplicable dans ce cadre.
Ce résultat montre que dans le modèle AFT, la variable de traitement Z agit directement sur le temps de survie : si un patient est traité (ou non), cela revient à étirer ou compresser le temps.
J'ai alors effectué deux fichiers de simulations, un Rmarkdown balayant plusieurs scénarios afin de tester les statistiques et essayer de dégager une tendance. Et un autre Rmarkdown où le but est de voir l'influence des HR lorsqu'il y a plusieurs mesures effectuées au cours du temps.

3.1 Premier notebook

Dans ce premier notebook comme dit plus haut, le but est de regarder les différents scénarios pouvant se produire avec des données réelles.
Nous effectuerons 2000 simulations selon des data frame contenant 200 individus dans chaque groupe classés en 4 strates équilibrées pré-définit. Les 4 strates peuvent être considérées comme un score de risque et sert à comparer les individus ayant le même risque.

Remarque : Le temps de compilation étant très long j'ai appris à paralléliser le code à l'aide des bibliothèques parallel, doParallel et foreach (disponible sur le CRAN) qui me serviront à utiliser plusieurs cœurs de l'ordinateur et faire des boucles "foreach" séparant le travail entre les différents cœurs. Je passe ici d'un temps d'exécution de 30 minutes à 2 minutes en utilisant 14 des 16 cœurs de mon ordinateur.

Nous avons commencé par effectuer des simulations selon deux situations. La première correspond au cas où les deux groupes suivent exactement la même distribution : chaque type de variable (outcome) suit la même loi avec les mêmes paramètres. Dans la seconde situation, le groupe T a des valeurs plus élevées que celles du groupe C.
Pour chacun de ces cas, nous avons testé plusieurs variantes. Nous avons fait varier le seuil τ en le fixant à 0 puis à 2 pour augmenter le nombre d'égalité. Nous avons aussi remplacé la variable continue normale par une variable discrète suivante une loi de Poisson. Enfin, nous avons modifié l'ordre des variables

observées : au lieu d'utiliser l'ordre tte/binaire/continue, nous avons testé l'ordre binaire/tte/continue.

Dans toute cette partie, nous nous focaliserons sur le WR et le WO car il est plus simple de les comparer étant donné qu'ils ont le même ordre de grandeur et que, comme nous l'avons montré à la fin de la section précédente, les WO et le NBT sont égaux à une transformation prête.
Dans ce notebook, les variables seront ordonnées comme ceci : premier outcome : survie, second outcome : binaire, troisième outcome : quantitatif. Des tableaux récapitulatifs seront situés aux début des sections où seront indiqué les lois, paramètres et ordre des outcomes considérés. Les données binaires seront simulées selon une distribution binomiale et les données quantitatives seront simulées selon une loi normale ou une loi de poisson.

3.1.1 Distributions identique dans les 2 groupes

Dans cette section, les distributions que nous allons évoquer seront choisis afin qu'il n'y ait pas de différence entre les groupes. Les priorités des critères de jugements sont données dans le tableau 5 ainsi que les différents cas de figures considéré à l'intérieur de ce scénario.
Nous verrons tout d'abord un premier cas où le seuil sera nul, puis un cas où le seuil sera augmenté à 2. Ensuite, nous changerons la loi de la variable quantitative, passant d'une loi continue normale à une loi de poisson donc discrète. Enfin, nous changerons l'ordre des outcomes en mettant en premier l'outcome binaire, en second l'outcome tte et en troisième l'outcome quantitatif continue.

Scénario	Fig.	outcome 1	outcome 2	outcome 3
$T \sim C\,;\, \tau = 0$	1a	$\lambda = 0.1,\ k = 2,\ \beta = 0$	$\mathcal{B}(0.5)$	$\mathcal{N}(3, 2)$
$T \sim C\,;\, \tau = 2$	1b	$\lambda = 0.1,\ k = 2,\ \beta = 0$	$\mathcal{B}(0.5)$	$\mathcal{N}(3, 2)$
Otc discret de Poisson	1c	$\lambda = 0.1,\ k = 2,\ \beta = 0$	$\mathcal{B}(0.5)$	$\mathcal{P}(3)$
Otc principal binaire	1d	$\mathcal{B}(0.5)$	$\lambda = 0.1,\ k = 2,\ \beta = 0$	$\mathcal{N}(3, 2)$

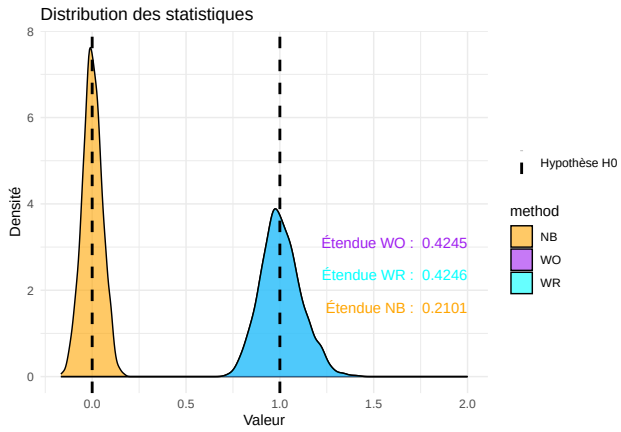
Table 5 – Tableau récapitulatif des distributions du premier scénario

La censure des variables tte suit une loi de Weibull $\mathcal{W}(2, 20)$ et a été choisie en fonction des données générées suivant le modèle de Cox.

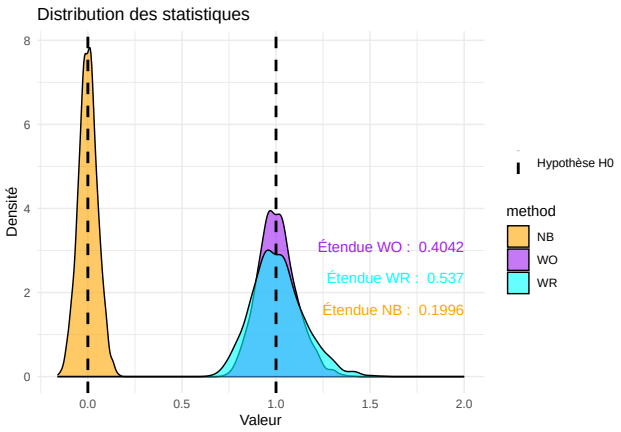
Table 6 – Tableau récapitulatif du taux de censure

scénario	T	C
$T \sim C\,;\, \tau = 0$	0.241 %	0.241 %
$T \sim C\,;\, \tau = 2$	0.241 %	0.241 %
Otc discret de Poisson	0.241 %	0.241 %
Otc principal binaire	0.241 %	0.241 %

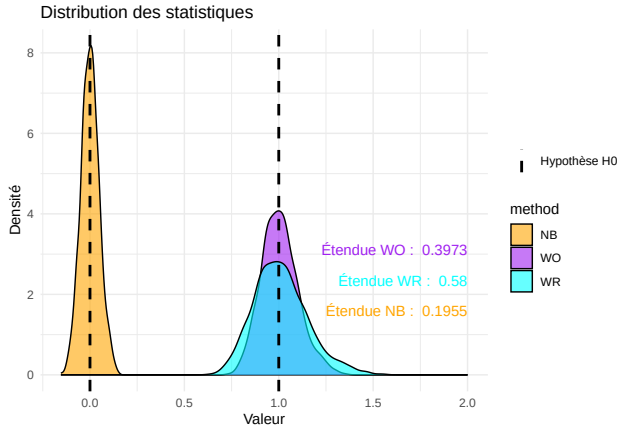
Dans le tableau 6, on remarque qu'il y a le même taux de censure pour tous les cas de figure de ce scénario. Ce résultat était attendu car les distributions sont identiques et que le paramètre β de la variable tte vaut 0 ne donnant donc pas de préférence à un groupe plutôt qu'à l'autre.



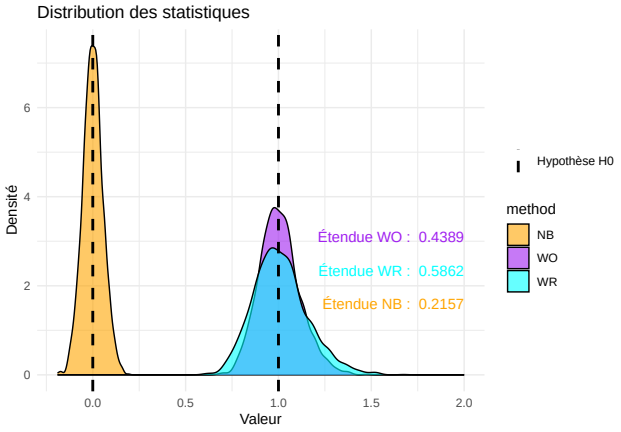
(a) $\tau = 0$



(b) $\tau = 2$



(c) Outcome discret de Poisson



(d) Outcome principal binaire

FIGURE 1 – Simulations suivant le premier scénario où les distributions sont égales

Ces graphiques représentent les distributions des statistiques sur les 2000 simulations
Sur ces graphiques, nous pouvons voir les hypothèses H_0 du NB valant 0 et des WR/WO valant 1. On y voit aussi les étendues des 3 statistiques, il s'agit de l'écart inter-quantile calculé à 95%, dans le cas où $\tau = 0$ sur la figure 1a on ne différencie pas le WR et le WO car il n'y a pas d'égalité mais nous commençons a voir des différence lorsque τ vaut 2 sur la figure 1b et où l'écart inter-quantile est d'environ 0.6 pour le WR et 0.43 pour le WO.

TABLE 7 – Simulations suivant le premier scénario où les distributions sont égales

	win	loose	tie
outcome 1 (tte)	782	787	8431
outcome 2 (binaire)	2110	2106	4215
outcome 3 (continue)	2110	2105	1
global	5002	4997	1

(a) $\tau = 0$

	win	loose	tie
outcome 1 (tte)	593	596	8811
outcome 2 (binaire)	2204	2201	5594
outcome 3 (continue)	977	973	8049
global	3775	3771	8049

(b) $\tau = 2$

	win	loose	tie
outcome 1 (tte)	593	596	8811
outcome 2 (binaire)	2204	2202	4404
outcome 3 (discret)	656	651	3097
global	6016	621	2354

(c) Outcome discret de Poisson

	win	loose	tie
outcome 1 (binaire)	2500	2500	5000
outcome 2 (tte)	297	298	4405
outcome 3 (continue)	977	973	2455
global	3774	3771	2455

(d) Outcome principal binaire

On remarque bien l'effet du changement de seuil sur les statistiques et cela se répercute sur l'écart inter-quantile et plus particulièrement sur le WR qui, ne prenant pas en compte les égalités, donne des valeurs plus variable bien que proche de l'hypothèse H_0. Nous remarquons sur les tableaux 7 représentant le nombre moyen de win/loose/tie que l'on retrouve dans les définitions des statistiques dans chacun des 4 cas considérés dans ce scénario que les égalités peuvent jouer un rôle central lors de l'estimation des statistiques, ce qui explique que les écarts inter-quantiles soient plus élevés pour le WR que pour le WO.

La fonction du package renvoi d'autres résultats intéressant, on y trouve le taux de censure moyen pour les 2 groupes et le taux de p-valeurs > 0.05 pour chaque statistiques montrant le taux de faux positif/faux négatif selon le scénario considéré.

Table 8 – Tableau récapitulatif du taux de p-valeur inférieur à 0.05

scénario	NBT	WR	WO
$T \sim C\,;\,\tau = 0$	0.046	0.048	0.048
$T \sim C\,;\,\tau = 2$	0.048	0.0495	0.0485
Otc discret de Poisson	0.0475	0.049	0.0485
Otc principal binaire	0.0525	0.0535	0.053

Dans le tableau 8, on peut voir le taux de p-valeur inférieur à 0.05 et on peut y voir que ce taux est équivalent pour tous les cas de figure de ce scénario et est assez logique au vu de ce qu'on avait été montré par Brunner et al. (2021). En effet, il avait montré que les p-valeur des statistiques restaient équivalente. On voit aussi bien qu'il y a environ 5% de chance de rejeter l'hypothèse H_0 à tord pour les 3 statistiques.

Nous allons maintenant considérer le scénario où les deux distributions n'auront pas les mêmes paramètres et donc où T est supérieur à C.

3.1.2 Groupe traité meilleur que le groupe contrôle

Dans cette section, les distributions que nous allons évoquer seront choisis afin de faire gaganr le groupe T. Les priorités des critères de jugements sont données dans le tableau 9 ainsi que les différents cas de figures considéré à l'intérieur de ce scénario.
Nous verrons tout d'abord un premier cas où le seuil sera nul, puis un cas où le seuil sera augmenté à 2. Ensuite, nous changerons la loi de la variable quantitative, passant d'une loi continue normale à une loi de poisson donc discrète. Enfin, nous changerons l'ordre des outcomes en mettant en premier l'outcome binaire, en second l'outcome tte et en troisième l'outcome quantitatif continue.

Scénario	Fig.	outcome 1	outcome 2	outcome 3
$T \gg C\,;\,\tau = 0$	2a	$\lambda = 0.1,\ k = 2,\ \beta = -2$	$\mathcal{B}_T(0.65)\,;\,\mathcal{B}_C(0.3)$	$\mathcal{N}_T(3,2)\,;\,\mathcal{N}_C(3,2)$
$T \gg C\,;\,\tau = 2$	2b	$\lambda = 0.1,\ k = 2,\ \beta = -2$	$\mathcal{B}_T(0.65)\,;\,\mathcal{B}_C(0.3)$	$\mathcal{N}_T(3,2)\,;\,\mathcal{N}_C(3,2)$
Otc discret de Poisson	2c	$\lambda = 0.1,\ k = 2,\ \beta = -2$	$\mathcal{B}_T(0.65)\,;\,\mathcal{B}_C(0.3)$	$\mathcal{P}_T(3)\,;\,\mathcal{P}_C(1)$
Otc principal binaire	2d	$\mathcal{B}_T(0.65)\,;\,\mathcal{B}_C(0.3)$	$\lambda = 0.1,\ k = 2,\ \beta = -2$	$\mathcal{N}_T(3,2)\,;\,\mathcal{N}_C(3,2)$

Table 9 – Tableau récapitulatif des distributions du second scénario

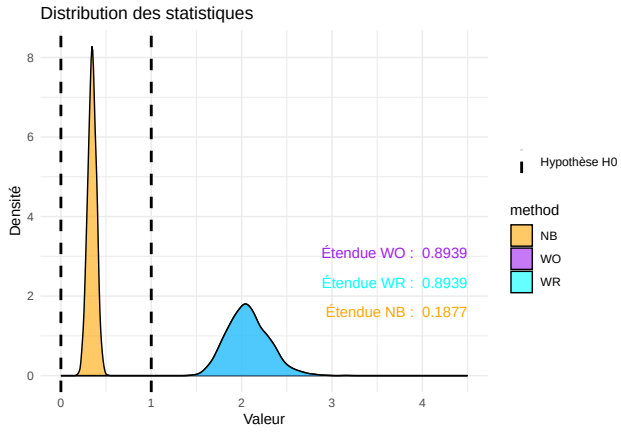
La loi de censure pour la variable tte est une loi de Weibull $\mathcal{W}(2, 20)$.

