Choix de la mesure de l'effet traitement dans les essais cliniques de patients infectés par le VIH

Xianmixikemaier Abulizi

▶ **To cite this version:**

Xianmixikemaier Abulizi. Choix de la mesure de l'effet traitement dans les essais cliniques de patients infectés par le VIH. Santé publique et épidémiologie. Sorbonne Université, 2019. Français. NNT : 2019SORUS446 . tel-02937883

HAL Id: tel-02937883

https://theses.hal.science/tel-02937883

Submitted on 14 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE DE DOCTORAT DE
SORBONNE UNIVERSITE

Spécialité

Epidémiologie clinique

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS : EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Présentée par

Mme **Xianmixikemaier ABULIZI**

Pour obtenir le grade de

DOCTEUR de SORBONNE UNIVERSITE

Sujet de la thèse :

Choix de la mesure de l'effet traitement dans les essais cliniques de patients infectés par le VIH

soutenue le 09/12/2019 devant le jury composé de :

Pr Laurence MEYER	Rapporteur
Pr Christian RABAUD	Rapporteur
Dr Valérie JOURNOT	Examinateur
Dr Florence TUBACH	Examinateur
Dr Raphaël PORCHER	Examinateur
Dr Philippe FLANDRE	Directeur de thèse

Sorbonne Université
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à MON : 01 42 34 68 41
Tél. pour les étudiants de MOO à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

LABORATOIRE DE RATTACHEMENT

INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique (IPLESP)
(Directeur : Fabrice CARRAT)

Equipe de recherche sur l'Epidémiologie clinique des maladies virales chroniques
(Responsable : Fabrice CARRAT)

Faculté de Médecine Sorbonne Université, Site Pitié-Salpêtrière
56, boulevard Vincent Auriol, 75013 Paris

A mes parents
A mon époux

REMERCIEMENTS

Je tiens tout d'abord à remercier Philippe Flandre, mon directeur de thèse, de m'avoir donné l'opportunité de réaliser cette thèse, de m'avoir épaulé et d'avoir été toujours très disponible pour m'aider au cours de ces quatre années de travail. Tes qualités humaines et scientifiques ainsi que tes remarques ont largement contribué à l'aboutissement de cette thèse. Je te prie de trouver ici le témoignage de ma profonde reconnaissance.

Un merci tout particulier à Madame Dominique Costagliola, je vous suis très reconnaissante de m'avoir accueilli au sein de votre structure et de m'avoir fourni les conditions matérielles et logistiques favorables au bon déroulement de ce travail. Veillez recevoir ici toute ma reconnaissance.

Je remercie sincèrement les rapporteurs de cette thèse, Laurence Meyer et Christian Rabaud, de me faire l'honneur de consacrer de leur temps précieux à la lecture de ce manuscrit. Je remercie aussi Valérie Journot, Florence Tubach et Raphaël Porcher d'avoir accepté d'examiner ce travail en participant à ce jury de thèse.

Je suis reconnaissante envers l'Ecole Doctorale Pierre Louis de Santé Publique - ED 393, pour avoir accepté favorablement mon projet de thèse et financé cette thèse en m'accordant un contrat doctoral pour les trois premières années. Je suis aussi reconnaissance à l'ANRS (Agence nationale de recherche sur le sida et les hépatites virales), qui a financé ma quatrième année de thèse en m'accordant une allocation de recherche.

Merci à mes amis doctorants de laboratoire avec qui j'ai partagé ces quatre ans d'aventure : Aliou, Sofia et Fabienne. Merci pour vos encouragements, vos précieux conseils et votre écoute m'ont permis de franchir ce cap.

Merci à Béatrice, Marianne et Sandra pour avoir toujours réglé rapidement tous les problèmes administratifs avec le sourire.

Merci tout spécialement Laurence Lièvre, pour sa disponibilité, son soutien moral et ces moments chaleureux passés en sa compagnie. J'exprime aussi toute ma reconnaissante à Hélène Roul pour la relecture du français de ce manuscrit.

Je remercie également tous mes collègues d'IPLESP, anciens et actuels, rencontrés pendant ces quatre années, pour votre sympathie et vos sourires qui ont rendu ce travail de thèse plus agréable : Murielle, Marguerite, Jonathan, Maxime, Valérie, Michelle, Hilma, Joséphine, Ophélia, Lydie, Hanna, Lambert, Rémy, Séverine.

Un grand merci aussi à tous ceux que j'aurais pu oublier.

Dédicaces

A mes chers parents, je dédie ce modeste travail,

A mon père, Ely, et ma mère, Roshane, qui à 8000 km ont toujours été près de moi et tiennent une place immense dans mon cœur. Merci de m'avoir inconditionnellement soutenu et encouragé à travers votre amour et votre confiance. Maman, une femme aussi adorable que toi je n'en connais pas, tu as toujours été là pour moi, et à aucun moment tu n'as cessé de me couvrir de ta tendresse. Papa, tu es une vraie école de la vie pour moi, je ne cesse d'apprendre avec toi.

A mon époux,

Muhter, pour ton irremplaçable et inconditionnel soutien et ta patience tout au long de cette aventure.

A ma sœur,

Mina, avec qui j'ai partagé cette aventure, toujours présente pour écarter les doutes, soigner les blessures et partager les joies. Merci pour tout ton amour.

A mes deux petits anges, mes filles

Deniz et Elise, votre venu au cours de thèse a été le plus beau cadeau pour moi. Je vous aime de tout mon cœur.

SOMMAIRE

RESUME……...9

ABSTRACT……11

VALORISATION DES TRAVAUX DE THESE ……...13

Article original publié 13

Communication affichée 13

OBJECTIFS ……...14

CHAPITRE I – INTRODUCTION GENERALE……..… 14

CHAPITRE II - METHODES……...……17

1. Le critère principal est mesuré à des temps fixes…….….17

2. Le critère principal est un délai…….……20

CHAPITRE III LES ESSAIS DU VIH…….25

1. Les essais du VIH…….…….25

2. Les essais de non-infériorité…….……28

3. Les essais de l'ACTG……32

3.1 L'essai ACTG 5202…….……32

3.1.1 La problématique et l'objectif de l'essai……..…32

3.1.2 Détermination de la borne de non-infériorité……...33

3.1.3 Résultats de l'essai……35

3.2 L'essai ACTG 5257…….....36

3.2.1 La problématique et l'objectif de l'essai……..…36

3.2.2 Détermination de la borne de non-infériorité…….……..37

3.2.3 Résultats de l'essai…….……..38

CHAPITRE IV CHOIX DE LA MESURE DE L'EFFET TRAITEMENT LORSQUE LES BORNES DE NON-INFERIORITE SONT ORIGINALEMENT DEFINIES PAR UNE DIFFERENCE ABSOLUE TRANFORMEE EN UNE DIFFERENCE RELATIVE A INFLUENCE LES RESULTATS DES ESSAIS CLINIQUES……...…41

1. Présentation de l'article……...…….41

2. Article intégral……..……...46

2.1 Appendice……...…….....56

2.2 Résultats supplémentaires……..……58

CHAPITRE V - L'UTILISATION DU TEMPS DE SURVIE MOYEN BORNE (RESTRICTED MEAN SURVIVAL TIME, RMST) COMME MESURE DE L'EFFET TRAITEMENT DANS LES ESSAIS CLINIQUES SUR LE VIH/SIDA : REANALYSE DE L'ESSAI ACTG5257……...……..……67

1. Présentation de l'article……...…67

2. Article intégral…….….71

2.1 Appendice……...……79

CHAPITRE VI - L'UTILISATION DU TEMPS DE SURVIE MOYEN BORNE (RMST) COMME MESURE DE L'EFFET TRAITEMENT DANS LES ESSAIS SUR LA PROPHYLAXIE PRE-EXPOSITION (PREP)……83

Manuscrit intégral……83

CHAPITRE VII – DISCUSSION GENERALE ET PERSPECTIVES…….99

1. Discussion des résultats……...…99

2. Enjeux futurs de la recherche……..……. …….104

BIBLIOGRAPHIE……..……..106

ANNEXES……..……111

RESUME

Différentes mesures de l'effet traitement sont disponibles. Les principales mesures sont la différence de risque, le risque relatif, et l'odds ratio. Il est bien connu en épidémiologie que ces mesures conduisent à des estimations différentes, elles sont moins étudiées dans le cadre des essais thérapeutiques. Par exemple, un essai récent (ACTG 5202) a montré des conclusions différentes selon la mesure de l'effet traitement utilisée. Le choix de cette mesure a de plus des implications très concrètes sur le choix du critère principal, le choix de l'analyse statistique, la taille d'échantillon à inclure et la détermination et l'interprétation de la borne de non-infériorité.

L'objectif de cette recherche est principalement d'étudier les différentes mesures de l'effet traitement pour d'une part comprendre et expliquer les différents résultats obtenus selon la mesure utilisée et d'autre part choisir la mesure la plus appropriée aux essais thérapeutiques du VIH. Ce choix est important car il implique la définition de nouvelles bornes de non-infériorité et détermine la taille d'échantillon pour ces essais. Une nouvelle mesure, la différence de temps de suivi sans événement (Restricted Mean Survival Time, RMST), est étudiée avec les mesures citées plus haut. L'étude de ces mesures est réalisée sur des données simulées et sur des données de deux essais de non-infériorité (équivalence) chez des patients naïfs de traitement (ACTG 5202 et ACTG 5257). Nous avons développé une collaboration avec une statisticienne des ACTG (Heather Ribaudo) pour l'analyse de l'ACTG 5257. Les méthodes sont des méthodes d'analyses statistiques en utilisant les logiciels SAS et R.

Globalement ce travail nous a permis d'améliorer la connaissance dans les choix méthodologiques lors de la construction d'un essai thérapeutique incluant des patients infectés par le VIH. Ce travail nous a permis en particulier 1) de mieux interpréter les résultats des essais thérapeutiques à partir des critères principaux utilisés, des mesures de l'effet traitement utilisées et des hypothèses de ces mesures; 2) de mieux comprendre les différences entre les essais

en fonction des différentes méthodologies utilisées ; 3) de proposer des recommandations sur l'utilisation des mesures de l'effet traitement en fonction des objectifs et des caractéristiques d'un essai mais aussi des hypothèses sous-jacentes aux choix méthodologiques.

Mots clefs : VIH ; Essais cliniques ; Mesures de l'effet traitement ; Recherche clinique ; Epidémiologie

ABSTRACT

Different measures of the treatment effect are available. The main measures are the risk difference, the relative risk, and the odds ratio. It is well known in epidemiology that these measures lead to different estimates, yet they are less studied in the context of therapeutic trials. For example, a recent trial (ACTG 5202) showed different conclusions according to the treatment effect measure used. The choice of this measure also has very concrete implications for the choice of the primary endpoint, the choice of the statistical analysis, the sample size to be included as well as the determination and interpretation of the non-inferiority margins.

The objective of this research is mainly to study the different measures of the treatment effect, on the one hand, to understand and explain the different results obtained according to the measure used, and, on the other hand, to choose the most appropriate measure for the HIV/AIDS clinical trials. This choice is important because it involves the definition of new non-inferiority margins and determines the sample size to be included for these trials. As a new measure, the difference in Restricted Mean Survival Time (RMST) is also studied with the measures mentioned above. The study of these measures is performed on simulated data and data from two non-inferiority (equivalence) trials in treatment-naïve patients (ACTG 5202 and ACTG 5257). We have developed collaboration with an ACTG statistician (Heather Ribaudo) for the analysis of ACTG 5257. The methods are statistical analysis methods using SAS and R software.

Overall, this work has allowed us to improve the knowledge of the methodological choices during the construction of a clinical trial including patients infected with HIV. In particular, this work allowed us to be able to 1) better interpret the results of the clinical trials based on the primary endpoint, the measures of the treatment effect used and the hypotheses of these measures; 2) better understand the differences between the trials according to the different methodologies used; 3) propose recommendations on the use of measures of treatment effect

according to the objectives and characteristics of a trial and provide underlying hypotheses for methodological choices.

Keywords: HIV; Clinical trials; Measures of treatment effect; Clinical research; Epidemiology

VALORISATION DES TRAVAUX DE THESE

Article original publié

Abulizi X, Flandre P. Choice of treatment-effect measures when noninferiority margins originally defined in absolute difference translated into relative difference influenced the results of clinical trials. *J Clin Epidemiol* 2018; 96:63-72. Doi:10.1016/j.jclinepi.2017.12.010.

Abulizi X, Ribaudo HJ, Flandre P. The use of the Restricted Mean Survival Time as a treatment measure in HIV/AIDS clinical trial: reanalysis of the ACTG A5257 trial. *J Acquir Immune Defic Syndr*. February 2019. doi:10.1097/QAI.0000000000001978

Communication affichée

Abulizi X, Ribaudo HJ, Flandre P. The use of the Restricted Mean Survival Time as a treatment measure in HIV/AIDS clinical trial: reanalysis of the ACTG A5257 trial. *26th Conference on Retroviruses and opportunistic Infection (CROI), Seattle, Washington, USA*: March 4 to 7, 2019, Washington, USA. Poster number 509.

OBJECTIFS

L'objectif de cette recherche est principalement d'étudier différentes mesures de l'effet traitement pour d'une part comprendre et expliquer les différents résultats obtenus selon la mesure utilisée et d'autre part choisir la mesure la plus appropriée aux essais cliniques du VIH.

CHAPITRE I - INTRODUCTION GENERALE

Les études cliniques et épidémiologiques visent à évaluer la relation entre les interventions et la survenue d'un événement. Les cliniciens souhaitent savoir non seulement s'il y a un lien entre une intervention donnée et la survenue d'un événement statistiquement significatif, mais également l'ampleur de cette relation. La mesure de l'effet traitement est une quantité qui résume l'association entre l'effet d'un traitement (d'une intervention thérapeutique au sens large) et le critère de jugement principal (la survenu d'un décès d'une progression d'une guérison ect…). Différents mesures de l'effet traitement comme la différence de risque, le risque relatif et l'odds ratio sont disponibles. Il est bien connu en épidémiologie que ces mesures conduisent à des estimations différentes (**Akobeng, 2005; Case, 2002; Walter, 2000**), mais elles ont été moins étudiées dans le cadre des essais thérapeutiques. Le choix de la mesure de l'effet traitement a de plus des implications très concrètes sur le choix du critère principal, le choix de l'analyse statistique, et la détermination et l'interprétation de la borne de non-infériorité.

La mesure la plus appropriée est aussi la mesure étant la plus robuste aux écarts des différentes hypothèses faites dans les considérations statistiques du protocole. En effet il est parfois difficile de fixer certaines hypothèses comme par exemple le taux de succès (ou échec)

dans chacun des bras de traitement. Ces choix déterminent le nombre de patients à inclure dans l'étude. Un écart trop grand entre le taux de succès observé et celui défini dans le protocole a des conséquences importantes sur la puissance de l'essai et donc sur les conclusions d'un essai. Ainsi, le choix de la mesure de l'effet a un impact important sur les essais. Un exemple est celui de l'essai ACTG 5202 détaillé ci-dessous et impliquant des patients infectés par la VIH (**Daar et al., 2011**). Dans cet essai on constate que le choix d'une différence de risques mesurée par une différence de Kaplan-Meier aurait conduit à une équivalence des bras de traitement dans l'essai alors que ceci n'a pas pu être mis en évidence dans l'analyse originale. Un tel exemple montre bien l'importance du choix de la mesure de l'effet sur les conclusions d'un essai qui ont par la suite un effet majeur sur les recommandations thérapeutiques.

Le domaine de recherche de cette thèse concerne la méthodologie statistique générale des essais quel que soit la pathologie même si les essais considérés dans ce document sont des essais du VIH. Mes travaux ont porté plus particulièrement sur la mesure de l'effet traitement dans le cas où le critère principal de l'essai est mesuré par la survenue d'un événement au cours du temps (décès, événement clinique ect…). Dans les essais du VIH c'est événement est le plus souvent un échec virologique. Les travaux ont particulièrement portés sur deux essais d'équivalence (non-infériorité) chez des patients naïfs de traitement : ACTG 5202 et ACTG 5257 (**Daar et al., 2011; Lennox et al., 2014**) mais les résultats de cette recherche pourront tout aussi bien s'appliquer pour des essais de switch ou d'intensification ou des essais chez des patients déjà pré-traités. La problématique se pose aussi pour d'autres pathologies que celles liées au VIH.

Ce manuscrit de thèse a été structuré sous la forme d'une « thèse-article ». Après une description des objectifs de cette thèse et une introduction générale visant à cadrer notre thématique de recherche, le deuxième chapitre se poursuit pour décrire les principales mesures de l'effet traitement et les méthodes statistiques mises en œuvre. La troisième partie est consacrée à la

présentation des essais cliniques de patients infectés par le VIH et particulièrement aux essais cliniques du VIH récents qui ont motivé nos travaux de recherche. Les quatrième et cinquième chapitres sont consacrés aux analyses empiriques menées dans le cadre de cette thèse. Ces chapitres commencent chacun par une présentation synthétique de l'article, suivie du texte intégral et des résultats supplémentaires pour répondre aux objectifs suivants :

- Etudier les performances de trois mesures de l'effet traitement : la différence de deux estimations de Kaplan-Meier à un temps donné (DR_{KM}), le rapport de risques (Hazard ratio, HR) estimé par le modèle de Cox, et la différence de temps de survie moyen borné sans événement (Restricted Mean Survival Time, RMST) - dans l'analyse d'essais randomisés utilisant un critère d'évaluation du délai jusqu'à l'événement et évaluer leur influence sur les résultats de l'essai,
- Comparer l'analyse basée sur la différence de la mesure du temps de survie moyen borné sans événement (RMST) et deux mesures de l'effet traitement (DR_{KM} et le rapport de risques) dans un essai récent d'équivalence du VIH, et étudier les performances et les caractéristiques d'une analyse basée sur la différence de RMST dans le contexte de risques proportionnels et non-proportionnels.

Celui présenté dans le quatrième chapitre a été publié dans la revue ***Journal of Clinical Epidemiology*** début 2018. Celui présenté dans le cinquième chapitre a été publié dans la revue ***Journal of Acquired Immune Deficiency Syndromes*** début 2019.

Le sixième chapitre se poursuit pour décrit un travail qui fait l'objet d'un article en cours de réalisation, présente la performance des analyses basées sur la différence de RMST dans le contexte des essais de la prophylaxie pré-exposition (PeEP). Le septième chapitre est consacré à une discussion générale des résultats obtenue dans le cadre de ce travail, et nous terminerons ce manuscrit par les conclusions et les perspectives et enjeux futurs qui font suite à ce travail de thèse.

CHAPITRE II - METHODES

Dans les études cliniques et épidémiologiques, différentes mesures de l'effet traitement sont disponibles. Néanmoins, il n'y a pas de consensus entre les chercheurs sur le choix de la mesure de l'effet traitement. Le choix de la mesure de l'effet la plus appropriée a fait l'objet d'un long débat, mais aucune référence générale n'a encore été élaborée pour une seule étude ainsi que pour les méta-analyses (**Akobeng, 2005; Papageorgiou et al., 2015; Walter, 2000**). Le choix de la mesure de l'effet revêt une importance primordiale lorsque différentes mesures conduisent à des conclusions différentes pour un essai donné.

1. Le critère principal est mesuré à des temps fixes

Les principales mesures de l'effet traitement à un temps fixe sont la différence de risque (DR), le risque relatif (RR), et l'odds ratio (OR). Pour plus de commodité, je soulignerai la situation où P_1 représente la proportion d'événements dans le groupe 1, et P_2 la proportion d'événements dans le groupe 2, mais nous pouvons facilement généraliser nos comparaisons aux autres conceptions. Alors les mesures de l'effet fréquemment utilisées sont les suivantes :

La différence de risque = DR = $P_1 - P_2$

Le risque relatif = RR = $\frac{P_1}{P_2}$

Odds ratio = OR = $\frac{P_1/(1-P_1)}{P_2/(1-P_2)}$

Par exemple, supposons que la proportion d'événements soit P_1=0.3 et P_2=0.4 dans le groupe 1 et le groupe 2 respectivement:

	Nombre de patients avec événement	Nombre de patients sans événement	Total
Groupe 1	30	70	100
Groupe 2	40	60	100

Ensuite nous avons :

DR = 0.3-0.4 = -0.1 ;

RR = 0.3/0.4 = 0.75 ;

OR = (0.3/0.7) / (0.4/0.6) = 0.64.

Ainsi, le calcul de DR est simplement la différence entre la proportion d'événements cumulée dans le groupe 1 et la proportion d'événements cumulée dans le groupe 2, il s'agit donc d'une mesure absolue. Le RR peut être calculé comme le rapport entre deux proportions d'événements. L'OR, également appelé le rapport des cotes, se définit comme le rapport de la cote d'un événement survenue au groupe 1, avec celle du même événement survenue au groupe 2. Le RR et OR sont des mesures relatives **(Bouyer, 2009; Rothman et al., 2008)**. Quand l'événement est peu fréquent, les estimations de RR et OR sont très proches. Maintenant, à titre d'exemple, nous considérons cette situation à partir du tableau ci-dessus : en supposant que l'exposition reste la même, le nombre de patients sans événement est triplé dans le groupe 1 et dans le groupe 2 :

	Nombre de patients avec événement	Nombre de patients sans événement	Total
Groupe 1	30	70+70+70	240
Groupe 2	40	60+60+60	220

Dans cette situation, la proportion (fréquence) d'événements dans le groupe 1 est de 30/240=0.125, la proportion d'événements dans le groupe 2 est de 40/220=0.180, le risque relatif est donc 0.125/0.182=0.68. Si l'on calcule l'OR on a (0.125/0.875) / (0.182/0.818)=0.64. Nous pouvons constater que le RR diminue de 0.75 à 0.68, en revanche, l'OR n'a pas changé. Cela est dû à la manière dont les deux mesures de l'effet sont calculées. En fait, la proportion d'événements avec laquelle nous calculons le RR dépend à la fois du nombre de patients avec événement et du nombre de patients sans événement, mais aussi du nombre d'individus exposés et non exposés ; tandis que l'OR, étant un indicateur de risque absolu au sein du groupe, est affecté par la répartition des expositions dans chaque groupe. Etant donné que la répartition d'exposition dans le groupe avec un nombre de patients sans événement plus important n'a pas changé par rapport au groupe d'origine, l'OR est resté inchangé **(Tripepi et al., 2007)**. Ainsi, quand l'événement étudié est rare, les estimations de RR et OR sont très proches, quand l'événement est très fréquent, les estimations de RR et OR sont éloignées.

En tant que mesure absolue, l'information donnée par la DR fournit, dans certaines situations, un résumé facile à interpréter cliniquement et une idée directe de ce qui se passe par rapport au RR ou à l'OR. La DR reflète la différence entre deux incidences cumulées jusqu'au temps T et l'analyse est réalisée quand tous les patients ont été suivis pour la même période. En tant que mesure relative, le RR est considéré comme une mesure relative de référence car il s'agit d'une mesure généralement bien comprise, mais dans certains cas, le RR ne peut pas être obtenu directement (comme dans une étude cas-témoins). L'OR est une estimation efficace pour l'association entre deux variables binaires dans de petits échantillons ou quand le phénomène est rare, mais l'OR est indéniablement la mesure la plus difficile à interpréter, il est donc probablement moins utile que la DR ou le RR pour communiquer le risque (**Newcombe, 2006**). Les mesures relatives ont la particularité de résumer deux quantités (la proportion

d'événements dans deux groupes) en une, donc le calcul des mesures relatives est toujours basé sur un risque de base. Cependant, cette caractéristique représente également leur principale faiblesse, à savoir que les mesures absolues sous-jacentes sont dissimulées et que les lecteurs ont tendance à surestimer l'effet d'un traitement lorsqu'il est présenté en termes relatifs, en particulier dans le cas où la maladie est rare dans la population source. Ainsi, un RR=2 pourrait correspondre à une différence de risque absolu de 2% versus 1%, ou à une différence de 60% versus 30%. Dans de nombreuses situations, les mesures absolues donnent une meilleure représentation de la situation réelle et, du point de vue du patient également, les mesures absolues donnent souvent des informations plus pertinentes (**Noordzij et al., 2017**).

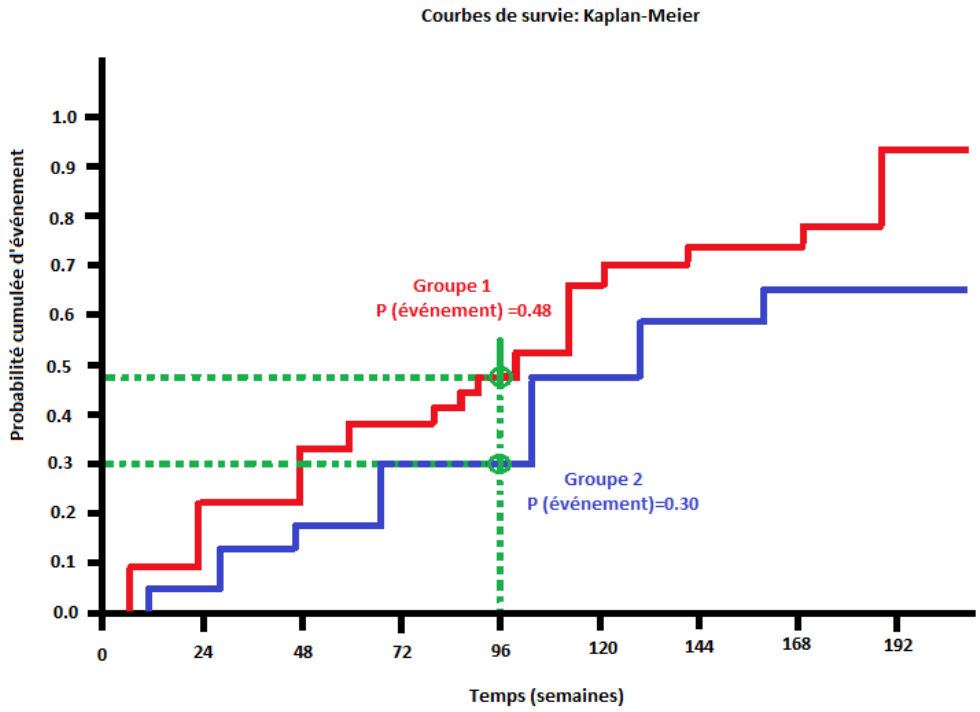
2. Le critère principal est un délai

Mon travail de thèse réalisé jusqu'à présent a particulièrement porté sur l'étude de mesures de l'effet traitement lorsque celui-ci est mesuré par la survenue d'un événement au cours du temps. Deux mesures classiques de l'effet traitement et une mesure introduite plus récemment ont été étudiées. Les deux mesures classiques sont la différence de deux estimations de Kaplan-Meier à un temps donné (DR_{KM}) et le rapport de risques (Hazard ratio, HR) estimé par le modèle de Cox. La nouvelle mesure est une différence de temps de survie moyen borné sans événement (Restricted mean survival time, RMST). Elles sont décrites plus bas.

2.1 Différence de risque fournie par deux estimations de Kaplan-Meier (DR_{KM})

La méthode de Kaplan-Meier permet d'obtenir une courbe de survie et d'estimer les probabilités cumulées d'événement à un temps donné. Les tests statistiques du type Logrank permettent de comparer des courbes de survie obtenues avec différents traitements. Cette méthode est la plus utilisée lorsqu'aucune hypothèse ne veut être faite sur la distribution des temps de survie. Ici la DR pourrait être estimée par la différence de deux probabilités cumulées

d'événement à un temps donné fournies par la courbe de Kaplan-Meier (DR_{KM}) comme présenté ci-dessous.



Dans cette courbe de Kaplan-Meier, la probabilité cumulée d'événement à la semaine 96 est de 0.48 dans le groupe 1 et de 0.30 dans le groupe 2. La DR_{KM} est donc de 0.18 (0.48-0.30). Un estimateur de la variance de cette différence permet d'estimer un intervalle de confiance à 95%.

Une différence de DR_{KM} avec DR est que la courbe de Kaplan-Meier peut prendre en compte des données censurées, ce qui intervient lorsqu'un patient disparaît d'une étude pour des raisons non liées à la conduite de celle-ci ou au traitement reçu (par exemple les patients qui déménagent), c'est-à-dire qu'on ne dispose plus de ses données avant l'apparition de l'événement. On parle de censure « non-informative ». Dans ce cas une hypothèse importante est que la raison du départ des patients de l'étude doit être indépendante de leur risque de l'événement. C'est-à-dire qu'à chaque temps, les patients censurés ont la même perspective de survie que ceux qui continuent d'être suivis. La deuxième notion associée aux données censurées est celle de censure non-informative a priori, lorsque les patients sont recrutés sur une

longue période de temps et que l'analyse est réalisée à un temps T donné, les participants recrutés le plus récemment sont censurés lors de l'analyse. C'est une différence importante par rapport à l'analyse de la DR car il ne peut pas y avoir de patients censurés dans l'analyse de la DR, un patient perdu de vue est soit considéré comme un échec soit exclu de l'analyse. En l'absence de censure, l'estimation de Kaplan-Meier au temps *t* dans un groupe est le nombre de patients ayant eu un événement entre le temps 0 et le temps *t* divisé par le nombre de patients randomisés dans ce groupe. Dans ce cas, la DR_{KM} au temps *t* est identique à la DR (P_2-P_1) au temps *t*.

La DR_{KM} fournit un résumé interprétable cliniquement des différences entre les groupes à un temps donné et ne nécessite pas d'hypothèse paramétrique pour la distribution des temps de survie dans les deux groupes. Un avantage de la DR_{KM} est qu'elle fournit des informations sur les risques absolus. Cependant, la DR_{KM} ne présente qu'une comparaison à un temps donné sans refléter toute la différence cumulée pour toute la durée d'une étude (**Uno et al., 2015**).

2.2 Le rapport de risques instantanés (Hazard ratio, HR)

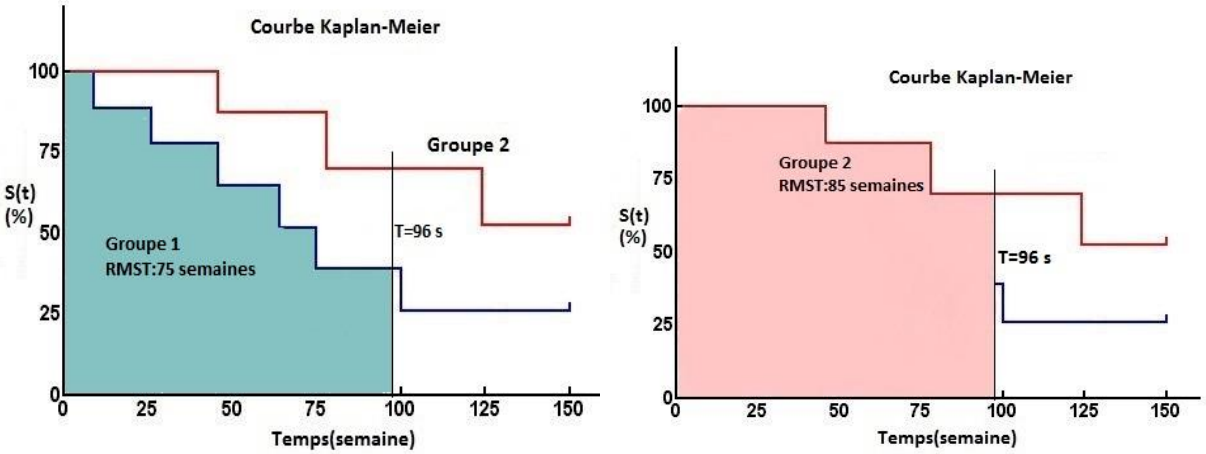
Le rapport de risques instantanés est le rapport des risques instantanés dans le groupe recevant le traitement 1 divisé par les risques instantanés dans le groupe recevant le traitement 2. Dans les études cliniques avec un critère de délai jusqu'à un événement, nous étudions la survenue d'un événement binaire en tenant compte de la dynamique temporelle de survenue. Comme décrit plus haut il existe des censures liées à l'interruption du suivi avant la survenue d'événement ou survenant durant la période de suivi, induites par des patients perdus de vue. On utilise le rapport de risques instantanés pour quantifier l'effet du traitement. Par exemple, si le risque instantané de l'événement dans le groupe 1 est de 0.05 et constant dans le temps,

et le risque instantané dans le groupe 2 est de 0.025 et également constant dans le temps, le rapport de risques instantanés est 0.025/0.05=0.5. On associe au rapport de risques instantanés un intervalle de confiance. L'interprétation de ce résultat est : le risque instantané de l'événement dans le groupe 2 est seulement la moitié de risque instantané de l'événement dans le groupe 1.

Le rapport de risques instantanés diffère des RR et OR décrits plus haut en ce que les RR et l'OR sont cumulatifs pour toute la durée d'une étude, en utilisant un critère défini, tandis que le rapport de risques instantanés représente les risques instantanés sur une période de l'étude. Dans la pratique, les risques instantanés sont généralement variables en fonction du temps donc le rapport de risques instantanés est une fonction du temps. Le rapport de risques instantanés est couramment estimé à l'aide d'un modèle à risques proportionnels de Cox **(Case, 2002; Hernán, 2010; Spruance et al., 2004)**. Ce modèle est une méthode de régression pour les données de survie à la fois puissante et flexible, elle fournit une estimation du rapport de risques et de son intervalle de confiance. Le rapport de risques instantanés qui est dérivé de ce modèle, constitue un résumé valide de la différence entre 2 distributions cumulatives lorsque l'hypothèse des risques proportionnels est respectée. Dans l'hypothèse de risques proportionnels, l'effet du traitement doit rester constant en fonction du temps. C'est une hypothèse forte du modèle à risques proportionnels de Cox, il faut donc au préalable vérifier que celle-ci est satisfaite. Comme le risque relatif et l'odds ratio, la précision des estimations de rapport de risques instantanés dépend principalement du nombre d'événements observés. **(Uno et al., 2015**).