TABLE 10 – Tableau récapitulatif du taux de censure

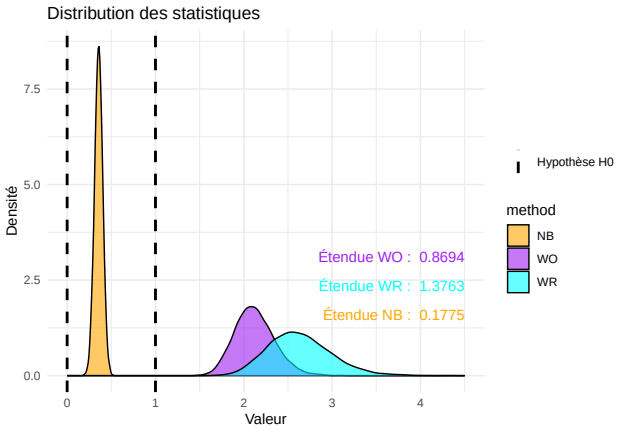
scénario	T	C
$T \sim C\,;\,\tau = 0$	0.549 %	0.242 %
$T \sim C\,;\,\tau = 2$	0.549 %	0.242 %
Otc discret de Poisson	0.549 %	0.242 %
Otc principal binaire	0.549 %	0.242 %

Sur la figure 2, on voit directement que les distributions des statistiques sont entièrement situées à droite des hypothèses H_0 respectives ce qui indique que le traitement est meilleur que le contrôle.
Sur la figure 2a les WO et WR sont confondues car il n'y a qu'un seul tie en moyenne (voir table 11a) et il n'y a donc pas de différence entre les statistiques, ce qui sera vérifié sur le tableau des résumé statistiques tab 49a.

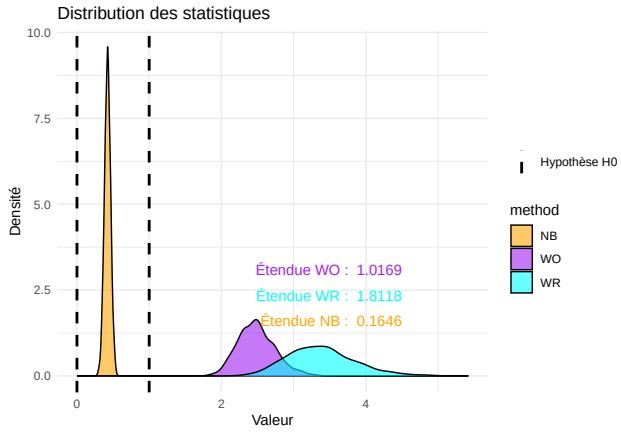
Maintenant, en regardant les autres figures, où le seuil est fixé à 2 pour tous les outcomes tte, continue et discret, on remarque une plus nette différence entre le WR et le WO. Le WO a une valeur plus faible que le WR et l'écart inter-quantile à 95% est plus faible ce qui démontre une plus faible variabilité du WO par rapport au WR. La valeur de la statistique du WO est plus faible que celle du WR ce qui implique que la différence entre le traitement et le contrôle est moins importante pour le Wo que pour le WR.



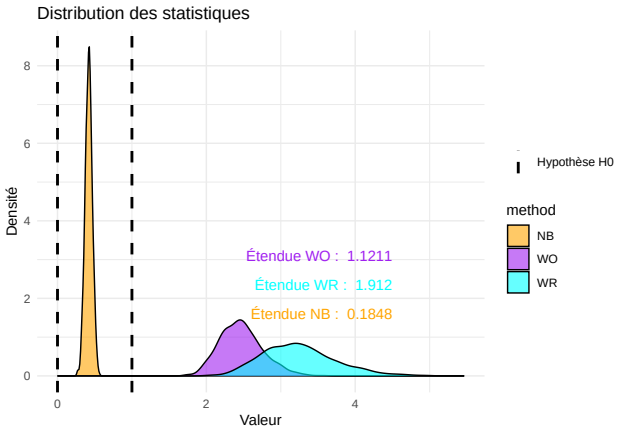
(a) $\tau = 0$



(b) $\tau = 2$



(c) Outcome discret de Poisson



(d) Outcome principal binaire

FIGURE 2 – Simulations suivant le second scénario où les distributions inégales

Les figures 2b et 2d sont ici intéressante à regarder étant donné que les distributions suivent les même lois mais avec des priorités différentes, la priorité maximale est pour l'outcome tte sur la figure 2b et

binaire pour la figure 2d. Or, on remarque une différence sur l'écart inter-quantile, lorsque l'outcome binaire est en premier, la variabilité est plus élevée (voir table **??**).
L'outcome de poisson semble aussi montrer une forte variabilité des valeurs du WR au vu de son écart inter-quantile qui monte à 1.937.

TABLE 11 – Nombre moyen de win/loose/tie par outcome suivant le second scénario où les distributions sont inégales

outcome	win	loose	tie
outcome1	1253	1256	7491
outcome2	3399	791	3301
outcome3	2073	1227	1
overall	6726	3274	1

(a) $\tau = 0$

outcome	win	loose	tie
outcome1	1064	939	7996
outcome2	3629	844	3523
outcome3	1096	442	1985
overall	5789	2226	1985

(b) $\tau = 2$

	win	loose	tie
outcome1	1095	919	7985
outcome2	3623	843	5534
outcome3	1347	278	2549
overall	6066	1790	2549

(c) Outcome discret de Poisson

outcome	win	loose	tie
outcome1	4538	414	4406
outcome2	470	414	3523
outcome3	1096	442	1985
overall	6104	1269	1985

(d) Outcome principal binaire

Les taux de censures sont égaux sur les simulations, nous voyons maintenant une différence entre les groupes, ce qui est dû au paramètre β valant -2 et donc favorisant le groupe traité. Ce résultat est logique au vu des données, la censure étant la même pour les deux groupes, comme nous favorisons le groupe traité, les valeurs seront plus haute et la censure s'opère donc plus pour le groupe T.

Comme nous le voyons sur le tableau 3, lorsqu'une variable est censurée, il y a plus de chance qu'elle soit noté comme "non-informative" et donc considéré comme "neutre" par la fonction win.stat ce qui explique le fort nombre d'égalité pour les outcome tte sur les tableaux 11.

TABLE 12 – Tableau récapitulatif du taux de p-valeur inférieur à 0.05

scénario	NBT	WR	WO
$T \sim C$; $\tau = 0$	1	1	1
$T \sim C$; $\tau = 2$	1	1	1
Otc discret de Poisson	1	1	1
Otc principal binaire	1	1	1

Lors de ce scénario, logiquement, nous rejetons l'hypothèse H_0 lors de 100% de nos simulations. Les distributions des deux groupes étant bien trop éloignées pour accepter l'hypothèse H_0 à 95%.

Ici, on ne remarque pas directement de différence entre le cas où $\tau = 2$ avec l'outcome tte en premier et le cas où l'outcome binaire est premier, bien que l'écart inter-quantile suggère l'inverse. On voit pourtant sur les tableaux 11 montrant le nombre moyen de win/loose/tie que ces nombres sont complètement différents ce qui s'explique par la construction des statistiques. En effet, comme vu sur le tableau 10 il y a beaucoup de censure (environ 55%) sur l'outcome tte ce qui provoque un grand nombre d'égalité et qui permet donc à la fonction win.stat de passer à l'outcome suivant (binaire dans le cas $\tau = 2$ et continue dans le cas où l'outcome binaire est premier). Or l'outcome binaire lui génère beaucoup moins d'égalité au vu des distributions marquée (voir table 9) et de la façon dont sont généré les égalité (voir tableau 1). Pourtant, les descriptions statistiques des distributions ne montrent pas de différences entre les deux.

Lorsque nous nous penchons sur la différence entre le WR et le WO, on voit une nouvelle fois une plus grosse variabilité des valeurs et des valeurs plus élevée pour le WR, dont, par exemple dans le cas où le seuil vaut 2, la médiane du WR est situé entre le 3ème quartile et le maximum pour le WO.

Lorsque l'on utilise un outcome discret, on remarque que le WR présente une variabilité plus importante que le WO. Cela se confirme notamment avec les distributions choisies pour cet outcome : l'écart entre les 1er et 3e quartiles, ainsi qu'entre les valeurs minimale et maximale, est plus marqué, ce qui suggère une dispersion plus forte des valeurs. En revanche, il n'est pas pertinent de comparer directement les situations où le troisième outcome est continu avec celles où il est discret, car la nature même des données diffère. Malgré cela, les p-valeurs obtenues sont similaires dans les deux cas, et elles mènent à la même conclusion concernant la comparaison entre le WR et le WO.

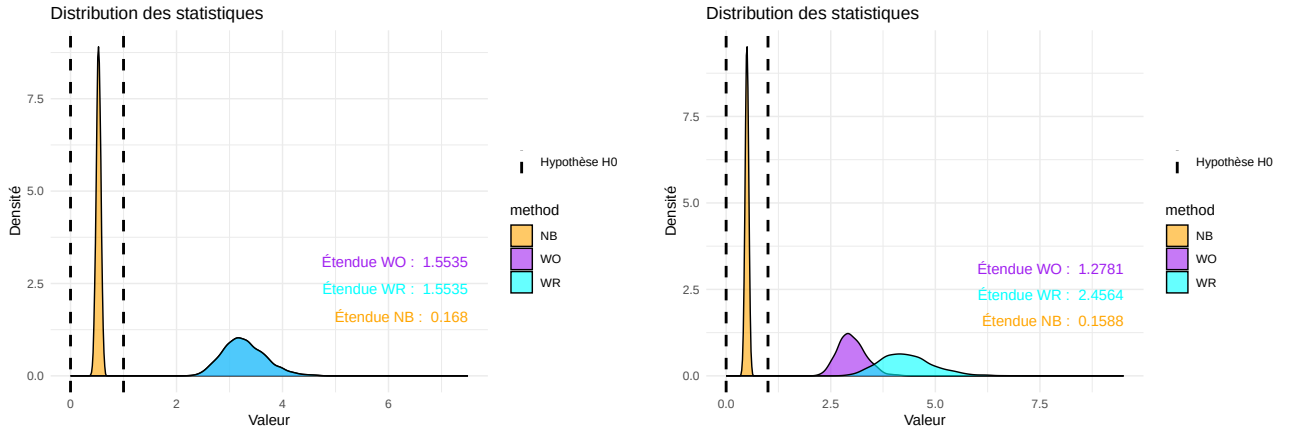
3.1.3 Grandes différences entre les groupes

Nous avons maintenant considéré un scénario où la différence entre les distributions du groupe T et du groupe C sont très différente en faveur de T. Le but ici est de confirmer la forte variabilité et les valeurs plus grandes du WR en considérant 2 cas de figure dont les paramètres des distributions seront regroupés dans le tableau 13. Dans le premier cas de figure, nous fixerons le seuil à 0 et l'ordre des outcomes sera tte/binaire/continue alors que dans le second cas, le seuil sera fixé à 2 et l'ordre des outcomes ne sera pas changé.

Scénario	Fig.	outcome 1	outcome 2	outcome 3
$\tau = 0$	3a	$\lambda = 0.1,\ k = 5,\ \beta = -4$	$\mathcal{B}_T(0.7)\,;\,\mathcal{B}_C(0.3)$	$\mathcal{N}_T(3,2)\,;\,\mathcal{N}_C(3,2)$
$\tau = 2$	3b	$\lambda = 0.1,\ k = 5,\ \beta = -4$	$\mathcal{B}_T(0.7)\,;\,\mathcal{B}_C(0.3)$	$\mathcal{N}_T(3,2)\,;\,\mathcal{N}_C(3,2)$

TABLE 13 – Tableau récapitulatif des distributions du quatrième scénario

La loi de la censure sera une Weibull de paramètre (2,25). Ici, nous avons fixé un $\beta = -4$ signifiant qu'il y a une grosse différence entre le groupe T et le groupe C en faveur de T. Ce β va influer sur les temps de survie qui seront plus grand dans le groupe T par rapport au groupe C. C'est pour cela qu'ils seront plus censurés. Le pourcentage de valeur censuré dans le groupe T est d'environ 41.6% alors que dans le groupe C il est d'environ 18.5%.



(a) $\tau = 0$ (b) $\tau = 2$

FIGURE 3 – Simulations suivant le quatrième scénario où les distributions sont fortement inégale en faveur de T

TABLE 14 – Nombre moyen de win/loose/tie par outcome suivant le quatrième scénario où les distributions sont fortement inégales en faveur de T

outcome	win	loose	tie
outcome1	895	897	8208
outcome2	4013	742	3453
outcome3	2735	718	1
overall	7643	2357	1

(a) $\tau = 0$

outcome	win	loose	tie
outcome1	768	689	8543
outcome2	4177	772	3594
outcome3	1557	70	1967
overall	6502	1531	1967

(b) $\tau = 2$

Au vu des tableaux 14, le scénario est bien respecté, dans le tableau 14a on a en moyenne 7643 wins, 2357 looses et 1 tie et lorsque l'on augmente le seuil pour avoir plus d'ex-aequo dans le tableau 14b le nombre d'ex-aequo augmente en moyenne à 1967, en diminuant fortement le nombre de looses et de wwins. Lorsque le seuil vaut 0, on ne voit pas de réelle différence entre le groupe T et le groupe C selon la variable tte mais cette différence se remarque plus lorsque le seuil vaut 2.

L'influence de cette différence se reflète parfaitement une fois de plus sur les figures 3 où l'on voit le WR s'étendre à 95% sur 2.4564 et, alors que le WO et le WR prennent des valeurs entre 1.33 et 3.022 lorsque le seuil est nul (voir tableaux 50), en présence d'égalité le WR prend ses valeurs entre 2.419 et 6.687 et le WO entre 1.871 et 3.869.

Nous avons aussi logiquement 100% de p-valeur supérieur à 0.05 pour les 2 cas de figure de ce scénario ce qui est attendu.

TABLE 15 – Tableau récapitulatif du taux de p-valeur inférieur à 0.05

scénario	NBT	WR	WO
$\tau = 0$	1	1	1
$\tau = 2$	1	1	1

Une nouvelle fois le WO semble plus utilisable que le WR en raison de sa variabilité et du fait que les valeurs sont plus faible.

3.1.4 Critères ne privilégiant pas tous le groupe T

Enfin, nous allons considérer un dernier scénario, cette fois-ci nous allons choisir d'avoir des outcomes en faveur de T et de C alors que jusque là les 3 outcomes étaient uniformément en faveur de T. Nous allons donc pour se faire, choisir 1 ou 2 outcome en faveur de C et faire varier la priorité de ces outcomes afin de balayer plusieurs cas de figure et voir l'influence de l'outcome principal.

On commence par un scénario où seule la variable continue (troisième outcome) est en faveur du groupe contrôle, tandis que les autres outcomes restent en faveur du groupe traité.
Ensuite, on considère un cas où la variable `tte` est elle aussi en faveur du groupe contrôle.
Enfin, on teste une configuration avec la variable continue (en faveur de C) en première priorité, une variable binaire (en faveur de T) en deuxième priorité, et la variable `tte` (en faveur de C) en troisième priorité, simulée avec un modèle de Cox.
Un résumé des distributions avec leur ordre plus précis est disponible sur le tableau 16.
On fixe le seuil des variables continue et tte à 1 (au lieu de 2) afin de réduire les cas d'égalité.

Scénario	Fig.	outcome 1	outcome 2	outcome 3
Cox - 2T1C	4	$\lambda = 0.1,\ k = 4,\ \beta = -3$	$\mathcal{B}_T(0.65)\,;\,\mathcal{B}_C(0.3)$	$\mathcal{N}_T(2,1)\,;\,\mathcal{N}_C(4,2)$
Cox - 1T2C - binaire	5	$\mathcal{B}_T(0.65)\,;\,\mathcal{B}_C(0.3)$	$\lambda = 0.1,\ k = 4,\ \beta = 3$	$\mathcal{N}_T(2,1)\,;\,\mathcal{N}_C(4,2)$
Cox - 1T2C	6	$\mathcal{N}_T(2,1)\,;\,\mathcal{N}_C(4,2)$	$\mathcal{B}_T(0.65)\,;\,\mathcal{B}_C(0.3)$	$\lambda = 0.1,\ k = 2,\ \beta = 2$

TABLE 16 – Tableau récapitulatif des distributions du cinquième scénario

Dans la colonne scénario du tableau 16, `2T1C` signifie qu'il y a 2 outcome en faveur de T et 1 en faveur de C et `1T2C` signifie qu'il y a 1 outcome en faveur de T et 2 en faveur de C. Le scénario Cox - 1T2C - binaire signifie que nous considérons des données tte simulé suivant un modèle de Cox où l'outcome principal est binaire et où 2 outcome favorise C et 1 favorise T.

Scénario	Loi de la censure	Censure T	Censure C
Cox - 2T1C	$\mathcal{W}(3,10)$	66.3%	43.2%
Cox - 1T2C - binaire	$\mathcal{W}(3,10)$	18.9%	43.2%
Cox - 1T2C	$\mathcal{W}(3,10)$	12.7%	43.2%

TABLE 17 – Censure des tte dans le cas d'un traitement à effets partiels

Les scénarios où l'outcome tte est en faveur du groupe contrôle se reflètent sur la censure, en effet lorsque le pourcentage de censure est plus élevé dans le groupe T cela signifie que l'outcome tte est en faveur du groupe T et lorsque le pourcentage de censure est moins élevé dans le groupe T cela signifie que le groupe C a été privilégié.

On commence par étudier un premier cas où la variable continue est en faveur du groupe C, mais elle n'occupe que la troisième priorité. L'outcome binaire, en deuxième priorité, est en faveur du groupe T, tandis que la variable `tte`, en première priorité, est également en faveur du groupe T. Cette variable `tte` est simulée à l'aide d'un modèle de Cox.
Ensuite, on considère un autre scénario où la variable `tte` est toujours générée avec un modèle de Cox. Cette fois-ci, l'outcome binaire devient la priorité principale et reste en faveur du groupe T. Les deux autres outcomes, `tte` et la variable continue, sont quant à eux en faveur du groupe C.
Enfin, on examine un dernier cas où l'outcome principal est une variable continue en faveur du groupe C. L'outcome binaire est une fois de plus en faveur du groupe T, tandis que la variable `tte`, en dernière priorité, est simulée avec un modèle de Cox, toujours en faveur du groupe C.

On remarque directement l'influence d'avoir l'outcome continue en faveur de C que ce soit sur la première figure ou sur la seconde. Sur la figure 4 on aura beaucoup de valeur inférieure à l'hypothèse nulle et cela se confirmera sur le tableau 51 où environ 1/4 des valeurs se situeront autour de l'hypothèse nulle. Le tableau 18 nous montre le nombre moyen de win/loose/tie selopn ce cas de figure et ce que l'on peut observer c'est qu'une nouvelle fois l'outcome tte ne différencie pas énormément le groupe T et le groupe C, alors que les 2 autres font bien la différence chacun selon leur préférence prédéfinies. Au vu du graphique et de ce tableau on peut voir l'importance de classer les outcome par ordre de priorité car même s'il y a une grosse différence entre le nombre de win et de loose selon le 3ème outcome, la distribution des 3 statistiques se situe quand même à droite de l'hypothèse nulle. On remarque aussi que l'écart entre le WR et le WO n'est pas significatif malgré la présence d'ex-aequo comme nous le suggère le tableau 51.

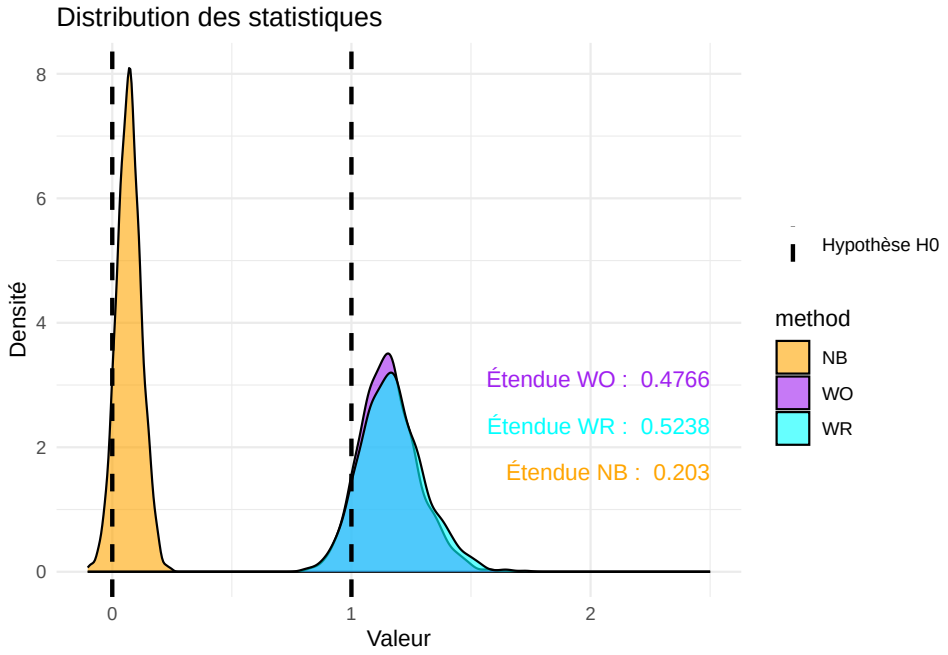


FIGURE 4 – Distribution continue en faveur de C

On remarque sur les figure 4 que les distributions des deux statistiques sont assez proche, nous n'avons plus comme précédemment des distributions bien distincte, cela est expliqué par le nombre de win/Looose/tie que l'on peut retrouver sur le tableau **??** au dessus. Nous observons que lorsqu'au global nous avons une différence moins marquée entre le nombre de win et de loose, le WO et le WR sont moins éloignés.

TABLE 18 – Comparaison des résultats par outcome lorsque les effets sont contrastés

outcome	win	loose	tie
endpoint1	1559	1476	6965
endpoint2	3161	735	3069
endpoint3	252	2063	754
overall	4972	4274	754

Pour nos 2 cas de figure étudiés précédemment, le troisième critère de jugement contrebalancé l'effet des 2 premier sur le nombre total de looses, le rendant plus proche du nombre de win. On obtient donc bien des statistiques préférant le groupe T au groupe C mais se rapprochant de l'hypothèse nulle.
Le WR semble donc être plus utilisable en pratique lorsque le nombre de win et le nombre de loose est proche bien qu'il rende des valeurs légèrement plus élevées comme nous pouvons le voir sur le résumé statistique de la figure 4 sur le tableau 51.
Nous allons aussi regarder le taux de p-valeurs ne rejetant pas l'hypothèse nulle. Il est important de noter que les tests à hypothèses sont bilatéraux et donc l'hypothèse alternative indique que la statistique rejette l'hypothèse nulle. Un test unilatéral aurait pu être effectué et dans ce cas la, l'hypothèse alternative aurait supposé qu'un groupe ou l'autre est meilleur que l'autre.

TABLE 19 – Tableau récapitulatif du taux de p-valeur inférieur à 0.05

scénario	NBT	WR	WO
$\tau = 0$	0.248	0.252	0.252

Lors du premier cas de figure, 24.8% des p-valeurs sont inférieur à 0.05 pour le NB contre 25.2% pour

le WO et le WR.

Nous allons maintenant regarder ce scénario où le premier outcome est binaire en faveur du groupe T, le second outcome sera en faveur de C est sera tte (simulé selon un modèle de Cox), enfin le troisième outcome sera continue et en faveur de C. Les paramètres des distributions utilisées se situe sur le tableau 16.

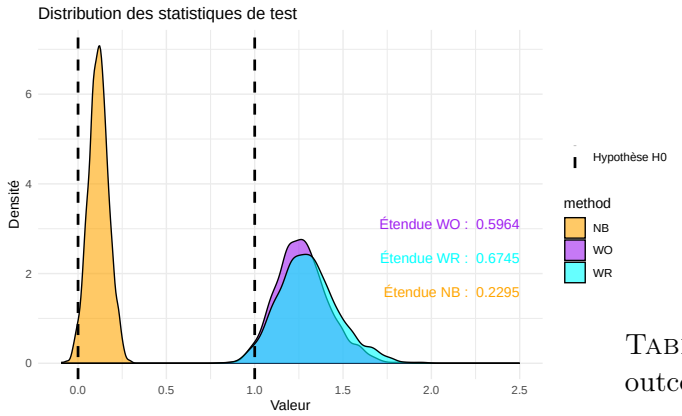


FIGURE 5 – Simulations suivant le cinquième scénario où un seul critère favorise le groupe traité

outcome	win	loose	tie
endpoint1	4538	1056	4406
endpoint2	237	264	3905
endpoint3	322	2623	960
overall	5097	3943	960

TABLE 20 – Nombre moyen de win/loose/tie par outcome – scénario avec effets opposés

Nous remarquons ici la même situation que précédemment sur la figure 4 en raison du scénario n'étant pas si différent, en effet, ici bien que le critère de jugement tte soit en faveur du groupe C comme le montre le paramètre $\beta > 0$, l'affectation des quantité win/loose/tie et dans ce cas "non-informatif" que vous pouvez retrouver sur le tableau 3, est beaucoup trop contraignant pour avoir une réelle différence.

Remarque : j'ai essayé d'augmenter fortement le paramètre d'échelle de la censure pour avoir peu d'individus censurer et se retrouver dans la première partie du tableau 3 en ayant environ 30% de censure dans le groupe C et 5% dans le groupe T mais un résultat identique a été observé bien que le nombre de T soit moins élevé, aucun réelle différence n'a été observé entre le nombre de win et de loose.
Nous allons alors vérifier ici l'influence de l'endpoint principal sur la statistique, car en effet il y a une plus grosse proportion de loose par rapport au win sur le troisième outcome que de win par rapport au loose dans le premier outcome. Pourtant la statistique est bien en faveur de T car ce qui nous importe dans notre cas c'est le global où le nombre de win est supérieur au nombre de loose.

Le WO et le WR prennent tout de même des valeurs comparable bien que la densité au niveau de la moyenne soit plus élevé pour le WO que pour le WR dû à la variabilité qu'apporte la non prise en compte du nombre d'égalité. L'écart inter-quantile à 95% nous montre une différence entre les deux assez faible (environ 0.08). Ce qui est confirmé par le tableau 52 où l'on peut voir le caractère légèrement plus variable du WR lorsque les outcomes se compensent entre eux, le minimum du WR étant plus faible et le maximum plus élevé que le WO.

TABLE 21 – Tableau récapitulatif du taux de p-valeur inférieur à 0.05

scénario	NBT	WR	WO
$\tau = 0$	0.49	0.495	0.495

Ici, 49% des p-valeurs sont inférieur à 0.05 sur le NB, 49.5% ne rejette pas l'hypothèse nulle à 95% pour le WO et le WR comme nous pouvons le voir dans le tableau 21.

Enfin nous allons passer à notre dernière partie de simulation sur ce notebook où seul le second critère de jugement sera en faveur du groupe T, l'outcome binaire ayant la seconde priorité. Le premier

outcome sera continue et favorisera C alors que le troisième outcome sera en faveur de C et sera de type tte d'abord simulé selon un modèle de Cox puis selon un modèle AFT. Le récapitulatif des paramètres utilisés se situe dans le tableau 16.

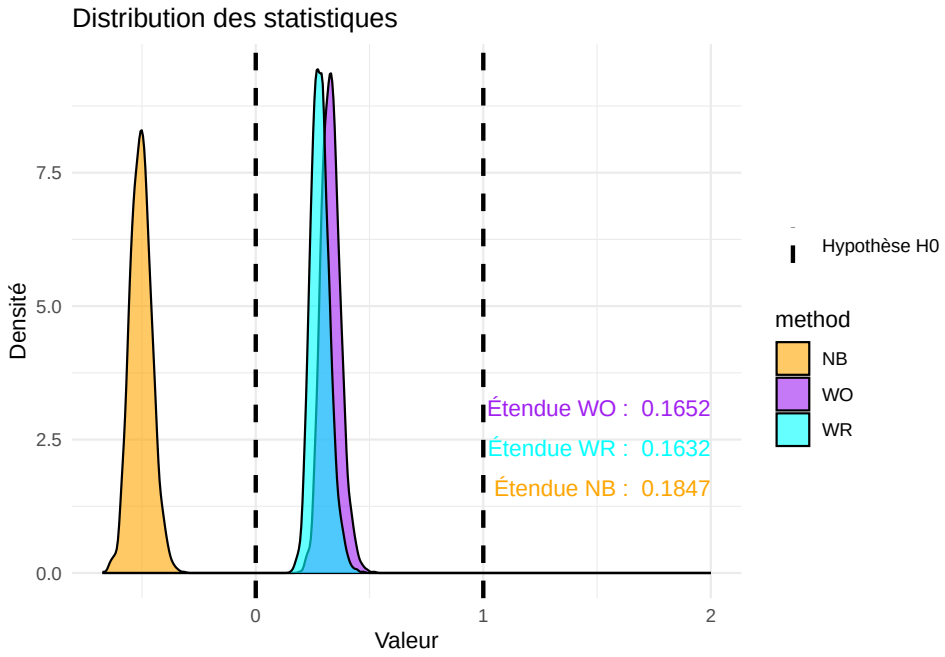


FIGURE 6 – Distribution des WR et WO pour un mélange d'effets positifs et négatifs du traitement

Nous voyons ici sur la figure 6 que le groupe C est bien privilégié par rapport au groupe T car les valeurs sont inférieur à celle de l'hypothèse nulle. On remarque aussi une différence plus notable entre le WO et le WR comme lorsque les distributions étaient bien en faveur de T. Le WR est plus éloigné de l'hypothèse nulle bien qu'autant étendue que le WO. La densité à la moyenne est égale mais la moyenne est plus faible que celle du WO ce qui est une fois de plus causé par la non-prise en compte du nombre de tie. Cette conclusion est valable pour les 2 cas de figure présentés ici.