2.3 Temps de survie moyen borné (Restricted Mean Survival Time, RMST)

La mesure du temps de survie moyen borné a été récemment proposée et consiste à estimer pour un temps T donné, le temps moyen passé sans survenue de l'événement que l'on considère comme critère principal. Techniquement il s'agit d'estimer l'intégrale sous une courbe de Kaplan-Meier comme présenté ci-dessous.



Il est important de spécifier un point temporel de troncature approprié (*tau*) pour calculer le RMST, qui doit être suffisamment longue pour évaluer ce que nous voulons. Notez que ce *tau* doit être inférieur au minimum du plus grand temps d'événement observé chez les sujets suivis. Lorsque le *tau* n'est pas spécifié, le minimum du plus grand temps d'événement observé chez les sujets est utilisé par défaut.

Par exemple, dans la figure ci-dessus, nous sommes intéressés à estimer le RMST à la semaine 96 (*tau* = 96 semaines). Pour une estimation de Kaplan-Meier faite jusqu'à la semaine 96, l'intégrale sous la courbe peut donner une mesure de temps de suivi moyen sans événement (RMST) de 75 semaines dans le groupe recevant le traitement 1. Ceci indique qu'un futur patient, issu de la même population et recevant le même traitement, suivi également pour une durée de 96 semaines passera en moyenne 75 semaines sans survenue de l'événement. La mesure d'un effet traitement est alors estimée par la différence entre deux moyennes estimées pour deux groupes de traitement. Ici, sachant que le RMST estimé est de 85 semaines dans le groupe 2, cette différence est de 10 semaines en faveur du traitement 2

par rapport au traitement 1, cela indique qu'un patient recevant le traitement 2 passera en moyenne 10 semaines de suivi de plus sans l'événement par rapport à un patient recevant le traitement 1. Un estimateur de la variance de cette différence a été proposé, il permet de tester statistiquement cette différence et d'estimer un intervalle de confiance (**Zhao et al., 2012**). Par exemple, une différence de RMST entre deux groupes de traitement peut être estimée à 10 semaines avec une variance conduisant à un intervalle de confiance de 7 à 15 semaines.

La mesure RMST a été principalement discutée en oncologie, pour de nombreux types de cancer (Trinquart et al., 2016, 2019a; Uno et al., 2014, 2015). L'analyse basée sur le RMST est considérée comme une mesure récapitulative robuste et interprétable cliniquement qui ne repose pas sur l'hypothèse de risques proportionnels et constitue un événement estimable même sous une forte censure (**Royston and Parmar, 2013, 2011; Uno et al., 2015; Zhao et al., 2012**). Cependant, la mesure nécessite également une pré-spécification du temps d'intérêt pour permettre aux chercheurs d'envisager la tentation de choisir un temps spécifique afin d'obtenir des résultats plus significatifs.

CHAPITRE III LES ESSAIS DU VIH

1. Les essais du VIH

La recherche clinique effectuée chez les patients infectés par le VIH vise à améliorer la prise en charge et le traitement des personnes vivant avec le VIH. Elle cherche aussi à prévenir

l'infection du VIH. La recherche clinique du VIH s'appuie sur des essais thérapeutiques, des études de cohortes et des études physiopathologiques. Un essai thérapeutique consiste à évaluer les effets de différentes stratégies de traitements antirétroviraux. L'évaluation peut porter sur de nouveaux médicaments ou associations de médicaments contre le VIH ou ses effets secondaires, de nouvelles façons de les administrer, ou de nouvelles modalités de traitement et de prise en charge. Le niveau de la charge virale plasmatique (ARN-VIH) des patients constitue le critère biomédical dominant dans le traitement du VIH.

Les essais thérapeutiques chez les patients infectés par le VIH utilisent le plus souvent deux critères principaux (**Flandre, 2014**). Le premier critère se défini par la proportion de patients ayant une charge virale inférieure à un seuil de détection (charge virale <50 copies/ml) après X semaines de traitement. Par exemple, un patient ayant une charge virale supérieure à 50 copies/ml à la semaine 48 (S48) est donc considéré comme un échec. Le plus souvent, un patient ayant quitté l'étude à S48 ou avant, ayant modifié ou arrêté définitivement son traitement à S48 ou avant, ou ayant une charge virale manquante à S48 sera considéré comme un échec. L'exclusion de ces patients dans l'analyse peut introduire des biais considérables car les données générées par ces patients avant leur sortie sont importantes pour la puissance et la validité de l'étude, donc c'est fortement déconseillé. Ce critère correspond à une proportion cumulée d'échecs jusqu'à S48 et non à une proportion d'échecs au temps S48, donc l'analyse est réalisée quand tous les patients ont été suivis durant la même période (48 semaines), et ce critère n'est pas purement virologique. Dans les essais du VIH, les analyses de ce critère sont souvent fondées sur une différence de risque (DR) qui est simplement la différence de 2 proportions d'événements jusqu'à S48 (**Gallant et al., 2016; Orkin et al., 2019; Venter et al., 2019**).

Le second critère se définit par un délai jusqu'à l'échec virologique après une certaine période de suivi. L'échec virologique est défini principalement par le fait de ne pas avoir une charge virale inférieure à un certain niveau après X semaines de traitement ou par un rebond de la charge virale au cours du suivi dans l'étude. Dans les essais du VIH qui utilisent ce critère, on définit une période de suivi après l'inclusion du dernier patient, par exemple 96 semaines. Si la période d'inclusion est aussi de 96 semaines et si le premier patient inclus n'est pas perdu de vue ou ne quitte pas l'essai, il sera suivi 192 semaines alors que le dernier patient sera suivi 96 semaines au maximum. On constate donc que dans ces essais les patients ne sont pas suivis durant la même période, et les patients perdus de vue ou les patients quittant l'essai pour des raisons non liées à la conduite de l'essai ou au traitement reçu avant l'apparition de l'échec seront considérés comme censurés dans l'analyse. Ces sont des données censurés non-informatives. Il peut y avoir également des données censurées non-informatives a priori comme décrit dans le chapitre II. L'analyse principale s'effectue en estimant un rapport de risques (Hazard ratio, HR) à l'aide du modèle de Cox. Dans le cas du délai jusqu'à un échec, la méthode de Kaplan-Meier est aussi largement utilisée pour fournir une estimation de la proportion cumulée d'échecs aux différents temps de suivi. Une différence entre deux estimations de Kaplan-Meier fournit donc une estimation d'une différence de risque (DR_{KM}) à un temps donné. Les essais du groupe AIDS Clinicat Trial Group (ACTG) utilisent traditionnellement ce type de critère **(Lennox et al., 2014; Riddler et al., 2008; Sax et al., 2009**).

Les cliniciens ont naturellement tendance à comparer les résultats des différents essais thérapeutiques incluant les mêmes combinaisons ou des combinaisons différentes. Or ces comparaisons sont difficiles en particulier du fait de la différence entre les critères principaux utilisés et la mesure de l'effet traitement choisie **(Case, 2002; Flandre, 2015; Walter, 2000)**. Par exemple, une analyse intermédiaire de l'essai ACTG 5202 a mis en évidence un risque accru

d'échec virologique chez les patients dont la charge virale plasmatique (ARN-VIH) était supérieure à 100 000 copies/mL et ayant reçu abacavir plus lamivudine (ABC/3TC) par rapport à ceux ayant reçu tenofovir plus emtricitabine (TDF/FTC) (**Sax et al., 2009**). Dans cet essai, l'analyse a inclus les données de 797 patients présentant un niveau de charge virale ARN-VIH au dépistage de 100 000 copies/mL ou plus, le critère principal d'évaluation de l'efficacité était le délai entre la randomisation et l'échec virologique, et l'analyse principale était basée sur le rapport de risques. Il s'agissait de la première étude montrant une différence d'efficacité entre ABC/3TC et TDF/FTC chez les patients avec un niveau de charge virale élevée. Cependant, une étude précédente avait montré une efficacité comparable entre ABC/3TC et TDF/FTC après 48 semaines de traitement, quel que soit le niveau de charge virale au dépistage des patients (la charge virale ARN-VIH < 100 000 copies/mL ou >= 100 000 copies /mL) (**Smith et al., 2009**), et un résultat similaire a été trouvé après 96 semaines de suivi. Dans cette étude, le critère principal d'évaluation était la proportion de patients dont la charge virale plasmatique est inférieure à 50 copies/ml après 48 semaines de traitement, les patients perdus de vue étaient considérés comme des échecs dans l'analyse. L'analyse principale était basée sur une différence de risque (DR) **(Smith et al., 2009)**. Certaines caractéristiques de base des patients et les définitions des critères principaux, étaient différentes entres les deux études **(Sax et al., 2009; Smith et al., 2009)**. Comme nous pouvons le constater, la différence entre les critères principaux utilisés et la mesure de l'effet traitement choisie rend la comparaison des combinaisons difficiles **(Ha et al., 2009)**.

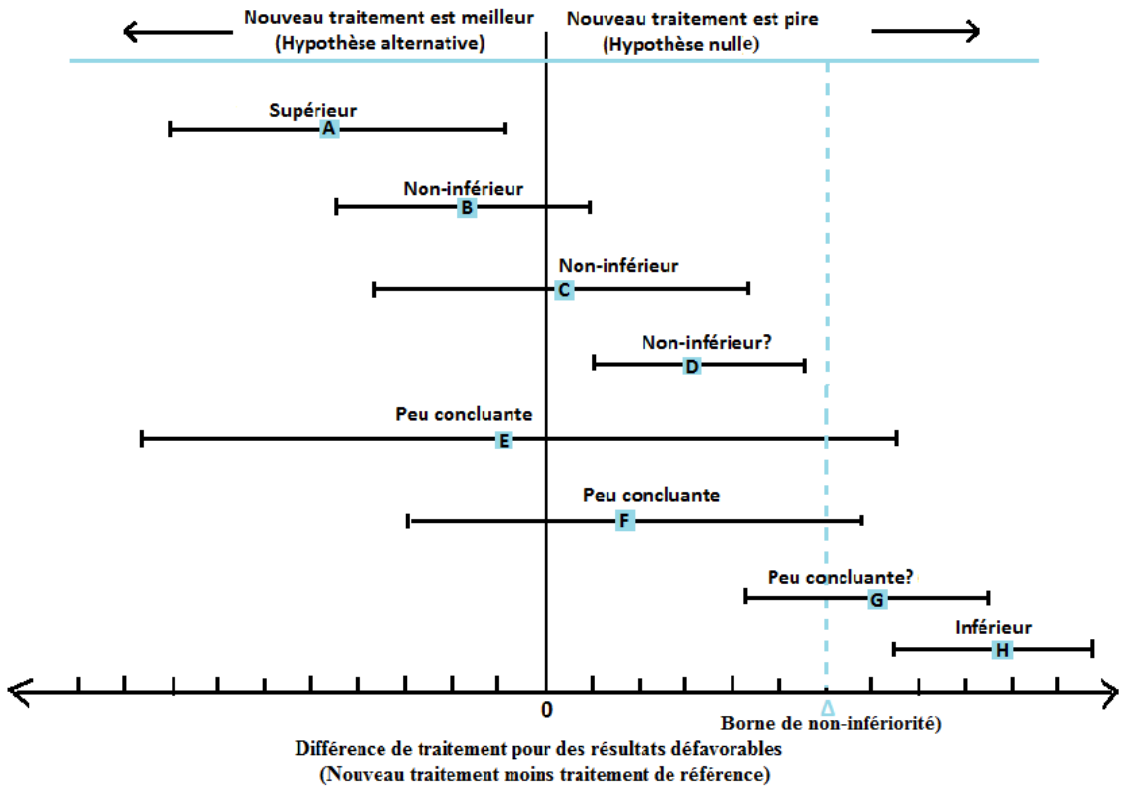
2. Les essais de non-infériorité

Depuis plusieurs années, les essais thérapeutiques du VIH ont la particularité d'utiliser quasiment exclusivement un design de non-infériorité ou d'équivalence **(Flandre, 2015, 2014; Hill and Sabin, 2008)**. Dans un essai de non-infériorité, l'objectif n'est pas de démontrer la supé-

riorité d'une nouvelle combinaison par rapport à un traitement de référence, mais de montrer que cette nouvelle combinaison n'est pas non-inférieure au traitement standard (**Fleming, 2008**). Si l'on accepte une certaine perte d'efficacité, ce nouveau traitement doit par ailleurs apporter un avantage par rapport au traitement de référence, par exemple une toxicité moindre, une résistance moindre, ou un coût réduit.

Dans un essai de non-infériorité, les hypothèses nulles et alternatives sont inversées car l'hypothèse nulle est l'infériorité du traitement expérimental par rapport au traitement standard, et l'alternative est la non-infériorité du traitement expérimental par rapport au traitement standard (**Blackwelder, 1982; Piaggio et al., 2006**). La non-infériorité est démontrée si l'efficacité du traitement expérimental n'est pas trop en deçà de celle du traitement standard. Ce résultat ('en deçà') se mesure par rapport à une borne de non-infériorité qui correspond donc à la perte maximale d'efficacité acceptée (**Fleming, 2008; Piaggio et al., 2006**). Le rejet de l'hypothèse nulle (rejet de l'infériorité du traitement expérimental) conduit à conclure à la non-infériorité. Cette non-infériorité s'entend par rapport à une borne ou une limite qui définit la perte maximale d'efficacité admise pour le nouveau traitement par rapport au traitement standard.

La figure ci-dessous présente les scénarios possibles de différence de traitement observée pour des résultats défavorables dans des essais de non-infériorité.



Dans la figure, les barres d'erreur indiquent les intervalles de confiance (IC) à 95% bilatéraux. La ligne pointillée bleue en x = Δ indique la borne de non-infériorité. La région à droite de x = Δ indique la zone d'infériorité. Dans le scénario A, si l'IC se situe entièrement à gauche de zéro, le nouveau traitement est supérieur. Dans les scénarios B et C, si l'IC se situe à gauche de Δ et comprend zéro, le nouveau traitement est non-inférieur, mais la supériorité n'est pas montrée. Dans le scénario D, si l'IC se situe entièrement à gauche de Δ et à droite de zéro, le nouveau traitement est non-inférieur au sens déjà défini, mais également inférieur au sens où une différence de traitement nulle est exclue. Cet IC indique la non-infériorité en ce sens qu'il n'inclut pas le Δ, mais le nouveau traitement est nettement pire que le traitement standard. Cette circonstance déroutante est rare, car elle nécessite un très grand échantillon. Cela peut également résulter d'une borne de non-infériorité trop large. Dans les scénarios E et F, si l'IC comprend Δ et zéro, la différence entre les deux traitements est insignifiante, mais la non-

infériorité n'est pas concluante. Dans le scénario G, si l'IC inclut Δ et se trouve entièrement à droite de zéro, la différence est statistiquement significative, mais le résultat n'est pas concluant quant à une éventuelle infériorité de magnitude Δ ou pire. Cet IC n'est pas concluant dans la mesure où il est toujours plausible que la différence de traitement réelle soit inférieure à Δ, mais que le nouveau traitement soit nettement pire que le traitement standard. Dans le scénario H, si l'IC est entièrement supérieur à Δ, le nouveau traitement est inférieur au traitement de référence **(Piaggio et al., 2006)**.

2.1 Les bornes de non-infériorité

Comme nous l'avons vu dans la figure ci-dessus, la signification de la borne de non-infériorité est importante : il s'agit de la plus grande perte d'efficacité acceptée par rapport au traitement de référence, par exemple, une borne relative de 10 % signifie que l'on considèrera le nouveau traitement comme « non-inférieur » tant que son efficacité ne sera pas inférieure de 10 % à celle du traitement de référence. La borne de non-infériorité est définie en fonction du choix de la mesure pour évaluer l'effet du traitement. Dans de nombreux essais thérapeutiques incluant des patients infectés par le VIH-1, le critère d'évaluation principal est le plus souvent la proportion de patients en succès (ou en échec) virologique à un temps donné, par exemple à la semaine 48 ("Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents.Department of Health and Human Service.," n.d.). Dans de tels essais, la DR, mentionnée plus haut, calculée sous la forme d'une simple différence de proportions, est très souvent utilisée comme mesure de l'effet traitement, et les bornes de non-infériorité autour de 10 % à 12 % ont été largement utilisées car cette borne est admise par la Food and Drug administration **(Mani et al., 2012)**. Dans les essais utilisant une analyse basée sur le rapport de risques, les bornes de non-infériorité sont souvent basées sur la DR en raison du nombre important d'essais précédents

utilisant la mesure DR. Ces bornes sont ensuite transformées en bornes pour une analyse conduisant à l'estimation de rapport de risques.

La borne de non-infériorité et de la mesure de l'effet du traitement sont des paramètres essentiels pour la conduite et l'analyse des essais thérapeutiques. Ces choix impactent en particulier le calcul de la taille d'échantillon à inclure dans l'essai.

3. Les essais de l'ACTG

Selon les informations divulguées dans le protocole de l'essai ACTG 5202, la discussion de l'essai a débuté vers 2005/2006 et a publié les résultats en 2011. Par conséquent, la discussion et les recommandations publiées de cette étude sont basées sur le contexte autour de 2011.

3.1 L'essai ACTG 5202

3.1.1 La problématique et l'objectif de l'essai

Les recommandations thérapeutiques pour le traitement initial du VIH-1 recommandent 2 inhibiteurs nucléosidiques de la transcriptase inverse(INTI) avec 1 inhibiteur non nucléosidique de la transcriptase inverse (INNTI), 1 inhibiteur de protéase (IP) ou 1 inhibiteur d'intégrase (II) ("Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents.Department of Health and Human Service.," n.d.) (Thompson and Aberg, 2010). Abacavir-lamivudine (ABC/3TC) et tenofovir –emtricitabine (TDF/FTC) sont des traitements INTI efficaces à prise unique quotidienne (**DeJesus et al., 2004; Gallant et al., 2006, 2004**). Le INNTI le plus utilisé est efavirenz (EFV), et atazanavir plus ritonavir (ATV/r) est l'un des IP le plus utilisé ("Pa-

nel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents.Department of Health and Human Service.," n.d.). Il existe peu de comparaisons entre les combinaisons à prise unique quotidienne pour le traitement initial du VIH-1. L'essai ACTG (AIDS Clinical Trial Group) 5202 est un essai de non-infériorité dont l'objectif est de comparer l'efficacité et la toxicité de ATV/r versus EFV pour chacune des deux combinaisons ABC/3TC et TDF/FTC, chez les patients naïfs de tout traitement. Un total de 1857 patients éligibles a été inclus dans l'étude entre septembre 2005 et novembre 2007 (**Daar et al., 2011**). Une comparaison pré-spécifiée de ATV/r et EFV avec les INTIs combinés (plan factoriel) a été effectuée car il n'y avait aucune preuve que l'effet de ces deux traitements différait en combinant avec les INTIs (p=0.65) **(Daar et al., 2011)**.

Le critère principal de l'essai était le délai jusqu'à un échec virologique. Ce critère était défini comme une charge virale confirmée >=1000 copies/ml à partir de la 16ème semaine et avant la 24ème semaine ou une valeur de charge virale >= 200 copies/ml à partir de la 24ème semaine. L'équipe de l'essai considérait comme acceptable une différence de probabilité d'échec virologique d'à peu près 6 % à la semaine 48 et de 10 % à la semaine 96 (cf. supplément (Sax et al., 2009)). Pour la semaine 96, les hypothèses de taux d'échec virologique étaient de 32 % dans le bras de référence et de 42 % dans un autre bras de traitement. Ceci conduisait à une différence absolue de 10 % (9.69 %) entre les deux bras de traitement. Ces hypothèses concernant le taux d'échec virologique étaient basées sur un autre essai ACTG utilisant le zidovudine-lamivudine plus l'efavirenz au moment de l'élaboration du protocole **(Gulick et al., 2004)**. Une taille d'échantillon de 450 participants par groupe a été calculée pour fournir une puissance statistique d'environ 90 % permettant de rejeter l'hypothèse nulle d'infériorité, c'est à dire de conclure à la non-infériorité.

3.1.2 Détermination de la borne de non-infériorité

Les bornes de non-infériorité sont définies en premier lieu en termes de différence de risque dans le protocole car elle a été utilisée dans de nombreux essais cliniques antérieurs en raison de sa simplicité **(Fischl et al., 2007; Gulick et al., 2006; Riddler et al., 2008)**.

Le critère principal étant un délai dans l'étude, l'analyse principale utilisait un modèle de Cox pour estimer les différents rapports de risques entre les bras de traitement. Les bornes de non-infériorité devaient donc être transformées en termes de rapport de risques. Pour réaliser cette transformation il faut supposer une distribution pour le temps d'échec virologique, et l'équipe d'ACTG 5202 a supposé une distribution exponentielle. L'hypothèse du modèle exponentiel est celle d'un taux ou d'un risque instantané constant au cours du temps. Si la probabilité cumulée de succès est *S(t)* au temps *t*, le risque instantané est λ, pour une distribution exponentielle nous avons :

$$S(t) = \exp(-\lambda t)\,, \qquad \lambda \in (0, \infty)$$

Le risque instantané λ est donc :

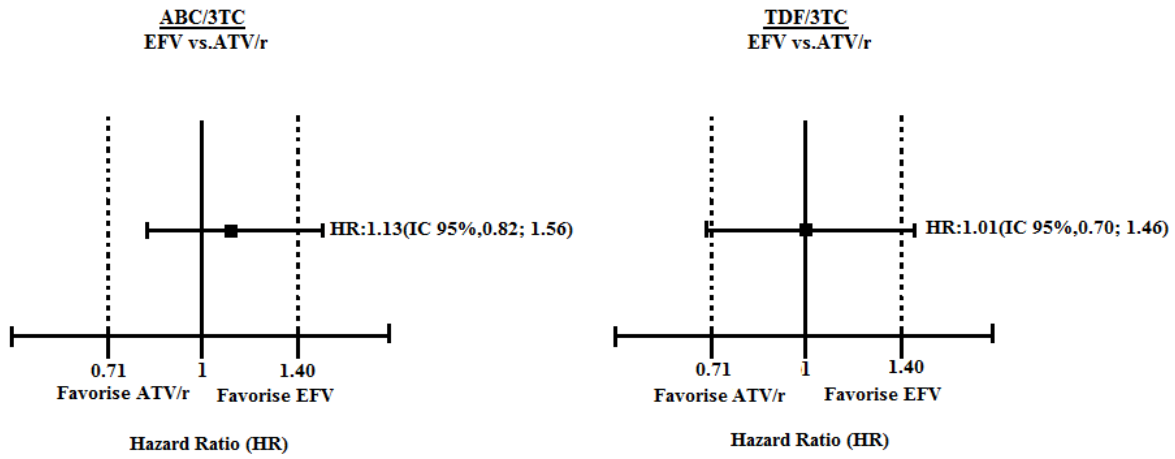
$$\lambda = -\frac{\log(S(t))}{t}$$

Ainsi, si l'on considère une probabilité d'échec de 32 % à la semaine 96, alors la probabilité de succès est de 68 %, et le risque instantané est de 0.004 (-log (0.68)/96). De la même façon pour une probabilité de 42 % le risque instantané est de 0.0056 (-log (0.58)/96). Les bornes d'équivalence pour un rapport de risques sont alors de 0.71 (0.004/0.0056) et 1.40 (0.0056/0.004).

En utilisant les hypothèses faites à la semaine 48 dans le protocole qui correspondent aux taux d'échec de 18 % et 24 %, les risques instantanés sont de 0.004 (-log (0.82)/48) et 0.056 (-log (0.76)/48) respectivement. A noter que l'on obtient les mêmes bornes d'équivalence pour un rapport de risques en utilisant les différentes hypothèses de taux d'échec faites à la semaine 48 et la semaine 96 dans le protocole. En d'autres termes, on constate que les bornes d'équivalence, pour les analyses fondées sur la différence de risque sont distinctes selon une analyse à la semaine 48 ou 96. Au contraire, pour les analyses fondées sur le HR les bornes sont les mêmes quel que soit le temps de l'analyse.

3.1.3 Résultats de l'essai

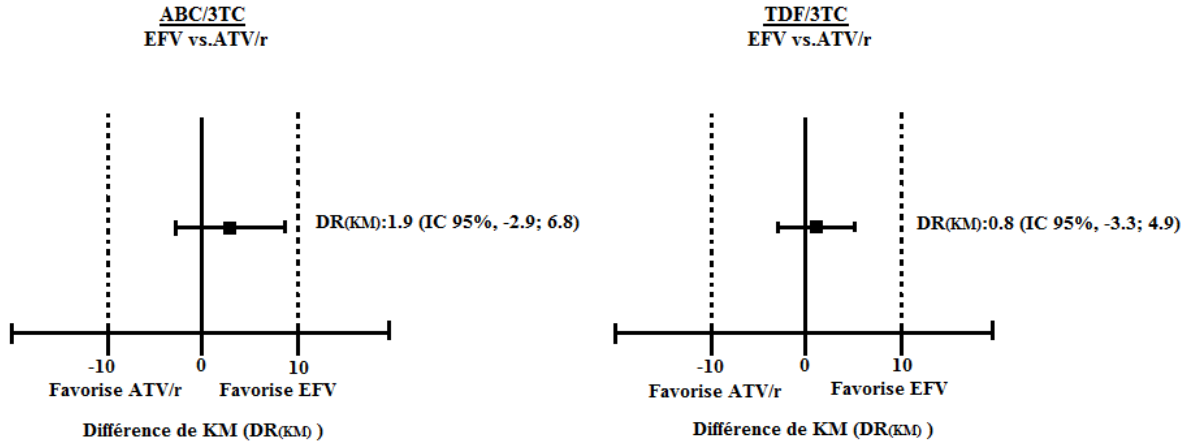
Comme indiqué plus haut l'analyse principale était fondée sur une estimation du rapport de risques par un modèle de Cox sur le suivi complet dans l'essai. La figure ci-dessous représente les estimations du rapport de risques par chaque comparaison.



Chez les patients randomisés recevant ATV/r ou EFV, le rapport de risques, avec EFV comme groupe de référence, était estimé à 1.13 (IC 95 %, 0.82—1.56) chez ceux recevant ABC/3TC et à 1.01 (IC 95 %, 0.70—1.46) chez ceux recevant TDF/FTC. Les deux limites supérieures des intervalles de confiance étant au-delà de la borne supérieure de 1.40, définie

plus haut, les auteurs ne pouvaient pas conclure à une équivalence entre ATV/r et EFV pour chacune des combinaisons.

Une estimation de la probabilité cumulée d'échec était également fournie dans chacun des bras au cours du suivi des patients par des courbes de Kaplan-Meier. La figure ci-dessous représente les différences des estimations de Kaplan-Meier (DR_{KM}) à la semaine 96 par chaque comparaison.



Ainsi, pour les patients recevant ABC/3TC la DR_{KM} était de 1.9 % (IC95 %, -2.9—6.8) et de 0.8 % (IC 95 %, -3.3—4.9) pour ceux recevant TDF/FTC, dans les deux cas en faveur d'EFV. On constate alors que les deux intervalles de confiance sont à l'intérieur des bornes d'équivalence +/- 10% (+/- 9.69 % exactement) définies dans le protocole à la semaine 96. L'analyse fondée sur la DR_{KM} conduit donc à accepter l'équivalence ce que ne fait pas l'analyse fondée sur le rapport de risques.

3.2 L'essai ACTG 5257

Selon les informations divulguées dans le protocole de l'essai ACTG 5257, la discussion de l'essai a débuté vers 2009 et a publié les résultats en 2014. Par conséquent, la discussion et les recommandations publiées de cette étude sont basées sur le contexte autour de 2014.

3.2.1 La problématique et l'objectif de l'essai

Les combinaisons antirétroviraux incluant efavirenz (EFV) sont recommandées comme schéma thérapeutique de première intention (ou première ligne) par l'Organisation Mondiale de la Santé (OMS) ("WHO | Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection," n.d.). Cependant, les femmes qui envisagent une grossesse, les patients présentant une résistance préexistante aux traitements inhibiteurs non nucléosidiques de la transcriptase inverse (INNTI) et ceux qui souffrent de troubles psychiatriques graves ne sont pas considérées comme des bons candidats pour un traitement basé sur l'EFV lorsque d'autres options sont disponibles. Pour mieux comprendre l'efficacité et la toxicité des traitements alternatifs aux combinaisons d'EFV, l'étude d'ACTG 5257 a été réalisée. Elle compare l'efficacité et la toxicité sur 96 semaines de traitement, chez des patients infectés par le VIH et naïfs de tout traitement, des trois combinaisons suivantes : ténofovir/emtricitabine plus atazanavir/ritonavir (TDF/FTC/ATV/r), ténofovir/mtricitabine plus raltégravir (TDF/FTC/RAL), et ténofovir/emtricitabine plus darunavir/ritonavir (TDF/FTD/DRV/r) (**Lennox et al., 2014**). C'est un grand essai de l'équipe ACTG qui suit l'étude ACTG 5202 décrite précédemment dans la partie 3.1.

Le critère principal virologique de l'essai ACTG 5257 était le même que celui défini dans l'essai ACTG 5202, soit une charge virale confirmée >=1000 copies/ml à partir de la 16ème semaine et avant la 24ème semaine ou une valeur de charge virale >= 200 copies/ml à partir de la 24ème semaine. Le critère principal de toxicité était le délai entre la randomisation et l'arrêt de traitement randomisé pour la toxicité. Un critère composite combinant critère virologique et toxicité était aussi analysé.

3.2.2 Détermination de la borne de non-infériorité

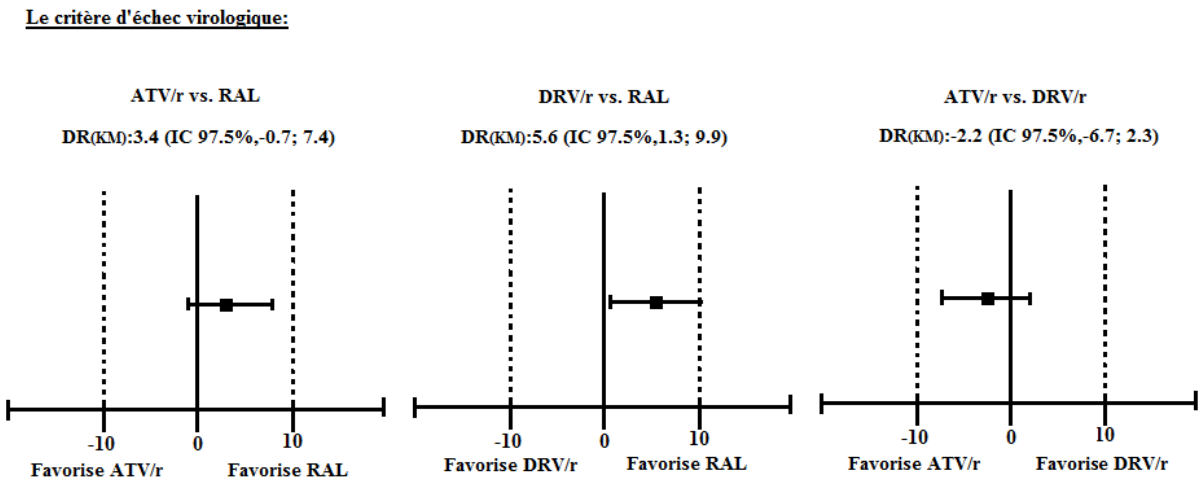
L'analyse principale de cet essai était basée sur des comparaisons de la différence entre les estimations de Kaplan-Meier (DR_{KM}) à la semaine 96 avec une équivalence acceptée si l'intervalle de confiance à 97.5 % était entièrement contenu dans la borne -/+10%. L'hypothèse de taux d'échec était de 25% dans le bras de référence à la semaine 96. Cette valeur est motivée par les données de l'étude CASTLE dans laquelle ATV/r plus TDF / FTC présentait un taux d'échec de 22% à la semaine 48 **(Molina et al., 2010)**. Avec cette hypothèse de taux d'échec, une taille d'échantillon de 600 participants par groupe a été ciblée pour fournir une puissance statistique d'environ 90 % pour montrer une équivalence dans les comparaisons par paires. Une approche similaire a été utilisée pour le critère de la toxicité, la méthode de Gray (**Gray, 1988**) était utilisée pour estimer l'incidence cumulative due à la présence de risques compétitifs (**Lennox et al., 2014**).

Les bornes d'équivalence pour une analyse alternative basée sur le rapport de risques à partir d'un modèle de Cox sont facilement calculées. Comme décrit précédemment dans la partie 3.1, en supposant une distribution exponentielle du délai jusqu'à un échec tel que défini dans le protocole, une probabilité d'échec virologique de 25 % (bras de référence) à la semaine 96 correspond à un risque instantané de 0.003 (-log (1-0.25)/96) et une probabilité de 35 % correspond à un risque instantané de 0.0045 (-log (1-0.35)/96). Pour l'échec virologique, la borne d'équivalence supérieure liée à une analyse basée sur le rapport de risques correspondant à la borne de +10 % pour la DR_{KM} est alors de 1.497 (0.0045/ 0.003). De même, la borne d'équivalence inférieure sur la base d'une probabilité d'échec de 15 % est de 0.56 (0.0017/0.003). Les bornes pour le critère de toxicité en supposant une probabilité d'échec de 10 % dans le bras de référence à la semaine 96 est de -/+7 %. De la même façon, les bornes pour une analyse basée sur le rapport de risques sont de 0.29 et 1.77. Dans le protocole ACTG

5257, il n'y a pas d'hypothèse concernant le niveau de taux d'échec pour le critère composite combinant l'échec virologique ou la toxicité.