TABLE 22 – Comparaison des résultats par outcome lorsque les effets sont contrastés

outcome	win	loose	tie
endpoint1	825	6717	2458
endpoint2	1114	260	1083
endpoint3	41	48	994
overall	1980	7006	994

La description du nombre moyen de win/loose/tie sur le tableau 22 suggère bien cette conclusion selon laquelle la différence entre le groupe C et le groupe T est importante en faveur du contrôle. Le WO au global est plus élevé que le WR, il semble donc plus nuancer.

La différence entre les statistiques est une nouvelle fois visible lors de ce scénario, c'est à dire lorsque le groupe C est privilégié. La différence peut sembler faible mais l'échelle est différente, le WO et le WR ne pouvant pas, en pratique[4], atteindre 0. Nous pouvons voir par exemple que 1/4 des valeurs du WR sont supérieur aux 3/4 des valeurs du WO environ.

4. En théorie c'est possible si le nombre de W valait 0 et qu'il n'y a pas de T c'est à dire qu'il y ait que des L.

TABLE 23 – Tableau récapitulatif du taux de p-valeur inférieur à 0.05

scénario	NBT	WR	WO
$\tau = 0$	1	1	1

Lorsque l'on se penche sur les tests d'hypothèse, 100% des p-valeurs rejettent l'hypothèse nulle, il y a bien une différence entre le traitement et le contrôle bien que cette fois-ci, cette différence est en faveur du contrôle.

3.1.5 Conclusion du premier notebook

Ce que nous pouvons retenir de ce premier notebook c'est que le WR semble, en pratique peu utilisable lorsque les groupes T et C sont trop distinct, la non-prise en compte des égalité rend sa statistique trop variable et pouvant donner des valeurs bien plus élevé que le WO se basant sur les même données. L'interprétation faite par le clinicien peut donc fortement différé quant à l'effet du traitement tout en restant sur la même conclusion (groupe T meilleur que groupe C ou inversement).
Cependant lorsque la différence entre les deux groupes est moins prononcée le WR peut être considéré tout en gardant en tête la variabilité de ce dernier.

Le choix du seuil influe aussi sur l'estimation des statistiques car il augmentera le nombre d'égalité et créera une différence entre les statistiques, ce choix est donc primordial.
L'outcome principal influe directement sur la décision prise par la statistique sur la préférence entre les groupes, c'est pourquoi il faut bien ordonner ces critères du plus grave au moins grave.

Toutefois si nous voulons nous baser uniquement sur la p-valeur, c'est à dire la probabilité de rejeter l'hypothèse nulle, nous avons vu très peu de différence entre le taux de p-valeur inférieur à 0.05 selon les 3 statistiques. Nous pouvons donc nous baser dessus quelque soit la statistique.

Je considère donc que l'utilisation du WO ou du NB est préférable quelque soit le cas car nuance plus la valeur de la statistique et que le choix de l'un ou l'autre est arbitraire et dépend de l'interprétation que le clinicien ou le médecin veut faire.

3.2 Second notebook

Nous allons maintenant nous concentrer sur les données time-to-event, nous aurons alors seulement deux outcome et tous les deux seront tte. L'idée ici est de comprendre la partie de gauche du graphique 7 de Verbeeck et al. (2023). Ici, nous cherchons à comprendre la raison pour laquelle le WR est préféré au WO (appelé dans l'article *Success Odds*) lorsqu'un modèle de Cox est utilisé pour simulé les données et qu'aucun seuil n'est appliqué à la comparaison entre les groupes. Nous voulons aussi savoir pourquoi lorsque les HR ne sont pas constant l'article préconise d'utiliser le WO.
La justification avancée par l'article à propos de l'utilisation du WR est qu'en présence d'un hazard ratio (HR) constant, le WR fournit une estimation équivalente (approximativement l'inverse du HR), tout en offrant une interprétation plus intuitive et cliniquement pertinente, notamment lorsqu'il s'agit de hiérarchiser plusieurs événements cliniques en fonction de leur gravité.

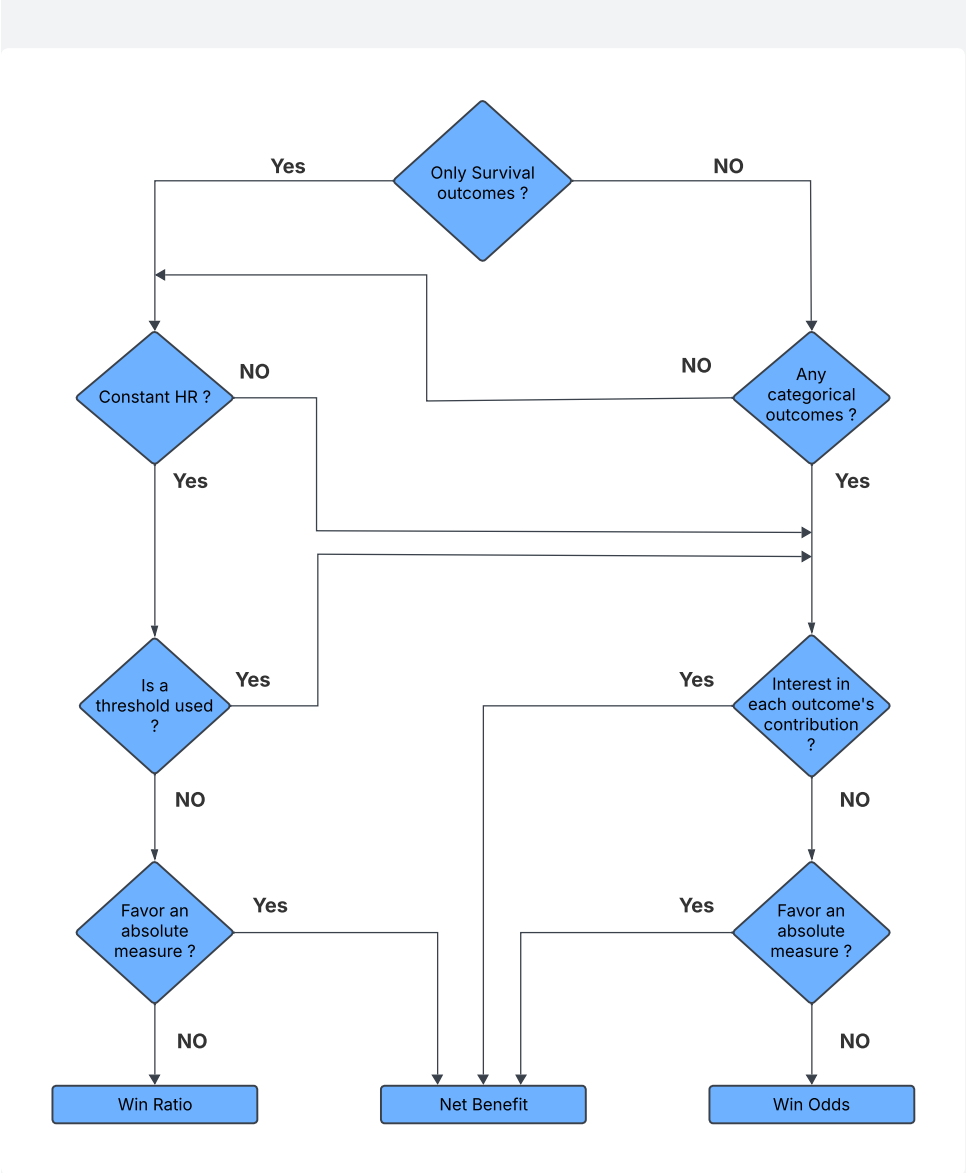


FIGURE 7 – Arbre de décision pour choisir une mesure selon Verbeeck et al. (2023)

Le losange *"Favor an absolute measure"* est choisi par le clinicien et est l'interprétation qu'il veut porter à la statistique. Et *"Interest in each outcome's contribution"* signifie que l'on veut savoir oui ou non quel outcome privilégie quel groupe, c'est pourquoi le NB est choisi car, comme vu dans la partie 2.2.1, le NB au global peut être vu comme la somme des NB dans chaque outcome.

Nous allons aussi mettre en œuvre une autre particularité des données tte, nous allons calculer les statistiques à différents temps pour évaluer l'effet des hasard constant ou pas sur les statistiques. Nous allons regarder nos observations en considérant 3 censures fixée au cours du temps et une censure aléatoire
Étant donné que nous générons nos données tte en mettant une censure fixe selon le temps d'étude, 4 cas peuvent être observé et seront présenté dans la figure 8.

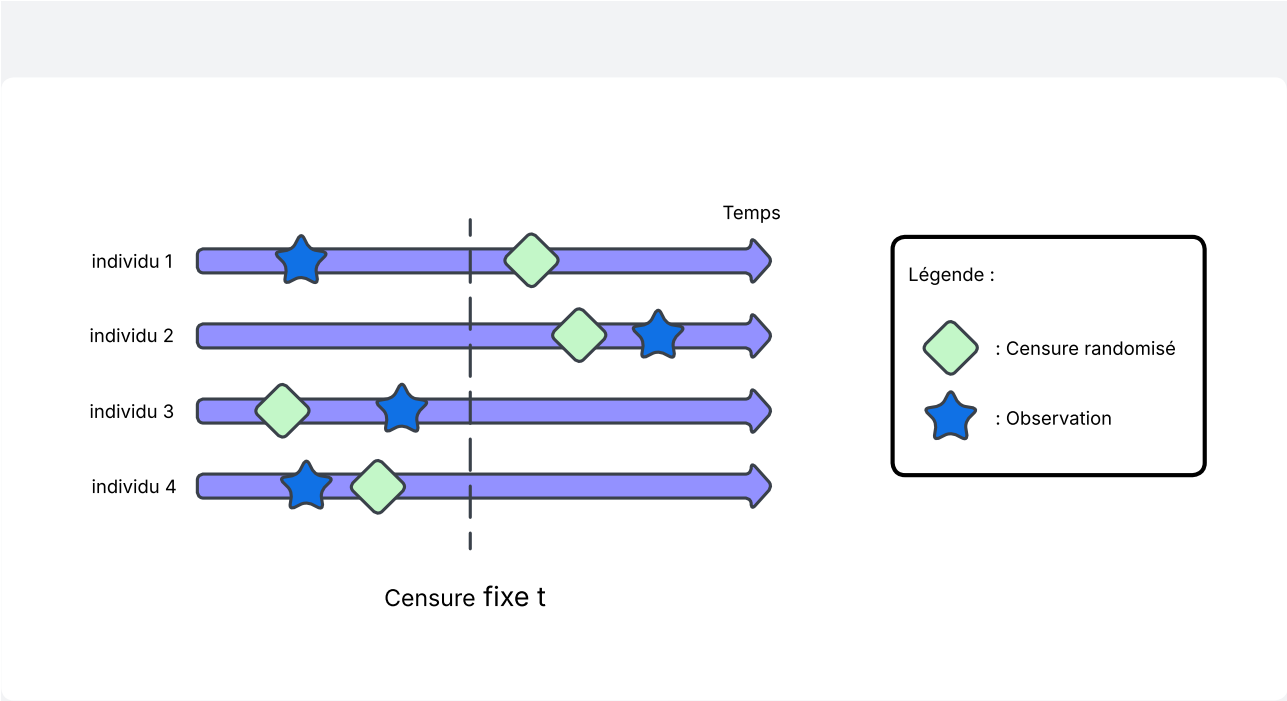


FIGURE 8 – Cas possible lorsque l'on a une censure fixe et une censure aléatoire

La censure aléatoire représente notre évènement et est simulé selon une loi de Weibull une nouvelle fois. La censure fixe elle, est simplement une censure que l'on a décidé de fixer et représenterait un suivi au cours de l'étude à t unité de temps (jour, mois ou année). Les différents cas sont montré dans la figure ce-dessus et sont expliqué ci-dessous.

1. L'individu 1 a eu son observation avec les 2 censures, il sera donc observé et ne sera pas censuré.
2. L'individu 2 a eu son observation après la censure fixe t, l'observation sera censuré et vaudra t.
3. L'individu 3 a eu sa censure aléatoire en premier, l'observation sera donc censuré est aura la valeur de sa censure.
4. L'individu 4 a eu son observation avant sa censure aléatoire et avant la censure fixe, l'observation ne sera donc pas censuré.

En terme mathématiques, voici ce que cela donne :
Soit Y notre observation, $Y = min(A, B)$, où A représente les temps que nous générons selon nos modèles, soit de Cox, soit AFT et B la censure aléatoire générée suivant une loi exponentielle.
On a alors $\Delta = \mathbb{I}(A \leq B)$, $\Delta = 1$ si l'observation est observée, c'est à dire que le temps simulé est plus élevé que le temps de censure généré.
On va alors posé $\tilde{Y}$ notre observation finale, c'est à dire en prenant en compte notre censure fixe t et sera donc $\tilde{Y} = min(Y, t)$ et la censure sera $\tilde{\Delta} = \Delta$ si $\tilde{Y} = Y$ mais $\tilde{\Delta} = 0$ si $\tilde{Y} = t$.

En résumé, notre observation sera censuré pour deux raisons, soit parce que l'observation se situe après la censure (individu 3) soit parce que l'observation se situe après la censure fixe (individu 2) alors que notre valeur sera non-censuré si la valeur est observé avant le temps de censure randomisé et avant la

censure fixe (individus 1 et 4).

Nous l'avons alors modélisé en prenant 2 critères de jugement de type *tte* où les 2 distributions sont d'abord simulé selon un modèle de Cox puis selon un modèle AFT. Nous ne considérerons pas de seuil afin de rester dans la branche de gauche du graphique de Verbeeck et al. (2023) et nous prendrons 3 temps de censures qui représenteront les temps de mesures effectuées tous les 5 mois en commençant 9 mois après le début de l'étude (mesure à 9 mois, 14 mois et 19 mois). Les deux distributions tte seront en faveur du groupe T en prenant un β négatif pour le modèle de Cox et positif pour le modèle AFT, augmentant l'effet de la covariable Z, "le patient a eu le traitement : oui/non".

Dans le cas du modèle de Cox, les temps de survie simulés suivent une loi de Weibull. En revanche, pour le modèle AFT, ils suivent une loi log-normale. Il n'est pas approprié d'utiliser une loi de Weibull pour le modèle AFT, car celle-ci constitue un cas particulier du modèle AFT où le hazard ratio est constant dans le temps.

Les distributions utilisées dans cette partie sont situées dans le tableau suivant :

Scénario	Outcome 1	Outcome 2
Cox	$\lambda = 0.1$, $k = 2$, $\beta = -3$	$\lambda = 0.12$, $k = 1.7$, $\beta = -3$
AFT	$\mu = 2$, $\sigma = 1.2$, $\beta = 1.5$	$\mu = 2.8$, $\sigma = 0.8$, $\beta = 1.5$

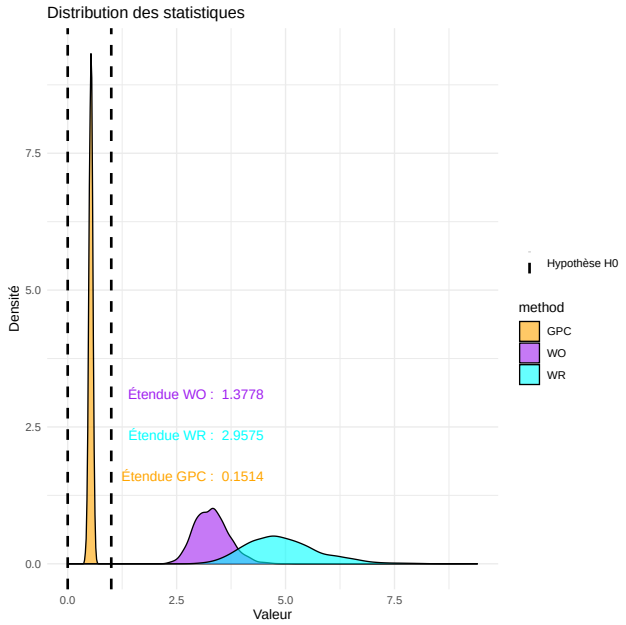
Table 24 – Distributions des outcomes du second notebook

Nous allons voir dans un premier temps le comportement des méthodes lorsque les données sont simulées selon le modèle Cox, puis nous verrons dans un second temps le comportement des méthodes lorsque les données sont simulées selon le modèle AFT. Enfin, nous verrons des graphiques que propose le package mais qui sont basé uniquement sur une seule simulation.

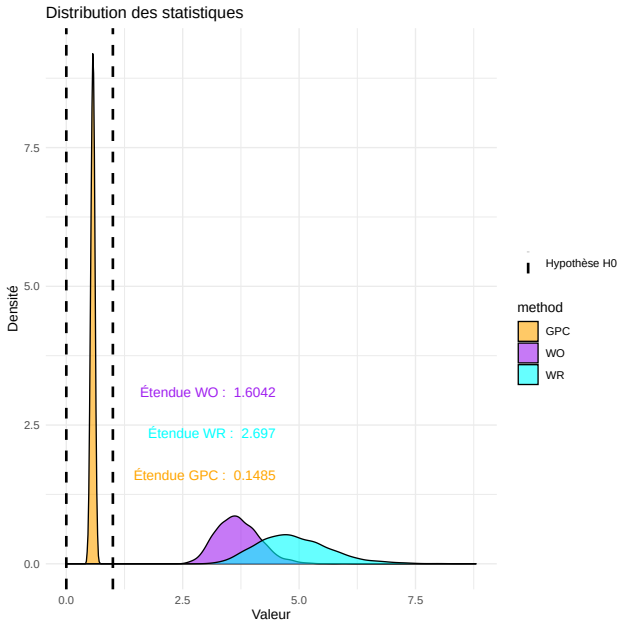
Cette fois, la censure aléatoire sera modélisée selon une loi exponentielle $\mathcal{E}(0.03)$ pour le scénario basé sur le modèle de Cox, et selon une loi $\mathcal{E}(0.08)$ pour celui correspondant au modèle AFT.

3.2.1 Données simulées selon le modèle de Cox

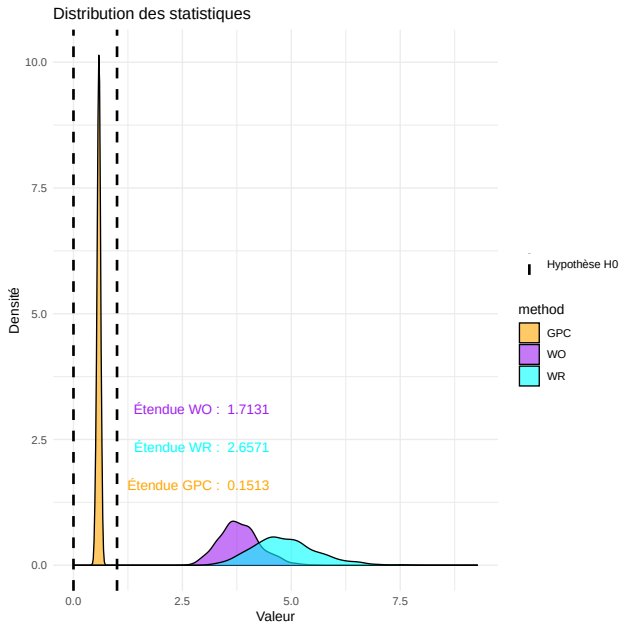
On remarque globalement, sans prendre en compte les temps de censure fixe que le WR est très variable, on le voit graphiquement en regardant la distribution en bleu clair en comparaison avec la distribution du WR en violet qui est lui moins variable et qui a une densité à la moyenne plus élevée. On le voit aussi en regardant l'écart inter-quantile valant, pour le WR, 2.9575 lorsque la censure fixe est au temps $t = 9$, 2.697 lorsque $t = 14$ et 2.6571 lorsque $t = 19$. Le WO est aussi un peu variable mais beaucoup moins quelque soit le temps de censure fixe comme le témoigne l'écart innter-quantile sur les figures 9. On voit aussi que les valeurs prises par le WO sont plus faible que celle du WR et l'écart est d'autant plus flagrant au temps $t = 9$ ce qui peut être expliqué par un taux de censure plus élevé lorsque l'on censure plus tôt. Nous pouvons voir dans les tableaux 26 les taux de censure par outcome et logiquement, on observe plus de censure au temps $t = 9$ et donc plus d'égalité et de variation pour le WR.



(a) $t = 9$



(b) $t = 14$



(c) $t = 19$

FIGURE 9 – Distribution des statistiques lorsque la censure est fixée aux temps indiqués

TABLE 25 – Comparaison des résultats par outcome lorsque la censure est fixé au temps t indiqué

	win	loose	tie
outcome1	4414	993	4593
outcome2	2294	396	1902
overall	6708	1390	1902

(a) $t = 9$

	win	loose	tie
outcome1	5060	1131	3809
outcome2	2138	369	1302
overall	7198	1500	1302

(b) $t = 14$

	win	loose	tie
outcome1	5308	1190	3502
outcome2	2138	369	1103
overall	7355	1542	1103

(c) $t = 19$

On remarque une augmentation croissante du nombre de win et de loose ainsi qu'une diminution du nombre de tie à mesure que la censure fixe est repoussée, cette différence intervient car plus la censure est basse, plus il y aura de tie. En repoussant la censure entre les temps $t = 9$ et $t = 14$ puis entre $t = 14$ et $t = 19$, on permet la comparaison de davantage de paire censurées, ce qui augmente le nombre de win et de loose.

Sur les tableaux 26, on distingue bien ce phénomène de censure repoussée car le taux d'observations censurées diminue dans les 2 groupes à mesure que la censure fixe est repoussée. On ne voit cependant pas beaucoup de différence entre le temps $t = 14$ et $t = 19$ ce qui est dû à la façon dont sont généré la censure aléatoire, c'est elle qui prendra le pas sur la censure fixe entre ces 2 temps (voir figure 8).

TABLE 26 – Comparaison des taux de censure pour T et C par outcome aux temps de censure fixe indiqué

	outcome1	outcome2
T	0.8387	0.8508
C	0.4704	0.4074

(a) $t = 9$

	outcome1	outcome2
T	0.7784	0.7930
C	0.3549	0.2981

(b) $t = 14$

	outcome1	outcome2
T	0.7784	0.7491
C	0.2961	0.2981

(c) $t = 19$

Pour l'instant, on ne voit pas de différence sur la conclusion que l'on peut apporter lorsque l'on prends des mesures régulières au cours du temps symbolisé par ces temps de censure fixe. Ce n'était pourtant pas ce que préconisait l'article comme nous pouvons le voir sur la figure 7. Nous avons pourtant respecté les conditions de l'article en ne prenant que des outcomes tte avec des HR constant et sans prendre de seuil. Tout ceci est logique étant donné que les HR sont constant au cours du temps, la statistique est donc la même. en revanche, nous avons des ex-aequo c'est pourquoi le WR est plus grand que le WO.

J'ai également fait varier le paramètre β afin de vérifier si les résultats précédents étaient uniquement dus à une forte différence entre les groupes. Les conclusions restent cependant identiques à celles de la partie 3.1.4 : lorsque la différence entre les groupes est faible — que ce soit en raison de préférences différentes selon les critères (dans la partie 3.1.4) ou ici à cause de la valeur de β —, le nombre de win et de loose est similaire. Par conséquent, la différence entre les scores WO et WR est moins marquée, bien que le score WR reste légèrement supérieur.
Nous allons alors regarder l'influence des HR non-constant lorsque l'on prends nos mesures à temps régulier dans la prochaine partie.

3.2.2 Données simulées selon le modèle AFT

Nous allons de nouveau considérer 3 temps de censure fixe, une première censure au temps $t = 9$ sur la figure 10a, une seconde censure au temps $t = 14$ sur la figure 10b et une dernière au temps $t = 19$ sur la figure 10c.

TABLE 27 – Comparaison des taux de censure pour T et C par outcome aux temps de censure fixe indiqué

	outcome1	outcome2
T	0.881	0.868
C	0.499	0.282

(a) $t = 9$

	outcome1	outcome2
T	0.882	0.749
C	0.399	0.200

(b) $t = 14$

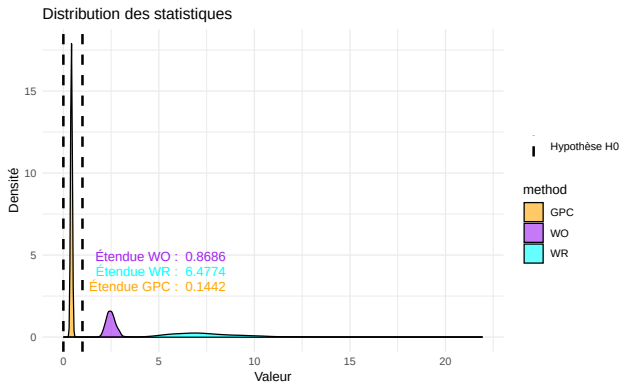
	outcome1	outcome2
T	0.760	0.660
C	0.350	0.174

(c) $t = 19$

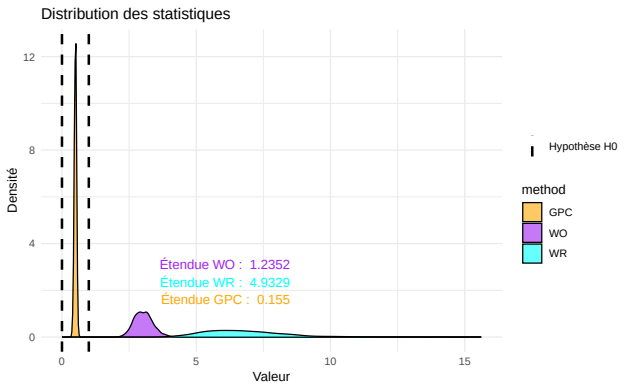
Nous allons tout d'abord regarder, dans les tableaux 27, les taux de censures aux différents temps pour chaque outcome. On observe une décroissance naturelle des taux de censure pour les deux outcomes à chaque temps de censure. On observe une décroissance naturelle des taux de censure en fonction du temps. Pour l'outcome principal,le taux de censure associé à T diminue de 0.881 à $t = 9$ à 0760 à $t = 19$, tandis que celui de C passe de 0.499 à 0.350. Cette tendance est similaire pour le second outcome. Ces taux confirment une amélioration de la complétude de l'information avec le temps, ce qui favorise les comparaisons nettes (win/loose) par rapport aux égalités, en accord avec les tendances observées dans le tableau 28.

On remarque une nouvelle fois la forte variabilité du WR, quel que soit le temps de censure fixe. Par exemple, à $t = 9$, l'écart interquartile du WR atteint 6.477, tandis que celui du WO reste bien plus modeste à 0.869. En parallèle, la distribution du WO évolue de manière plus régulière au fil du temps : à $t = 9$, la moitié des valeurs (entre le 1er et le 3ème quartile) se situe entre 2.321 et 2.643 ; à $t = 14$, entre 2.787 et 3.262 ; et à $t = 19$, entre 3.043 et 3.567. Ces résultats sont en cohérence avec les observations du tableau 28, qui montre une augmentation du nombre moyen de win plus marquée que celle des loose au fil du temps. Ce phénomène s'explique par l'amplification de l'effet différentiel entre les groupes T et C induit par des hazards non proportionnels.

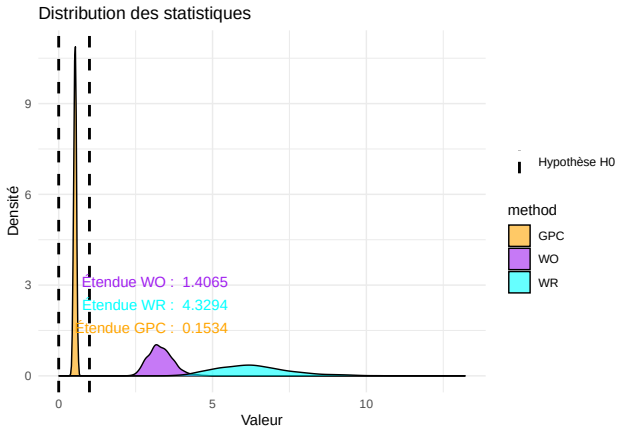
La forte variabilité observée dans le WR, en particulier aux temps précoces (par exemple une valeur maximale de 22 à $t = 9$), s'explique principalement par la manière dont cette statistique est calculée : contrairement au WO ou au NB, le WR ne prend pas en compte les cas de tie, ce qui amplifie mécaniquement sa variabilité en présence de comparaisons peu informatives ou de censure importante. Cette variabilité n'est donc pas nécessairement révélatrice d'une sensibilité particulière à la violation de l'hypothèse des hazards proportionnels.



(a) $t = 9$



(b) $t = 14$



(c) $t = 19$

FIGURE 10 – Distribution des statistiques lorsque la censure est fixée aux temps indiqués

Lorsque les hazards ne sont pas constants, on observe une forte variabilité du WR, notamment à des temps précoces de censure. Par exemple, à $t = 9$, le WR peut atteindre des valeurs extrêmes allant jusqu'à 22, tandis que le WO ne dépasse pas 4 au même temps de suivi. Cette variabilité marquée du WR ne résulte pas nécessairement d'une sensibilité à la non-proportionnalité des risques, mais plutôt de sa méthode de calcul : contrairement au WO, le WR ne prend pas en compte les cas de tie, ce qui tend à accentuer les valeurs extrêmes lorsque les comparaisons sont déséquilibrées ou que la censure reste élevée. Ainsi, la variabilité du WR dans ce contexte reflète davantage les propriétés de la statistique elle-même que les caractéristiques sous-jacentes du modèle de survie.