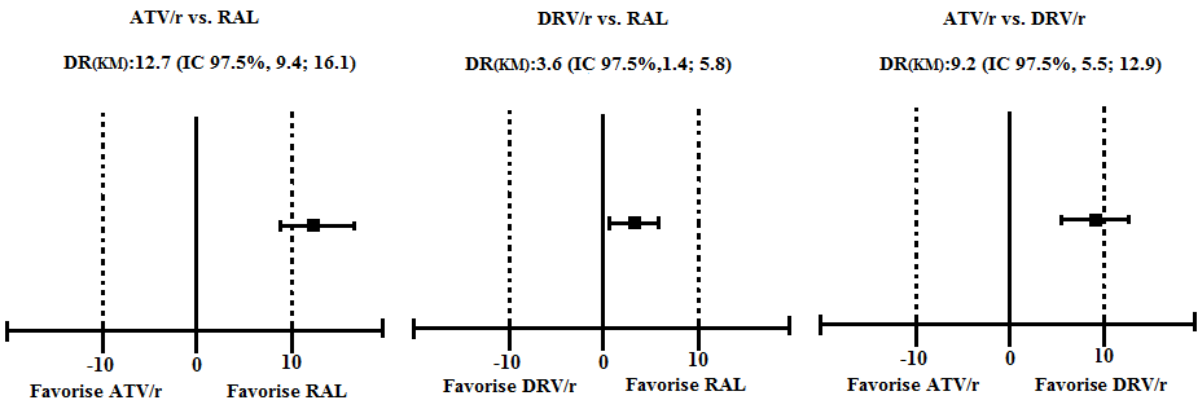
3.2.3 Résultats de l'essai

La figure ci-dessous représente les estimations de la DR_{KM} en fonction du critère utilisé.



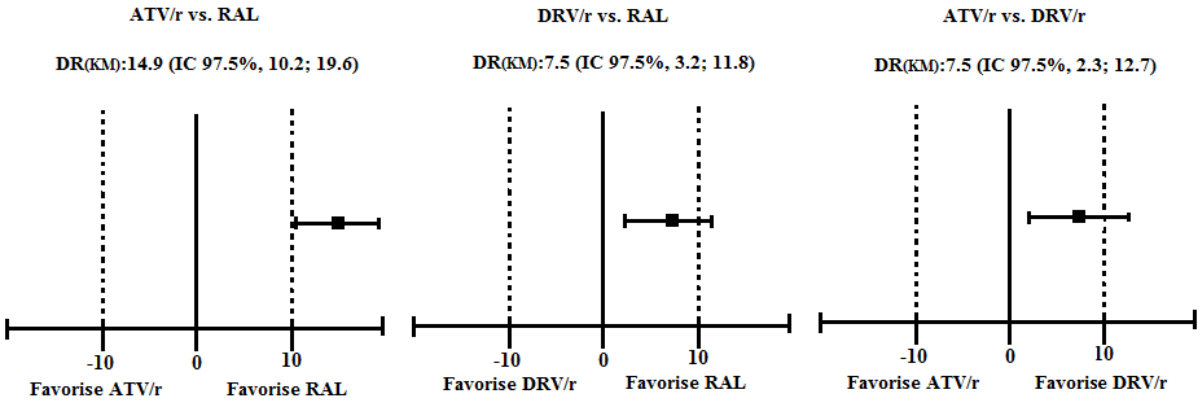
La probabilité cumulée d'échec virologique estimée par la courbe de Kaplan-Meier à la semaine 96 était de 12.6 % dans le groupe recevant TDF/FTC/ATV/r, de 9.0 % dans le groupe TDF/FTC/RAL et de 14.9 % dans le groupe recevant le TDF/FTC/DRV/r. Pour toutes les comparaisons de traitement par paires, les intervalles de confiance à 97.5 % se situaient dans la limite d'équivalence prédéfinie de -/+10 %, démontrant l'équivalence des trois régimes par rapport à ce critère d'échec virologique.

Le critère de tolérance:



Pour le critère principal de tolérance associé à la toxicité, l'équivalence est démontrée entre le groupe recevant TDF/FTC/RAL et TDF/FTC/DRV/r, alors que TDF/FTC/ATV/r a entraîné une augmentation de 12.7 % (97.5 % IC, 9.4 %—16.1 %) par rapport au TDF/FTC/RAL et une augmentation de 9.2 % (97.5 % IC, 5.5 %—12.9 %) par rapport au TDF/FTC/DRV/r au niveau de la discontinuation due à la toxicité du traitement.

Le critère combiné (échec virologique/tolérance):



Pour le critère composite, TDF/FTC/ATV/r était inférieur à TDF/FTC/RAL de 14.9 % (97.5 % IC, 10.2 %—19.6 %) et à TDF/FTC/DRV/r de 7.5 % (97.5 % IC, 2.3 %—12.7 %). TDF/FTC/DRV/r était inférieur à TDF/FTC/RAL de 7.5 % (97.5 % IC, 3.2 %—11.8 %).

CHAPITRE IV - CHOIX DE LA MESURE DE L'EFFET TRAITEMENT LORSQUE LES LIMITES DE NON-INFERIORITE SONT ORIGINALEMENT DEFINIES PAR UNE DIFFERENCE ABSOLUE TRANFORMEE EN UNE DIFFERENCE RELATIVE A INFLUENCE LES RESULTATS DES ESSAIS CLINIQUES

1. Présentation de l'article

Comme nous l'avons vu dans le chapitre précédent, dans les essais cliniques randomisés avec un événement dépendant du temps, différentes mesures de l'effet traitement peuvent être utilisées pour comparer l'efficacité ou la tolérance de deux ou plusieurs traitements. La plupart des essais cliniques randomisés incluant des patients infectés par le VIH sont des essais de non-infériorité et la borne de non-infériorité est définie en fonction du choix de la mesure. Dans ces études, l'effet du traitement est souvent estimé par la DR et le rapport de risques à partir d'un modèle de risques proportionnels de Cox. Comme présenté précédemment dans le chapitre 2, la DR est une mesure de l'effet absolu, et dans les essais cliniques avec un événement dépendant du temps, cette mesure peut être évaluée par la différence des probabilités d'échec cumulées à un instant donné estimé par les courbes de Kaplan-Meier (DR_{KM}) et les bornes de non-infériorité autour de 10-12 % ont été largement utilisées. Dans les essais utilisant une analyse basée sur le rapport de risques, la borne de non-infériorité est souvent basée sur la DR et transformée ultérieurement en rapport de risques sous une hypothèse de la distribution de survie. Le rapport de risques est une mesure de l'effet relatif. Un essai récent (ACTG 5202) incluant des patients infectés par le VIH a montré des conclusions discordantes entre les analyses basées sur ces deux mesures de l'effet traitement.

Cette étude ACTG 5202 nous a motivé à comparer dans le présent article les résultats des analyses basées sur le rapport de risques, DR_{KM}, et Δ-RMST et à expliquer la différence entre les résultats obtenus pour ces trois mesures sur les données de l'essai ACTG 5202. Les données de cet essai ne sont pas publiques donc la reconstruction de données individuelles des

patients de l'ACTG 5202 a tout d'abord été réalisée. En effet, j'ai utilisé d'abord une image au format pdf de la courbe de Kaplan-Meier publiée dans l'article original ((Bormann, n.d.)**logiciel Digitizelt**) ainsi qu'un tableau du nombre de patients à risque à certains temps de l'essai. Ces données ont ensuite été analysées par un algorithme écrit sous R (voir **ANNEXES**) et qui permet de récupérer les données pour chaque individu (**Guyot et al., 2012**).

Comme nous l'avons vu dans chapitre 3 il s'agit d'un essai d'équivalence portant sur la comparaison ATV/r versus EFV pour chacune des deux comparaisons ABC/3TC et TDF/FTC (**Daar et al., 2011**). L'analyse principale était fondée sur une estimation du HR par un modèle de Cox sur le suivi complet dans l'essai. Chez les patients randomisés recevant ATV/r ou EFV, le rapport de risques, avec EFV comme groupe de référence, était estimé à 1.13 (IC 95 %, 0.82—1.56) chez ceux recevant ABC/3TC et à 1.01 (IC 95 %, 0.70—1.46) chez ceux recevant TDF/FTC. Les deux limites supérieures des intervalles de confiance étant au-delà de la borne supérieure de 1.40, définie dans le chapitre 3 ; les auteurs ne pouvaient pas conclure à une équivalence entre ATV/r et EFV pour chacune des combinaisons. La non acceptation de l'équivalence pour la comparaison ATV/r versus EFV dans le groupe ABC/3TC est surprenante car il n'y a aucun effet traitement visible (HR=1.01). Cette conclusion est due à une variance de l'estimateur plus large que prévue.

Les différences des estimations de Kaplan-Meier (DR_{KM}) à la semaine 96 dans les groupes EFV et ATV/r pour chacune des combinaisons ABC/3TC et TDF/FTC étaient aussi fournies. Ainsi, pour les patients recevant ABC/3TC la DR_{KM} était de 1.9 % (IC 95 %, -2.9—6.8) et de 0.8 % (IC 95 %, -3.3—4.9) pour ceux recevant TDF/FTC, dans les deux cas en faveur de EFV. On constate alors que les deux intervalles de confiance sont à l'intérieur des bornes d'équivalence-/+10 % définies dans le protocole à la semaine 96. L'analyse fondée sur la DR_{KM} conduit donc à accepter l'équivalence ce que ne fait pas l'analyse fondée sur le rapport de risques.

J'ai réalisé dans mon article une troisième analyse de ces données basée sur la différence de RMST (Δ-RMST). Dans les données de l'essai ACTG 5202, pour un suivi de 96 semaines, la mesure de RMST était de 87.1 semaines pour le bras EFV/ABC/3TC et de 86.6 semaines pour le bras ATV/TDF/3TC. La différence estimée était de -0.53 semaine (IC 95 %, -3.61 à 2.55). Pour la comparaison EFV/TDF/FTC versus ATV/TDF/FTC la différence était de -1.03 semaine (IC 95 %, -3.58 à 1.51). Pour cet essai, les bornes d'équivalence pour une différence de RMST sont -5.48 à +5.48 semaines.

Les résultats de cette analyse montrent que, pour les deux comparaisons EFV vs ATV les conclusions basées sur la DR_{KM} et la différence de RMST sont identiques et auraient conduit à une équivalence des bras de traitement. A l'inverse les conclusions basées sur le rapport de risques ont conduit à un rejet de l'équivalence dans la publication principale de cet essai. Mon analyse montre donc une différence majeure entre ces analyses sur la base de bornes d'équivalence équivalentes.

La seconde partie de cet article est consacré à expliquer pourquoi ces analyses conduisent à des conclusions différentes. Trois hypothèses pouvant expliquer ces différences ont été étudiées.

- La première hypothèse est que les différences sont dues à la nature même des mesures (mesure relative pour le rapport de risques et mesure absolue pour la DRKM et la différence de RMST).

- La seconde hypothèse est liée à la nature de la distribution des temps d'événements supposée dans le protocole. En effet, pour traduire un pourcentage d'échec à la semaine 96 en une fonction de risque instantané, une distribution exponentielle a été choisie. Cette distribution suppose un risque instantané constant dans le temps. Or les résultats de l'essai montraient un nombre plus important d'échecs virologiques au dé-

but du suivi (jour 0—semaine 48) que lors de la seconde partie du suivi (semaine 48—semaine 96) ce qui indiquerait une violation potentielle de l'hypothèse des risques instantanés constants.

- La troisième hypothèse est liée à la surestimation du taux d'échec à la semaine 96 dans le protocole. En effet, l'équipe d'ACTG 5202 a déjà indiqué qu'un faible taux d'échec virologique était observé (11-17 %) dans leurs données par rapport à l'hypothèse (32 %).

Des simulations ont été réalisées pour évaluer l'impact de ces trois hypothèses. Brièvement, mes résultats ont montré un impact plus important d'une surestimation ou d'une sous-estimation du niveau d'échec sur les résultats d'un essai qu'un choix inapproprié de distribution. Mes simulations indiquaient clairement qu'une analyse basée sur le rapport de risques avait une forte probabilité de ne pas démontrer l'équivalence dans le cas d'une surestimation du taux d'échec dans le protocole par rapport aux deux autres analyses (résultats observés pour l'essai ACTG 5202). A l'inverse, lors d'une sous-estimation du taux d'échec, les analyses basées sur la DR_{KM} et la différence de RMST avaient une forte probabilité de ne pas démontrer l'équivalence. Même si l'effet est moindre, le choix d'une distribution inadéquate augmente ou diminue les probabilités de conclure à l'équivalence selon les mesures choisies. Ainsi un autre résultat indiquait que lorsque la distribution exponentielle est la distribution correspondant aux données observées, c'est l'analyse basée sur la DR_{KM} qui conduit à inclure le plus petit nombre de patients lorsque le taux d'échec est faible (10 %) ou modéré (30 %) à égalité avec le rapport de risques lorsque le taux est élevé (50 %).

Ce travail a fait l'objet d'un article publié dans la revue *Journal of Clinical Epidemiology* en 2018 :

Abulizi X, Flandre P. Choice of treatment-effect measures when nonionferiority margins originally defined in absolute difference translated into relative difference influenced the results of clinical trials. J Clin Epidemiol 2018; 96:63-72. Doi:10.1016/j.jclinepi.2017.12.010.

Journal of Clinical Epidemiology 96 (2018) 63−72

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Choice of treatment-effect measures when noninferiority margins originally defined in absolute difference translated into relative difference influenced the results of clinical trials

Xianmixikemaier Abulizi*, Philippe Flandre

Sorbonne Universités, UPMC Univ Paris 06, INSERM, Institut Pierre Louis d'épidémiologie et de Santé Publique (IPLESP UMRS 1136), F75013, Paris, France

Accepted 11 December 2017; Published online 28 December 2017

Abstract

Objectives: The aim of this study was to investigate the use of three treatment-effect measures in the analysis of randomized trials using a time-to-event endpoint and assess their influence on the results.

Study Design and Setting: A recent equivalence trial showed discordant results with the use of different measures. Different hypotheses may explain such discordant results including a mistaken hypothesized distribution of time to failure and an overestimation of failure rates in the protocol. In a simulation study, we investigate different situations in comparing analyses based hazard ratio (HR), risk difference estimated by Kaplan−Meier curves, and difference in restricted mean survival time. We also compared these three analyses on genuine data from a recent equivalence trial.

Results: In the equivalence trial, two analyses would have concluded equivalence, whereas the original analysis based on HR estimate did not declare equivalence. Results of our simulation study indicate little to moderate differences between the three analyses when the true distribution of time to failure is different to the hypothesized distribution. The main discordant results are found when failure rates have been overestimated or underestimated regardless of the distribution.

Conclusion: With the distributions investigated, differences between analyses based on different measures are much more driven by mistaken hypothesized failure rates than by the shape of the distribution of time to failure. © 2018 Elsevier Inc. All rights reserved.

Keywords: HIV; Effect measures; Time-to-event endpoint; Survival analysis; Simulation study

1. Introduction

Many randomized clinical trials used a time-to-event primary endpoint to compare efficacy or safety of two or more treatments. Clinicians and investigators would like to summarize and represent their data as simply as possible but also want that the effect measure depict the significance of the intervention being studied in the clinically useful way. Different effect measures can be employed to report results of such studies including the relative risk (RR) and the risk difference (RD). The former is a measure of relative effect, whereas the latter is a measure of absolute effect [1−5]. Quite different in nature, there is not a clear consensus between investigators concerning the choice of effect measure. The selection of the most appropriate effect measure has been long debated, but no general guidelines have yet been developed for a single study as well as for meta-analyses [2,6−8]. The choice of the effect measure is of paramount importance when different measures lead to different conclusions for a given trial.

In clinical trials with a time-dependent event, the treatment effect is often estimated by the hazard ratio (HR) from a Cox proportional hazards model. The RD could be assessed by the difference of cumulative failure probabilities at a given time estimated by Kaplan−Meier curves (RD_{KM}). The RD_{KM} provides a clinically interpretable summary of differences between groups at a specific time and does not require a parametric assumption for the outcome distribution in the two groups, but it presents only a comparison at a given time point without reflecting the whole difference [9]. An HR is a simple and generally well-understood measure; however, the magnitude of the HR depends on the length of follow-up and which may

Conflict of interest: All authors declare no conflict of interest.

* Corresponding author. Tel.: +33-(0)1-42-16-42-96; fax: +33-(0)1-42-16-42-67.

E-mail address: xian.abulizi@iplesp.upmc.fr (X. Abulizi).

https://doi.org/10.1016/j.jclinepi.2017.12.010
0895-4356/© 2018 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Incorrect hypotheses of failure rates and distribution of time to failure have an impact on the performance of analyses based on hazard ratio, risk difference estimated by Kaplan−Meier curves (RD_{KM}), and restricted mean survival time difference in trials using a time-to-event primary endpoint.

What this adds to what was known?

- The influence of incorrect hypotheses of failure rates and distribution of time to failure might not be negligible in noninferiority trials.
- In the distributions investigated, differences between analyses based on different measures are much more driven by mistaken hypothesized failure rates than by the shape of the distribution of time to failure.

What is the implication and what should change now?

- Caution is needed in the choice of effect measures in trials when the hypothesis of failure rate or distribution of time to failure is incorrect because these might have an impact on their results.
- Our recommendation is to use the RD_{KM} analysis.

have a built-in selection bias [10]. In addition, HR may also be difficult to interpret when the proportional hazards assumption is violated [9]. The restricted mean survival time (RMST) has recently been proposed to estimate average event-free time from randomization to a suitable follow-up time point *t*. The measure is an estimate of the integral under the survival curve [9,11−13]. In general, the Kaplan−Meier method is used to estimate such a curve. RMST difference (Δ-RMST) provides a stable estimate of the difference and it is valid under any distribution of the time to event in the treatment groups of which the proportional hazard assumption is far from correct. The measure, however, also needs prespecification of the time point of interest so investigators could envisage a temptation to choose a specific time point so as to obtain more significant results [12].

Most randomized controlled trials including HIV-infected patients are noninferiority trials. In a noninferiority trial, null and alternative hypotheses are reversed because the null hypothesis is inferiority of the experimental treatment over the standard, and the alternative is noninferiority of the experimental over the standard treatment [14,15]. The noninferiority is demonstrated if the efficacy of the experimental treatment is not unacceptably worse than that of the standard based on a prestated margin of noninferiority [15,16]. Rejecting the null hypothesis (rejecting the inferiority of the experimental treatment) leads to conclude noninferiority. The noninferiority margin is defined according to the choice of the measure for measuring the treatment effect. In many trials including HIV-1−infected patients, the primary endpoint is most often the proportion of patients in success (or failure) at a given week, for example, week 48 [17]. In such trials, the RD, computed as a simple difference in proportions, is the summary effect measure, and noninferiority margins around 10−12% have been widely used. In trials using an HR analysis, the noninferiority margin is often based on RD due to the large number of previous trials using the RD measure and then subsequently translated into the HR metric. To carry out this translation, an assumption of the survival distribution is required. For simplicity, the exponential distribution considering a constant hazard assumption is employed.

A recent trial including HIV-infected patients showed discordant conclusions between analyses based on these two effect measures. Indeed, results based on RD_{KM} at week 96 would have led to demonstrate equivalence, whereas results based on HR did not conclude to such equivalence [18,19]. Different hypotheses may explain such a discordance including (i) an inherent difference between analyses based on RD_{KM} and HR, (ii) an overestimation of failure rates in the protocol, and (iii) a distribution of failure times that does not follow a simple exponential distribution. In this work, we aim to investigate the reasons that lead to such discordant conclusions in the use of different effect measures. We then compare analyses based on HR, RD_{KM}, and Δ-RMST as summary effect estimates in a simulation study when sample sizes are computed from an HR-based analysis. The equivalence trial introduced above is also analyzed using the effect measure based on Δ-RMST.

2. Materials and methods

2.1. Motivating example

The AIDS Clinical Trials Group (ACTG) 5202 was an equivalence study comparing atazanavir + ritonavir (ATV/r) vs. efavirenz (EFV) for each of the two combinations abacavir/lamivudine (ABC/3TC) and tenofovir/emtricitabine (TDF/FTC) [18]. The primary efficacy endpoint was the time from randomization to virologic failure. The study team considered that a difference in the probability of virologic failure by week 96 of approximately 10% was acceptable for equivalence (see the Supplementary Appendix available with reference [20]). Assuming a probability of virologic failure of 31.9% by week 96 in the reference arm, a rate difference of 10% implies a probability of failure rate by week 96 in the other regimen of 41.6% (an exact difference of 9.7%). In other words, equivalence

margins for RD_{KM} estimates by week 96 are (−9.7% to 9.7%). Assuming that the time to virologic failure follows an exponential distribution at a rate of 31.9% by week 96 corresponds to an HR of 0.004 for the reference arm [−log (1−0.319)/96] and a rate of 41.6% by week 96 to an HR of 0.0056 for the other regimen arm. Then margins for the HR-based analysis are 0.71 (0.004/0.0056) and 1.40 (0.0056/0.004), respectively [18]. A sample size of 450 subjects in each arm was computed to provide approximately a 90% statistical power of rejecting the null hypothesis of inferiority (probability that the 95% two-sided confidence interval [CI] upper bound for the HR will be less than 1.40), that is, concluding noninferiority.

Among the patients randomized in the ATV/r or EFV groups, with EFV as reference, the estimated HR was 1.13 (95% CI, 0.82−1.56) in the group receiving ABC/3TC and 1.01 (95% CI, 0.70−1.46) in the group receiving TDF/FTC [18]. Neither 95% CIs met the prespecified equivalence boundaries of 0.71 and 1.40. Differences in Kaplan−Meier estimates by week 96 were also provided [18]. In particular, RD_{KM} by week 96 was 1.9% (95% CI, −2.9 to 6.8) for patients randomized in the ABC/3TC group and was 0.8% (95% CI, −3.3 to 4.9) for patients randomized in the TDF/FTC group. We noted that both RD_{KM} CIs are within the equivalence margin of ±9.7% hypothesized in the protocol to compute the HR bounds [19]. Thus analyses based on RD_{KM} would have led to equivalence for both comparisons, whereas HR-based analyses did not conclude to equivalence.

At least three hypotheses may explain such discordance. First, the difference is due to the nature of the two measures, one is a relative measure, whereas the other is an absolute measure. Second, the overestimation of the failure rate in the protocol would explain such a difference between the two measures. Indeed, computation of the sample size was based on a failure rate of 32%, whereas the observed rates were 11−17% [18]. Third, the time to virologic failure did not follow an exponential distribution that would lead to different results according to the measure used. Indeed, more virologic failures were observed before week 48 than after week 48 in violation of the constant hazard assumption. For example, in the ABC/3TC/EFV group, 72 virologic failures were observed over the full follow-up, 52 failure occurred before week 48, 11 failures occurred between weeks 48 and 96, and 9 failures occurred beyond week 96. Similar findings are found in the other treatment groups suggesting that the instantaneous risk may be not constant along the follow-up time.

2.2. Alternative analysis and ACTG 5202 data reconstruction

To reanalyze the ACTG 5202 with an RMST difference, we need to compute corresponding bounds for such an analysis. Using the same parameters than those defined in the ACTG 5202 protocol, we simulated data sets to estimate the upper limit that provides a probability that the 95% two-sided CI upper bound for the Δ-RMST is approximately 90% (see below). Based on 5,000 replications, the upper bound for a difference in RMST was 5.48 weeks.

We have reconstructed ACTG 5202 individual-patient-level data by applying an algorithm developed by Guyot et al [21]. Briefly, we used the DigitizeIt software to measure the time and survival probability co-ordinates on the Kaplan−Meier curves [22]. We extracted the numbers of patients at risk and the total numbers of events, when available. The data were then input into an algorithm on the basis of iterative numerical methods to solve the inverted Kaplan−Meier equations [21]. More details can be found in Guyot [21]. We also test the proportional hazards assumption in comparing ATV/r vs. EFV in both combinations ABC/3TC and TDF/FTC using a statistic based on scaled Schoenfeld residuals [23].

Those reconstructed data were then reanalyzed with HR, RD_{KM}, and Δ-RMST as effect measures. The data were also analyzed using two parametric models: exponential and Weibull models. Those parameters were used to generate 1,000 data sets mimicking the ACTG 5202 data. Each simulated data set was analyzed according to the three different measures to compute the proportion of data set leading to conclude noninferiority.

2.3. Simulation study

We conducted two series of simulation studies to compare analyses based on HR, RD_{KM}, and Δ-RMST and to assess their influence on the results of trial. In our simulation study, we consider that an exponential distribution for the time to failure is assumed in the protocol and that the computation of the power and sample size is based on an HR analysis.

2.3.1. Simulation study 1

The first series investigate the performance of the three measures when an exponential or a Weibull distribution is used to generate the data. We investigate low (10%), moderate (30%), and high (50%) failure rates for the reference arm. As in the ACTG 5202, we consider ±10% noninferiority bounds for the RD_{KM} analysis by week 96. As above, bounds for the HR-based analysis are computed from failure rates using corresponding hazards from an exponential distribution. For example, a 10% failure rate corresponds to an HR of 0.0011 in the reference arm, and a 20% failure rate in the other arm corresponds to an HR of 0.00232, providing an upper bound of 2.12 (0.00232/0.0011) for an HR-based analysis. Bounds for analyses based on Δ-RMST are computed by simulation after computation of the sample size (see below). Bounds are provided in the Appendix Table 1.

We determine the sample size corresponding to low, moderate, and high failure rates from an HR-based

analysis. Each sample size is computed to give approximately 90% probability of concluding noninferiority (probability of 90% that the 95% two-sided CI upper limit for the HR estimate will be less than the corresponding upper bound value). The above calculations make the following additional assumptions: (1) accrual takes 96 weeks after a uniform distribution, (2) follow-up continues for 96 weeks after the last entry, (3) time to lost to follow-up from entry follows also an exponential distribution, (4) 10% lost to follow-up at 96 weeks, and (5) equal allocation between the two arms. Sample sizes are provided in the Appendix Table 2.

Once sample sizes were obtained, bounds for RMST-based analyses were computed by simulation. Those bounds correspond to the mean difference in RMST between the two arms at week 96 based on 5,000 replications. For example, assuming a 10% failure rate in the reference arm and then a 20% failure rate in the other arm, a mean of 4.42 weeks was found for the difference in RMST by week 96 (see Appendix Table 1).

Once upper bounds and sample sizes were computed, we then generated a time-to-event outcome for each subject according to two different distributions for the time to failure: exponential and Weibull. For each subject, we generated a random number from a standard uniform distribution: $u \sim U(0,1)$. Half of the subject is in the reference arm, the other half in the other arm. We generated a survival or event time for each subjects as follows: $-\log(u)/\lambda^{(1/p)}$. For the exponential distribution, $P = 1$ and two values of P were investigated to explore monotonic decreasing ($P = 0.75$) or increasing ($P = 1.5$) hazard function from a Weibull distribution (see the Appendix). The hazard λ takes different values to provide the expected probability of failure rate by week 96 in the two arms. Thus data are generated according to the proportional hazards assumption. For a given level of failure rate in the reference arm (low, moderate, or high), data were generated with an increasing probability of failure in the other arm, providing by week 96 a difference in probability from 0% to 7%. For example, with a low failure rate, the probability of failure in the reference arm is 10%, whereas the probability of failure in the other arm varies from 10% to 17%.

2.3.2. Simulation study 2

The second series investigates the impact of an underestimation or overestimation of the probability of failure in a protocol. We consider a design assuming a 20% probability of failure by week 96 in the reference arm. Upper bounds and sample sizes are computed in a similar way than previously (see Appendix Tables 3 and 4). Survival data were also generated as previously with a sample size corresponding to a 20% failure rate by week 96. Investigating an overestimation (underestimation) of failure rates in the protocol, we simulated data providing a 10% (30%) probability of failure rate by week 96 in the reference arm. Data were simulated in a similar way than previously to provide difference in the probability of failure by week 96 between the two arms varying from 0% to 7%. Thus, investigating overestimation, the probability of failure by week 96 in the reference arm was 10% and varies from 10% to 17% in the other arm. Investigating underestimation, the probability of failure by week 96 in the reference arm was 30% and varies from 30% to 37% in the other arm.

In both simulation studies, for each situation we simulated 1,000 data sets. Within each simulated data set, we estimated the HR using a Cox proportional hazards model censoring the data by week 96, RD_{KM} estimate by week 96 and Δ-RMST by week 96 from Kaplan−Meier survival curves between the two treatment groups. Contrary to the ACTG 5202 study, in our simulation studies only the follow-up up to week 96 was used to estimate the HR. Proportions of simulated trials that did not conclude noninferiority (did not reject the null hypothesis) were computed using corresponding upper bounds. These analyses were done using a discrete time corresponding to study evaluations in clinical settings. Discrete times were used to mimic the ACTG 5202 study. The observed time span in the simulated study was divided into a series of fine intervals so that there is a separate interval for each time of failure. All events occurring in the certain interval are fixed at the largest time point of that interval. For example, all events occurring before week 16 had their time to failure at week 16, and for all events occurring between weeks 16 and 24, had their time to failure at week 24, and so on. The discrete times to failure were at week 16 and every 12 weeks thereafter. Computations were performed using the SAS statistical software, version 9.4.

3. Results

Estimates of HR and RD_{KM} based on original or reconstructed data of the ACTG 5202 are similar (Table 1). Both 95% CI of Δ-RMST between EFV and ATV/r for ABC/3TC and TDF/FTC met equivalence bounds. Therefore, all analyses based on RD_{KM} and Δ-RMST would have led to equivalence for both comparisons, whereas analyses based on HR estimates did not meet the prespecified criteria for equivalence. Testing the proportional hazards assumption using a score test based on scaled Schoenfeld residuals leads to $P = 0.45$ in the ABC/3TC combination and $P = 0.20$ in the TDF/3TC combination. Then we cannot reject the proportional hazards assumption for both comparisons.

Parameters of the analysis of the reconstructed ACTG 5202 data with both exponential and Weibull models are given in the Appendix Table 5. Of note, the shape parameter of the Weibull distribution corresponds to a decreasing hazard for all treatment regimens ($P \simeq 0.65$). Results based on 1,000 simulated data using those parameters estimates are displayed in Table 2. In the ABC/3TC group, we found that the probability to conclude equivalence using an

Table 1. Summary of the analyses in the ACTG 5202 study and analyses based on reconstructed data. With efavirenz as the reference in all comparisons

Analysis according to the three treatment-effect measures	Equivalence bound	ABC/3TC		Conclusion	TDF/FTC		Conclusion
		EFV	ATV/r		EFV	ATV/r	
HR (95% CI)	(0.71–1.40)	1.13 (0.82–1.56) 1.14 (0.83–1.56)[a]		Not equivalence	1.01 (0.70–1.46) 1.01 (0.70–1.45)[a]		Not equivalence
K-M Difference at W96 (95% CI)	(−9.69 to 9.69)	1.9 (−2.9 to 6.8) 1.88 (−2.94 to 6.71)[a]		Equivalence	0.8 (−3.3 to 4.9) 0.82 (−3.24 to 4.88)[a]		Equivalence
RMST difference at W96 (95% CI)	(−5.48 to 5.48)	−0.53 (−3.61 to 2.55)[a]		Equivalence	−1.03 (−3.58 to 1.51)[a]		Equivalence

Abbreviations: EFV, efavirenz; ATV/r, atazanavir + ritonavir; ABC/3TC, abacavir/lamivudine; TDF/FTC, tenofovir/emtricitabine; HR, hazard ratio; K-M difference, risk difference in Kaplan–Meier estimates; RMST difference, difference in restricted mean survival time (by week); CI, confidence interval.

[a] Analyses based on reconstructed ACTG 5202 data.

HR-based analysis was around 24–27%, whereas it was greater than 90% with an analysis based on RD_{KM} and Δ-RMST, whatever the survival distribution. Similar findings were found with a piecewise exponential distribution of time to failure (data not shown). Similar results were observed in the TDF/FTC groups although with smaller differences between analyses.

Results of the first series of the simulation study are displayed in Fig. 1. The y-axis provides the probability of not demonstrating noninferiority; that is, not rejecting inferiority. The power of rejecting the null hypothesis of inferiority corresponds to no difference in failure rates between the two arms and corresponds to one minus that probability. In Fig. 1A the distribution of failure time is exponential that corresponds to the distribution both assumed to compute sample sizes and boundaries and to simulate the data. The main differences between the three analyses were obtained for low (panel A) failure rates. For low failure rates, analyses based on RD_{KM} (HR) provided lower (higher) probabilities of not demonstrating noninferiority. For moderate and high failure rates in the reference arm, all analyses provided almost similar results and differences between analyses decreased with increasing failure rates differences. Performances of the three analyses when the time to failure follows a Weibull distribution are shown in Fig. 1B and C. Much larger differences between the three analyses were found compared with previously, especially with low failure rates and an increasing hazard (Fig. 1C). Analyses based on Δ-RMST provided higher (lower) probabilities of not concluding noninferiority with decreasing (increasing) hazard. With low failure rates and an increasing hazard our results show large differences between the three analyses (Fig. 1C panel A). Thus, for a 6% difference in failure rate by week 96, the difference in probability of not demonstrating noninferiority is 41.6% (78.3% vs. 36.7%) between an HR-based analysis and an RMST-based analysis and 26.7% (63.4% vs. 36.7%) between an RD_{KM}-based analysis and an RMST-based analysis. For moderate and high failure rates, HR-based and RD_{KM}-based analyses provided similar results, whereas RMST-based analysis provides a much lower probability of not demonstrating noninferiority.

From Fig. 1A, we can determine the relative sample size required by the three analyses. For example, when $p_1 = p_2 = 10\%$ (low failure rates), the probability of not demonstrating noninferiority is lower with an RD_{KM}-based analysis than with an RMST-based analysis, and the latter is lower than with an HR-based analysis. Therefore, the sample size required for a given statistical power is lower with an RD_{KM}-based analysis than with an RMST-based analysis, and the latter is lower than with an HR-based analysis (Table 3). For all levels of failure rate, a larger sample size is required for the RMST-based analysis.

Overestimation and underestimation of assumed failure rates provide marked differences between the three analyses (Fig. 2). Not only differences between analyses are much important than previously, but the HR-based analysis provided results far from what has been computed in terms of power. Probabilities of not demonstrating noninferiority are much higher with HR-based analyses in the case of an overestimation of failure rates (Fig. 2 left panel), whatever the distribution of time to failure although differences are more pronounced with an increasing hazard. Thus, when there is no difference between the two arms ($p_1 = p_2 = 10\%$), the sample size has been computed to

Table 2. Probability that the 95% two-sided CI upper margin for the estimated HR will be <1.40 for full follow-up, RD_{KM} will be <9.69% and RMST difference will be <5.48 weeks at week 96, with sample size of 450 subjects per arm: results of simulation based on the ACTG 5202 parameters

Distribution model	HR <1.40	RD_{KM} <9.69%	RMST difference <5.48 weeks
ABC/3TC group			
Exponential	24.80%	98.80%	99.90%
Weibull	27.00%	97.00%	99.30%
TDF/FTC group			
Exponential	40.50%	99.10%	99.60%
Weibull	38.60%	99.90%	99.90%

Abbreviations: ABC/3TC, abacavir/lamivudine; TDF/FTC, tenofovir/emtricitabine; HR, hazard ratio; RD_{KM}, risk difference estimated by Kaplan–Meier curves; RMST difference, difference in restricted mean survival time (by week); CI, confidence interval.

ABC/3TC group = EFV + ABC/3 TC vs. ATV/r + ABC/3TC.

TDF/FTC group = EFV + TDF-FTC vs. ATV/r + TDF-FTC.

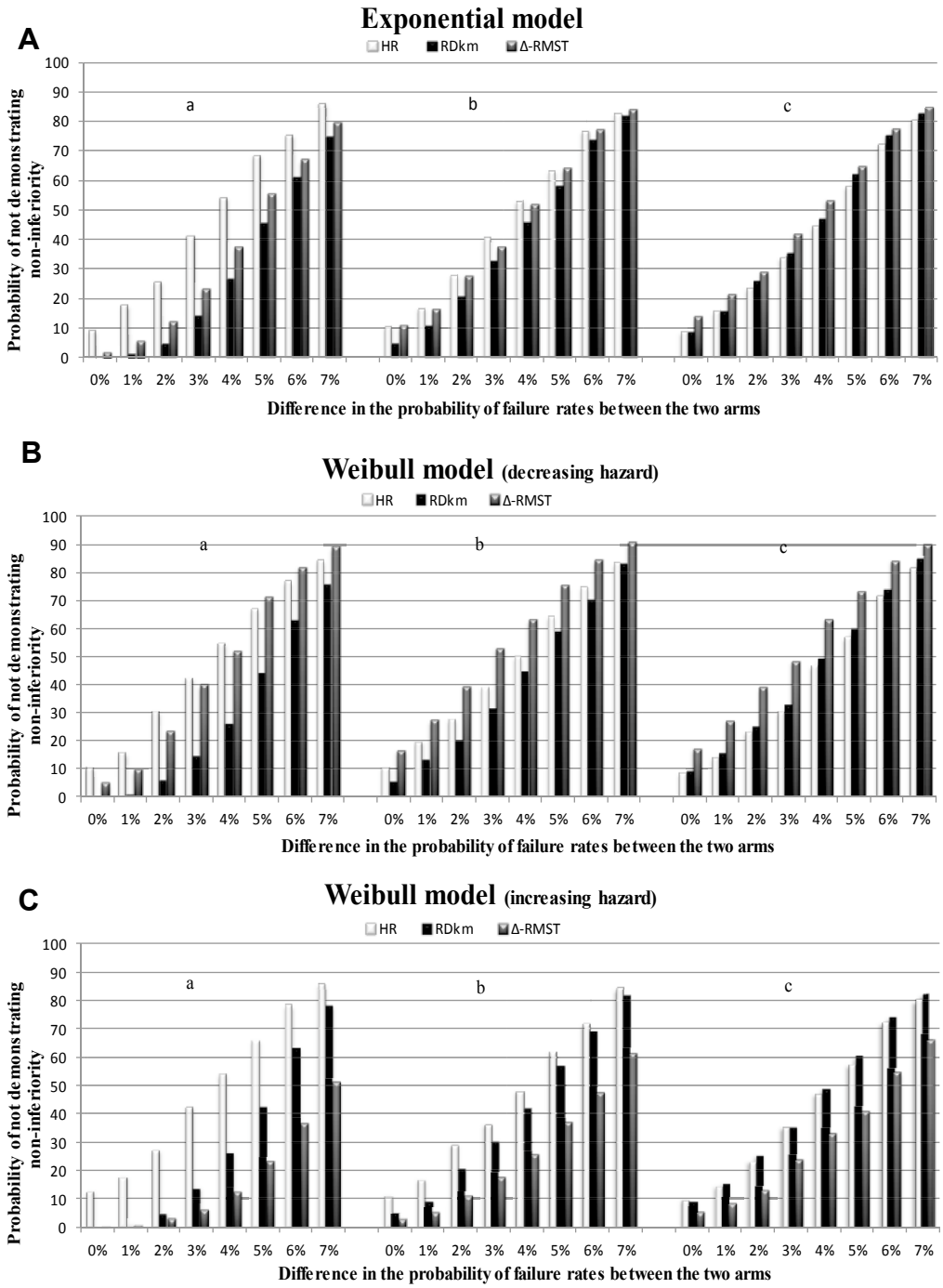


Fig. 1. Probability of not demonstrating noninferiority according to three effect measures used. In the figure, the probability of failure in the reference arm is 10% (panel A), 30% (panel B), and 50% (panel C), respectively. HR, hazard ratio; RD_{KM}, risk difference in Kaplan–Meier estimates (by week); Δ-RMST, difference in restricted mean survival time estimates (by week).

provide a 90% power with an HR-based analysis, whereas the simulation study provides a power of approximately 65% (Fig. 2A). Note that the ACTG 5202 study with an observed decreasing hazard and overestimation of the failure rate in the protocol corresponds to Fig. 2B left panel. Inversely, analyses based on RD_{KM} and Δ-RMST provided higher probabilities of not declaring noninferiority in the case of an underestimation of failure rates. For each difference in the probability of failure rates, we also computed the mean and the standard deviation of the probability of not demonstrating noninferiority (Appendix Table 6). The results indicated that analyses based on RD_{KM} provided much lower standard deviations compared with analyses based on the two other measures.

Table 3. Relative sample size for the three analyses with an assumed exponential distribution for the time to failure according to failure rates

Distribution of time to failure	Failure rates		
	Low	Moderate	High
Exponential	$N_{KM} < N_{RMST} < N_{HR}$	$N_{KM} < N_{HR} \approx N_{RMST}$	$N_{HR} \approx N_{KM} < N_{RMST}$

N_{KM}, N_{HR}, N_{RMST}: simple size required for a given statistical power with a RD_{KM}, HR, RMST-based analyses, respectively.

4. Discussion

The objective of this work was to investigate differences between analyses based on HR, RD estimated by Kaplan–Meier curves and RMST difference in a genuine equivalence trial and in a simulation study. The HR-based analysis was considered as the reference analysis to compute sample sizes. This work was motivated by a recent equivalence trial, leading to different conclusions with the use of HR estimates or RD_{KM} [18,19]. Our results based on reconstructed data showed that an RMST-based analysis would also have led to a different conclusion that the trial's primary analysis is based on HR estimate. Such differences between effect measures need to be highlighted and more widely understood because they potentially explain differences in the results observed between trials using different measures. Of course, conclusions of a trial must be based on the analysis defined in the protocol. Similar findings, however, based on different measures reinforce conclusions of a trial especially when noninferiority margins are described according to different measures.

With the distributions and shape parameters investigated, our results showed that violation of the assumption of constant hazard rate had less impact in the difference observed between the analyses than mistaken levels of failure rates. Nevertheless, the difference due to the distribution shape increase with low failure rates and a stronger increasing hazard. In particular, an overestimation of the failure rates would likely lead to not demonstrating noninferiority with an HR-based analysis even with no or little difference between the two arms. Inversely, the power of demonstrating noninferiority is much larger with an HR-based analysis than with the two other analyses in a situation of underestimation. From the different situations of our simulation studies, it appears that the RD_{KM} analysis provides more stable results to departure from the hypotheses of levels of failure rates or incorrect distribution of the time to virologic failure (see Appendix Table 6). Although other simulation studies can be carried out involving other situations, our results lead to recommend the RD_{KM} analysis. This recommendation is valid only in our setting where a noninferiority margin for an RD measure is translated into an HR measure. Our recommendation does not cover other situations where original margins given for HR or RMST analyses would be transformed into RD measures. Further studies are required to investigate those latter situations. One difficulty, however, is to transform margins from a relative measure into margins for absolute measures.

In general, ACTG used HR-based analyses though RD_{KM} was used in a recent study [19,24]. ACTG trials are very well designed and have main influence in the international guidelines for the use of antiretroviral agents. Most other trials, however, conducted by research groups or pharmaceutical companies used an RD at week 48 as primary efficacy analysis estimated by a simple difference of proportions [17]. Using results of those trials in designing a new study with an HR-based analysis, one needs to translate failure rates at a given time point into hazards assuming a statistical distribution of time to failure. For simplicity, the exponential distribution is widely used. In the ACTG 5202 study, results in the TDF/FTC group was particularly puzzled because there was no difference between EFV and ATV/r as indicated by an HR point estimate of 1.01 (95% CI, 0.70–1.46). Both lower and upper limits of the 95% CI were beyond corresponding bounds. This clearly indicates a much larger variance than expected for the log HR parameter in the proportional hazards model. The larger variance is due to a lower number of events than expected resulting in an overestimation of the hypothesized failure rates. Our simulation study with an overestimation of the assumed failure rates confirmed that finding. The ACTG 5202 team already pointed out that the low rate of virologic failure observed (11–17%) in their data compared with the hypothesized one (32%) could explaine that equivalence was not declare [18]. Overestimation of failure rates or underestimation of success rates, has been previously described in HIV noninferiority trials, leading to overpowered trials [25]. We think that this point also influenced the design of a recent study because the ACTG 5257 trial used a RD_{KM}-based analysis to investigate equivalence between three antiretroviral regimens [24]. Results of that study showed virologic failure rates of 9% to 15%, whereas the assumed failure rate was 25%. Again there is an overestimation of the assumed failure rate although smaller than the one in the ACTG 5202 study. We may suspect that an HR-based analysis in the ACTG 5257 study would likely not conclude to equivalence, whereas equivalence was demonstrated with RD_{KM}-based analyses [24].

Testing the proportional hazards assumption in the ACTG 5202 data does not show violation of that assumption. Consequently, this cannot be used to rule out the HR although it is known that those tests have a low power to detect a violation of the proportional hazards assumption with a high level of censoring. Parametric distribution

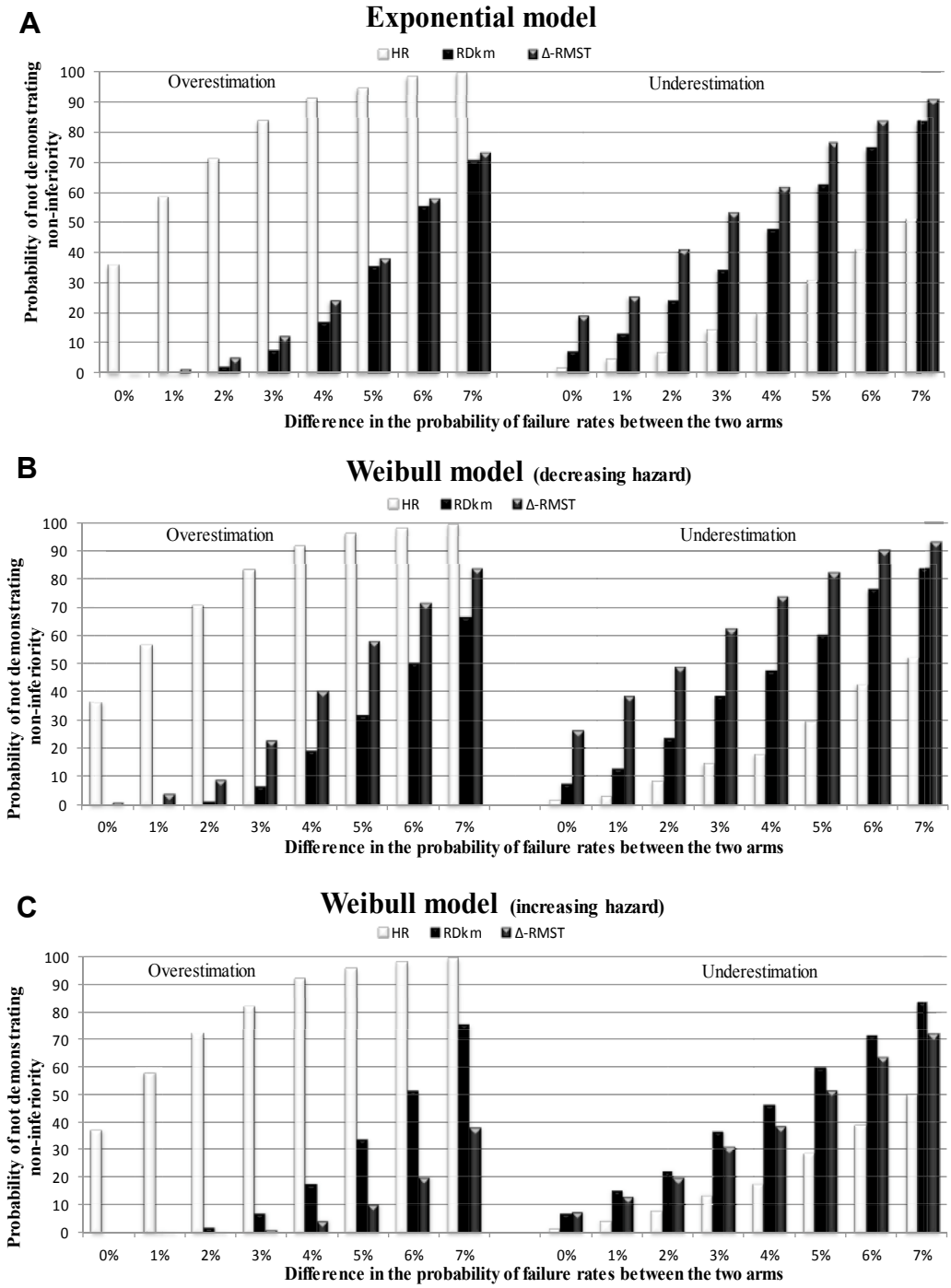


Fig. 2. Probability of not demonstrating noninferiority according to the three effect measures used. In the figure, the sample size was computed with a probability of 20%, whereas the probability of failure is 10% in reference arm and varies from 10% to 17% in the other arm (overestimation panel), and 30% in the reference arm and varies from 30% to 37% in the other arm (underestimation panel). HR, hazard ratio; RD_{KM}, risk difference in Kaplan−Meier estimates (by week); Δ-RMST, difference in restricted mean survival time estimates (by week).

could be envisaged instead of the Kaplan−Meier method to estimate RD. However, we did not explore such methods because Kaplan−Meier curves are widely used in trials with time-to-event primary outcome avoiding parametric assumptions. Of note, in the presence of competing risks, the Kaplan−Meier is not recommended and cumulative incidence functions can be used in this setting [26]. Our results are limited to randomized trials and further investigation should be conducted in the observation cohort context.

Noninferiority trials present specific difficulties in the design, analysis, and interpretation [15]. In superiority

trials, the null hypothesis (H_0) is that there is no difference between treatments and the alternative is that they differ [15]. Investigators want to demonstrate a difference (superiority) between a new experimental treatment and the placebo or standard treatment (rejecting H_0). As explained previously, the power of noninferiority trial is computed with no difference in failure rates between treatment arms although alternative designs have been discussed [16]. Our simulation study indicates that, if failure rates are correctly hypothesized, RD_{KM} analysis requires lower sample sizes compared with other analyses with all levels of failure rates (Table 3). In noninferiority trials, a type I error is the erroneous acceptance of an inferior experimental treatment, that is, we choose the experimental treatment when the standard is actually better, whereas a type II error is the erroneous rejection of a truly noninferior treatment, we choose the standard therapy when the experimental is essentially just as good [14]. The y-axis in the figures corresponds to the type II error, in particular, when there is no difference between failure rates in the two arms. Overestimation of failure rates lead to an inflation of the type II error with an HR-based analysis, whereas underestimation lead to overpowered trials.

Previous studies have also provided empirical evidence regarding the comparison between effects measures of time-to-event outcomes in clinical trials. Seruga et al. [27] compared the difference in survival probabilities at specific time points with the difference in RMST, and provided that difference in RMST is less dependent on the shape of survival curves and might be advantageous in the absolute treatment-effect investigations. Deeks [28] already showed that in 551 systematic reviews of clinical trials with binary outcomes, the RR and odds ratio models are on average more consistent than RD model. Trinquart et al. [29] compared the treatment effects measured by the HR and RMST difference (and ratio) in oncology randomized trials and point out that treatment-effect measures on the basis of RMST analysis yielded more conservative estimates than HR. To our knowledge, our study is the first to compare empirically the probabilities of not demonstrating noninferiority of HR, RD estimated by Kaplan−Meier curves (RD_{KM}), and RMST difference in the context of different failure rate and different time to failure distribution hypotheses.

A limitation of our work is that other distributions of failure could have been investigated, maybe leading to slightly different findings. However, both distributions and shape parameters used in our simulation studies explore common situations of constant, monotonic increasing, and monotonic decreasing hazards. The use of piecewise exponential distribution with monotonic increasing or decreasing hazard lead to similar results than those obtained with the Weibull distribution (data not shown).

5. Conclusions

According to the results of our simulation study, we showed that discordance between analyses based on the three effect measures are much more driven by mistaken hypothesized failure rates than by the shape of the distribution of time to failure. When an HR-based analysis has been designed in the protocol, an overestimation of failure rates would lead to increase the probability of not concluding noninferiority even when there is no difference between the two arms. Inversely, underestimation of assumed failure rates would tend to declare noninferiority even in the presence of a subsequent difference between treatment failure rates. When margins are originally defined for RD measures, analyses based on RD_{KM} could be preferable because results with that measure are less influenced by mistaken levels of failure rates leading to more stable findings.

Acknowledgments

Funding: X.A. was supported by a Ph.D fellowship from the French Ministry of Higher Education and Research.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.12.010.

References

[1] Tripepi G, Jager KJ, Dekker FW, Wanner C, Zoccali C. Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat'. Kidney Int 2007;72:789−91.
[2] Walter S. Choice of effect measure for epidemiological data. J Clin Epidemiol 2000;53:931−9.
[3] Austin PC. Different measures of treatment effect for different research questions. J Clin Epidemiol 2010;63:9−10.
[4] Schechtman E. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat−which of these should we use? Value Health 2002;5:431−6.
[5] Elbarbary M. Understanding and expressing "Risk". J Saudi Heart Assoc 2010;22:159−64.
[6] Akobeng AK. Understanding measures of treatment effect in clinical trials. Arch Dis Child 2005;90:54−6.
[7] Case LD. Interpreting measures of treatment effect in cancer clinical trials. Oncologist 2002;7:181−7.
[8] Papageorgiou SN, Tsiranidou E, Antonoglou GN, Deschner J, Jäger A. Choice of effect measure for meta-analyses of dichotomous outcomes influenced the identified heterogeneity and direction of small-study effects. J Clin Epidemiol 2015;68:534−41.
[9] Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. Ann Intern Med 2015;163:127.
[10] Hernán MA. The hazards of hazard ratios. Epidemiology 2010;21:13−5.
[11] Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the

proportional hazards assumption is in doubt. Stat Med 2011;30: 2409−21.
[12] Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 2013;13.
[13] Zhao L, Tian L, Uno H, Solomon SD, Pfeffer MA, Schindler JS, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. Clin Trials 2012;9:570−7.
[14] Blackwelder WC. "Proving the null hypothesis" in clinical trials. Control Clin Trials 1982;3:345−53.
[15] Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006; 295:1152.
[16] Fleming TR. Current issues in non-inferiority trials. Stat Med 2008; 27:317−32.
[17] Panel on antiretroviral guidelines for adults and adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Department of Health and Human Service. Available at http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf. Accessed November 30, 2017.
[18] Daar ES, Tierney C, Fischl MA, Sax PE, Mollan K, Budhathoki C, et al. Atazanavir plus ritonavir or efavirenz as part of a 3-drug regimen for initial treatment of HIV-1. Ann Intern Med 2011;154: 445−56.
[19] Flandre P. Choice of effect measure in HIV randomized trials. AIDS 2015;29:2057−60.
[20] Sax PE, Tierney C, Collier AC, Fischl MA, Mollan K, Peeples L, et al. Abacavir−lamivudine versus tenofovir−emtricitabine for initial HIV-1 therapy. N Engl J Med 2009;361:2230−40.
[21] Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 2012;12:9.
[22] Bormann I. DigitizeIt - Plot digitizer software. Digitize graphs, charts and math data. Available at http://www.digitizeit.de/. Accessed November 7, 2017.
[23] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515−26.
[24] Lennox JL, Landovitz RJ, Ribaudo HJ, Ofotokun I, Na LH, Godfrey C, et al. Efficacy and tolerability of 3 nonnucleoside reverse transcriptase inhibitor−sparing antiretroviral regimens for treatment-naive volunteers infected with HIV-1: a randomized, controlled equivalence trial. Ann Intern Med 2014;161:461.
[25] Flandre P. Some current issues in the design of HIV noninferiority trials. AIDS 2014;28:1921−9.
[26] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999;94:496.
[27] Seruga B, Pond GR, Hertz PC, Amir E, Ocana A, Tannock IF. Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. Ann Oncol 2012;23:2977−82.
[28] Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21:1575−600.
[29] Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol 2016;34:1813−9.

2.1 Appendices

General definitions: Exponential distribution, Weibull distribution:),,. Note that λ=instantaneous hazard, t=time to event, p=Weibull shape parameter. If P is the probability of failure, then for exponential distribution , for Weibull distribution. Monotonic decreasing hazard (with p=0.75) and monotonic increasing hazard (with p=1.50) were explored for Weibull distribution in the simulations scenarios.

Appendix Table 1. Upper bounds for HR, RD_{KM} and Δ-RMST analyses for the simulation study 1.

Failure rate at week 96 (corresponding hazard)		Upper bounds		
Reference arm	Other arm	HR	RDKM	Δ-RMST*
10% (0.0010975)	20% (0.0023244)	2.12	10%	4.42
30% (0.0037154)	40% (0.0053211)	1.43	10%	4.97
50% (0.00722)	60% (0.0095447)	1.32	10%	5.63

**in weeks computed by simulation*

Appendix Table 2. Sample size required to achieve 90% probability that the 95% two-sided CI upper bound for the HR will be < 2.12 (or 1.43, or 1.32) assuming the virolgoic failure prbability in both arms by week 96 is 10% (or 30%, or 50%).

virologic failure by week 96 in both arms	Bound of HR	Sample size	Pr (95%CI upper bound < HR)
10%	2.12	800 (400/arm)	89.85%
30%	1.43	1100 (550/arm)	89.80%

50%	1.32	1140 (570/arm)	90.10%

Appendix Table 3. Upper bounds for HR, RD_{KM} and Δ-RMST analyses for the simulation study 2.

Failure rate at week 96 (corresponding hazard)		Upper bound		
Reference arm	Other arm	HR	RD_{KM}	Δ-RMST*
20% (0.0023244)	30% (0.0037154)	1.60	10%	4.57

**in weeks computed by simulation*

Appendix Table 4. Sample size required to achieve 90% probability that the 95% two-sided CI upper bound for the HR will be < 1.60 assuming the virologic failure prbability in both arms by week 96 is 20%.

virologic failure by week 96 in both arms	Bound of HR	Sample size	Pr (95%CI upper bound < HR)
20%	1.60	1000 (500/arm)	90.2%

Appendix Table 5.

	ABC/3TC		TDF/FTC	
Distribution model	EFV	ATV/r	EFV	ATV/r
Exponential	λ=0.0011324	λ=0.0012657	λ=0.0008179	λ=0.0008116
Weibull	λ=0.0079839	λ=0.0074349	λ=0.0043321	λ=0.0059642
	p=0.6143261	p=0.6531252	p=0.6755844	p=0.6123324

λ=corresponding instantaneous risk during full follow up.

p=corresponding Weibull shape parameters.

Appendix Table 6.

Difference in the probability of failure rates between the two arms. Mean (standard deviation) of the probability of not demonstrating non-inferiority from results of Figure 2.

Analyses	0%	1%	2%	3%	4%	5%	6%	7%
HR	19.4 (19.0)	31.1 (29.2)	39.7 (34.6)	48.8 (37.5)	55.3 (39.9)	62.9 (35.9)	69.7 (31.3)	75.5 (26.5)
RD_{KM}	3.9 (4.2)	7.3 (7.6)	12.9 (11.8)	22.2 (16.1)	32.8 (16.0)	47.5 (14.9)	63.4 (12.1)	77.3 (7.4)
Δ-RMST	9.3 (11.2)	14.1 (15.4)	21.1 (20.1)	30.8 (23.7)	40.8 (24.9)	53.0 (26.3)	64.7 (24.7)	75.3 (20.0)

2.2 Résultats supplémentaires

Dans le présent article seuls les résultats de la distribution exponentielle et de deux distributions de Weibull sont présentés. Pour conclure que la discordance entre les analyses basées sur les trois mesures d'effet est beaucoup plus influencée par des taux d'échec hypothétiques erronés que par la distribution du temps d'échec inadéquate, des simulations avec une distribution de Weibull avec un risque fortement décroissant (p=0.50) et croissant (p=1.75), et une distribution exponentielle par trois intervalles qui développe 4 scénarios en fonction de différents risques ont également été réalisées pour explorer des situations avec le plus des distributions possibles, et les résultats sont fournis dans les **Figures 1** - **4**.

Les résultats correspondant à la première série d'études de simulation sont présentés dans les **Figures 1** et **2**. Les performances des trois analyses lorsque le temps d'échec suit une distribu-

tion de Weibull avec un risque fortement décroissant (p=0.50) et croissant (p=1.75) sont illustrées dans la **Figure 1**. Dans le cas d'une distribution Weibull avec un risque fortement décroissant p=0.50 (croissant p=1.75), les trois analyses fournissent des résultats similaires à ceux obtenus précédemment avec une distribution de Weibull avec un risque décroissant p=0.75 (croissant p=1.50). Les performances des trois analyses lorsque le temps d'échec suit une distribution exponentielle par trois intervalles sont présentées dans la **Figure 2**. A noter que le temps d'échec suit une distribution avec un risque décroissant (croissant) monotone dans la **Figure 2-A** (**Figure 2-B**). Les résultats de trois analyses fournis sont cohérents avec les résultats de la distribution Weibull avec un risque décroissant (croissant). La **Figure 2-C** présente les résultats de trois analyses avec un risque parabolique, assez intéressant, les résultats sont cohérents avec une distribution avec un risque décroissant ; La **Figure2–D** présente les résultats de trois analyses avec un risque parabolique-inversé, et les résultats sont cohérents avec une distribution avec un risque décroissant lorsque le taux d'échec est faible et modéré.

Les résultats correspondant à la deuxième série d'études de simulation sont présentés dans les **Figures 3** et **4**. Dans le cas de surestimation ou de sous-estimation de taux d'échec, les trois analyses ont fourni des résultats cohérents avec ceux obtenus précédemment avec différentes distributions du temps d'échec en fonction de risques décroissant ou croissant, comme présenté ci-dessus.

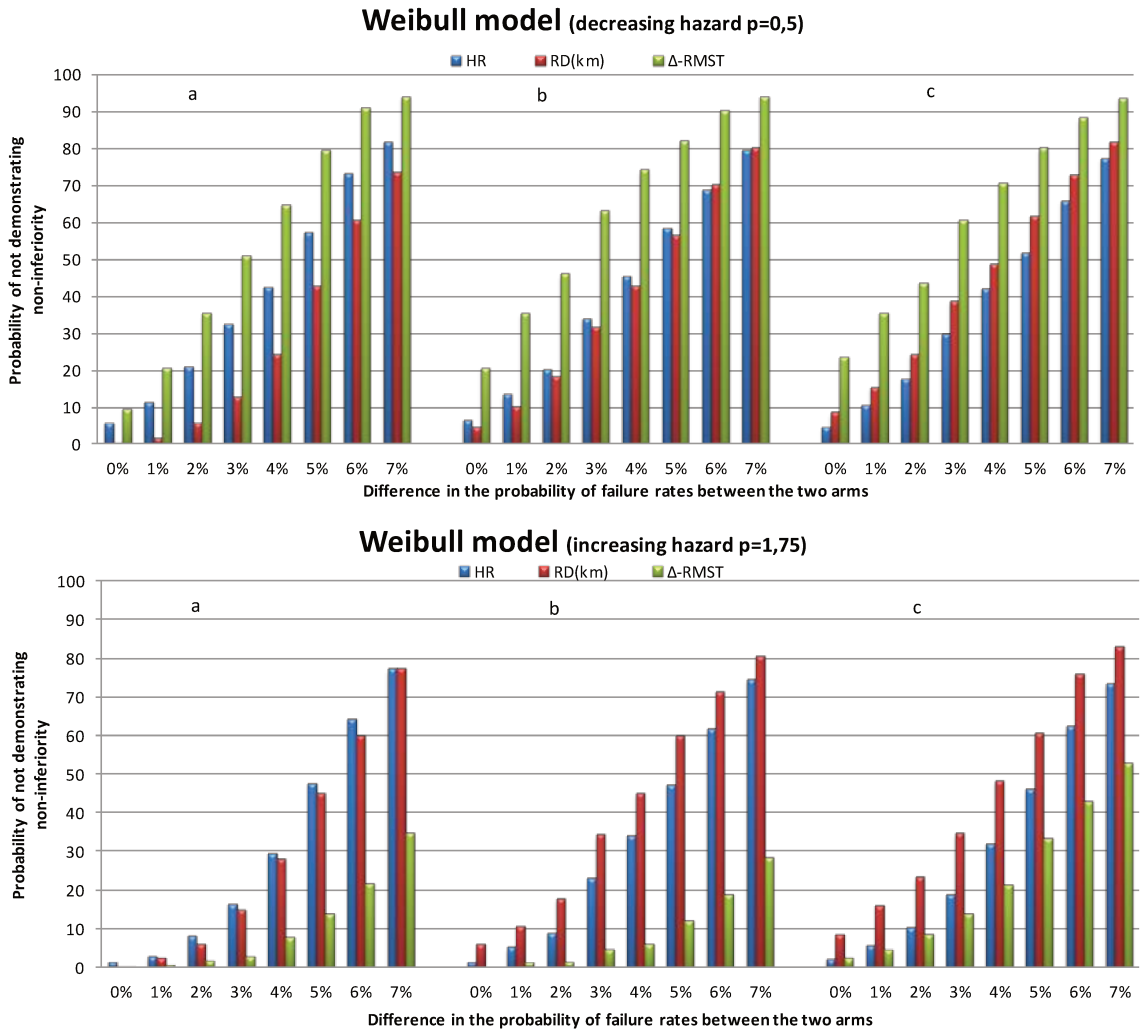


Figure 1. Probabilité de ne pas démontrer la non-infériorité selon les trois mesures d'effet traitement utilisées à partir d'une distribution Weibull avec risques instantanés (λ) monotone décroissante (p=0.5) ou croissante (p=1.75). Dans la figure, la probabilité d'échec dans le bras de référence est respectivement de 10 % (panneau a), 30 % (panneau b) et 50 % (panneau c). HR, le rapport de risques ; RD(km), la différence de risque dans les estimations de Kaplan-Meier ; Δ-RMST, la différence entre les estimations de temps de survie moyen borné (par semaine).