TABLE 28 – Comparaison des résultats par outcome lorsque la censure est fixé au temps t indiqué

	win	loose	tie
outcome1	4294	649	5058
outcome2	390	23	4644
overall	4684	671	4644

(a) $t = 9$

	win	loose	tie
outcome1	4877	827	5058
outcome2	626	48	3622
overall	5503	875	3622

(b) $t = 14$

	win	loose	tie
outcome1	5095	908	3998
outcome2	772	71	3155
overall	5866	978	3155

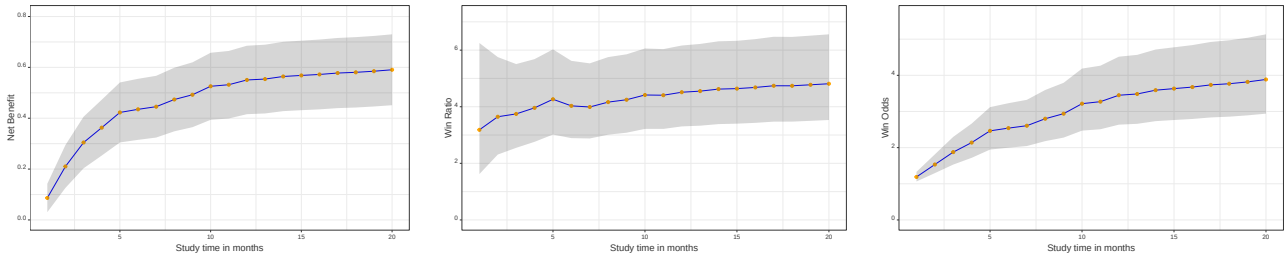
(c) $t = 19$

Lorsque l'on se penche sur les proportions de win, loose et tie selon les outcomes et les temps de cen-

sure, on observe des résultats cohérents qui renforcent les constats tirés des graphiques. Le tableau 28 présente ces proportions aux différents temps. Le nombre de win augmente avec le temps de censure passant de 4684 à $t = 9$, à 5503 au temps $t = 14$ et à 5866 au temps $t = 19$. Parallèlement, le nombre de losses et de ties évolue de manière cohérente, avec une diminution du nombre de tie et une légère augmentation du nombre de loose.

3.2.3 Graphiques sur base d'une simulation

Le package WINS propose aussi des graphiques lorsque tous les outcomes sont de type time-to-event, ces graphiques sont effectués uniquement sur 1 simulations mais permettent de voir l'évolution de la statistique choisie au cours d'un temps donné (ici nous allons du temps $t = 0$ au temps $t = 20$). On compare donc l'évolution des statistiques en prenant les mesures à chaque temps ce qui est intéressant pour regarder l'influence des HR constant.
Plusieurs analyses ont été regardés en changeant la seed, les résultats étant toujours similaire seule une simulation sera donc montrée.



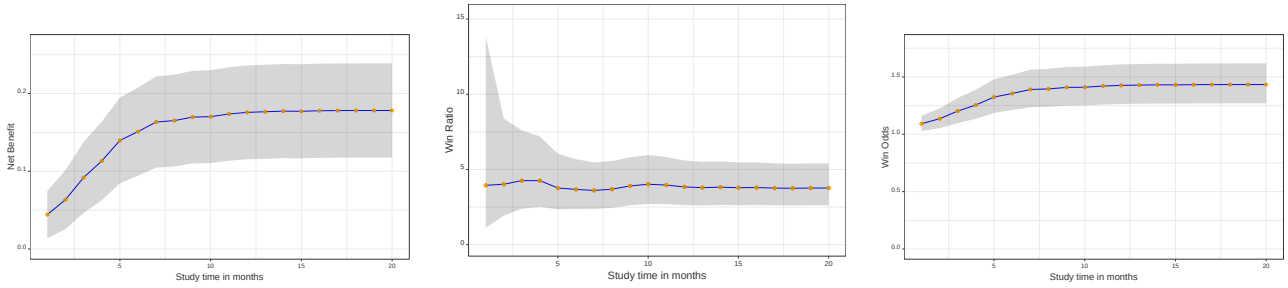
(a) Evolution du Net Benefit (b) Evolution du win Ratio (c) Evolution du win Odds

FIGURE 11 – Données issues d'un modèle de Cox

Les graphiques 11 présentent l'évolution des statistiques au cours du temps lorsque les données de type tte sont simulées selon un modèle de Cox, c'est-à-dire sous l'hypothèse de hazard ratio (HR) constant. On observe une croissance progressive des courbes au fil du temps, en lien avec l'augmentation du nombre d'événements observés due à la diminution de la censure, ce qui permet davantage de comparaisons entre les groupes.

Aucune oscillation notable ni rupture de pente brutale n'est visible. Les trois statistiques étudiées (Net Benefit, win Odds et win Ratio) suivent une tendance croissante similaire. Seule une légère baisse du win Ratio est observable entre les temps $t = 5$ et $t = 6$, sans impact sur la conclusion globale concernant la supériorité du traitement sur le contrôle.

On note également une différence d'échelle entre le WO et le WR, ce dernier prenant des valeurs systématiquement plus élevées tout au long de la période d'observation.



(a) Evolution du Net Benefit (b) Evolution du win Ratio (c) Evolution du win Odds

FIGURE 12 – Données issues d'un modèle AFT

Les graphiques 12 illustrent les mêmes statistiques mais dans un contexte où l'hypothèse de proportionnalité des risques est violée (HR non constant). Malgré cette non-proportionnalité, aucune variation brutale ou anomalie n'est détectée dans l'évolution temporelle des courbes. Le NB semble se stabiliser à partir de $t = 9$, tandis que le WR atteint un plateau autour de $t = 7$. Une interprétation similaire peut être faite pour le WO, puisque cette statistique est une simple transformation monotone du NB.

Ainsi, les résultats suggèrent que ces statistiques sont robustes à la violation de l'hypothèse de HR constant, et donc relativement insensibles à la non-proportionnalité des risques. Cela renforce leur pertinence dans des contextes où cette hypothèse est susceptible d'être remise en cause, comme dans le cas d'effets différés ou de traitements à efficacité variable dans le temps.

3.2.4 Conclusion du second notebook

Sur base de ces graphiques et des simulations précédentes, on peut penser qu'il n'y a pas d'influence des risques proportionnels sur l'évaluation des valeurs des statistiques. Le graphique proposé par Verbeeck et al. (2023) ne semble donc pas bon et la même conclusion que le premier notebook peut être apportée : le WO ou le NB sont plus appropriés, quels que soient les cas de figure que nous pourrions rencontrer avec des données réelles. La différence se fera effectivement sur l'intérêt à utiliser une mesure absolue ou s'il y a un intérêt à examiner la statistique outcome par outcome.

Il est également important de souligner que la conclusion de l'article ne repose pas sur des simulations extensives couvrant divers scénarios, mais uniquement sur un jeu de données particulier, ce qui limite la portée générale de son analyse. Par ailleurs, le fait que le WR soit égal à l'inverse du HR lorsque l'hypothèse de risques proportionnels est respectée ne justifie pas en soi le choix du WR comme mesure de référence, puisque cette équivalence mathématique n'implique pas nécessairement une meilleure pertinence statistique ou clinique dans les cas concrets. Le choix de la statistique doit donc davantage se fonder sur sa robustesse, son interprétabilité et son adaptabilité aux différents contextes de données, plutôt que sur une simple relation formelle avec le HR.

Cette robustesse est cohérente avec la nature même de ces statistiques, qui sont toutes fondées sur des comparaisons deux à deux entre individus et une classification des paires en win, loose ou tie. Elles ne reposent pas sur une modélisation du taux de risque instantané, mais uniquement sur les ordres relatifs des temps d'événement. Par conséquent, seules les différences de temps observées entre individus influencent la mesure, et non la forme précise de la fonction de risque.

Il est toutefois important de noter que dans ces simulations, les distributions de survie ne se croisent pas au cours du temps. Si des scénarios de ce type avaient été envisagés (par exemple avec un effet traitement initialement négatif puis positif), ces croisements auraient probablement eu un impact direct sur les proportions de win/loss/tie, et donc sur les statistiques étudiées.

4 Données réelles

Je vais alors mettre en œuvre ces méthodes sur un jeu de données réelles issues de l'étude `BICAR − ICU` de Jung et al. (2023). Cette étude sur 28 jours vise à savoir si le bicarbonate de sodium a un effet positif sur des personnes en réanimation respiratoire ayant un pH faible (pH $\leq$ 7.20).

4.1 Contexte d'étude

La thérapie par bicarbonate de sodium pour l'acidémie métabolique sévère (pH faible dans le corps) est controversée, et son efficacité sur les critères de jugement n'a pas été évaluée de manière rigoureuse comme stipulé dans l'article.

Les médecins ont alors mené un essai multicentrique, ouvert, randomisé, contrôlé en phase III (voir 2 pour le détail des phases) dans 26 unités de soins intensifs adultes en France. Les patients présentaient une acidémie métabolique sévère (pH $\leq$ 7.20, bicarbonate $\leq$20mmol/L et Pa$CO_2 \leq$ 45 mmHg) ont été répartis au hasard pour recevoir soit un traitement conventionnel (groupe Contrôle) soit un traitement conventionnel plus du bicarbonate de sodium visant à maintenir un pH $\geq$ 7.3 (groupe Traité). 389 patients ont été inclus, 194 dans le groupe Traité et 195 dans le groupe contrôle.

Lors de la première étude, le critère de jugement principal composite était la mortalité à 28 jours toute cause confondue et la présence d'une défaillance d'organe nécessitant une dialyse.

Plusieurs analyses avaient été effectuées suivant le type de variable, les variables catégorielles composant le critère de jugement principal composé ont été comparées à l'aide d'un test du χ^2 ajusté ainsi qu'une régression logistique multivariée.
Globalement, aucune différence significative n'avaient été observés sur le critère de jugement (p-valeur = 0.41) mais pour le sous-groupe de personnes ayant un insuffisance rénal aiguë de stade 2 ou 3, le traitement par bicarbonate a significativement réduit le critère de jugement principal (p-valeur = 0.046).

4.2 Étude

Dans ce contexte d'étude, j'ai pu étudier l'effet du bicarbonate de sodium sur les patients ayant une acidémie sévère. Le critère de jugement principal composite a été revu, et maintenant il est composé de 4 variables que je vais classer par priorité, le décès, le recours à la dialyse, le nombre de jours sans ventilation mécanique et le nombre de jours sans vasopresseurs.

La **dialyse** permet de remplacer artificiellement les reins lorsque ceux-ci ne fonctionnent plus correctement (insuffisance rénale aiguë), elle permet d'éliminer les déchets, d'évacuer l'excès d'eau dans le corps et de corriger les déséquilibre électrolytique (potassium ou bicarbonate).
La **ventilation mécanique** consiste à assister ou à remplacer la respiration d'un patient lorsque ses poumons ne sont plus capable d'assurer une oxygénation suffisante ou d'éliminer correctement le CO_2. Il existe deux types principaux, la ventilation invasive où un tube est inséré dans la trachée et la ventilation non-invasive où on utilise un masque.
Les **vasopresseurs** sont des médicaments qui contractent les vaisseaux sanguins pour augmenter la pression artérielle.

Mes 4 critères de jugement composant le critère de jugement principal composite sont de type binaire pour le décès et la dialyse et quantitatif discret pour les jours sans ventilation mécanique et les jours sans vasopresseurs. Lorsque nous ferons la comparaison pour les variables quantitatives, des valeurs supérieures seront préférés, c'est à dire qu'une paire sera win si le patient du groupe T a des valeurs plus grande que le patient du groupe C selon ces critères quantitatifs et le seuil τ fixé par le médecin est nul.

Le critère de jugement principal composite repose sur 4 critères : deux sont binaire (décès et recours à la dialyse) et deux sont quantitatifs discrets (nombre de jours sans ventilation mécanique et nombre de jours sans vasopresseurs). Pour ces variables quantitatives, des valeurs plus élevées sont préférées. Une paire est considérée comme gagnante si le patient du groupe T présente une valeur supérieure à celle du patient du groupe C, avec un seuil τ fixé à 0 par le professeur Molinari.

Mon étude sera dans un premier temps non stratifiée et dans un second temps, 3 stratifications seront effectuées selon l'âge (≤ 65 ans ; ≥ 65 ans), la septis[5] (oui/non) et un score AKIN *Acute Kidney Injury Network*, un outil de classification de l'insuffisance rénale aiguë (stade 0-1 contre stade 2-3).
Les critères de jugements regroupé dans le critère de jugement composite et les différentes stratifications ont été choisis par un des médecins chargé de l'étude.
Comme dit précédemment, une étude a déjà été faite et le seul résultat significatif a été sur les patients ayant un score AKIN de stade 2-3.

4.2.1 Analyse descriptive du jeu de données

Il y avait 6 NA, 5 dans le groupe T et 1 dans le groupe C j'ai supprimé les individus. A partir de cela, nous avons 194 patients dans le groupe contrôle et 189 dans le groupe traité.

Les analyse se portent sur 28 jours

TABLE 29 – Répartition des variables binaires et stratification selon le groupe

Groupe	0	1
C (Contrôle)	90	104
T (Traitement)	107	82

(a) Décès

Groupe	0	1
C (Contrôle)	94	100
T (Traitement)	123	66

(b) Dialyse

Groupe	< 65 ans	≥ 65 ans
C (Contrôle)	94	100
T (Traitement)	89	100

(c) Âge à la randomisation

Groupe	Sans sepsis	Avec sepsis
C (Contrôle)	79	115
T (Traitement)	70	119

(d) Sepsis

Groupe	AKIN 0–1	AKIN 2–3
C (Contrôle)	104	90
T (Traitement)	101	88

(e) Score AKIN

Les tableaux 29 présentent la répartition des événements binaires et des variables de stratification selon les groupes d'étude.

La répartition des décès montre une mortalité légèrement plus importante dans le groupe contrôle (104 décès) par rapport au groupe traitement (82 décès), ce qui pourrait indiquer un effet favorable du traitement (voir tableau 29a). De même, le recours à la dialyse est plus fréquent dans le groupe contrôle (100 patients) que dans le groupe traitement (66 patients), suggérant une meilleure préservation de la fonction rénale chez les patients traités (tableau 29b).

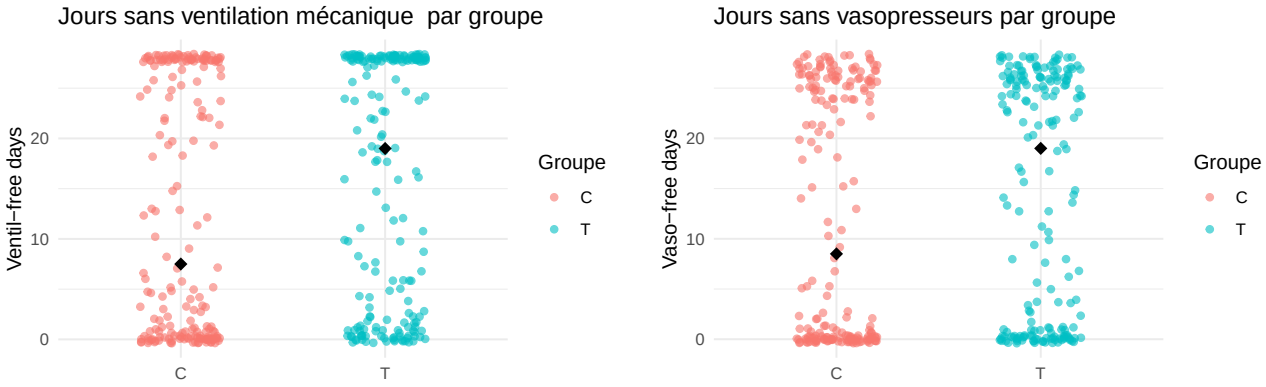
Concernant la présence de sepsis à l'inclusion, sur le tableau 29d, les deux groupes sont globalement équilibrés, avec 115 cas dans le groupe contrôle et 119 dans le groupe traitement. La distribution par tranche d'âge est également homogène, avec une répartition proche entre les patients de moins de 65

5. Réaction extrême et généralisée de l'organisme à une infection

ans et ceux de 65 ans et plus dans les deux groupes.