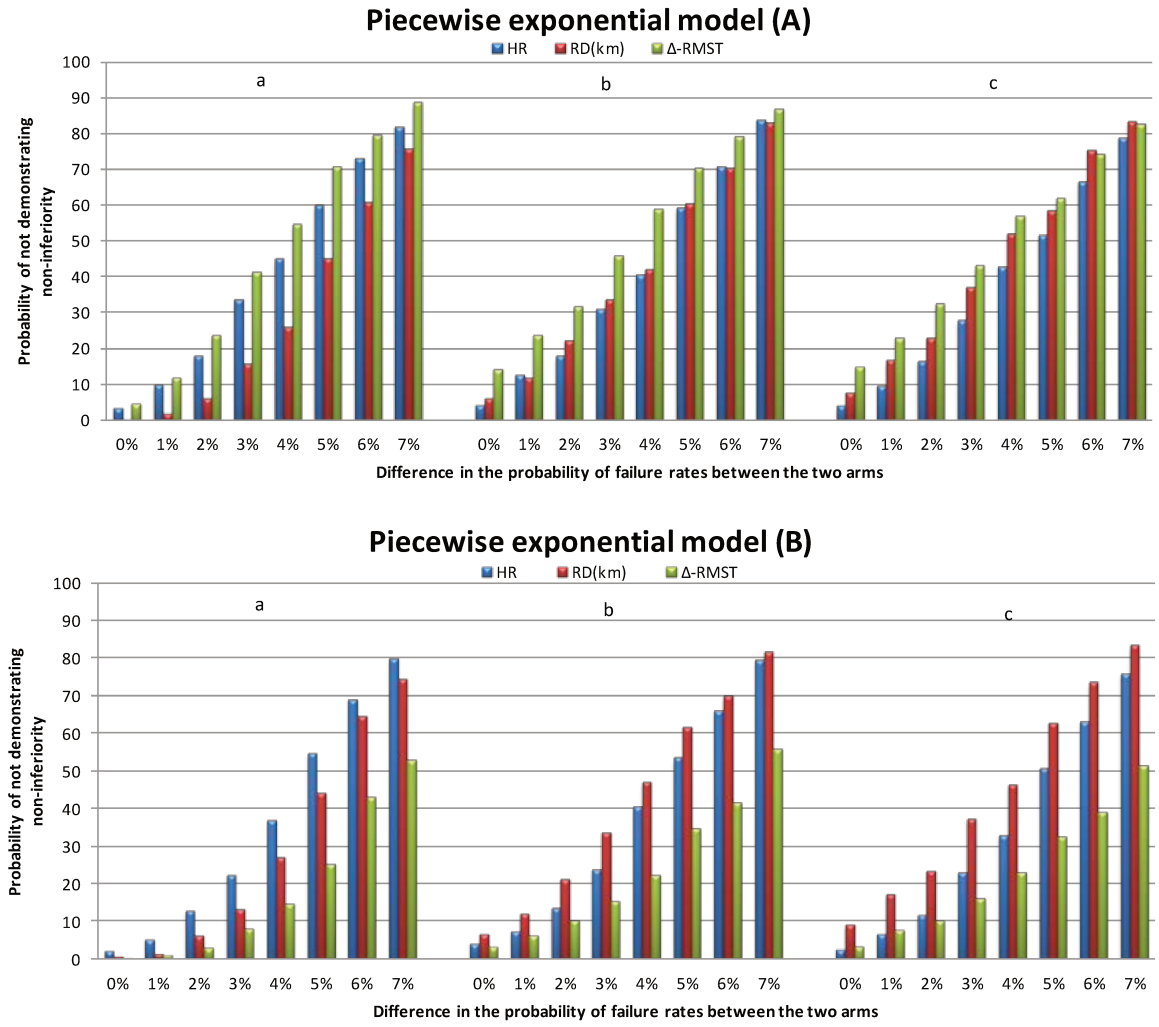

Piecewise exponential model (A)
HR
RD(km)
Δ-RMST
a
b
c
Probability of not demonstrating non-inferiority
0
10
20
30
40
50
60
70
80
90
100
0% 1% 2% 3% 4% 5% 6% 7%
Difference in the probability of failure rates between the two arms
Piecewise exponential model (B)
HR
RD(km)
Δ-RMST
a
b
c
Probability of not demonstrating non-inferiority
0
10
20
30
40
50
60
70
80
90
100
0% 1% 2% 3% 4% 5% 6% 7%
Difference in the probability of failure rates between the two arms

Figure 2.

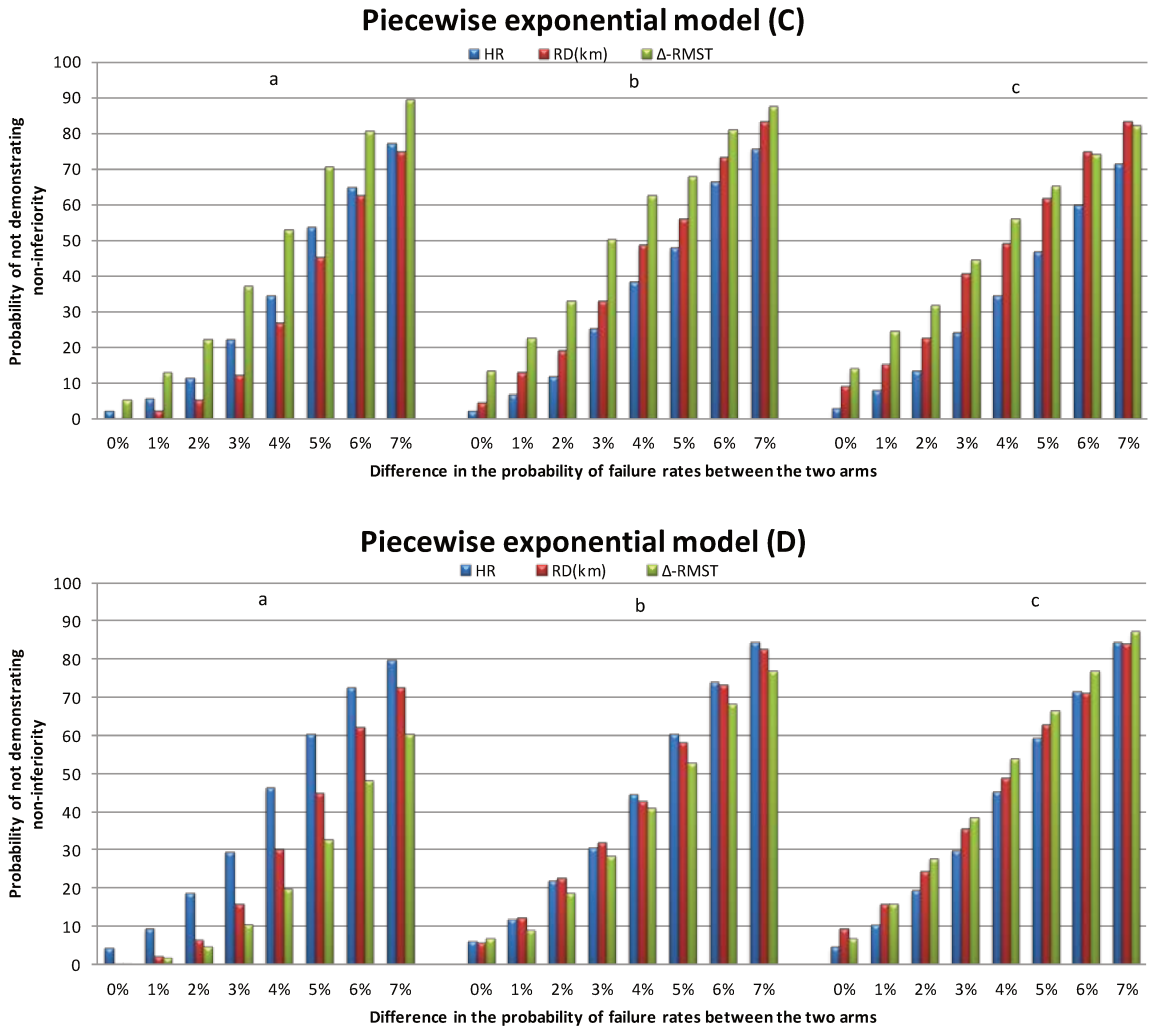


Figure 2 (suite).

Figure 2. Probabilité de ne pas démontrer la non-infériorité selon les trois mesures d'effet traitement utilisées à partir d'une distribution exponentielle par trois intervalles. Le premier intervalle de temps (T1) est à partir du jour 0 jusqu'à la semaine 48 ; le deuxième intervalle de temps (T2) est à partir de la semaine 48 jusqu'à la semaine 96 ; le troisième intervalle de temps (T3) est à partir de la semaine 96 jusqu'à la fin d'étude. Le risque instantané pendant T1 est λ1, pendant T2 est λ2 et pendant T3 est λ3. Si le risque instantané constant pour une distribution exponentielle pour un suivi complet est λ, alors dans la panneau A, λ1=2/3 λ, λ2=λ3=1/3λ; dans la panneau B, λ1=1/3λ, λ2=λ3=2/3λ; dans la panneau C, λ1=2/3λ, λ2=1/3λ, λ3=2/3λ (risque parabolique); dans la panneau D, λ1=1/3 λ, λ2=2/3λ; λ3=1/3λ (risque parabo-

lique-inversé). Dans la figure, la probabilité d'échec dans le bras de référence est respectivement de 10 % (panneau a), 30 % (panneau b) et 50 % (panneau c). HR, le rapport de risques ; RD(km), la différence de risque dans les estimations de Kaplan-Meier ; Δ-RMST, la différence entre les estimations de temps de survie moyen borné (par semaine).

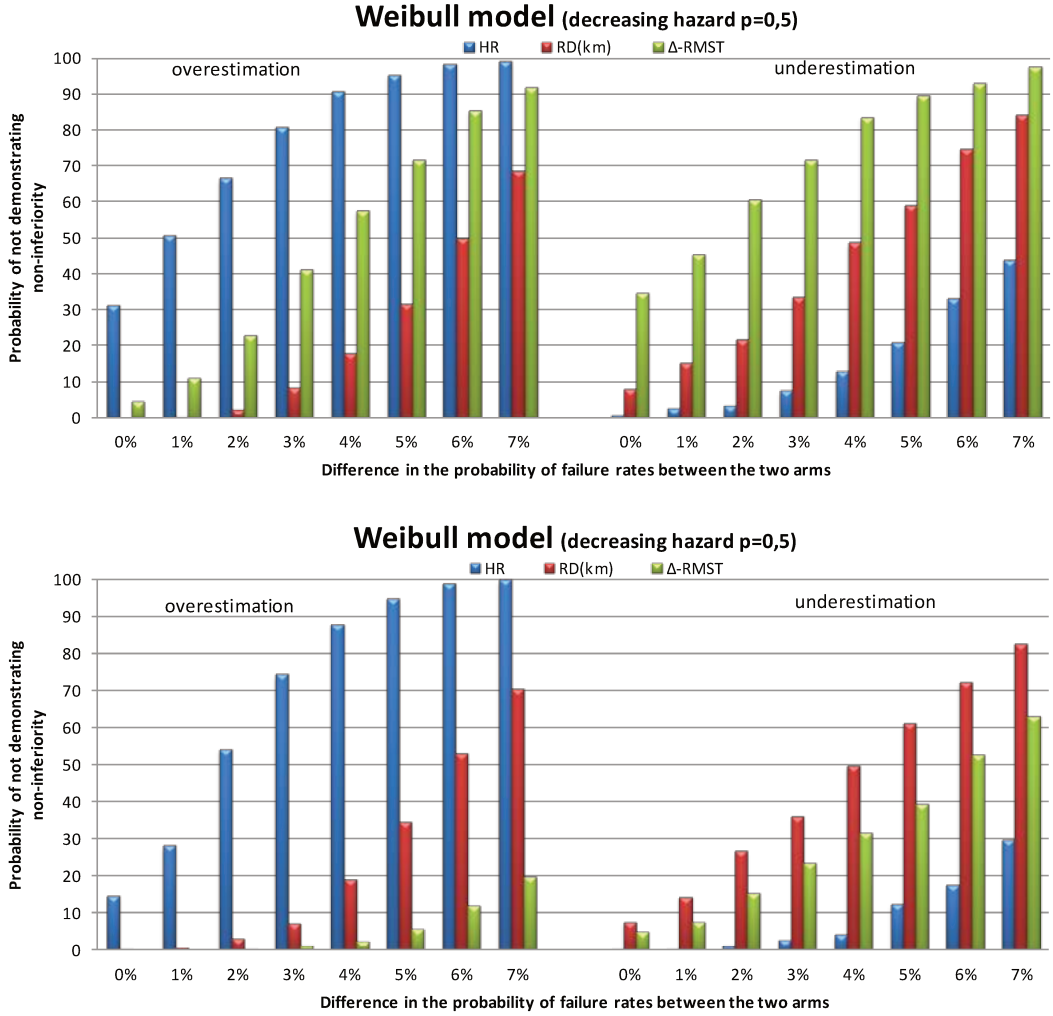
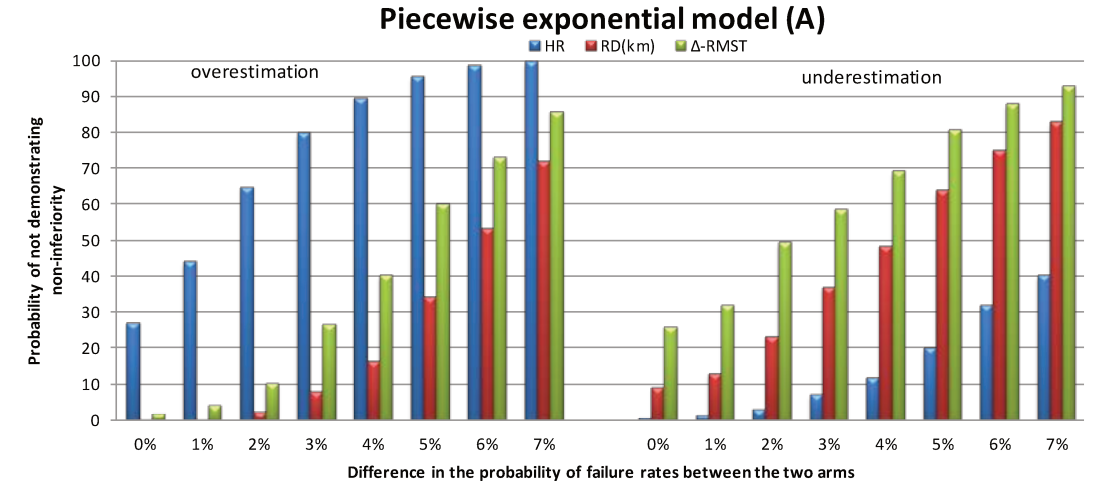
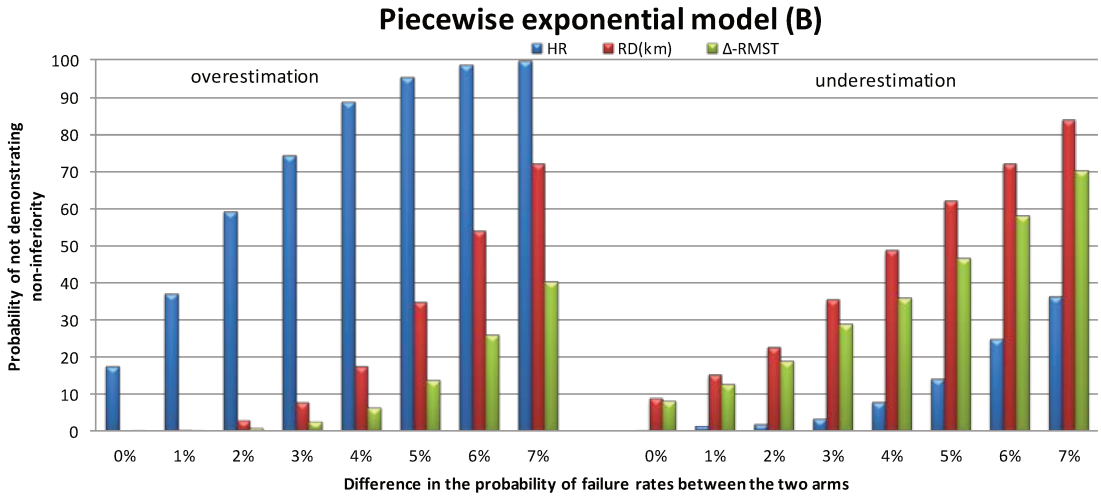


Figure 3. Probabilité de ne pas démontrer la non-infériorité selon les trois mesures de l'effet traitement utilisées à partir d'une distribution Weibull avec risques instantanés (λ) monotone décroissante (p=0.5) ou croissante (p=1.75). Dans la figure, la taille d'échantillon a été calculée avec une probabilité d'échec de 20 %, alors que la probabilité d'échec est de 10 % dans le bras de référence et varie de 10 % à 17 % dans l'autre bras (panneau overestimation), et de 30 % dans le bras de référence, et varie de 30% à 37% dans l'autre bras (panneau underesti-

mation). HR, le rapport de risques ; RD(km), la différence de risque dans les estimations de Kaplan-Meier ; Δ-RMST, la différence entre les estimations de temps de survie moyen borné (par semaine).





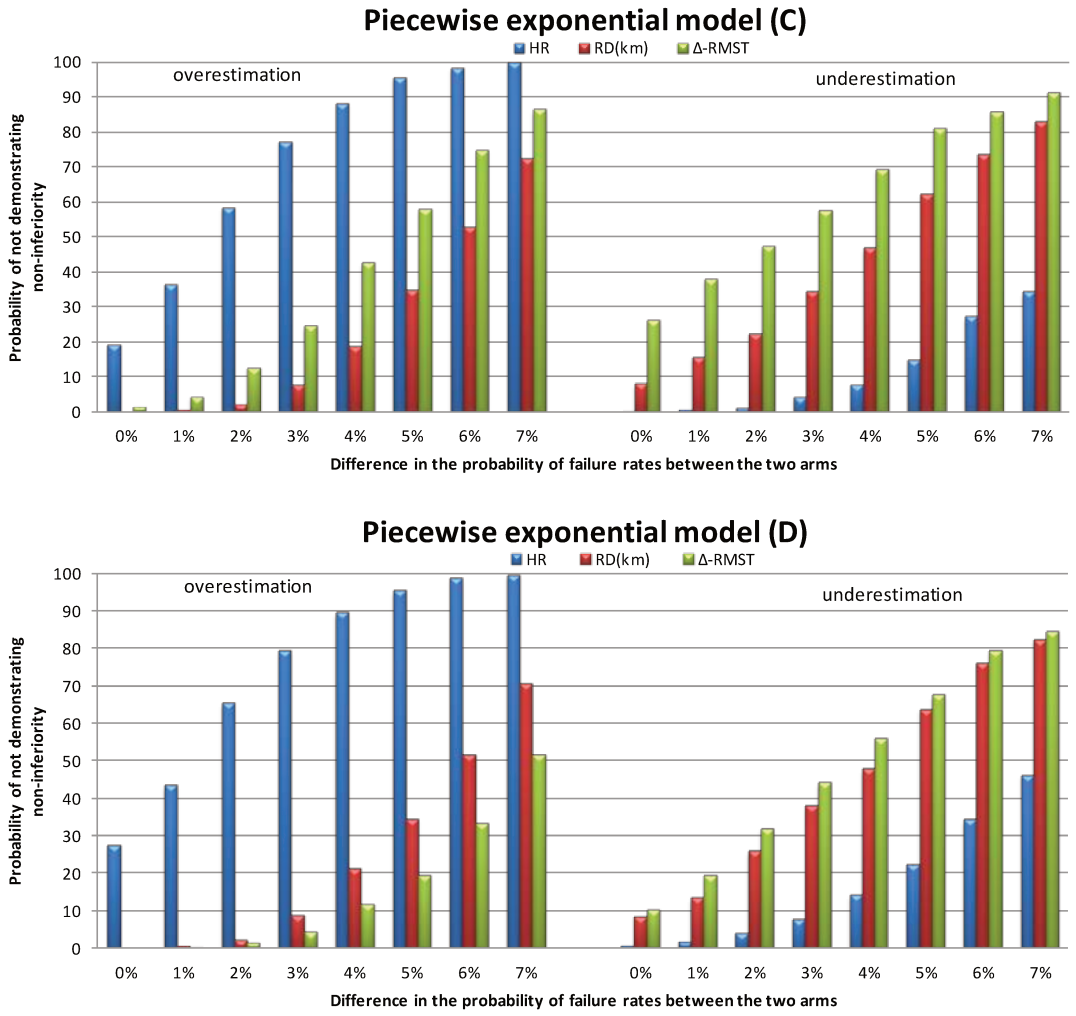


Figure 4. Probabilité de ne pas démontrer la non-infériorité selon les trois mesures de l'effet traitement utilisées à partir d'une distribution exponentielle par trois intervalles. Le premier intervalle de temps (T1) est à partir du jour 0 jusqu'à la semaine 48 ; le deuxième intervalle de temps (T2) est à partir de la semaine 48 jusqu'à la semaine 96 ; le troisième intervalle de temps (T3) est à partir de la semaine 96 jusqu'à la fin d'étude. Le risque instantané pendant T1 est λ1, pendant T2 est λ2 et pendant T3 est λ3. Si le risque instantané constant pour une distribution exponentielle pour un suivi complet est λ, dans le panneau A, λ1=2/3 λ, λ2=λ3=1/3λ ; dans le panneau B, λ1=1/3λ, λ2=λ3=2/3λ; dans le panneau C, λ1=2/3λ, λ2=1/3λ, λ3=2/3λ (risque parabolique); dans le panneau D, λ1=1/3 λ, λ2=2/3λ; λ3=1/3λ (risque parabolique-inversé). Dans la figure, la taille d'échantillon a été calculée avec une

probabilité d'échec de 20 %, alors que la probabilité d'échec est de 10 % dans le bras de référence et varie de 10 % à 17 % dans l'autre bras (panneau overestimation), et de 30 % dans le bras de référence, et varie de 30 % à 37 % dans l'autre bras (panneau underestimation). HR, le rapport de risques ; RD(km), la différence de risque dans les estimations de Kaplan-Meier ; Δ-RMST, la différence entre les estimations de temps de survie moyen borné (par semaine).

CHAPITRE V - L'UTILISATION DU TEMPS DE SURVIE MOYEN BORNE (RESTRICTED MEAN SURVIVAL TIME, RMST) COMME MESURE DE L'EFFET TRAITEMENT DANS LES ESSAIS CLINIQUES DU VIH/SIDA: REANALYSE DE L'ESSAI ACTG5257

1. Présentation de l'article

Comme nous l'avons vu dans le chapitre 3, dans les essais thérapeutiques sur les patients infectés par le VIH, le principal critère d'évaluation de l'efficacité est souvent la proportion d'avoir une charge virale inférieure à un seuil de détection (charge virale < 50 copies/mL) à la semaine X ou un délai jusqu'à l'échec virologique après X semaines de traitement. Pour les essais utilisant ce dernier critère, l'analyse s'effectue en estimant un rapport de risques à l'aide du modèle de Cox, ou une différence entre deux estimations de la proportion cumulée d'échec aux différents temps de suivi fournie par la courbe de Kaplan-Meier (DR_{KM}). Nous avons discuté dans mon premier article du fait que l'utilisation du rapport de risques et de la DR_{KM} conduit à des conclusions différentes lorsque le taux d'échec est sur- ou sous-estimé dans le protocole et lorsque la distribution des événements est différente de celle supposée dans le protocole. La robustesse et l'interprétation du rapport de risques estimé par le modèle de Cox sont conditionnées à l'hypothèse des risques proportionnels durant la période étudiée, or les risques non-proportionnels sont détectés plus souvent dans de nombreux essais cliniques impliquant la recherche d'une autre mesure d'effet. L'analyse basée sur la mesure de RMST est considérée comme une mesure d'effet robuste et n'implique pas l'hypothèse de risques proportionnels.

Dans le cadre de cette étude, j'ai appliqué les analyses basées sur la DR_{KM}, le rapport de risques et Δ-RMST aux données de l'essai ACTG 5257 et j'ai mis en pratique d'une analyse des essais du VIH par la mesure de RMST. A la différence de mon premier article, en collaboration avec les ACTG, les données de l'ACTG 5257 ont été mises à ma disposition par Hea-

ther Ribaudo la statisticienne de l'essai. Comme nous l'avons vu dans le chapitre 3, cet essai comparait l'efficacité et la toxicité, chez des patients infectés par le VIH et naïfs de tout traitement, des trois combinaisons suivantes: FTC/TDF/ATV/r, FTC/TDF/RAL et FTC/TDF/DRV/r. Le critère virologique principal était le même que celui défini dans l'essai ACTG 5202. Un critère de toxicité et un critère composite combinant critère virologique et toxicité étaient aussi analysés. L'équivalence entre les traitements 2 à 2 était déterminée par l'intervalle de confiance de la DR_{KM} pour chacun des 3 critères (virologique, toxicité et combiné) à la semaine 96. L'équivalence était acceptée si l'intervalle de confiance était entièrement dans la borne -/+10 %. La potentielle discordance entre les analyses basées sur le rapport de risques et sur la DR_{KM} est disponible dans l'article d'ACTG 5257 même si celle-ci n'est pas mise en évidence dans leurs résultats. On peut penser que cette discordance a été largement discutée par les statisticiens et les cliniciens des ACTG car l'analyse choisie pour l'ACTG 5257 était basée sur la DR_{KM} et non plus sur un rapport de risques.

La première partie de mon article s'intéresse à la comparaison des types d'analyses citées plus haut et aux raisons éventuelles des différences observées. La seconde partie se concentre sur la mise en pratique d'une analyse basée sur la mesure de RMST et de ses conséquences dans le cadre des risques proportionnels et non-proportionnels. Les résultats entre les 3 analyses ont encore montré des discordances principales entre les DR_{KM}, le rapport de risques et la différence de RMST. L'hypothèse de risques proportionnels a été testée par des modèles paramétriques et notre analyse montrait que les risques n'étaient pas proportionnels dans l'essai ACTG 5257 pour un bon nombre de comparaisons. Les différences obtenues entre les conclusions basées sur le rapport de risques et sur la DR_{KM} sont sans doute dues aux risques non-proportionnels.

La mesure de RMST propose un résumé simple et cliniquement compréhensible de l'effet traitement et elle ne nécessite pas l'hypothèse des risques proportionnels. Ce dernier avantage

a souvent été mis en avant pour justifier son utilisation lorsque le modèle de Cox ne peut pas être employé. Néanmoins, dans les essais cliniques les risques sont considérés comme proportionnels car il est difficile d'envisager d'autres situations permettant un calcul simple du nombre de sujets nécessaires. De la même façon, la distribution des temps d'événements est considérée comme exponentielle car il est difficile d'envisager d'autres distributions. Dans la seconde partie de l'article, en utilisant la différence de RMST comme mesure de l'effet traitement, j'ai étudié l'impact d'une distribution de risque instantané non constant (distribution de Weibull) pour les temps d'événement dans des situations de risques proportionnels et non-proportionnels. Les deux situations correspondant à un risque instantané qui augmente ou diminue au cours du temps ont été étudiées. Les résultats ont montré que la puissance pour démontrer l'équivalence et la probabilité de démontrer à tort l'équivalence (taux de faux positifs) entre les deux traitements étaient très largement modifiées dans le cas de risques non-proportionnels.

Une grande difficulté de la mise en place d'une mesure de l'effet traitement basée sur une différence de RMST est la détermination de bornes de non-infériorité acceptables par les cliniciens. Dans cet essai les gains en termes de RMST peuvent paraître faibles ou modestes. Pour le critère virologique les différences de RMST varient entre 1 et 4 semaines pour des suivis de 96 semaines. Pour les deux autres critères, toxicité et combiné, les gains sont de 2 à 19 semaines. Une analyse rétrospective sur d'autres essais est aussi discutée dans mon article et montre aussi des gains faibles.

Ce travail a fait l'objet d'un article publié dans la revue *Journal of Acquired Immune Déficience Syndromes* en 2019.

Abulizi X, Ribaudo HJ, Flandre P. The use of the Restricted Mean Survival Time as a treatment measure in HIV/AIDS clinical trial: reanalysis of the ACTG A5257 trial. *J Acquir Immune Defic Syndr*. February 2019. doi:10.1097/QAI.0000000000000197

The Use of the Restricted Mean Survival Time as a Treatment Measure in HIV/AIDS Clinical Trial: Reanalysis of the ACTG A5257 Trial

Xianmixikemaier Abulizi, MPH,[a] Heather J. Ribaudo, PhD,[b] and Philippe Flandre, PhD[a]

Background: The restricted mean survival time (RMST) measures have not been used as primary measure of efficacy in HIV/AIDS clinical trials. In this work, we aim to compare analysis based on the difference in RMST (Δ-RMST) measure and 2 other treatment-effect measures in a recent HIV equivalence trial, and to investigate the performance and characteristics of Δ-RMST–based analysis in a simulation study.

Setting and Methods: We reanalyzed a recent HIV equivalence trial (ACTG A5257 trial) with hazard ratio and Δ-RMST, and then compared the results with the original analysis based on risk difference estimated by Kaplan–Meier curves (RD_{KM}). In a simulation study, we investigated the performance and operating characteristics of Δ-RMST–based analysis in the setting of nonproportional hazards (PH) ratio.

Results: In the ACTG A5257 trial, analyses based on Δ-RMST globally led to similar conclusions as the published finding based on RD_{KM}. By contrast, analyses based on hazard ratio provided some discordant equivalence conclusions compared both with the initial analyses based on RD_{KM} and the Δ-RMST. Results of simulation study indicate that the violation of the PH assumption has an impact on Δ-RMST–based analysis regarding the probability of declaring equivalence.

Conclusions: Although the RMST-based analysis is an alternative measure of efficacy in HIV/AIDS, clinical trials such an analysis can be strongly impacted by departures from the PH assumption.

Key Words: HIV, restricted mean survival time, equivalence trials, simulation study

(*J Acquir Immune Defic Syndr* 2019;81:44–51)

INTRODUCTION

In HIV treatment clinical trials, the primary efficacy estimand is often either the proportion of HIV-1 RNA below 50 copies/mL at week 48 or a time-to-event outcome measuring treatment/virologic failure. For the latter endpoint, Kaplan–Meier survival curves and Cox proportional hazards (PH) model are the statistical methods used to analyze such data. With the Kaplan–Meier method, the efficacy could be measured by comparing survival curves at a prespecified time point. Such a risk difference in cumulative probabilities at a given time, estimated by the difference in Kaplan–Meier (RD_{KM}), corresponds to the difference in proportions in a clinical trial using a time-independent outcome. Using a Cox PH model, efficacy is quantified by the hazard ratio (HR) estimate.[1] Both measures have been widely used and discussed for superiority and noninferiority trials.[2] It is well known from epidemiological data that measures of treatment effect may lead to different conclusions and interpretations.[3] A recent example demonstrated that in a large clinical trial, the use of RD_{KM} and HR also leads to different conclusion.[4,5] In particular, an underestimation or overestimation of the hypothesized failure rates in the study design has an important impact on the probability of demonstrating noninferiority.[4]

The use of the Cox model implies the well-known PH assumption.[6] However, increasing non-PH have been often detected in randomized trials implying the search for other effect measures.[7] An alternative outcome measure at a specific time has recently been proposed, the restricted mean survival time (RMST).[8,9] The RMST is simply the population average of the amount of event-free survival time experienced during the period of follow-up. In each randomized group, the RMST can be easily estimated by the area under the KM curve up to a given time point. For example, suppose in the group receiving treatment A, the RMST estimated is 80 weeks for a prespecified time point of 96 weeks. One expects a typical patient treated with treatment A to be event-free for 80 weeks of the 96 weeks of follow-up. The treatment effect can be measured either as difference (Δ-RMST) or a ratio of RMST.

Received for publication August 31, 2018; accepted January 3, 2019.

From the [a]INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique (IPLESP) équipe Epidémiologie clinique des maladies virales chroniques, Paris, France; and [b]Center for Biostatistics in AIDS Research, Harvard T.H. Chan School of Public Health, Boston, MA.

X.A. is supported by a PhD fellowship from the French Ministry of Education and Research. H.J.R. is supported by National Institute of Allergy and Infectious Diseases (NIAID) grantUM1AI68634 for the ACTG Statistics and Data Management Center. A5257 was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number UM1 AI068634, UM1 AI068636, and UM1 AI106701. Bristol-Myers Squibb, Merck Inc., Janssen Therapeutics, and Gilead Sciences provided the study medications (atazanavir, raltegravir, darunavir, and tenofovir-emtricitabine, respectively). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors have no funding or conflicts of interest to disclose.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.jaids.com).

Correspondence to: Philippe Flandre, PhD, INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique (IPLESP), CS 81393 - 75646, Cedex 13 Paris, France (e-mail: philippe.flandre@iplesp.upmc.fr).

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

The RMST-based analysis is considered as a robust and clinically interpretable summary measure that does not rely on the PH assumption and is estimable even under heavy censoring.[2,8–10] RMST measures have been mainly used in oncology including many cancer types.[7] To our knowledge, RMST has not been used as primary measure of efficacy in HIV/AIDS clinical trials. In this work, we aim to compare analysis based on RD_{KM}, HR, and Δ-RMST as a summary effect estimates in the ACTG 5257 study. We further investigate the performance and operating characteristics of Δ-RMST–based analysis in the context of proportional and non-PH in a limited simulation study. We also discussed the use of such measure in the context of HIV trials.

MATERIALS AND METHODS

ACTG 5257 Study and Alternative Analysis

The AIDS Clinical Trials Group (ACTG) 5257 study compared the virologic efficacy and tolerability over 96 weeks of combination regimens that include tenofovir/emtricitabine (FTC/TDF) plus atazanavir/ritonavir (ATV/r), raltegravir (RAL), or darunavir/ritonavir (DRV/r) for treatment of antiretroviral-naive participants (ClinicalTrial.gov: NCT00811954).[11] Between May 2009 and June 2011, a total of 1809 eligible participants from 57 sites were enrolled in the study. Participants were aged 18 years or older with HIV-1 RNA greater than 1000 copies/mL without resistance to nucleoside reverse transcriptase inhibitors or protease inhibitors, and receiving care in the United States and Puerto Rico. The primary efficacy endpoint was time from randomization to virologic failure defined as a confirmed HIV-1 RNA level greater than 1000 copies/mL at or after 16 weeks and before 24 weeks from randomization or greater than 200 copies/mL at or after 24 weeks.[11] The primary tolerability endpoint was the time from randomization to discontinuation of the randomized regimen component for toxicity.[11] The primary efficacy analysis was based on pairwise comparisons of difference in Kaplan–Meier estimates (RD_{KM}) by week 96 with equivalence accepted if the 97.5% CI was wholly contained within ±10%. Assuming rates of virologic failure and loss to follow-up of 25% and 12%, respectively, a sample size of 600 participants per group was targeted to provide approximately a 90% statistical power to show equivalence in pairwise regimen comparisons. A similar approach was used for the tolerability outcome, although the RD_{KM} estimated used the method of Gray[12] to estimate cumulative incidence in the presence of the competing risks. Details and results can be found in Lennox et al.[11] Participant-level data sets for the present analysis were provided by the ACTG.

Equivalence bounds for an alternative analysis based on HR from a Cox PH model and based on Δ-RMST were computed assuming exponential virologic failure distributions and ACTG 5257 design characteristics. For virologic endpoint, the corresponding equivalence bounds for an HR-based analysis are 0.56–1.50, for the tolerability endpoint bounds are 0.29–1.77, and for the combined endpoint 0.63–1.43. For an analysis based on Δ-RMST, the bounds are −5.18 to +5.43 weeks, and bounds are −3.44; +3.53 weeks, and −5.30; +5.57 weeks for the virologic endpoint, the tolerability endpoint, and the combined endpoint, respectively. Details of the calculation are given in the Appendix, Supplemental Digital Content, http://links.lww.com/QAI/B283.

The analysis based on Δ-RMST has been suggested as an alternative to the use of the HR when the PH assumption is not valid. Several lack-of-fit tests have been introduced to investigate the PH assumption. In this work, we used a test based on the Schoenfeld residuals to investigate such an assumption[13] of ACTG 5257 data. We also investigated whether the hazard function is constant over time, ie, the exponential model was a plausible model to compute the HR bounds. We used the Lagrange multiplier χ^2 for scale test to investigate such a hypothesis.

Simulation Study

To investigate the performance of Δ-RMST–based analyses, with proportional and non-PH, we conducted a limited simulation study. We investigate the situation where the exponential distribution, hypothesized in the design, is used to compute the bounds for a Δ-RMST–based analysis, whereas the underlying failure distribution derives from a Weibull model (see Figures 1 and 2, Appendix, Supplemental Digital Content, http://links.lww.com/QAI/B283). To root our base parameters for the simulation study, Weibull parametric models were fit for each randomized group of ACTG 5257 data. These gave estimates of the Weibull shape parameter (k) between 0.7 and 1.1 (see Appendix Table 1, Supplemental Digital Content, http://links.lww.com/QAI/B283). These shape parameters were then used to simulate proportional and non-PH for hypothetical pairwise comparisons (see Figures 1 and 2, Supplemental Digital Content, http://links.lww.com/QAI/B283). To simulate a clinical trial data set, time-to-event outcomes were generated with a sample size of 600 per arm using an exponential ($k = 1$) or a Weibull distribution ($k \neq 1$). We investigate the situation where the failure rate is 25% in both groups (statistical power of the study) and a 25% failure rate for the reference group and 35% failure rate for the other group. Within each data set, we then estimated the Δ-RMST by week 96 from Kaplan–Meier survival curves between the pairwise regimen groups with confidence interval. We replicated 5000 such data sets. The proportion of the 5000 simulated trials that conclude equivalence was then computed using Δ-RMST bounds previously derived (−5.18 to 5.43, as derived in 2.1). The objective of our simulation study was then to investigate the performance of a Δ-RMST analysis when both the distribution of the time-to-event data do not follow the exponential distribution hypothesized in the study design and in the case of non-PH. All analyses and simulations were performed using SAS 9.4.

Reanalysis of Recent HIV Trials With Δ-RMST Measure

Finally, we performed a reanalysis of 4 recent trials[1,14–16] for the comparison of initial treatment of HIV-1 infection with the measure of Δ-RMST. In the absence of original data sets, we reconstructed individual patient data by applying an algorithm developed by Guyot et al[17] for each trial. Briefly, we used the Digitizelt software to measure the time and survival probability coordinates on the Kaplan–Meier

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

curves.[18] We extracted the numbers of patients at risk and the total numbers of events, when available. The data were then input into an algorithm on the basis of iterative numerical methods to solve the inverted Kaplan–Meier equations. More details can be found in Guyot.[17] Once the data of trials are reconstructed, we reanalyzed each trial and compared the treatment effect estimated by the Δ-RMST that by the original analysis.

RESULTS

ACTG 5257 Study and Alternative Analysis

Table 1 summarizes estimates of RD_{KM} as presented in the final A5257 publication,[11] as well as alternative analyses based on HR with full follow-up and Δ-RMST estimated based on Kaplan–Meier curves at week 96. We remark that, based on RD_{KM} analyses, equivalence is demonstrated in all comparisons for virologic endpoint and one comparison for tolerability endpoint, and equivalence is not demonstrated in any comparison for combined endpoint. Based on Δ-RMST analyses, equivalence is demonstrated in 2 comparisons for virologic endpoint and one comparison for tolerability endpoint, again, equivalence is not demonstrated in any comparison for combined endpoint. Overall, these results are consistent with the results based on RD_{KM} analyses. By contrast, equivalence is demonstrated in analyses based on HR only in one comparison (virologic endpoint) of the 9 comparisons. Estimated RMST for each endpoint and randomized groups can be found in the Table 2. For example, the RMST for the virologic failure endpoint in the ATV/r arm is approximately 89.2 weeks. The interpretation of the RMST is the following: if we treat future patients with ATV/r with characteristics consistent with the A5257 study population, the average time spent without virologic failure over 96 weeks would be approximately 89.2 weeks. Considering the virologic endpoint and the comparison between ATV/r and RAL, the interpretation of the analysis based on Δ-RMST is that the average difference in time spent without virologic failure would be approximately 2.3 weeks in favor of the RAL treatment. That represents only 2.4% of the 96 weeks of follow-up. Across all comparisons, the maximum gain for the virologic failure endpoint is 3.7 weeks for the comparison of DRV/r versus RAL in favor of the RAL, representing only 3.8% of the 96 weeks of follow-up. For the tolerability and combined endpoints, only the DRV/r versus RAL comparison leads to equivalence for the Δ-RMST analysis; this is consistent with the original RD_{KM} analyses. Again, by contrast, the HR-based analysis does not lead to an equivalence conclusion in this case.

For the comparison ATV/r versus RAL, the differences in the estimates for the tolerability endpoint and for the combined endpoint illustrate an interesting feature of these measures. The RD_{KM} and Δ-RMST estimates for both endpoints are approximately of same magnitude, RD_{KM} = 12.7% (Δ-RMST = 7.9 weeks) for the tolerability endpoint and 14.9% (9.5 weeks) for the combined endpoint. By contrast, HR estimates are quite different. The HR is 12.7 for the tolerability endpoint compared with only 2.4 for the combined endpoint. Such a difference is explained by the

TABLE 1. Summary of the Analyses in the ACTG 5257 Study[11] and Alternative Analyses

		Comparison ATV/r vs. RAL			Comparison DRV/r vs. RAL			Comparison ATV/r vs. DRV/r		
Method	Equivalence Bounds	Point Estimate	97.5% CI	Equivalence	Point Estimate	97.5% CI	Equivalence	Point Estimate	97.5% CI	Equivalence
Virologic endpoint										
RD_{KM}	±10%	3.4%	−0.6 to 7.4	Yes	5.6%	1.4 to 9.8	Yes	−2.2%	−6.7 to 2.2	Yes
HR	0.56; 1.50	1.14	0.82 to 1.58		1.40	1.0 to 1.93		0.81	0.6 to 1.10	Yes
Δ-RMST	−5.18; +5.43	2.3 wk	-0.08 to 4.7	Yes	3.7	1.1 to 6.2		−1.4	−4.1 to 1.4	Yes
Tolerability endpoint										
RD_{KM}	±7%	12.7%	9.4 to 16.1		3.6%	1.4 to 5.8	Yes	9.2%	5.5 to 12.9	
HR	0.29; 1.77	12.7	5.6 to 29.1		4.07	1.7 to 9.9		3.14	2.0 to 4.9	
Δ-RMST	−3.44; +3.53	7.9 wk	5.5 to 10.3		1.6	0.1 to 3.0	Yes	6.4	3.8 to 9.0	
Combined endpoint (virologic or tolerability endpoint)										
RD_{KM}	±10%	14.9%	10.2 to 19.6		7.5%	3.2 to 11.8		7.5%	2.3 to 12.7	
HR	0.63; 1.43	2.4	1.7 to 3.0		1.55	1.1 to 2.1		1.46	1.1 to 1.9	
Δ-RMST	−5.30; +5.57	9.5 wk	6.3 to 12.6		4.1	1.4 to 6.8		5.4	1.9 to 8.8	

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

TABLE 2. RMST Estimate for Each Endpoint and Randomized Groups of ACTG 5257 Study

Endpoint	Arm	RMST by Week 96 (wk)
Virologic failure	ATVr	89.2
	RAL	91.5
	DRVr	87.8
Tolerability endpoint	ATVr	87.1
	RAL	95.0
	DRVr	93.4
Combined endpoint	ATVr	81.5
	RAL	90.9
	DRVr	86.9

distribution of the events in each arm and the very low hazard of events in the RAL group.

Of note, visual inspection of the cumulative incidence curves and formal testing suggest violation of the PH assumption for some comparisons (P = 0.003, Fig. 1, Appendix Table 2, Supplemental Digital Content, http://links.lww.com/QAI/B283). Although this does not impact the published trial results based on RD_{KM}, it follows that in many comparisons, the interpretation of the HR estimate is difficult, and that such an analysis is not appropriate. Of note, for the virologic failure endpoint, the PH assumption is likely violated for the ATV/r versus RAL comparison (P = 0.04) and the DRV/r versus RAL comparison (P = 0.06). These 2 comparisons correspond to discordance between the HR-based analysis and the RD_{KM} analysis. Inversely, the PH assumption is not violated for the ATV/r versus DRV/r comparison (P = 0.88) that corresponds to a concordance of the 3 analyses.

Furthermore, investigation of the constant hazard rate assumption suggests that the exponential distribution for the time-to-event data may not be valid for the tolerability and combined endpoints in the ATV/r group and for the combined endpoint in the DRV/r (see Appendix Table 3, Supplemental Digital Content, http://links.lww.com/QAI/B283).

Simulation Study

Results of our simulation study are displayed in Tables 3 and 4. Of note, hazard rates are proportional when the shape parameter (k) is the same in both randomized groups. Agreement between the design and the data generation regarding the time-to-event distribution corresponds to both exponential distributions (k = 1). Considering an exponential distribution in both groups per the design assumptions, the expected probability of concluding equivalence (power of the study) for the Δ-RMST analysis is 93.6% for the pairwise comparison (Table 3). When survival curves deviate from the exponential distribution, but the PH assumption remains valid (k = 0.7, 0.9 and 1.1 in both groups), the probability of concluding equivalence decreases with monotonic decreasing hazard function (k = 0.7 and k = 0.9) and increases with monotonic increasing hazard function (k = 1.1). When the PH assumption is not valid, with a shape parameter that is greater in one group, the probability of concluding equivalence is markedly decreased (Table 3).

With 96-week failure rates of 25% and 35% in the 2 study arms, under the assumed design (survival curves for both groups follow exponential distributions), the probability of incorrectly concluding noninferiority (the false positive rate) for the Δ-RMST analysis is 5.4% for the pairwise comparison (Table 4). When the PH assumption is valid but the survival curves deviate from the exponential distribution (k= 0.7, 0.9, and 1.1 in both groups), the false positive rate decreases with monotonic decreasing hazard function (k = 0.7 and k = 0.9) and increases with monotonic increasing hazard function (k = 1.1). When the PH assumption is not valid, if the shape parameter is highest in the group with the highest failure rate, the false positive rate is markedly increased (Table 4). Inversely, the false positive rate is close to 0 when the group with the highest failure rate has the lowest shape parameter.

Reanalysis of Recent HIV Trials With Δ-RMST Measure

Table 5 summarizes the results of reanalysis of 4 recent trials for the comparisons of initial treatment of HIV-1 infection with the Δ-RMST measure. For most comparisons, we remark that the gain in Δ-RMST is low. For example, in the recent NEAT study comparing DRV/r + RAL with TDF/FTC + RAL in antiretroviral-naive adults, the Δ-RMST is 2.4 weeks (95% CI: −5.19 to 0.34) in favor of the latter group for 96 weeks of follow-up.

DISCUSSION

The aim of our study was to compare analyses based on RD_{KM}, HR, and Δ-RMST as a summary effect estimates in ACTG 5257 study and most specifically to investigate the performance and operating characteristics of Δ-RMST–based analysis in the context of HIV trials. It is important to understand the meaning of these 3 different measures. The RD_{KM} is an absolute risk difference based on a cumulative number of events at a given time. Without censoring, it corresponds exactly to a simple difference in proportion failing at a fixed time point. The HR compares the instantaneous rate (hazard) of an event in a treatment group with that in another group. Using the Cox's model for its estimation implies that the hazard in a treatment group remains proportional to the hazard the other group at any point during the study period. Such an assumption seems very stringent, but it has been shown that tests to detect violation of the proportional assumption have low statistical power.[19] RMST is a more recent measure that can be interpreted as the mean event-free time up to a prespecified important time point.[20] The RMST difference means gain or loss in the event-free time because of treatment versus reference during this period. RMST difference is valid and interpretable whether or not the PH assumption is violated.

As expected, analyses of Δ-RMST globally led to similar conclusions as the published findings of A5257 based

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

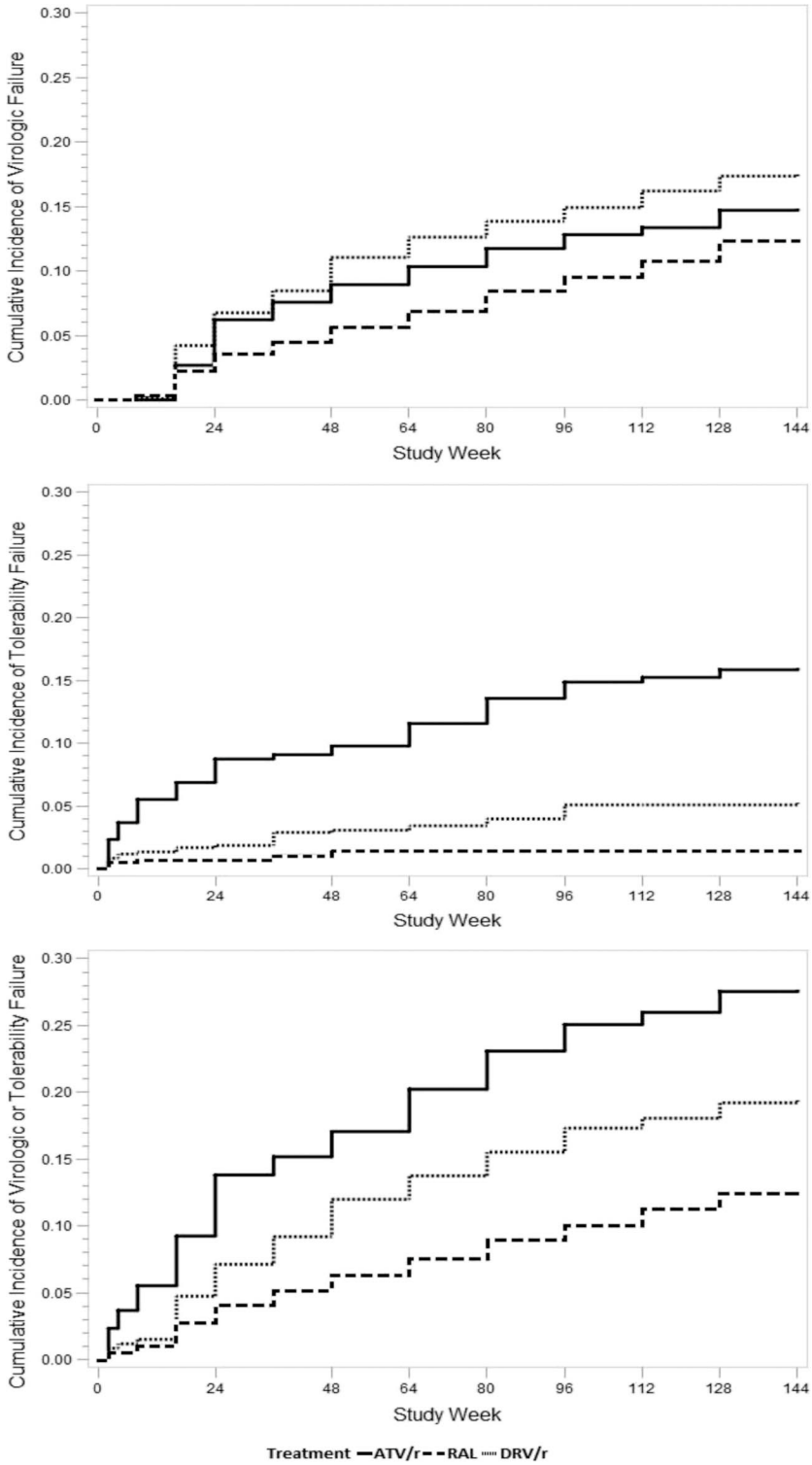


FIGURE 1. Cumulative incidence of the different endpoints based on ACTG 5257 data.

on RD_{KM}. By contrast, analyses based on HR provided some discordant equivalence conclusions compared both with the initial analysis based on RD_{KM} and the Δ-RMST. In particular, considering the virologic failure endpoint, analyses based on RD_{KM} at week 96 led to demonstrating equivalence for all pairwise comparisons, whereas except for the comparison ATV/r versus DRV/r, results based on HR did not conclude to such equivalence. It is worth to note that globally, even in the setting of very minimal and clinically insignificant group differences estimated with RD_{KM} and Δ-RMST, inference based on HR still did not reach the equivalence conclusion.

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

TABLE 3. Probability of Concluding Equivalence (Power of the Study) When Time-to-Event Outcome Follow an Exponential Distribution (Shape Parameter = 1) or a Weibull Distribution (Shape Parameter ≠ 1) Comparing 2 Groups Both With 25% Failure Rates

		25% Failure Rate			
		Shape Parameter (k)			
	Shape Parameter (*k*)	0.7	0.9	1.0	1.1
25% failure rate	0.7	81.5	66.5	51.2	38.3
	0.9		89.4	86.8	77.4
	1.0			**93.6**	89.9
	1.1				95.2

In the ACTG 5257 protocol, a probability of virologic failure of 25% by week 96 was assumed in study design, whereas results of study showed virologic failure rates of 9%–15%. It has been indicated in our recent study that when a HR-based analysis has been designed in the protocol, an overestimation of failure rates when defining equivalence bounds can lead to increase in the probability of not concluding equivalence even when there is no clinically relevant difference between the 2 regimens.[4] This issue was a motivating factor in the choice of the RD_{KM} estimand for A5257, and our results of alternative analyses of ACTG 5257 data confirmed this finding.

Apart from discordant conclusions, large differences in estimates between the analyses can be found according to the endpoint. In particular, in the comparison of ATV/r versus RAL, for the tolerability endpoint, the week 96 RD_{KM} and Δ-RMST estimates give a 12.7% higher failure probability and 7.9 weeks lower expected time without tolerability failure with ATV/r with a high corresponding HR estimate of 12.7. Although for the combined endpoint of virologic or tolerability failure, the 96-week RD_{KM} and Δ-RMST estimates are relatively unchanged (14.9% and −9.5 weeks, respectively), yet with this endpoint, there is a much lower HR estimate of 2.4. Such a discrepancy is explained by nonconstant hazard function, non-PH ratio and the very low failure rate in the RAL group. In this case when treatment effects are measured with HR rather than RD_{KM} or Δ-RMST, a larger variability in the estimates could be observed. Thus, the HR-based analyses showed much more variability than the RD_{KM} and Δ-RMST in the analysis of ACTG 5257 data.

TABLE 4. Probability of Incorrectly Concluding Equivalence (False Positive Rate) When Time-to-Event Outcome Follow an Exponential Distribution (Shape Parameter = 1) or a Weibull Distribution (Shape Parameter ≠ 1) Comparing 2 Groups With 35% and 25% Failure Rates

		35% Failure Rate			
		Shape Parameter (k)			
	Shape Parmeter (*k*)	0.7	0.9	1.0	1.1
25% failure rate	0.7	1.6	19.6	38.9	61.3
	0.9	0.1	3.9	12.2	27.7
	1.0	0.1	1.6	**5.4**	15.2
	1.1	0.0	0.7	2.5	9.2

Our simulation study focused on the performance of Δ-RMST–based analysis in the context of nonproportional and nonconstant hazards. RD_{KM} and HR-based analyses were not evaluated because the RD_{KM} estimate will not change because the difference in the rate of failure at week 96 is the same whatever the time-to-event distribution (see Figures 1 and 2, Appendix, Supplemental Digital Content, http://links.lww.com/QAI/B283), and the HR estimate is not appropriate in the case of non-PH. The RMST analysis has been introduced as a promising alternative in the case of non-PH ratio. Our limited simulation study, however, shows that discordance between assumptions in the study design and observed data may have a great impact on the results based on Δ-RMST analyses. In particular, in certain situations violation of the PH assumption markedly increase the probability of falsely concluding noninferiority. Then, as for a HR-based analysis, departures from the hypotheses made in the study design have an important impact on the study conclusions. This makes sense intuitively and reflects that early advantages of one regimen are being offset by later advantages of the second regimen. In our more extreme situation, the probability to conclude noninferiority with an RMST-based analysis is 38.3%, whereas the design is based on a 93.6% probability. Then, although the RMST difference is valid under non-PH assumption, findings are greatly altered in such a situation.

The results of some recent studies suggest estimating the treatment effects by RMST-based measure in randomized trials with time-to-event outcomes.[2,8–10] To our knowledge, this measure has not been used as primary measure of efficacy in HIV/AIDS clinical trials. Based on the illustration in Table 5, one could find that compared with HR, it is straightforward to interpret Δ-RMST from a clinical perspective by quantifying the absolute survival difference that the clinicians and patients could comprehend. Meanwhile, before being used in HIV trials, clinicians should think about what bounds are clinically acceptable to declare a treatment equivalent or noninferior to the reference arm in a Δ-RMST–based analysis. For example, in the NEAT study in Table 5, the Δ-RMST estimate is 2.4 weeks with 95% CI of −5.2 at 0.3 weeks. This estimate provide a high degree of confidence that, on average, through 96 weeks of follow-up, the DRV/r + RAL group would be free of events no more than 5.2 weeks less than the TDF + FTC + DRV/r group that represents only 5.4% of the 96 weeks of follow-up. Thus, clinicians should think about if a difference of 5.2 weeks of 96 is a clinically acceptable bound of noninferiority or equivalence. We have seen that the determination of such a bound is not easy as being dependent of the time-to-event distribution.

CONCLUSIONS

According to the results of alternative analyses of ACTG 5257 study, we showed that analyses based on Δ-RMST globally led to similar conclusions as the

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

TABLE 5. Summary of the Original Analyses and Alternative Analyses by Δ-RMST From Reconstructed Data of the Trials for the Comparison of Initial Treatment of HIV-1 Infection[1,14–16]

Author, yr	Study Team	Objective	Original Analysis	Δ-RMST Estimates
Raffi F et al,[14] 2014	NEAT001	To assess the 96 wk efficacy and safety of an NtRTI-sparing regimen vs. standard regimen	Difference in Kaplan–Meier estimate by week 96	During 96 weeks of follow-up, 2.4 wk (95% CI: −0.3 to 5.2), $P = 0.09$
	ANRS143		4.0% (95% CI: −0.8 to 8.8)	
			Noninferiority threshold = 9%	
Daar ES et al,[1] 2011	ACTG 5202	To compare the efficacy, safety, and tolerability of atazanavir plus ritonavir vs. efavirenz for each of the 2 combinations ABC/3TC and TDF/FTC	By week 96, for comparison in ABC/3TC group:	During 96 weeks of follow-up, for comparison in ABC/3TC group: 0.5 weeks (95% CI: −2.6 to 3.6) for comparison in TDF/FTC group: 1.0 weeks (95% CI: −1.5 to 3.6) noninferiority threshold for Δ-RMST: 95% upper CI ≤ 5.48
			HR 1.13 (95% CI: 0.82–1.56) for comparison in TDF/FTC group:	
			HR 1.01 (95% CI: 0.70 to 1.46) noninferiority threshold for HR: 95% upper CI ≤ 1.40	
Gulick RM et al, 2006	ACTG 5095	To compare the safety/efficacy of 3-drug vs. 4-drug regimen for the initial treatment of HIV-1 infection	By week 156, HR 0.95 (97.5% CI: 0.69 to 1.33), $P = 0.73$	During 156 weeks of follow-up, −1.6 weeks (97.5% CI: −10.3 to 7.0), $P = 0.67$
Gulick RM et al, 2004	ACTG 5095	To compare the efficacy of triple-nucleoside regimen vs. efanvirenz-containing regimen for the initial treatment of HIV-1 infection	By week 32, 82 (21%) of 382 subjects in the triple-nucleoside group, 85(11%) of those in the combined efavirenz group had virologic failure, $P < 0.001$	During 32 weeks of follow-up, −1.1 weeks (95% CI: −1.7 to 0.5), $P < 0.001$

Δ-RMST, difference in restricted mean survival time, estimated by week; ABC/3 TC, abacavir/lamivudine; ATV/r, atazanavir+rotonavir; CI, confidence interval; EFV, efavirenz; HR, hazard ratio; TDF/FTC, tenofovir/emtricitabine.

published findings of A5257 based on RD_{KM}. By contrast, analyses based on HR provided some discordant equivalence conclusions compared both with initial analysis based on RD_{KM} and the Δ-RMST. Such discordance is mainly explained by violation of the PH assumption and by an underestimation or overestimation of the hypothesized failure rates in the study design.[4] Although there are some advantages to use the RMST, further discussion is needed with clinicians involved in the HIV/AIDS field to determine which estimand best informs the clinical question and, as applicable, to suggest equivalence bounds. We also show that the power and false positive rate provided by RMST-based analyses are sensitive to departures from the PH assumption.

ACKNOWLEDGMENTS

ACTG 5257 Study Acknowledgements. The authors gratefully acknowledge the efforts of the ACTG 5257 protocol team, the participating Clinical Trial Units of the ACTG, and the study participants.

ACTG 5257 Study Team: Jeffrey L. Lennox (Chair), Judith S. Currier (Co-Chair), Raphael J. Landovitz (Vice chair) Ighovwerha Ofotokun (Vice chair), Catherine Godfrey (DAIDS Clinical Representative), Jennifer Rothenberg, Heather J. Ribaudo, Lumine H. Na, Bernadette Jarocki, Alex Benns, Lynette Perdue, Daniel R. Kuritzkes, Manish Sagar, Francesca Aweeka, Michael M. Lederman, Todd T. Brown, Susan E. Cohn, Grace A. McComsey, Susan Pedersen, Carmen M. Irizarry, George Bishopric, Bryan P. Baugh, Randi Y. Leavitt, James F. Rooney, and Daniel Seekins.

REFERENCES

1. Daar ES, Tierney C, Fischl MA, et al. Atazanavir plus ritonavir or efavirenz as part of a 3-drug regimen for initial treatment of HIV-1A randomized trial. *Ann Intern Med.* 2011;154:445–456.
2. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol.* 2014;32:2380–2385.
3. Walter S. Choice of effect measure for epidemiological data. *J Clin Epidemiol.* 2000;53:931–939.
4. Abulizi X, Flandre P. Choice of treatment effect measures when noninferiority margins originally defined in absolute difference translated into relative difference influenced the results of clinical trials. *J Clin Epidemiol.* 2017;96:63–72.
5. Flandre P. Choice of effect measure in HIV randomized trials. *AIDS.* 2015;29:2057–2060.
6. Hernán MA. The hazards of hazard ratios. *Epidemiology.* 2010;21:13–15.
7. Trinquart L, Jacot J, Conner SC, et al. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol.* 2016;34:1813–1819.
8. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med.* 2011;30:2409–2421.
9. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol.* 2013;13:152.
10. Uno H, Wittes J, Fu H, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Ann Intern Med.* 2015;163:127.
11. Lennox JL, Landovitz RJ, Ribaudo HJ, et al. Efficacy and tolerability of 3 nonnucleoside reverse transcriptase inhibitor–sparing antiretroviral regimens for treatment-naive volunteers infected with HIV-1: a randomized, controlled equivalence trial. *Ann Intern Med.* 2014;161:461.
12. Gray RJ. A class of K-Sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat.* 1988;16:1141–1154.
13. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika.* 1994;81:515–526.

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

14. Raffi F, Babiker AG, Richert L, et al. Ritonavir-boosted darunavir combined with raltegravir or tenofovir–emtricitabine in antiretroviral-naive adults infected with HIV-1: 96 week results from the NEAT001/ANRS143 randomised non-inferiority trial. *Lancet.* 2014;384:1942–1951.
15. Gulick RM, Ribaudo HJ, Shikuma CM, et al. Three-vs four-drug antiretroviral regimens for the initial treatment of HIV-1 infection: a randomized controlled trial. *JAMA.* 2006;296:769–781.
16. Gulick RM, Ribaudo HJ, Shikuma CM, et al. Triple-nucleoside regimens versus efavirenz-containing regimens for the initial treatment of HIV-1 infection. *N Engl J Med.* 2004;350:1850–1861.
17. Guyot P, Ades AE, Ouwens MJ, et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol.* 2012;12:9.
18. Bormann I. DigitizeIt—Plot digitizer software. Digitize graphs, charts and math data. Available at: http://www.digitizeit.de/. Accessed November 7, 2017.
19. Austin PC. Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model. *J Stat Comput Simul.* 2018;88:533–552.
20. Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. *JAMA Cardiol.* 2017;2:1179–1180.

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.

Copyright © 2019 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

2.1 Appendices

Equivalence bounds for an alternative analysis based on HR from a Cox proportional hazards model are easily computed. Assuming an exponential distribution for the time to event as defined in the protocol, a probability of virologic failure of 25% (reference arm) by week 96 corresponds to a hazard of 0.0030 (-log(1-0.25)/96) and of 35% to a hazard of 0.0045. For the virologic failure, the upper equivalence-bound for an HR-based analysis corresponding to the +10% limits for inference on RD_{KM} is then 1.497 (0.0045/0.003). Similarly, the lower bound on the basis of 15% failure rate is 0.56 (0.0017/0.003). The bound for the tolerability endpoint assuming a 10% failure rate in the reference arm by week 96 is -/+ 7%. In a similar way, the bounds for the HR-based analysis are 0.29 and 1.77. In the A5257 protocol, there is no assumption concerning the level of failure rate for the combined endpoint, virologic or tolerability failure. To illustrate our purpose, we consider a failure rate of 30% for the combined endpoint in the reference arm by week 96 and a -/+10% bound for the RD_{KM}–based analysis. The corresponding bounds for an HR-based analysis are 0.63 and 1.43.

The bounds correspond to the mean difference in Restricted Mean Survival Time (RMST) between the two regimens being compared at weeks 96 could be estimated with the integral under the Kaplan-Meier curve up to week 96. Assuming an exponential distribution for the time to event, a probability of virologic failure of 25% (hazard 0.0030) and 35% (hazard 0.0045) at week 96 lead to a RMST estimate of 83.43 weeks ((-1/0.0030)*(exp(-0.0030*96)-1)) and 78.00 weeks ((-1/0.0045)*(exp(-0.0045*96)-1)), and a probability of 15% (hazard 0.0017) at week 96 lead to a RMST estimate of 88.6 weeks ((-1/0.0017)*(exp(-0.0017*96)-1)). Thus the equivalence-bounds for a Δ-RMST analysis corresponding to the -/+10% limits for Δ-RMST are then -5.18 to 5.43 weeks. In a similar way, bounds are [-3.44; +3.53] weeks for the tolerability endpoint and [-5.30; +5.57] weeks for the combined endpoint.

Appendix Table 1. Results of fitting a Weibull model to the ACTG 5257 data. We use the following parameterization for the Weibull model $S(t)=\exp((-\lambda t)^k)$.

Endpoint	Arm	Weibull shape parameter, *k* (95% CI)
Virologic failure	ATVr	0.87 (0.72 to 1.05)
	RAL	1.07 (0.87 to 1.31)
	DRVr	0.89 (0.74 to 1.05)
Tolerability endpoint	ATVr	0.73 (0.63 to 0.83)
	RAL	0.85 (0.71 to 1.03)
	DRVr	0.91 (0.78 to 1.07)
Combined endpoint	ATVr	0.75 (0.67 to 0.84)
	RAL	0.91 (0.78 to 1.07)
	DRVr	0.83 (0.72 to 0.95)

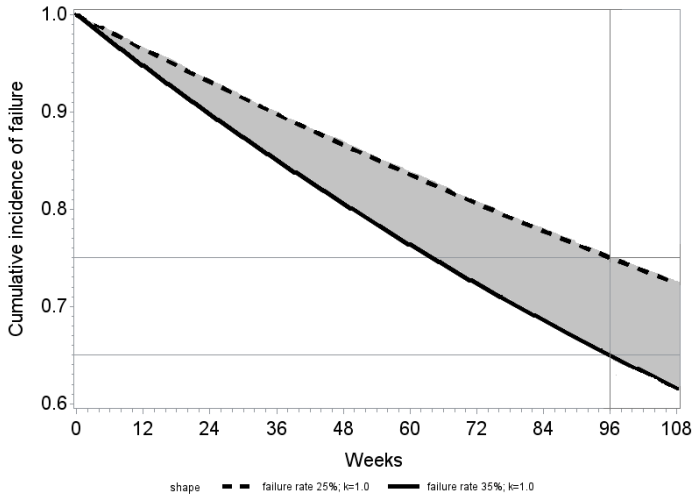
Appendix Table 2. Test of the proportional hazards assumption.