Enfin, la répartition du score AKIN, visible sur le tableau 29e indicateur de l'atteinte rénale aiguë, est semblable entre les groupes, avec des proportions comparables pour les stades 0–1 et 2–3. Ces observations confirment une bonne comparabilité initiale des groupes sur les principales variables cliniques et de stratification.

Sur le graphique 13b on voit la distribution du nombres de jours sans vasopresseurs pour chaque individus des groupes T et C. Visuellement, la dispersion semble similaire entre les groupes mais le point noir représentant la médiane est plus élevée dans le groupe T que dans le groupe C. On remarque que la plupart des patients passant soit peu de temps avec des vasopresseurs (inférieur à 3 jours) soit passent la plupart de l'étude sans (supérieur à 22 jours). Cette même analyse peut être portée sur la figure 13a



(a) Analyse de ventil_free_day par groupe

(b) Analyse de vaso_free_day par groupe

FIGURE 13 – Analyse univariée de 2 critères de jugements

Les variables *jours sans ventilation mécanique* et *jours sans vasopresseurs* présentent une distribution asymétrique dans les deux groupes (Figure 13). La médiane est plus élevée dans le groupe traité, ce qui pourrait traduire un effet bénéfique du bicarbonate de sodium. Des tests descriptifs sont présentés dans le tableau 30, où l'ont y voit la moyenne, la médiane et l'écart-type des deux variables quantitatives selon leur groupe.

TABLE 30 – Statistiques descriptives des variables quantitatives par groupe

Variable	**Groupe T**	**Groupe C**
Jours sans ventilation mécanique	Moyenne = 15.90 Médiane = 19 Écart-type = 12.11	Moyenne = 12.98 Médiane = 7.5 Écart-type = 12.44
Jours sans vasopresseurs	Moyenne = 14.45 Médiane = 19 Écart-type = 11.82	Moyenne = 12.44 Médiane = 8.5 Écart-type = 12.21

On remarque une forte différence entre les médianes selon les groupes, pour la variable *jours sans ventilation mécanique* par exemple, la médiane est beaucoup plus élevée dans le groupe T (19) que dans le groupe C (7.5) mais les moyennes sont plus proche (Groupe T : 15.9 ; Groupe C : 12.98).

Ces premières analyses suggèrent des tendances en faveur du traitement, notamment sur les variables *décès* et *dialyse*. Nous allons maintenant quantifier ces différences à l'aide des méthodes de comparaison par paire généralisée sur l'ensemble des patients, puis par stratification.

4.2.2 Étude non-stratifiée

Pour cette étude non-stratifié, la méthode sera non-appariée, il y aura donc $194 \times 195 = 37830$ paires sur lesquels les méthodes seront appliquées.

Table 31 – Résumé des résultats selon les critères

	win	**loose**	**tie**
Décès	11128	7380	19322
Dialyse	5664	3080	10578
Ventil_free_day	1937	1910	6731
Vaso_free_day	1800	1925	3006
Global	20529	14295	3006

Sur le tableau 31 on y voit le nombre de win/loose/tie au total lorsque l'étude est non-stratifiée. On constate que l'effet du bicarbonate de soude est bénéfique pour les outcomes **décès** et **dialyse** qui sont les plus prioritaires. En revanche il ne semble pas apporter de différence pour l'outcome **Ventil_free_day** et présente un plus fort nombre de loose par rapport au nombre de win pour l'outcome **Vaso_free_day** . Ce qui nous amène à penser que le traitement est globalement bénéfique ce qui sera confirmer dans le tableau 32 où les p-valeurs indiqueront l'effet significatif du traitement car les valeurs sont supérieures à l'hypothèse nulle.

Table 32 – Résumé global de l'étude non-stratifié

Métrique	**Valeur**	**CI_lower**	**CI_upper**	**z_score**	**p_val**
WR	1.4361	1.1202	1.8410	2.8560	0.0043
NB	0.1700	0.0521	0.2880	2.8252	0.0047
WO	1.4097	1.1135	1.7848	2.8529	0.0043

Ici, sur le tableau 32 on voit que les les 3 statistiques rejettent l'hypothèse nulle, leur p-valeur est inférieur à 0.05 et la valeur de l'hypothèse nulle ne situe pas dans l'intervalle de confiance à 95%. L'effet du bicarbonate de sodium parait donc bénéfique sur les patients étant en réanimation respiratoire et ayant une acidémie métabolique sévère.

Le choix de la statistique pour la conclusion de l'étude non-stratifiée sera donc apporter par le clinicien/médecin en fonction de l'interprétation qu'il voudra donner et s'il y a un intérêt à regarder chaque critère de jugement.

4.2.3 Étude stratifié selon l'âge

Nous allons maintenant stratifier selon l'âge en commençant par une étude stratifiée où nous aurons un résultat au global puis nous créerons deux sous-tableau composé des individus des deux strates afin d'avoir un résultat par strates ce qui nous permettra de porter une conclusion non pas globale mais spécifique aux individus ayant moins de 65 ans et ceux ayant plus de 65 ans.

Table 33 – Résultats des statistiques pour les données stratifiées selon l'âge

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.4868	1.1579	1.9091	3.1096	0.0019
NB (Net Benefit)	0.1859	0.0672	0.3045	3.0695	0.0021
WO (win Odds)	1.4566	1.1488	1.8468	3.1056	0.0019

On remarque sur le tableau 33 que les p-valeurs montrent encore l'effet significatif du traitement au bicarbonate par rapport au traitemetn contrôle, de plus on ne voit pas beaucoup de différence avec le tableau 32 avec les valeurs des statistiques et les interprétations sous-jacentes.
Nous allons alors voir si ce résultat global est aussi vérifié à l'intérieur de chaque strates et donc si l'effet positif du traitement par bicarbonate l'est quelque soit l'âge du patient. C'est pourquoi nous allons regarder maintenat ces analyses sur les personnes ayant moins de 65 ans.

Table 34 – Comparaisons pour le sous-ensemble de données comportant uniquement les personnes ayant moins de 65 ans

Critère	win	Lose	tie
Décès	2436	1581	4260
Dialyse	1215	738	2307
Ventil_free_day	297	392	1618
Vaso_free_day	632	562	424
Global	4580	3273	424

Tout d'abord, sur le tableau 34 on voit le descritif de win/loose/tie par critère de jugement composé, le traitement est bénéfique par rapport aux deux premiers critères de jugement comme lors de l'analyse globale, mais cette fois les conclsuions semblent différentes pour les deux autres, le critère de ventilation mécanique est légèrement perdant par rapport au groupe contrôle mais le critère de vasopresseur est légèrement gagnant.
Mais en regardant le tableau 35 on remarque que les p-valeur sont supérieures à 0.05, l'hypothèse nulle appartient aux intervalles de confiance montrant l'effet non-significatif du traitement pour l'ensemble de la sous-population ayant moins de 65 ans.
Il va falloir alors regarder la sous-population de personnes ayant plus de 65 ans afin de vérifier que l'effet du traitement est bien significatif pour cette tranche d'âge.

Table 35 – Résultats des statistiques pour le sous-ensemble de données comportant uniquement les personnes ayant moins de 65 ans

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.3993	0.9792	1.9997	1.8447	0.0651
NB (Net Benefit)	0.1579	-0.0114	0.3273	1.8275	0.0676
WO (win Odds)	1.3750	0.9800	1.9293	1.8430	0.0653

Sur le tableau regroupant le descriptif des win/loose/tie pour le sous-groupe d'individus ayant plus de 65 ans, on retroue des résultats similaires au tableau 32, le traitement semble être bénéfique pour les trois premiers critères de jugements mais pas pour le dernier, le nombre de jours sans vasopresseurs.

Table 36 – Comparaisons pour le sous-ensemble de données comportant uniquement les personnes ayant plus de 65 ans

Critère	win	Lose	tie
Décès	3038	1938	5024
Dialyse	1733	804	2487
Ventil_free_day	743	578	1166
Vaso_free_day	281	385	500
Global	5795	3705	500

Cette fois-ci, on note que l'effet du traitement est significatif pour le sous-ensemble d'individus ayant plus de 65 ans, la p-valeur étant de l'ordre de 10^{-2}

Table 37 – Résultats des statistiques pour le sous-ensemble de données comportant uniquement les personnes ayant plus de 65 ans

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.5641	1.1039	2.2162	2.5157	0.0119
NB (Net Benefit)	0.2090	0.0435	0.3745	2.4746	0.0133
WO (win Odds)	1.5284	1.0977	2.1283	2.5116	0.0120

En résumé, lorsque l'on stratifie selon l'âge, on observe un effet significatif du traitement par rapport au traitement de contrôle (ancien traitement) mais cet effet dépend de l'âge du patient, s'il a moins de 65 ans, on ne voit pas de différence significative entre le traitement de contrôle et le traitement au bicarbonate de sodium. L'analyse stratifiée cache cette information, l'effet global lorsque l'analyse est stratifiée est en fait moyenné sur des effets différents selon les sous-groupe de population.
L'étude doit donc tenir compte de cette intéraction, l'effet n'est pas homogène selon l'âge.

4.2.4 Étude stratifié selon la sepsis

Nous allons donc voir une étude stratifiée selon la sepsis, de la même façon que pour la stratification selon l'âge, trois analyses seront effectuées, une étude stratifiée et deux études selon le sous-groupe de population ayant oui ou non une sepsis.
Pour rappel, une sepsis survient quand une infection déclenche une réaction excessive du système immunitaire ce qui peut entraîner une chutte de pression artérielle ou une dysfonction des organes comme les reins, le foie ou les poumons.

Table 38 – Résultats des statistiques globales pour la stratification selon la sepsis

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.5005	1.1607	1.9397	3.0975	0.0020
NB (Net Benefit)	0.1890	0.0678	0.3103	3.0557	0.0022
WO (win Odds)	1.4662	1.1505	1.8686	3.0929	0.0020

Lorsque l'analyse est stratifiée selon la sepsis, le résultat montre la significativité du traitement, l'effet est positif selon les trois statistiques. Cependant, au vu de la stratification selon l'âge, on a vu qu'il fallait se méfier du résultat global stratifié et qu'une analyse plus approfondie en prennant les sous-groupe d'individus ayant la sepsis ou non est primordiale.
Nous allons donc commencer par prendre en compte uniquement les individus ayant eu une sepsis et regarder le nombre de win/loose/tie par critère de jugement. A noter qu'il y a 237 individus dans la catégorie des personnes ayant eu une sepsis contre seulement 151 n'ayant pas eu de sepsis.

Table 39 – Comparaisons pour les patients avec sepsis

Critère	win	Lose	tie
Décès	4225	2646	7151
Dialyse	2152	1197	3802
Ventil_free_day	820	869	2113
Vaso_free_day	367	541	1205
Global	7564	5253	1205

Le traitement a une nouvelle fois le même effet sur les premiers outcomes mais ici c'est intéressant de regarder les deux autres critères car nous nous plaçons sur une catégorie de personne ayant eu une sepsis qui peut donc entraîner un disfonctionnement du foie causant des problèmes directs sur nos critères de jugements composite expliquant le nombre élevé d'égalité. Parmi les personnes ayant eu une

sepsis, la prise de vasopresseurs est donc plus importante chez les patients suivant le traitement, c'est un résultat que nous retrouvons depuis le début de l'étude mais semble plus accrue sur ces personnes. Le résultat global semble aller en faveur du traitement comme nous le voyons dans le tableau 40 indiquant que le traitement est 1.41 fois meilleur que le contrôle selon le WO (plus ou moins 0.33694). Les p-valeurs vont dans cette direction, étant de l'ordre de 2×10^{-2}.

Table 40 – Résultats des statistiques pour les patients avec sepsis

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.4399	1.0443	1.9855	2.2242	0.0261
NB (Net Benefit)	0.1704	0.0186	0.3221	2.1999	0.0278
WO (win Odds)	1.4107	1.0413	1.9110	2.2215	0.0263

Nous avons donc vu que le traitement était meilleur que le contrôle pour les personnes ayant une sepsis, nous allons voir dans le sous-groupe de personne n'en n'ayant pas eu, un sous-groupe composé de 151 individus.
Ce que nous remarquons c'est que cette fois, les personnes n'ayant pas eu de sepsis passent plus de temps sans vasopresseurs lorsqu'ils ont eu le traitement au bicarbonate par rapport à ceux ne l'ayant pas eu et plus généréralement, le traitement semble utile pour tous les critères de jugements principaux.

Table 41 – Comparaisons pour les patients sans sepsis

Critère	win	Lose	tie
Décès	1638	1120	2772
Dialyse	790	294	1688
Ventil_free_day	284	189	1215
Vaso_free_day	547	357	311
Global	3259	1960	311

D'après le tableau 42, le traitement est 1.614 fois meilleur que le contrôle lorsque le patient n'a pas subi de sepsis, le traitement a donc un effet significatif sur le patient d'après les trois statistiques.

Table 42 – Résultats des statistiques pour les patients sans sepsis

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.6628	1.1080	2.4953	2.4550	0.0141
NB (Net Benefit)	0.2349	0.0433	0.4265	2.4034	0.0162
WO (win Odds)	1.6140	1.1003	2.3676	2.4491	0.0143

Ce que nous pouvons retenir de cette section c'est la validité du traitement par rapport au contrôle lorsque nous stratifions selon la sepsis et à l'intérieur des sous-populations ce qui dénote une absence d'intéraction forte entre le traitement testé et la présence d'une sepsis contrairement à l'âge. Ceci confirme l'intérêt du traitement dans la pratique clinique.

4.2.5 Étude stratifié selon le score AKIN

Nous allons enfin voir une stratification selon le score *AKIN*, un outil de classification de l'insuffisance rénale aiguë. Nous ferons alors une analyse stratifiée avant de faire cette même analyse sur les sous-groupes de populations. Il y a 206 patients étant de stade 0-1 contre 182 patients dans les stades 2-3.

Table 43 – Résultats des statistiques globales

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.4683	1.1418	1.8881	2.9935	0.0028
NB (Net Benefit)	0.1792	0.0604	0.2979	2.9572	0.0031
WO (win Odds)	1.4366	1.1329	1.8217	2.9895	0.0028

L'effet semble significatif lorsque nous stratifions, les p-valeur vont dans ce sens et les valeurs montre que le traitement au bicarbonate de sodium est plus efficace que l'ancien traitement.

Table 44 – Comparaisons pour le sous-groupe de stade 0-1 pour le score AKIN

Critère	win	Lose	tie
Décès	2726	2408	5269
Dialyse	1228	651	3390
Ventil_free_day	544	402	2444
Vaso_free_day	941	951	552
Global	5439	4412	552

Chez les patients présentant une atteinte rénale légère ou nulle, les résultats du tableau 45 montrent une tendance en faveur du traitement bien que les effets estimés ne soient pas significatifs. Le WO est estimé à 1.2191 (Intervalle de confiance à 95% : 0.8887-1.6723 et p-valeur de 0.2194) et pour le NB la valeur estimée est de 0.2848 (IC 95% : -0.0593-0.2568 et p-valeur de 0.2209). Ces valeurs, bien que supérieur à l'hypothèse nulle ne peuvent pas permettre de conclure d'un effet significatif du traitement.