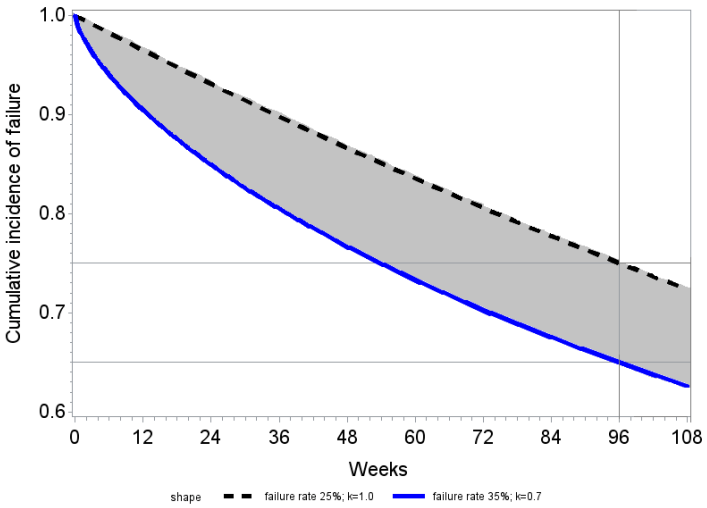
Endpoint	ATV/r vs RAL	DRV/r vs RAL	ATV/r vs DRV/r
Virologic failure	p=0.04	p=0.06	p=0.88
Tolerability endpoint	p=0.12	p=0.08	p=0.37
Combined endpoint	p=0.003	p=0.17	p=0.06

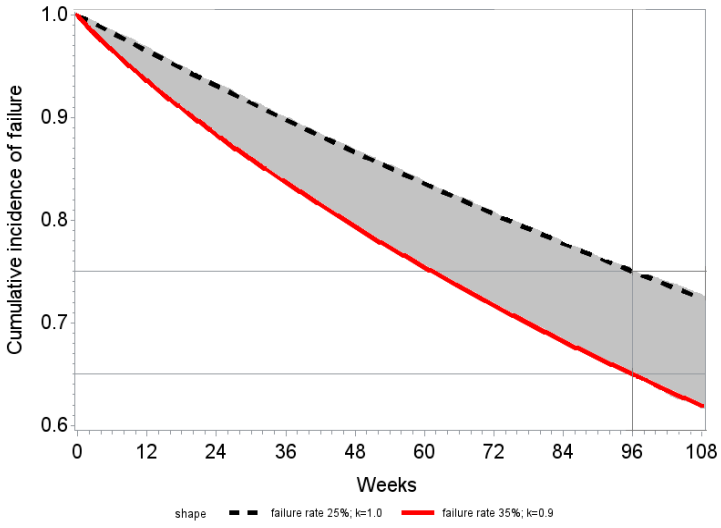
Appendix Table 3. Test that the hazard function is constant over time (Lagrange Multiplier Chisquare for Scale).

Endpoint	ATV/r	DRV/r	RAL
Virologic failure	p=0.17	p=0.19	p=0.50
Tolerability endpoint	p<0.0001	p=0.29	p=0.13
Combined endpoint	p<0.0001	p=0.014	p=0.26

Figure 1. Illustration of the simulation study where in the reference group (failure rate 25% at week 96) the distribution of time-to-event outcome is exponential (shape parameter k=1) whereas in the other group (failure rate 35%) the distribution is exponential (k=1.0) or Weibull (k=0.7, 0.9 and 1.1).







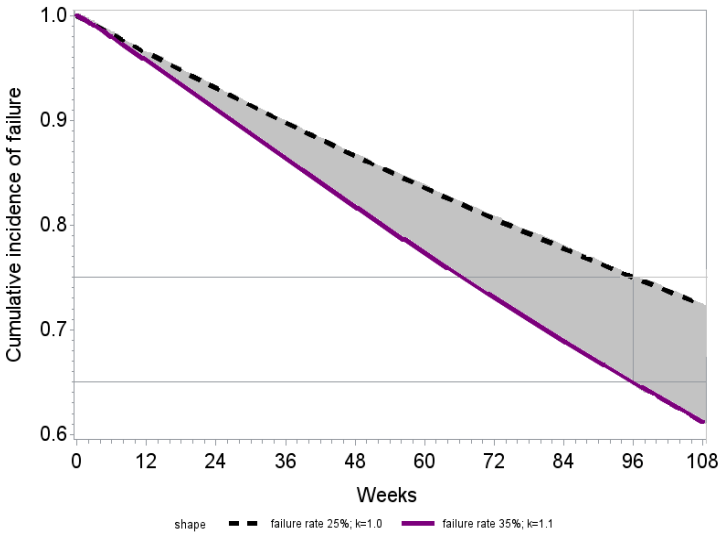
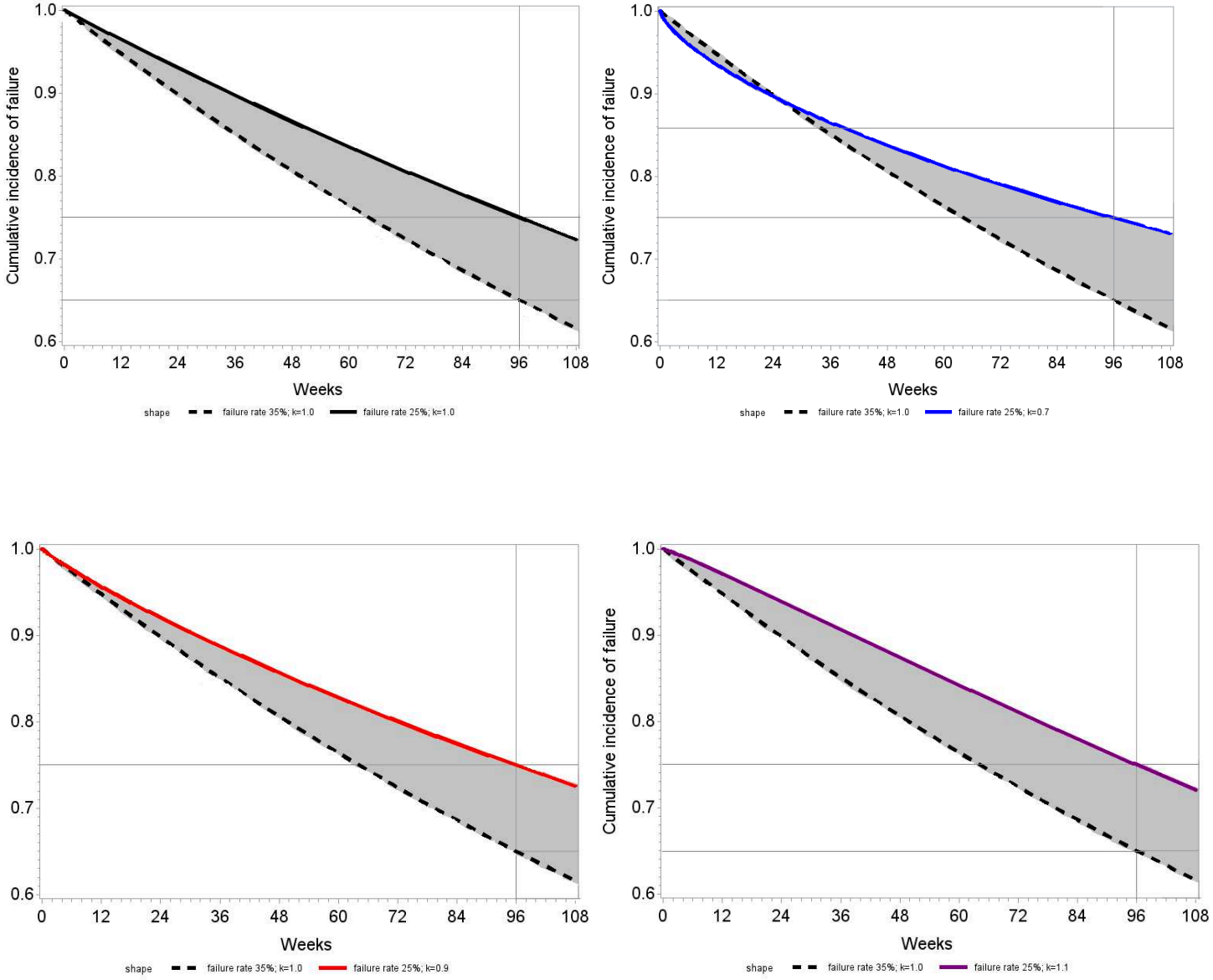


Figure 2. Illustration of the simulation study where in the reference group failure rate is 25% at week 96 with two different distributions (exponential with k=1.0 and Weibull with k=0.7, 0.9 and 1.1) whereas in the other group (failure rate 35%) the distribution is exponential (shape parameter k=1).



CHAPITRE VI – L'UTILISATION DU TEMPS DE SURVIE MOYEN BORNE (RMST) COMME MESURE DE L'EFFET TRAITEMENT DANS LES ESSAIS SUR LA PROPHYLAXIE PRE-EXPOSITION (PREP)

Introduction

La prévention de l'infection par le VIH passe par différents outils (préservatif, traitement post exposition, traitement antiviral des personnes infectées etc.) dont la combinaison permet une adaptation à différents groupes à risque. La prophylaxie pré-exposition (PrEP) est l'un de ces outils ; elle consiste à administrer préventivement, avant un rapport sexuel, un traitement anti-rétroviral pour éviter la survenue d'une infection ; elle s'adresse à des populations régulièrement exposées.

Les essais sur la PrEP visent à évaluer l'effet de la prise d'antirétroviraux en prophylaxie de l'infection à VIH chez les individus à haut risque, ils s'intéressent à un événement dépendant du temps : le délai jusqu'à une contamination par le VIH. Ces essais comparent le plus souvent un groupe placebo à un groupe recevant du tenofovir (TDF) seul ou combiné avec de l'emtricitabine (TDF-FTC). Le critère principal est le nombre de patients ayant contracté l'infection pendant le suivi des patients avec souvent une analyse fondée sur un rapport de taux d'incidence de nouvelles infections. Cette analyse est équivalente à une estimation du rapport de risques par un modèle à risques proportionnels. De même que pour le rapport de risques, le rapport de taux d'incidence est peu interprétable lorsque les taux d'incidence ne sont pas constants au cours du temps. Ces essais se sont beaucoup intéressés aux résultats en fonction du niveau d'adhérence des patients traités. Une question posée concerne la persistance ou non de l'adhérence au cours du temps. Si ce niveau n'est pas constant cela conduira à une grande difficulté dans l'interprétation du rapport de taux d'incidence.

L'analyse basée sur la mesure du temps de survie moyen borné (Restricted Mean Survival Time, RMST) est considérée comme une alternative à l'utilisation du rapport de risques lors-

que l'hypothèse des risques proportionnels (PH) n'est pas valide. Cette mesure a été principalement utilisée en oncologie (Saluja et al., 2019; Trinquart et al., 2019b, 2016), mais elle n'a pas été utilisée comme principale mesure de l'effet traitement dans les essais sur la PrEP. L'objectif de cette étude est de ré-analyser les données de cinq essais sur la PrEP par la mesure de différence de RMST (Δ-RMST) et d'étudier les performances d'une analyse basée sur une telle mesure dans les essais sur la PrEP.

Matériel et méthodes

Les essais sur la PrEP

L'étude iPrEx **(Grant et al., 2010)** est un essai randomisé qui a évalué la protection de TDF-FTC par voie orale contre l'acquisition de l'infection à VIH chez 2499 hommes ou femmes transgenres séronégatifs ayant des rapports sexuels avec des hommes. Les sujets provenaient de 6 sites : Pérou, Équateur, Afrique du Sud, Brésil, Thaïlande et États-Unis. La période d'étude était de juin 2007 à décembre 2009. Il a été déterminé dans le protocole que l'observation de 85 cas d'infection à VIH aurait conduit à une puissance d'au moins 80 % avec un niveau d'alpha unilatéral de 0,05 pour rejeter une hypothèse nulle d'efficacité de 30 % ou moins. Les sujets de l'étude ont été suivis pendant 3324 personnes-années (médiane, 1,2 ans; maximum, 2,8 ans). Parmi ces sujets, 100 ont été infectés au cours du suivi (36 dans le groupe FTC–TDF et 64 dans le groupe placebo), indiquant une réduction de 44 % de l'incidence du VIH (IC à 95 %, 15 à 63; p = 0,005). Ainsi, l'équipe a conclu à un effet protecteur de TDF-FTC contre l'acquisition de l'infection par le VIH chez les sujets.

L'étude FEM-PrEP **(Van Damme et al., 2012)** est un essai randomisé qui a évalué l'efficacité de TDF-FTC dans la prévention de l'acquisition du VIH chez 2120 femmes séronégatives au Kenya, en Afrique du sud et en Tanzanie. La période d'étude était du 11 juin 2009 au 15 avril 2011. L'étude était conçue pour tester l'hypothèse selon laquelle TDF-FTC en prise unique

quotidienne est efficace à plus de 30 % dans la prévention de l'acquisition du VIH. Pour tester cette hypothèse, l'équipe de recherche a calculé que 72 cas d'infection fourniraient une puissance de 90%, avec un niveau alpha unilatéral de 0,025. Au total, 33 infections à VIH sont survenues dans le groupe TDF – FTC (taux d'incidence de 4,7 pour 100 personnes-années) et 35 infections dans le groupe placebo (taux d'incidence de 5,0 pour 100 personnes-années), pour un rapport de risques estimé pour le groupe TDF-FTC de 0,94 (IC à 95 %, 0,59 à 1,52; p = 0,81). L'équipe de FEM-PrEP a conclu que la prophylaxie avec TDF–FTC ne réduisait pas significativement le taux d'infection par le VIH.

L'étude Partners PrEP **(Baeten et al., 2012)** est un essai randomisé de traitement antirétroviral à utiliser comme prophylaxie avant exposition chez 4747 couples hétérosexuels sérodiscordants du Kenya et de l'Ouganda. La période d'étude était de juin 2008 à novembre 2010. L'équipe de l'étude a calculé que 147 cas de séroconversion VIH-1 pour chaque comparaison (TDF versus placebo et TDF – FTC versus placebo) fourniraient une puissance de 80 %, avec un niveau d'alpha unilatéral de 0,025, permettant de détecter une réduction relative de 60 % sur l'incidence de l'infection par le VIH, avec la borne inférieure de l'intervalle de confiance à 95 % excluant une diminution relative de l'incidence de 30 % (hypothèse nulle). Parmi les participants, le partenaire VIH séronégatif de chaque couple a été randomisé à l'un des trois schémas thérapeutiques suivants : TDF en prise unique quotidienne, TDF – FTC en prise unique quotidienne ou placebo correspondant - et suivi jusqu'à 36 mois. Au total, 82 infections à VIH sont survenues chez des participants séronégatifs au cours du suivi, 17 dans le groupe TDF (incidence, 0,65 pour 100 personnes-années), 13 dans le groupe TDF-FTC (incidence, 0,50 pour 100 personnes-années), et 52 dans le groupe placebo (incidence, 1,99 pour 100 personnes-années), ce qui indique une réduction relative de 67 % de l'incidence du VIH avec le TDF (IC à 95 % , 44 à 81; p <0,001) et de 75 % avec le TDF – FTC (IC à 95 %, 55 à

87; p <0,001). L'équipe de Partners PrEP a donc conclu que ces deux traitements réduisent significativement l'incidence du VIH chez les couples hétérosexuels sérodiscordants.

L'étude TDF2 **(Thigpen et al., 2012)** est un essai randomisé qui a évalué la prise unique quotidienne de FTC-TDF dans la prévention de l'infection par le VIH chez 1219 hommes et femmes séronégatifs au Botswana. La période d'étude était de mars 2007 à octobre 2009. L'hypothèse principale de l'étude était que le TDF-FTC, par rapport au placebo, réduirait l'incidence d'infection par le VIH d'au moins 65 %, avec une borne inférieure prédéfinie pour l'intervalle de confiance à 95 % de 10 %. Au total, 1219 hommes et femmes (45,7 % de femmes) ont été randomisés pour recevoir soit le TDF-FTC, soit un placebo en prise unique quotidienne, et ont été suivis 1 563 personnes-années (médiane: 1,1 an; maximum, 3,7 ans). Au total, 33 participants ont été infectés au cours de l'étude (9 dans le groupe TDF-FTC et 24 dans le groupe placebo ; soit 1,2 et 3,1 infections pour 100 personnes-années respectivement), l'efficacité du TDF-FTC était de 62,2 % (IC à 95 %, 21,5 à 83,4; p = 0,03). L'équipe de TDF2 a conclu que la prophylaxie quotidienne par le TDF-FTC a permis de prévenir l'infection à VIH chez les adultes hétérosexuels sexuellement actifs.

L'étude ANRS IPERGAY **(Molina et al., 2015)** est un essai randomisé qui a évalué l'utilisation de la prophylaxie TDF-FTC dans la prévention de l'infection à VIH chez 400 hommes séronégatifs ayant des relations sexuelles anales non protégées avec des hommes en France et au Canada. La période d'étude était du 22 février 2012 au 23 octobre 2014. L'équipe de l'étude a calculé que 64 cas de séroconversion du VIH fourniraient une puissance de 80 % pour détecter une réduction relative de 50 % de l'incidence de l'infection par le VIH dans le groupe TDF-FTC, par rapport au groupe placebo, avec un niveau d'alpha bilatéral de 0,05. Tous les participants ont été suivis pendant une médiane de 9,3 mois (écart interquartile de 4,9 à 20,6). Au total, 16 infections à VIH-1 sont survenues pendant le suivi, 2 dans le groupe TDF-FTC (incidence, 0,91 pour 100 personnes-années) et 14 dans le groupe placebo (inci-

dence, 6,60 pour 100 personnes-années), ce qui indique une réduction relative pour le groupe TDF-FTC de 86 % (IC à 95%, 40 à 98; p = 0,002). L'équipe a donc conclu que l'utilisation de TDF-FTC avant et après une activité sexuelle offrait une protection contre l'infection à VIH chez les hommes ayant des rapports sexuels avec des hommes.

Il est à noter que, cette analyse a été réalisée sur demande du comité indépendant de surveillance à la suite de la publication des résultats de l'essai PROUD **(McCormack et al., 2016)**, et que donc la différence entre le nombre attendu de contaminations VIH (n=64) et observé (n=16) ne correspond pas à une surestimation du risque, mais à un arrêt prématuré de l'essai pour efficacité, avec un suivi moins long et un nombre de participants inclus plus faible que prévu, ce qui explique –en partie- l'écart entre nombre d'événements observé et attendu.

Ré-analyse des essais sur la PrEP avec la mesure Δ-RMST

Nous avons procédé à une nouvelle analyse de cinq essais sur la PrEP présentés ci-dessous par la mesure de Δ-RMST. En l'absence des jeux de données originaux, nous avons reconstruit les données individuelles des patients en appliquant un algorithme développé par Guyot et ses collègues et le logiciel Digitizelt comme décrit dans le Chapitre IV **(Guyot et al., 2012; Bormann, n.d.)**. Une fois les données des essais reconstituées, nous avons procédé à une nouvelle analyse de chaque essai et comparé l'effet du traitement estimé par la Δ-RMST avec celui de l'analyse originale.

Test de l'hypothèse des risques proportionnels

L'analyse basée sur Δ-RMST a été suggérée comme alternative à l'utilisation du rapport de risques lorsque l'hypothèse des risques proportionnels (PH) n'est pas respectée **(Uno et al., 2015)** . Cependant, notre étude récente a montré que dans les essais cliniques sur le VIH/SIDA, une telle analyse peut être fortement affectée par des écarts par rapport à l'hypothèse de PH **(Abulizi et al., 2019)**. Dans ce travail, nous avons utilisé le test basé sur les résí-

dus de Schoenfeld (même test que nous avons utilisé pour l'analyse de l'essai ACTG 5257) pour étudier une telle hypothèse **(Grambsch and Therneau, 1994)** sur les données de ces cinq essais sur la PrEP.

Calcul de la taille de l'échantillon basé sur la mesure Δ-RMST

Nous avons utilisé le package *SSRMST* (Voir **ANNEXE**) pour calculer la taille d'échantillon et la puissance de l'étude en utilisant la mesure Δ-RMST. En résumé, nous avons d'abord extrait les paramètres d'hypothèse tels que le taux d'incidence, le nombre de nouvelles infections, la puissance ciblée et l'estimation sur la réduction relative du risque. Ces paramètres sont disponibles dans la partie "Méthodes" des publications originales. Ensuite sur la base de ces paramètres, nous avons calculé les paramètres de forme et d'échelle de la distribution en tant que fonction de survie. Puis basé sur le concept de chaque essai, nous avons déterminé le temps d'intérêt cliniquement pertinent pour calculer le RMST. Enfin, lorsque tous ces paramètres ont été spécifiés, le package *SSRMST* calcule la taille d'échantillon et la puissance de l'étude en utilisant la Δ-RMST comme mesure de l'effet. Plus de détails peuvent être trouvés dans la vignette du package *SSRMST* qui aide à comprendre non seulement comment calculer les RMST, mais aussi le groupe d'avantages associés à l'utilisation de mesures basées sur RMST **(Vignette for "SSRMST", n.d.)**.

Le package *SSRMST* est disponible dans logiciel R version 3.4.0.

Résultats

Le **Tableau 1** dans la page suivante résume les caractéristiques et les résultats publiés de cinq essais sur la PrEP décrits ci-dessus.

Tableau 1. Résumé des caractéristiques et des résultats des essais de PrEP.

Etude	Nombre de patients randomisés (N)	Considérations statistiques		Résultats	
		Nombre d'infection	Puissance d'étude	Nombre d'infection	P-value
iPrEX (Grant RM et al. 2010)	N (total)=2499 N (TDF-FTC)=1251 N (placebo)=1248	85	80%	Total=100 Groupe TDF-FTC=36 Groupe Placebo=64	0.005
FEM-PrEP (Van Damme et al.2012)	N (total)=2120 N (TDF-FTC)=1062 N (placebo)=1058	72	90%	Total=68 Groupe TDF-FTC=33 Groupe Placebo=35	0.81
Partners PrEP (Baeten JM et al.2012)	N (total)=4747 couples N (TDF)=1584 couples N (TDF-FTC)=1579 couples N (placebo)=1584 couples	147/ comparaison	80%	Total=82 Groupe TDF=17 Groupe TDF-FTC=13 Groupe Placebo=52	TDF vs. Placebo :<0.001 TDF-FTC vs. Placebo :<0.001
TDF2 (Thigpen MC et al.2012)	N (total)=1219 N (TDF-FTC)=611 N (placebo)=608	Non communiqué dans la publication	80%	Total=33 Groupe TDF-FTC=9 Groupe Placebo=24	0.03
ANRS IPERGAY (Molina JM et al. 2015)	N (total)=414 N (TDF-FTC)=199 N (placebo)=201	64	80%	Total=16 Groupe TDF-FTC=2 Groupe Placebo=14	0.002

Le **Tableau 2** résume les résultats de l'analyse initiale dans la publication finale, ainsi que des analyses alternatives basées sur Δ-RMST avec un temps d'intérêt cliniquement pertinent pour chaque étude. Nous pouvons également trouver dans le tableau le RMST estimé pour chaque critère d'évaluation et chaque groupe randomisé. Nous remarquons que, basé sur les analyses de Δ-RMST, les mêmes conclusions d'efficacité de traitement que l'analyse initiale sont retrouvées pour ces 5 études. Par exemple, dans l'étude iPrEx (**Grant et al., 2010**), l'analyse initiale était basée sur un rapport de taux d'incidence de nouvelles infections du VIH, et une réduction de 44 % [IC à 95 % : 15-63, p=0.005] sur l'incidence du VIH-1 dans le groupe re-

cevant TDF-FTC a été retrouvée, indiquant que cette réduction est statistiquement significative. A titre d'analyse alternative, l'estimation de Δ-RMST pour cette comparaison est de 3.4 semaines [IC à 95 % : 1.5-5.3, p<0.001] pour un suivi de 132 semaines, l'interprétation de cette estimation est la suivante : si la prophylaxie pré-exposition avec TDF-FTC ou placebo est appliquée aux futurs individus à haut risque présentant des caractéristiques conformes à la population de l'étude iPeEx, pour un suivi de 132 semaines, la différence moyenne de temps passé sans infection par le VIH serait d'environ 3,4 semaines en faveur du traitement TDF-FTC, et cette différence est statistiquement significative.

Les résultats du test de Scheoenfeld indiquent que l'hypothèse des risques proportionnels n'est pas violée pour aucune des comparaisons dans ces 5 études de PrEP (p > 0.05 pour toutes les comparaisons). Ces résultats ne sont pas fournis dans le tableau.

Tableau 2. Résumé des analyses originales et des analyses alternatives par Δ-RMST à partir des données reconstituées des essais de PrEP.

Etude	Analyse originale	Estimation de Δ-RMST
iPrEX (Grant RM et al. 2010)	Réduction de l'incidence du VIH: FTC-TDF vs. placebo 44% (IC à 95%, 15 à 63, p=0.005) HR: 0.56 (IC à 95%, 0.37 à 0.85, p=0.005)	Temps=132 semaines FTC-TDF: 128.3 semaines Placebo: 125.0 semaines Δ-RMST: 3.4 [IC à 95%, 1.5 à 5.3, p<0.001]
FEM-PrEP (Van Damme et al.2012)	TDF-FTC vs. placebo: HR: 0.94 (IC à 95%: 0.59 à 1.52, p=0.81)	Temps=52 semaines TDF-FTC: 50.9 semaines placebo: 50.7 semaines Δ-RMST: 0.1 [IC à 95%, -0.5 à 0.7, p=0.70]
Partners PrEP (Baeten JM et al.2012)	Réduction de l'incidence du VIH: TDF vs. Placebo : 67% (IC à 95%, 44 à 81, p<0.001), HR: 0.33 (IC à 95%: 0.19 à 0.56, p<0.001) FTC-TDF vs. placebo 75% (IC à 95%, 55 à 87, p<0.001) HR: 0.25 (IC à 95%: 0.13 à 0.45, p<0.001)	Temps=36 mois TDF: 35.7 mois TDF-FTC: 35.7 mois placebo: 35.1 mois TDF vs. placebo Δ-RMST: 0.6 [IC à 95%, 0.4 à 0.9, p<0.000] TDF-FTC vs. Placebo Δ-RMST: 0.7 [IC à 95%, 0.4 à 1.0, p<0.000]
TDF2 (Thigpen MC et al.2012)	Réduction de l'incidence du VIH: FTC-TDF vs. placebo 62% (IC à 95%, 22 à 83, p=0.03)	Temps=170 semaines TDF-FTC: 166.5 semaines placebo: 161.9 semaines Δ-RMST: 4.6 [IC à 95%, 0.6 à 8.7, p=0.02]
ANRS IPERGAY (Molina JM et al. 2015)	Réduction de l'incidence du VIH: TDF-FTC vs. Placebo : 86% (IC à 95%, 40 à 98; p=0.002)	Temps=24 mois TDF-FTC: 23.9 mois placebo: 22.6 mois Δ-RMST: 1.4 [IC à 95%, 0.6 à 2.2, p<0.001]

HR: le rapport de risques;

Le Tableau 3 représente la taille de l'échantillon estimée dans le protocole basé sur l'analyse de rapport du taux d'incidence, ainsi que la taille de l'échantillon estimée par l'analyse basée sur Δ-RMST. Par exemple, dans le protocole de l'étude FEM-PrEP **(Van Damme et al., 2012)**, 3900 sujets sont ciblés (incluant 72 événements) pour atteindre 90 % de puissance de l'étude afin de détecter une diminution relative de l'incidence de l'infection par le VIH-1 de

70 %. Cependant, pour une analyse basée sur la Δ-RMST, seulement 2496 sujets (incluant 71 événements) seraient nécessaires pour atteindre une puissance de 90 %, ce qui signifie une diminution relative de la taille de l'échantillon de 36 % (**Tableau 3**). Nous pouvons remarquer que, pour une puissance statistique donnée, la taille de l'échantillon requise est moindre avec une analyse basée sur la Δ-RMST dans les 5 essais du tableau.

Tableau 3. Taille de l'échantillon requise dans les essais sur la PrEP, sur la base de l'analyse Δ-RMST

Etude	**Taille de l'échantillon nécessaire en fonction de l'hypothèse du protocole**	**Taille de l'échantillon nécessaire sur la base de l'analyse Δ-RMST**	**Réduction de la taille de l'échantillon basé sur l'analyse Δ-RMST**
iPrEX (Grant RM et al. 2010)	Evénement=85 Puissance=80% Taille de l'échantillon totale=2499**	Evénement=75 Puissance=80% Taille de l'échantillon totale=2200	**12.0%**
FEM-PrEP (Van Damme et al.2012)	Evénement=72 Puissance=90% Taille de l'échantillon totale=3900	Evénement=71 Puissance=90% Taille de l'échantillon totale=2496	**36.0%**
Partners PrEP (Baeten JM et al.2012)	Evénement=147/comparaison Puissance=80% Taille de l'échantillon totale =4700	Evénement=118/comparaison Puissance=80% Taille de l'échantillon totale=3750	**20.0%**
TDF2 (Thigpen MC et al.2012)	Evénement=non communiqué Puissance=80% Taille de l'échantillon totale=1000	Evénement=58 Puissance=80% Taille de l'échantillon totale=780	**22.0%**
ANRS IPERGAY (Molina JM et al. 2015)	Evénement=64 Puissance=80% Taille de l'échantillon totale=1900	Evénement=75 Puissance=80% Taille de l'échantillon totale=1716	**9.7%**

***: Taille de l'échantillon effectivement utilisée dans l'étude (cette hypothèse n'est pas communiquée dans la publication finale)*

Discussion

Le but de notre étude était d'étudier et comparer avec les résultats originaux la performance des analyses basées sur la Δ-RMST dans le contexte des essais de la PrEP. La plupart des essais cliniques sur la PrEP sont conçus et analysés avec un rapport de taux d'incidence ciblé afin d'évaluer l'effet préventif du traitement. Par convention, le rapport du taux d'incidence est généralement considéré comme la fonction d'incidence dans le groupe de traitement divisée par celle du groupe placebo, les valeurs <1 représentant un effet de traitement positif. Par exemple dans le protocole de l'étude ANRS IPERGAY **(Molina et al., 2015)**, le choix du rapport de taux d'incidence ciblé est de 0,50, ce qui implique une réduction relative sur le taux d'incidence de 50 % avec la prophylaxie TDF-FTC après la randomisation. Selon un calcul standard de la taille de l'échantillon sur la base du test Logrank, environ 1900 sujets (incluant 64 événements) sont estimés pour atteindre une puissance de 80 % afin de détecter un tel effet de traitement à un niveau d'alpha bilatéral de 5 %. Comme pour une estimation du rapport de risques avec le modèle de Cox, pour qu'un rapport de taux d'incidence unique ait une interprétation claire, nous devons supposer que le risque proportionnel (PH) de l'effet du traitement est valable, au moins approximativement. Cependant, lorsque l'hypothèse de PH échoue, il est trompeur de rapporter l'effet du traitement sur le rapport de taux d'incidence estimé.

De nombreux chercheurs ont proposé d'estimer le RMST, exprimant l'effet du traitement en tant que Δ-RMST entre les bras randomisés à une période de suivi appropriée car le RMST est une mesure robuste qui est valable quelle que soit la distribution du temps d'événement dans les groupes de traitement **(Kim et al., 2017; Royston and Parmar, 2013; Trinquart et al., 2016; Uno et al., 2015, 2014)**. Bien que le RMST soit considéré comme une mesure non fondée sur une hypothèse des PH et même interprétable lorsque celle-ci est loin d'être correcte **(Uno et al., 2015)**, les résultats de notre étude récente suggèrent que le fait de s'écarter

de l'hypothèse de PH peut avoir une forte influence sur l'analyse basée sur le RMST (Abulizi et al., 2019).

Comme prévu, sur la base des données reconstituées, les analyses de Δ-RMST ont globalement abouti aux mêmes conclusions que les conclusions publiées de ces 5 études de PrEP (**Tableau 2**), indiquant que Δ-RMST est une mesure alternative prometteuse. A noter que les résultats du test de Scheoenfeld indiquent que l'hypothèse des PH n'est violée pour aucune des comparaisons dans ces 5 études. Dans ce contexte, les résultats d'analyse basés sur la Δ-RMST dans ces 5 études nous ont donné l'impression que lorsque l'hypothèse de PH est approximativement satisfaite, le test de l'hypothèse nulle basé sur la Δ-RMST a des caractéristiques de fonctionnement similaires à celles du test de Log-rank. Plus précisément, le niveau de signification des deux tests semble similaire.

D'après les résultats de la taille de l'échantillon et de la puissance d'étude calculés à l'aide du package *SSRMST* (**Tableau 3**), nous pouvons remarquer que la taille de l'échantillon estimée pour une puissance donnée sur la base de Δ-RMST est relativement inférieure à celle estimée sur la base du rapport de taux d'incidence. Ceci explique que dans le **Tableau 2**, dans certaines études les niveaux significatifs de test pour les analyses basées sur Δ-RMST sont très forts malgré des différences relativement faibles dans le RMST, car une taille d'échantillon plus importante que nécessaire pourrait conduire à un essai surpuissant. Par exemple, dans l'étude Partners PrEP **(Baeten et al., 2012)**, sur la base d'une analyse sur le rapport de taux d'incidence, environ 4900 couples (incluant 147 événements pour chaque comparaison) sont estimées pour atteindre une puissance de 80 % afin de détecter une réduction relative sur le taux d'incidence de 60 % avec le traitement. Sur la base de cette hypothèse, l'essai a été mené avec une taille d'échantillon de 4747 couples et les résultats ont montré une réduction relative de l'incidence du VIH-1 de 67 % (IC à 95 %, 44 à 81, p <0,001) avec TDF et de 75 % (IC à 95%, 55 à 87, p <0,001) avec TDF-FTC. Les résultats de l'analyse alternative avec Δ-RMST

ont montré pour un suivi de 36 mois, la Δ-RMST est de 0.6 mois (IC à 95 %,0.4 à 0.9, p<0.001) pour la comparaison TDF versus placebo, est de 0.7 mois (IC à 95 %,0.4 à 1.0, p<0.001) pour la comparaison TDF-FTC versus placebo. Nous remarquons que ces différences sont relativement faibles par rapport à un suivi de 36 mois, mais les niveaux significatifs de test sont très forts, car basés sur la même hypothèse que le protocole, notre estimation sur la base d'une analyse de Δ-RMST montre que seulement une taille d'échantillon de 3750 couples (incluant 118 événements pour chaque comparaison) est nécessaire pour obtenir la même puissance d'étude. Lorsque le nombre de sujets réellement inclus dans l'étude est supérieur au nombre nécessaire, la variance d'analyse est réduite et donc les intervalles de confiance de l'estimation sont plus étroits, de sorte que la valeur p est plus significative.