Table 45 – Résultats pour le sous-groupe de stade 0-1 pour le score AKIN

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.2328	0.8829	1.7214	1.2286	0.2192
NB (Net Benefit)	0.0987	-0.0593	0.2568	1.2241	0.2209
WO (win Odds)	1.2191	0.8887	1.6723	1.2281	0.2194

Ces résultats suggèrent que, dans cette population à faible risque rénal, le traitement n'a pas démontré un effet statistiquement détectable, bie que l'effet observé aille dans le sens d'un bénéfice.
Nous allons voir dans le sous-groupe ayant un risque plus sévère d'insuffisance rénale si le résultat que l'on obtient est similaire.

Table 46 – Comparaisons pour le sous-groupe de stade 2-3 pour le score AKIN

Critère	win	Lose	tie
Décès	2793	1287	3840
Dialyse	1359	540	1941
Ventil_free_day	560	603	778
Vaso_free_day	143	169	466
Global	4855	2599	466

On voit une grosse différence du nombre de win suivant le groupe, une différence plus grande que lorsque le score AKIN était de 0 ou de 1. Un effet significatif est donc attendue et c'est ce que confirmera le tableau 47. Les effets sont plus marqués et significatifs, d'après le WO, le nouveau traitement est 1.7966 fois meilleur que l'ancien traitement (IC 95% : 1.2537-2.5747 et p-valeur de 0.0014) confirmant un effet robuste du traitement sur la sous-population.

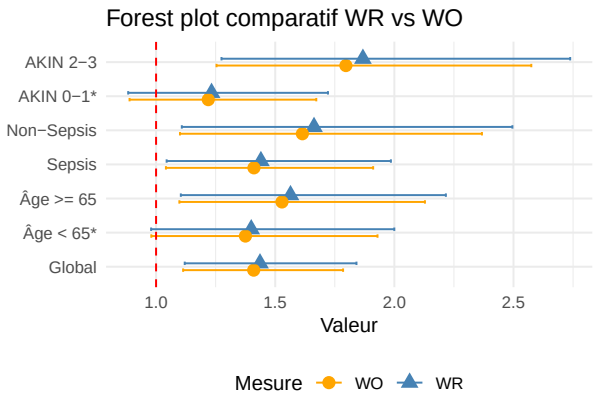
TABLE 47 – Résultats pour le sous-groupe de stade 2-3 pour le score AKIN

Statistiques	Valeur	CI_lower	CI_upper	z_score	p_val
WR (win Ratio)	1.8680	1.2745	2.7379	3.2034	0.0014
NB (Net Benefit)	0.2848	0.1049	0.4648	3.1031	0.0019
WO (win Odds)	1.7966	1.2537	2.5747	3.1914	0.0014

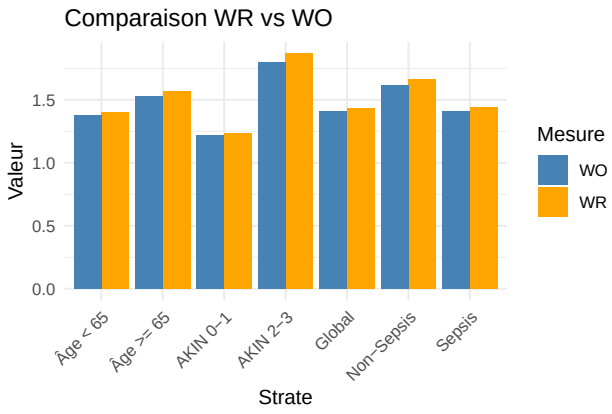
En comparant ces deux groupe, on observe une hétérogénéité de l'effet du traitement selon le score AKIN. Le traitement semble d'autant plus efficace que l'atteinte rénale est sévère, ce résultat était attendu au vu des anciennes analyses sur le sujet. Le traitement bénéfierait particulièrement aux patients ayant une atteinte rénale sévère.

4.3 Conclusion

Les résultats suggèrent que le traitement est globalement bénéfique, même si cet effet n'est pas systématiquement significatif dans tous les sous-groupes analysés. Par exemple, les patients âgés de moins de 65 ans ainsi que ceux présentant une atteinte rénale légère ou absente (score AKIN 0–1) ne montrent pas d'effet statistiquement significatif du traitement comme le montre les figures 14, bien que les valeurs des estimations restent au-dessus de l'hypothèse nulle. Cela pourrait indiquer un bénéfice potentiel non détecté par manque de puissance statistique ou une réelle moindre efficacité dans ces populations.



(a) Forest plot des WO et WR



(b) Barplot des WR et WO

FIGURE 14 – Analyse comparative des WO et WR selon les différentes analyses

En regardant les graphiques 14, on voit que les valeurs apportées par le WR sont légèrement supérieure à celles du WO tout en apportant la même conclusion, les strates *AKIN 0-1* et *Âge < 65 ans* ne sont pas significatives tout en gardant des valeurs supérieure à l'hypothèse nulle.
En examinant plus finement les différents critères de jugement, on constate que le traitement présente un effet favorable clair pour les deux premiers critères (décès et recours à la dialyse), avec un nombre de win systématiquement supérieur au nombre de loose dans la majorité des sous-groupes. En revanche, les résultats concernant le nombre de jours sans ventilation mécanique sont plus nuancés : le traitement ne semble pas supérieur à l'ancien, sauf chez les patients sans atteinte rénale sévère (AKIN 0–1) et chez les sujets âgés de plus de 65 ans. Cela suggère que, pour ce critère, l'effet du traitement dépend fortement du profil du patient.

Concernant la prise de vasopresseurs, l'effet du traitement apparaît plus hétérogène. Un bénéfice est observé dans certaines sous-populations comme les patients âgés de plus de 65 ans ou ceux présentant une sepsis, mais les résultats sont moins cohérents chez les patients plus jeunes ou sans sepsis. Cela pourrait traduire une interaction complexe entre l'état clinique du patient, l'usage de vasopresseurs, et l'efficacité du traitement étudié. Un approfondissement clinique avec un spécialiste est donc nécessaire

pour mieux interpréter ces disparités.

5 Conclusion

En conclusion, les trois méthodes étudiées — le *Win Ratio*, le *Win Odds* et le *Net Benefit* — aboutissent à des conclusions similaires lorsqu'il s'agit de tester une hypothèse nulle, notamment en termes de significativité statistique. Cela signifie que, pour une simple décision de rejet ou non de l'hypothèse nulle fondée sur la p-valeur, chacune de ces méthodes peut être utilisée de manière équivalente.
Cependant, dès lors que l'on souhaite quantifier la différence entre les groupes, des distinctions importantes apparaissent. En particulier, le *Win Ratio* tend à surestimer les effets observés par rapport au *Win Odds*. Cette surestimation s'explique par le fait que le *Win Ratio* ne prend pas en compte les égalités entre paires, ce qui entraîne une plus grande variabilité et des valeurs souvent exagérées. À l'inverse, le *Win Odds*, en intégrant ces égalités dans son calcul, fournit une estimation plus modérée et plus stable.
Le choix entre le *Win Odds* et le *Net Benefit* dépend alors de l'objectif de l'analyse : le *Win Odds* propose une mesure absolue de l'effet global, tandis que le *Net Benefit* permet une interprétation séparée par critère clinique (*outcome*), ce qui peut s'avérer pertinent dans des contextes où les priorités médicales varient selon les événements étudiés.
Enfin, il est important de souligner que le WO et le NB ont l'avantage de ne pas reposer sur l'hypothèse des risques proportionnels. Elles peuvent ainsi être utilisées de manière robuste pour analyser des données de type *time-to-event*, même dans des situations où cette hypothèse classique n'est pas vérifiée.

6 Annexe

Dans cette partie, nous aurons les tableaux correspondant aux descriptions statistiques des différents scénario du premier notebook. Les tableaux sont ordonné comme dans le rapports, les tableaux 48d représente les description statistiques des figures 1c'est à dire du premier scénario selon lequel les distributions sont identiques.

TABLE 48 – résumé statistiques du premier scénario où les distributions sont égales

Statistique	NBT	WR	WO
Min.	-0.157	0.729	0.729
1er Qu.	-0.0359	0.931	0.931
Médiane	-0.00045	0.999	0.999
Moyenne	-0.0000775	1.01	1.01
3e Qu.	0.0364	1.08	1.08
Max.	0.172	1.42	1.42

(a) $\tau = 0$

Statistique	NBT	WR	WO
Min.	-0.192	0.583	0.678
1er Qu.	-0.0351	0.908	0.932
Médiane	-0.00065	0.998	0.999
Moyenne	0.00046	1.01	1.01
3e Qu.	0.0341	1.10	1.07
Max.	0.186	1.67	1.46

(b) $\tau = 2$

Statistique	NBT	WR	WO
Min.	-0.172	0.582	0.707
1er Qu.	-0.0344	0.900	0.934
Médiane	-0.00105	0.997	0.998
Moyenne	0.000427	1.01	1.01
3e Qu.	0.0334	1.11	1.07
Max.	0.160	1.63	1.38

(c) Outcome discret de Poisson

Statistique	NBT	WR	WO
Min.	-0.193	0.580	0.676
1er Qu.	-0.0356	0.906	0.931
Médiane	-0.00055	0.998	0.999
Moyenne	0.000531	1.01	1.01
3e Qu.	0.035	1.10	1.07
Max.	0.192	1.70	1.47

(d) Outcome principal binaire

Les tableaux 49 représente les résumé statistiques du second scénario selon lequel le groupe T est meilleur que le groupe C dont les distribution des statistiques sont sur la figure 2.

TABLE 49 – résumé statistiques du second scénario où les distributions sont inégales

Statistique	NBT	WR	WO
Min.	0.180	1.44	1.44
1er Qu.	0.313	1.91	1.91
Médiane	0.346	2.06	2.06
Moyenne	0.345	2.07	2.07
3e Qu.	0.378	2.22	2.22
Max.	0.519	3.16	3.16

(a) $\tau = 0$

Statistique	NBT	WR	WO
Min.	0.200	1.68	1.50
1er Qu.	0.326	2.38	1.97
Médiane	0.356	2.60	2.11
Moyenne	0.356	2.63	2.12
3e Qu.	0.388	2.86	2.27
Max.	0.511	4.32	3.09

(b) $\tau = 2$

Statistique	NBT	WR	WO
Min.	0.28	2.16	1.78
1er Qu.	0.395	3.05	2.31
Médiane	0.424	3.35	2.47
Moyenne	0.423	3.38	2.48
3e Qu.	0.452	3.66	2.65
Max.	0.559	5.42	3.54

(c) Outcome discret de Poisson

Statistique	NBT	WR	WO
Min.	0.2001	1.682	1.500
1er Qu.	0.3260	2.380	1.967
Médiane	0.3563	2.605	2.107
Moyenne	0.3563	2.632	2.123
3e Qu.	0.3876	2.855	2.266
Max.	0.5109	4.324	3.089

(d) Outcome principal binaire

les tableaux 50 représente les résumé statistiques du troisième scénario.

Table 50 – résumé statistiques du quatrième scénario où les distributions sont fortement inégales en faveur de T

Statistique	NBT	WR	WO
Min.	0.1415	1.330	1.330
1er Qu.	0.2950	1.837	1.837
Médiane	0.3297	1.984	1.984
Moyenne	0.3302	2.004	2.004
3e Qu.	0.3658	2.154	2.154
Max.	0.5027	3.022	3.022

(a) $\tau = 0$

Statistique	NBT	WR	WO
Min.	0.3034	2.419	1.871
1er Qu.	0.4199	3.513	2.447
Médiane	0.4508	3.892	2.642
Moyenne	0.4505	3.969	2.665
3e Qu.	0.4822	4.384	2.862
Max.	0.5892	6.687	3.869

(b) $\tau = 2$

Les tableaux suivant seront ceux du dernier scénario du premier notebook

Table 51 – Résumé statistique des mesures – cas de critères mixtes

Statistique	NBT	WR	WO
Min	-0.10270	0.8003	0.8137
1er qu.	0.03555	1.0799	1.0737
Médiane	0.07005	1.1642	1.1507
Moyenne	0.06982	1.1712	1.1568
3e qu.	0.10433	1.2530	1.2330
Max	0.24520	1.7301	1.6497

Table 52 – Résumé statistique des mesures – cas de critères mixtes

Statistique	NBT	WR	WO
Min	-0.0983	0.8011	0.821
1er qu.	0.0743	1.1802	1.161
Médiane	0.1122	1.2833	1.253
Moyenne	0.1115	1.2939	1.260
3e qu.	0.1483	1.3944	1.348
Max	0.2997	1.9634	1.856

Table 53 – Résumé statistique des mesures – cas de critères mixtes

Statistique	NBT	WR	WO
Min	-0.6742	0.1545	0.1946
1er qu.	-0.5372	0.2529	0.3011
Med	-0.5052	0.2793	0.3287
Moy	-0.5042	0.2818	0.3310
3e qu.	-0.4737	0.3073	0.3572
Max	-0.3140	0.4721	0.5221

Bibliographie

Références

Brunner, Edgar, Vandemeulebroecke, Marc & Mütze, Tobias (2021). win odds : an adaptation of the win ratio to include ties. *Statistics in Medicine*, 40(14), 3367–3384.

Buyse, Marc, Saad, Everardo D., Peron, Julien, Chiem, Jean-Christophe, De Backer, Mickaël & Cantagallo, Eva & Ciani, Oriana (2021). The Net Benefit of a treatment should take the correlation between benefits and harms into account. *Journal of Clinical Epidemiology*, 137, 148–158.

Buyse, Marc (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29(30), 3245–3257.

Dong, Gaohong, Hoaglin, David C., Qiu, Junshan, Matsouaka, Roland A., Chang, Yu-Wei, Wang, Jiuzhou & Vandemeulebroecke, Marc (2020). The win ratio : on interpretation and handling of ties. *Statistics in Biopharmaceutical Research*.

Finkelstein, Dianne M. & Schoenfeld, David A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*, 18(11), 1341–1354.

Gehan, Edmund A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2), 203–224.

Jung, Boris and Huguet, Helena and Molinari, Nicolas and Jaber, Samir (2023). Sodium bicarbonate for the treatment of severe metabolic acidosis with moderate or severe acute kidney injury in the critically ill : protocol for a randomised clinical trial (BICARICU-2), au *British Medical Journal Publishing Group*, 13(8), e073487.

McCoy, C. Eric (2018). Understanding the use of composite outcomes in clinical trials. *Western Journal of Emergency Medicine*, 19(4), 631.

Pocock, Stuart J and Ariti, Cono A and Collier, Timothy J and Wang, Duolao(2012). The win ratio : a new approach to the analysis of composite outcomes in clinical trials based on clinical priorities. *European heart journal*, 33(2), 176–182

Verbeeck, Johan and De Backer, Mickaël and Verwerft, Jan and Salvaggio, Samuel and Valgimigli, Marco and Vranckx, Pascal and Buyse, Marc and Brunner, Edgar (2023) Generalized pairwise comparisons to assess treatment effects : JACC review topic of the week *American College of Cardiology Foundation Washington DC*, 82(13), 1360–1372.