Bien que le RMST soit une mesure établie de l'avantage en termes de survie qui facilite l'interprétation des résultats en fournissant une mesure statistiquement significative, la sensibilisation de cette mesure est limitée. L'interprétation d'une mesure de l'effet traitement est basée sur une combinaison du raisonnement statistique et du jugement clinique. En d'autres termes, pour les estimations statistiquement significatives, nous devons également déterminer si elles sont cliniquement acceptables. Comme nous l'avons vu précédemment, selon nos résultats sur la base de Δ-RMST dans le **Tableau 2**, pour certaines études l'estimation de Δ-RMST est statistiquement très significative malgré des différences relativement très faibles dans le RMST. Par exemple dans l'étude iPrEx **(Grant et al., 2010)** la Δ-RMST estimée est de 3.4 semaines, qui ne représentent que 2,6 % des 132 semaines de suivi. Sachant que cette estimation est statistiquement très significative, les cliniciens devraient se demander si cette valeur est cliniquement significative aussi et applicable aux futurs individus à haut risque.

Conclusions

La mesure Δ-RMST offre une interprétation simple et statistiquement significative lorsque l'hypothèse de PH est respectée et peut entraîner des réductions considérables de la taille de l'échantillon dans les essais sur la PrEP. Bien que les mesures basées sur le RMST puissent être davantage prises en compte dans la conception et l'analyse d'essais de la PrEP avec un critère d'évaluation dépendant du temps, nous devons nous demander si le bénéfice clinique est également significatif lors de l'interprétation d'une telle mesure.

Références

Abulizi, X., Ribaudo, H.J., Flandre, P., 2019. The Use of the Restricted Mean Survival Time as a Treatment Measure in HIV/AIDS Clinical Trial: Reanalysis of the ACTG A5257 Trial. J. Acquir. Immune Defic. Syndr. 1999 81, 44–51. https://doi.org/10.1097/QAI.0000000000001978

Baeten, J.M., Donnell, D., Ndase, P., Mugo, N.R., Campbell, J.D., Wangisi, J., Tappero, J.W., Bukusi, E.A., Cohen, C.R., Katabira, E., Ronald, A., Tumwesigye, E., Were, E., Fife, K.H., Kiarie, J., Farquhar, C., John-Stewart, G., Kakia, A., Odoyo, J., Mucunguzi, A., Nakku-Joloba, E., Twesigye, R., Ngure, K., Apaka, C., Tamooh, H., Gabona, F., Mujugira, A., Panteleeff, D., Thomas, K.K., Kidoguchi, L., Krows, M., Revall, J., Morrison, S., Haugen, H., Emmanuel-Ogier, M., Ondrejcek, L., Coombs, R.W., Frenkel, L., Hendrix, C., Bumpus, N.N., Bangsberg, D., Haberer, J.E., Stevens, W.S., Lingappa, J.R., Celum, C., 2012. Antiretroviral Prophylaxis for HIV Prevention in Heterosexual Men and Women. N. Engl. J. Med. 367, 399–410. https://doi.org/10.1056/NEJMoa1108524

Bormann, I., n.d. DigitizeIt - Plot Digitizer Software. Digitize graphs, charts and math data. [WWW Document]. URL http://www.digitizeit.de/ (accessed 11.7.17).

Grambsch, P.M., Therneau, T.M., 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81, 515–526. https://doi.org/10.1093/biomet/81.3.515

Grant, R.M., Lama, J.R., Anderson, P.L., McMahan, V., Liu, A.Y., Vargas, L., Goicochea, P., Casapía, M., Guanira-Carranza, J.V., Ramirez-Cardich, M.E., Montoya-Herrera, O., Fernández, T., Veloso, V.G., Buchbinder, S.P., Chariyalertsak, S., Schechter, M., Bekker, L.-G., Mayer, K.H., Kallás, E.G., Amico, K.R., Mulligan, K., Bushman, L.R., Hance, R.J., Ganoza, C., Defechereux, P., Postle, B., Wang, F., McConnell, J.J., Zheng, J.-H., Lee, J., Rooney, J.F., Jaffe, H.S., Martinez, A.I., Burns, D.N., Glidden, D.V., 2010. Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men. N. Engl. J. Med. 363, 2587–2599. https://doi.org/10.1056/NEJMoa1011205

Guyot, P., Ades, A.E., Ouwens, M.J., Welton, N.J., 2012. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med. Res. Methodol. 12, 9.

Kim, D.H., Uno, H., Wei, L.-J., 2017. Restricted Mean Survival Time as a Measure to Interpret Clinical Trial Results. JAMA Cardiol. 2, 1179–1180. https://doi.org/10.1001/jamacardio.2017.2922

McCormack, S., Dunn, D.T., Desai, M., Dolling, D.I., Gafos, M., Gilson, R., Sullivan, A.K., Clarke, A., Reeves, I., Schembri, G., Mackie, N., Bowman, C., Lacey, C.J., Apea, V., Brady, M., Fox, J., Taylor, S., Antonucci, S., Khoo, S.H., Rooney, J., Nardone, A., Fisher, M., McOwan, A., Phillips, A.N., Johnson, A.M., Gazzard, B., Gill, O.N., 2016. Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. The Lancet 387, 53–60. https://doi.org/10.1016/S0140-6736(15)00056-2

Molina, J.-M., Capitant, C., Spire, B., Pialoux, G., Cotte, L., Charreau, I., Tremblay, C., Le Gall, J.-M., Cua, E., Pasquet, A., Raffi, F., Pintado, C., Chidiac, C., Chas, J., Charbonneau, P., Delaugerre, C., Suzan-Monti, M., Loze, B., Fonsart, J., Peytavin, G., Cheret, A., Timsit, J., Girard, G., Lorente, N., Préau, M., Rooney, J.F., Wainberg, M.A., Thompson, D., Rozenbaum, W., Doré, V., Marchand, L., Simon, M.-C., Etien, N., Aboulker, J.-P., Meyer, L., Delfraissy, J.-F., 2015. On-Demand Preexposure Prophylaxis in Men at High Risk for HIV-1 Infection [WWW Document]. http://dx.doi.org/10.1056/NEJMoa1506273. https://doi.org/10.1056/NEJMoa1506273

Royston, P., Parmar, M.K., 2013. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med. Res. Methodol. 13. https://doi.org/10.1186/1471-2288-13-152

Saluja, R., Everest, L., Cheng, S., Cheung, M., Chan, K.K.W., 2019. Assessment of Whether the American Society of Clinical Oncology's Value Framework and the European Society for Medical Oncology's Magnitude of Clinical Benefit Scale Measure Absolute or Relative Clinical Survival Benefit: A Meta-analysis of Randomized Clinical Trials. JAMA Oncol. https://doi.org/10.1001/jamaoncol.2019.0818

SSRMST: vignettes/vignette-ssrmst.Rmd [WWW Document], n.d. URL https://rdrr.io/cran/SSRMST/f/vignettes/vignette-ssrmst.Rmd (accessed 4.4.19).

Thigpen, M.C., Kebaabetswe, P.M., Paxton, L.A., Smith, D.K., Rose, C.E., Segolodi, T.M., Henderson, F.L., Pathak, S.R., Soud, F.A., Chillag, K.L., Mutanhaurwa, R., Chirwa, L.I., Kasonde, M., Abebe, D., Buliva, E., Gvetadze, R.J., Johnson, S., Sukalac, T., Thomas, V.T., Hart, C., Johnson, J.A., Malotte, C.K., Hendrix, C.W., Brooks, J.T., 2012. Antiretroviral Preexposure Prophylaxis for Heterosexual HIV Transmission in Botswana. N. Engl. J. Med. 367, 423–434. https://doi.org/10.1056/NEJMoa1110711

Trinquart, L., Bill-Axelson, A., Rider, J.R., 2019. Restricted Mean Survival Times to Improve Communication of Evidence from Cancer Randomized Trials and Observational Studies. Eur. Urol. https://doi.org/10.1016/j.eururo.2019.04.002

Trinquart, L., Jacot, J., Conner, S.C., Porcher, R., 2016. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. J. Clin. Oncol. 34, 1813–1819. https://doi.org/10.1200/JCO.2015.64.2488

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., Wei, L.-J., 2014. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J. Clin. Oncol. 32, 2380–2385. https://doi.org/10.1200/JCO.2014.55.2208

Uno, H., Wittes, J., Fu, H., Solomon, S.D., Claggett, B., Tian, L., Cai, T., Pfeffer, M.A., Evans, S.R., Wei, L.-J., 2015. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. Ann. Intern. Med. 163, 127. https://doi.org/10.7326/M14-1741

Van Damme, L., Corneli, A., Ahmed, K., Agot, K., Lombaard, J., Kapiga, S., Malahleha, M., Owino, F., Manongi, R., Onyango, J., Temu, L., Monedi, M.C., Mak'Oketch, P., Makanda, M., Reblin, I., Makatu, S.E., Saylor, L., Kiernan, H., Kirkendale, S., Wong, C., Grant, R., Kashuba, A., Nanda, K., Mandala, J., Fransen, K., Deese, J., Crucitti, T., Mastro, T.D., Taylor, D., 2012. Preexposure Prophylaxis for HIV Infection among African Women. N. Engl. J. Med. 367, 411–422. https://doi.org/10.1056/NEJMoa1202614

CHAPITRE VII – DISCUSSION GENERALE ET PERSPECTIVES

1. Discussion des résultats

1.1 Choix de la mesure de l'effet traitement lorsque les bornes de non-infériorité sont originalement définies par une différence absolue transformée en une différence relative a influencé les résultats des essais cliniques

L'objectif de ce travail était d'étudier les différences entre les analyses basées sur le rapport de risques, la DR estimée par les courbes de Kaplan-Meier (DR_{KM}) et la différence de RMST (Δ-RMST) dans un essai d'équivalence réelle et dans une étude de simulation. L'analyse basée sur le rapport de risques a été considérée comme l'analyse de référence pour calculer la taille des échantillons. Ce travail a été motivé par un essai récent d'équivalence, conduisant à des conclusions différentes avec l'utilisation d'estimations de rapport de risque ou RD_{KM} **(Daar et al., 2011; Flandre, 2015)**. Nos résultats basés sur des données reconstituées ont montré qu'une analyse basée sur Δ-RMST aurait également conduit à une conclusion différente de l'analyse principale de l'essai basée sur une estimation de rapport de risques. De telles différences entre les mesures d'effet doivent être soulignées et mieux comprises, car elles expliquent potentiellement les différences de résultats observés entre les essais utilisant différentes mesures. Bien entendu, les conclusions d'un essai doivent être basées sur l'analyse définie dans le protocole. Des conclusions similaires, toutefois, basées sur différentes mesures renforcent les conclusions d'un essai, en particulier lorsque les bornes de non-infériorité sont basées sur des hypothèses équivalentes.

Avec les distributions et les paramètres de forme étudiés, nos résultats ont montré que la violation de l'hypothèse de risque constant avait moins d'impact sur la différence observée entre les analyses que l'hypothèse erronée sur le taux d'échec. Néanmoins, la différence due à la forme de la distribution augmente avec des taux d'échec faibles et un risque croissant. En par-

ticulier, une surestimation des taux d'échec conduirait probablement à ne pas démontrer la non-infériorité avec une analyse basée sur le rapport de risques même avec une faible différence entre les deux bras. En revanche, la probabilité de démontrer la non-infériorité est beaucoup plus grande avec une analyse basée sur le rapport de risques qu'avec les deux autres analyses dans une situation de sous-estimation des taux d'échec.

Il ressort des différentes situations de nos études de simulation que l'analyse basée sur la RD_{KM} fournit des résultats plus stables qui s'écartent des hypothèses de niveaux de taux d'échec ou d'une distribution incorrecte du temps d''échec. Bien que d'autres études de simulation puissent être réalisées avec d'autres situations, nos résultats nous ont conduits à recommander l'analyse RD_{KM} dans le cadre de ce travail. Cette recommandation n'est valide que dans notre contexte où les bornes de non-infériorité pour une mesure de DR sont traduites en une mesure de rapport de risques. Notre recommandation ne couvre pas les autres situations dans lesquelles les bornes initiales données pour les analyses de rapport de risques ou RMST seraient transformées en mesures de DR. Des études complémentaires sont nécessaires pour étudier ces dernières situations.

En général, l'équipe de l'ACTG a utilisé des analyses basées sur le rapport de risques alors que la RD_{KM} a été utilisé dans une étude récente : l'essai ACTG 5257 **(Lennox et al., 2014)** . Les essais ACTG sont très bien conçus et ont une influence principale sur les directives internationales relatives à l'utilisation d'agents antirétroviraux. L'équipe ACTG 5202 a déjà indiqué dans la publication finale que le faible taux d'échec virologique observé (11,17%) dans leurs données par rapport à l'hypothèse présumée (32%) pourrait expliquer que l'équivalence n'a pas été déclarée **(Daar et al., 2011)**. La surestimation des taux d'échec ou la sous-estimation des taux de succès a déjà été décrite dans des essais de non-infériorité du VIH, conduisant à des essais surpuissants **(Flandre, 2014)**. Nous pensons que ce point a également influencé la conception de l'étude ACTG 5257 car cet essai a utilisé une analyse basée sur le

RD_{KM} pour rechercher l'équivalence entre trois schémas thérapeutiques antirétroviraux **(Lennox et al., 2014)**. Les résultats de cette étude ont montré que le taux d'échec présumé est également surestimé, que cette surestimation soit moins important que l'étude ACTG 5202. On peut donc penser qu'une analyse basée sur le rapport de risques dans l'étude ACTG 5257 ne conclurait probablement pas à l'équivalence, alors que l'équivalence a été démontrée avec analyses basées sur RD_{KM} **(Lennox et al., 2014)**.

Une limite de notre travail est que d'autres distributions d'échecs auraient pu être étudiées, conduisant peut-être à des résultats légèrement différents. Cependant, les distributions et les paramètres de forme utilisés dans nos études de simulation explorent des situations courantes de risques constants, croissants monotones et décroissants monotones. L'utilisation de la distribution exponentielle par intervalle avec un risque croissant ou décroissant monotone conduit à des résultats similaires à ceux obtenus avec la distribution de Weibull (les résultats sont présentés en tant que contenu supplémentaire au Chapitre IV).

Pour conclure, d'après les résultats de notre étude par simulation, nous avons montré que la discordance entre les analyses basées sur les trois mesures d'effet est beaucoup plus influencée par des taux d'échec hypothétiques erronés que par la forme de la distribution du temps d'échec. Lorsqu'une analyse basée sur le rapport de risques a été conçue dans le protocole, une surestimation des taux d'échec augmenterait la probabilité de ne pas conclure à la non-infériorité, même en l'absence de différence entre les deux bras. Inversement, une sous-estimation des taux d'échec présumés tendrait à déclarer la non-infériorité même en présence d'une différence ultérieure entre les taux d'échec du traitement. Lorsque les bornes sont définies à l'origine pour la DR, les analyses basées sur RD_{KM} pourraient être préférables, car les résultats obtenus avec cette mesure sont moins influencés par des niveaux erronés de taux d'échec menant à des résultats plus stables.

1.2 L'utilisation du temps de survie moyen borné (RMST) comme mesure de l'effet traitement dans les essais cliniques sur le VIH/SIDA : ré-analyse de l'essai ACTG 5257

L'objectif de cette étude était de comparer les analyses basées sur DR_{KM}, le rapport de risques et Δ-RMST en tant que mesures d'effet traitement dans l'étude ACTG 5257 et plus particulièrement d'étudier les performances et les caractéristiques de fonctionnement de l'analyse basée sur Δ-RMST dans le contexte des risques non-proportionnels et non constant dans une étude de simulation. Comme prévu, les analyses de Δ-RMST ont globalement abouti à des conclusions similaires à celles des conclusions publiées de l'A5257 basées sur RD_{KM}. En revanche, les analyses basées sur le rapport de risques ont fourni des conclusions discordantes sur l'équivalence comparées à la fois avec l'analyse initiale basée sur la RD_{KM} et la Δ-RMST. Il est intéressant de noter que globalement, même dans le cas de différences de groupe estimées avec RD_{KM} et Δ-RMST sont très minimes et cliniquement non significatives, l'inférence basée sur le rapport de risques ne parvient toujours pas à la conclusion d'équivalence, et nos résultats d'analyses alternatives des données de l'ACTG 5257 ont confirmé cette hypothèse.

Notre étude de simulation s'est concentrée sur la performance de l'analyse basée sur Δ-RMST dans le contexte des risques non-proportionnels et non constants. Les analyses basées sur le RD_{KM} et le rapport de risques n'ont pas été évaluées, car l'estimation du RD_{KM} ne changera pas en raison de la différence du taux d'échec à la semaine 96 est la même, quelle que soit la distribution du temps d'échec (voir les figures 1-4, appendice au Chapitre V) , et l'estimation de rapport de risques n'est pas appropriée dans le cas des risques non proportionnels. L'analyse de RMST a été présentée comme une alternative prometteuse dans le cas des risques non proportionnels. Notre étude de simulation montre cependant que la discordance entre les hypothèses de la conception de l'étude et les données observées peut avoir un impact important sur les résultats basés sur les analyses de Δ-RMST. En particulier, dans certaines

situations, la violation de l'hypothèse de risques proportionnels augmente considérablement la probabilité de non-infériorité erronée. Ensuite, bien que l'estimation de RMST soit valide dans le contexte de risques non proportionnels, les résultats sont grandement altérés dans une telle situation.

En bref, selon les résultats d'analyses alternatives de l'étude ACTG 5257, nous avons montré que les analyses basées sur le Δ-RMST aboutissaient globalement à des conclusions similaires à celles des résultats publiés de l'A5257 basés sur RD_{KM}. En revanche, les analyses fondées sur le rapport de risques ont fourni des conclusions discordantes sur l'équivalence. Une telle discordance s'explique principalement par la violation de l'hypothèse de risque proportionnels et par surestimation des taux d'échec présumés dans la conception de l'étude (**Abulizi and Flandre, 2018**). Bien qu'il soit avantageux d'utiliser le RMST comme mesure de l'effet traitement,, des discussions supplémentaires sont nécessaires avec les cliniciens dans le domaine de VIH / SIDA pour déterminer quelle estimation est le mieux adapté à la question clinique et, le cas échéant, pour suggérer des bornes d'équivalence. Nous montrons également que la puissance et le taux de faux positifs fournis par les analyses basées sur RMST sont sensibles aux écarts par rapport à l'hypothèse de risques proportionnels.

1.3 L'utilisation du temps de survie moyen borné (RMST) comme mesure de l'effet traitement dans les essais sur la prophylaxie pré-exposition (PrEP)

Comme nous l'avons discuté dans le Chapitre VI, nos résultats montrent que la mesure Δ-RMST offre une interprétation simple et significative dans le cas où l'hypothèse de risques proportionnels est respectée et peut entraîner des réductions considérables de la taille de l'échantillon dans les essais cliniques sur la PrEP. Cependant, lorsque les mesures basées sur le RMST sont prises en compte dans la conception et l'analyse d'essais de la PrEP,

l'interprétation de cette mesure de l'effet traitement devrait baser sur une combinaison des considérations statistiques et cliniques.

2. Enjeux futures de la recherche

L'objectif principal de cette recherche est de comparer différentes mesures de l'effet traitement afin de choisir la plus appropriée dans les essais cliniques de patients infectés par le VIH. Ce travail montre que les différences entre les analyses basées sur différentes mesures sont beaucoup plus affectées par des taux d'échec hypothétiques erronés que par la forme de la distribution du temps d'l'échec. Autrement dit, les résultats montrent que la supériorité d'une ou l'autre mesure dépend de la validité des hypothèses qui sont faites sur les données mais ces résultats permettront dans des futurs essais de faire un choix éclairé.

Un autre résultat important qui ressort de ce travail concerne la nouvelle mesure largement recommandée dans les essais cliniques de non-infériorité, la mesure de temps de survie moyen borné (RMST). Bien que l'analyse basée sur le RMST soit une mesure alternative de l'efficacité dans les essais du VIH / SIDA, une telle mesure peut être fortement affectée par un écart par rapport à l'hypothèse de risques proportionnels. En outre, l'interprétation de cette mesure de l'effet traitement doit être effectuée en collaboration avec les cliniciens afin que les résultats soient plus appropriés et pertinente sur le plan clinique.

Globalement le domaine de recherche de ce travail concerne la méthodologie statistique des essais du VIH/SIDA, les résultats s'appliquent essentiellement à des essais du VIH qui doivent permettre d'améliorer la connaissance dans les choix méthodologiques lors de la construction d'un essai thérapeutique incluant des patients infectés par le VIH. Ces résultats permettra en particulier de mieux interpréter les résultats des essais cliniques sur le VIH à partir des critères principaux utilisés, des mesures de l'effet traitement utilisées et des hypothèses de

ces mesures, et de proposer des recommandations sur l'utilisation des mesures de l'effet en fonction des objectifs et des caractéristiques d'un essai.

BIBLIOGRAPHIE

Abulizi, X., Flandre, P., 2018. Choice of treatment-effect measures when noninferiority margins originally defined in absolute difference translated into relative difference influenced the results of clinical trials. J Clin Epidemiol 96, 63–72. https://doi.org/10.1016/j.jclinepi.2017.12.010

Abulizi, X., Ribaudo, H.J., Flandre, P., 2019. The Use of the Restricted Mean Survival Time as a Treatment Measure in HIV/AIDS Clinical Trial: Reanalysis of the ACTG A5257 Trial. J. Acquir. Immune Defic. Syndr. 1999 81, 44–51. https://doi.org/10.1097/QAI.0000000000001978

Akobeng, A.K., 2005. Understanding measures of treatment effect in clinical trials. Archives of Disease in Childhood 90, 54–56. https://doi.org/10.1136/adc.2004.052233

Austin, P.C., 2010. Different measures of treatment effect for different research questions. J Clin Epidemiol 63, 9–10. https://doi.org/10.1016/j.jclinepi.2009.07.006

Baeten, J.M., Donnell, D., Ndase, P., Mugo, N.R., Campbell, J.D., Wangisi, J., Tappero, J.W., Bukusi, E.A., Cohen, C.R., Katabira, E., Ronald, A., Tumwesigye, E., Were, E., Fife, K.H., Kiarie, J., Farquhar, C., John-Stewart, G., Kakia, A., Odoyo, J., Mucunguzi, A., Nakku-Joloba, E., Twesigye, R., Ngure, K., Apaka, C., Tamooh, H., Gabona, F., Mujugira, A., Panteleeff, D., Thomas, K.K., Kidoguchi, L., Krows, M., Revall, J., Morrison, S., Haugen, H., Emmanuel-Ogier, M., Ondrejcek, L., Coombs, R.W., Frenkel, L., Hendrix, C., Bumpus, N.N., Bangsberg, D., Haberer, J.E., Stevens, W.S., Lingappa, J.R., Celum, C., 2012. Antiretroviral Prophylaxis for HIV Prevention in Heterosexual Men and Women. N. Engl. J. Med. 367, 399–410. https://doi.org/10.1056/NEJMoa1108524

Blackwelder, W.C., 1982. “Proving the null hypothesis” in clinical trials. Controlled clinical trials 3, 345–353.

Bormann, I., n.d. DigitizeIt - Plot Digitizer Software. Digitize graphs, charts and math data. [WWW Document]. URL http://www.digitizeit.de/ (accessed 11.7.17).

Bouyer, J., 2009. Épidémiologie: principes et méthodes quantitatives. Lavoisier.

Case, L.D., 2002. Interpreting Measures of Treatment Effect in Cancer Clinical Trials. The Oncologist 7, 181–187. https://doi.org/10.1634/theoncologist.7-3-181

Daar, E.S., Tierney, C., Fischl, M.A., Sax, P.E., Mollan, K., Budhathoki, C., Godfrey, C., Jahed, N.C., Myers, L., Katzenstein, D., others, 2011. Atazanavir plus ritonavir or efavirenz as part of a 3-drug regimen for initial treatment of HIV-1A randomized trial. Annals of internal medicine 154, 445–456.

Deeks, J.J., 2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Statistics in Medicine 21, 1575–1600. https://doi.org/10.1002/sim.1188

DeJesus, E., Herrera, G., Teofilo, E., Gerstoft, J., Buendia, C.B., Brand, J.D., Brothers, C.H., Hernandez, J., Castillo, S.A., Bonny, T., Lanier, E.R., Scott, T.R., 2004. Abacavir versus Zidovudine Combined with Lamivudine and Efavirenz, for the Treatment of Antiretroviral-Naive HIV-Infected Adults. Clin Infect Dis 39, 1038–1046. https://doi.org/10.1086/424009

Elbarbary, M., 2010. Understanding and expressing “Risk.” J Saudi Heart Assoc 22, 159–164. https://doi.org/10.1016/j.jsha.2010.04.002

Fine, J.P., Gray, R.J., 1999. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association 94, 496. https://doi.org/10.2307/2670170

Fischl, M.A., Collier, A.C., Mukherjee, A.L., Feinberg, J.E., Demeter, L.M., Tebas, P., Giuliano, M., Dehlinger, M., Garren, K., Brizz, B., Bassett, R., 2007. Randomized open-

label trial of two simplified, class-sparing regimens following a first suppressive three or four-drug regimen. AIDS 21, 325–333. https://doi.org/10.1097/QAD.0b013e328011ddfa

Flandre, P., 2015. Choice of effect measure in HIV randomized trials: AIDS 29, 2057–2060. https://doi.org/10.1097/QAD.0000000000000820

Flandre, P., 2014. Some current issues in the design of Hiv noninferiority trials. Aids 28, 1921–1929. https://doi.org/10.1097/QAD.0000000000000369

Fleming, T.R., 2008. Current issues in non-inferiority trials. Statistics in Medicine 27, 317–332. https://doi.org/10.1002/sim.2855

Gallant, J.E., Daar, E.S., Raffi, F., Brinson, C., Ruane, P., DeJesus, E., Johnson, M., Clumeck, N., Osiyemi, O., Ward, D., Morales-Ramirez, J., Yan, M., Abram, M.E., Plummer, A., Cheng, A.K., Rhee, M.S., 2016. Efficacy and safety of tenofovir alafenamide versus tenofovir disoproxil fumarate given as fixed-dose combinations containing emtricitabine as backbones for treatment of HIV-1 infection in virologically suppressed adults: a randomised, double-blind, active-controlled phase 3 trial. Lancet HIV 3, e158-165. https://doi.org/10.1016/S2352-3018(16)00024-2

Gallant, J.E., Pozniak, A.L., McColl, D., Toole, J.J., 2006. Tenofovir DF, Emtricitabine, and Efavirenz vs. Zidovudine, Lamivudine, and Efavirenz for HIV. The New England Journal of Medicine 10.

Gallant, J.E., Staszewski, S., Pozniak, A.L., DeJesus, E., Suleiman, J.M.A.H., Miller, M.D., Coakley, D.F., Lu, B., Toole, J.J., Cheng, A.K., Group, for the 903 S., 2004. Efficacy and Safety of Tenofovir DF vs Stavudine in Combination Therapy in Antiretroviral-Naive Patients: A 3-Year Randomized Trial. JAMA 292, 191–201. https://doi.org/10.1001/jama.292.2.191

Grambsch, P.M., Therneau, T.M., 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81, 515–526. https://doi.org/10.1093/biomet/81.3.515

Grant, R.M., Lama, J.R., Anderson, P.L., McMahan, V., Liu, A.Y., Vargas, L., Goicochea, P., Casapía, M., Guanira-Carranza, J.V., Ramirez-Cardich, M.E., Montoya-Herrera, O., Fernández, T., Veloso, V.G., Buchbinder, S.P., Chariyalertsak, S., Schechter, M., Bekker, L.-G., Mayer, K.H., Kallás, E.G., Amico, K.R., Mulligan, K., Bushman, L.R., Hance, R.J., Ganoza, C., Defechereux, P., Postle, B., Wang, F., McConnell, J.J., Zheng, J.-H., Lee, J., Rooney, J.F., Jaffe, H.S., Martinez, A.I., Burns, D.N., Glidden, D.V., 2010. Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men. N. Engl. J. Med. 363, 2587–2599. https://doi.org/10.1056/NEJMoa1011205

Gray, R.J., 1988. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. The Annals of Statistics 16, 1141–1154. https://doi.org/10.1214/aos/1176350951

Gulick, R.M., Ribaudo, H.J., Shikuma, C.M., Lalama, C., Schackman, B.R., Meyer, W.A., Acosta, E.P., Schouten, J., Squires, K.E., Pilcher, C.D., 2006. Three-vs four-drug antiretroviral regimens for the initial treatment of HIV-1 infection: a randomized controlled trial. Jama 296, 769–781.

Gulick, R.M., Ribaudo, H.J., Shikuma, C.M., Lustgarten, S., Squires, K.E., Meyer III, W.A., Acosta, E.P., Schackman, B.R., Pilcher, C.D., Murphy, R.L., 2004. Triple-nucleoside regimens versus efavirenz-containing regimens for the initial treatment of HIV-1 infection. New England Journal of Medicine 350, 1850–1861.

Guyot, P., Ades, A.E., Ouwens, M.J., Welton, N.J., 2012. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC medical research methodology 12, 9.

Ha, B., Liao, Q.M., Dix, L.P., Pappa, K.A., 2009. Virologic response and safety of the abacavir/lamivudine fixed-dose formulation as part of highly active antiretroviral therapy: analyses of six clinical studies. HIV Clin Trials 10, 65–75. https://doi.org/10.1310/hct1002-65

Hernán, M.A., 2010. The Hazards of Hazard Ratios: Epidemiology 21, 13–15. https://doi.org/10.1097/EDE.0b013e3181c1ea43

Hill, A., Sabin, C., 2008. Designing and interpreting HIV noninferiority trials in naive and experienced patients. AIDS 22, 913–921. https://doi.org/10.1097/QAD.0b013e3282f5556d

Kim, D.H., Uno, H., Wei, L.-J., 2017. Restricted Mean Survival Time as a Measure to Interpret Clinical Trial Results. JAMA Cardiol. 2, 1179–1180. https://doi.org/10.1001/jamacardio.2017.2922

Lennox, J.L., Landovitz, R.J., Ribaudo, H.J., Ofotokun, I., Na, L.H., Godfrey, C., Kuritzkes, D.R., Sagar, M., Brown, T.T., Cohn, S.E., McComsey, G.A., Aweeka, F., Fichtenbaum, C.J., Presti, R.M., Koletar, S.L., Haas, D.W., Patterson, K.B., Benson, C.A., Baugh, B.P., Leavitt, R.Y., Rooney, J.F., Seekins, D., Currier, J.S., 2014. Efficacy and Tolerability of 3 Nonnucleoside Reverse Transcriptase Inhibitor–Sparing Antiretroviral Regimens for Treatment-Naive Volunteers Infected With HIV-1: A Randomized, Controlled Equivalence Trial. Annals of Internal Medicine 161, 461. https://doi.org/10.7326/M14-1084

Mani, N., Murray, J., Gulick, R.M., Josephson, F., Miller, V., Miele, P., Strobos, J., Struble, K., 2012. Novel clinical trial designs for the development of new antiretroviral agents. AIDS 26, 899–907. https://doi.org/10.1097/QAD.0b013e3283519371

Molina, J.-M., Andrade-Villanueva, J., Echevarria, J., Chetchotisakd, P., Corral, J., David, N., Moyle, G., Mancini, M., Percival, L., Yang, R., Wirtz, V., Lataillade, M., Absalon, J., McGrath, D., 2010. Once-Daily Atazanavir/Ritonavir Compared With Twice-Daily Lopinavir/Ritonavir, Each in Combination With Tenofovir and Emtricitabine, for Management of Antiretroviral-Naive HIV-1-Infected Patients: 96-Week Efficacy and Safety Results of the CASTLE Study: JAIDS Journal of Acquired Immune Deficiency Syndromes 53, 323–332. https://doi.org/10.1097/QAI.0b013e3181c990bf

Molina, J.-M., Capitant, C., Spire, B., Pialoux, G., Cotte, L., Charreau, I., Tremblay, C., Le Gall, J.-M., Cua, E., Pasquet, A., Raffi, F., Pintado, C., Chidiac, C., Chas, J., Charbonneau, P., Delaugerre, C., Suzan-Monti, M., Loze, B., Fonsart, J., Peytavin, G., Cheret, A., Timsit, J., Girard, G., Lorente, N., Préau, M., Rooney, J.F., Wainberg, M.A., Thompson, D., Rozenbaum, W., Doré, V., Marchand, L., Simon, M.-C., Etien, N., Aboulker, J.-P., Meyer, L., Delfraissy, J.-F., 2015. On-Demand Preexposure Prophylaxis in Men at High Risk for HIV-1 Infection [WWW Document]. http://dx.doi.org/10.1056/NEJMoa1506273. https://doi.org/10.1056/NEJMoa1506273

Newcombe, R.G., 2006. A deficiency of the odds ratio as a measure of effect size. Statistics in Medicine 25, 4235–4240. https://doi.org/10.1002/sim.2683

Noordzij, M., van Diepen, M., Caskey, F.C., Jager, K.J., 2017. Relative risk versus absolute risk: one cannot be interpreted without the other. Nephrol Dial Transplant 32, ii13–ii18. https://doi.org/10.1093/ndt/gfw465

Orkin, C., Squires, K.E., Molina, J.-M., Sax, P.E., Wong, W.-W., Sussmann, O., Kaplan, R., Lupinacci, L., Rodgers, A., Xu, X., Lin, G., Kumar, S., Sklar, P., Nguyen, B.-Y., Hanna, G.J., Hwang, C., Martin, E.A., DRIVE-AHEAD Study Group, 2019. Doravirine/Lamivudine/Tenofovir Disoproxil Fumarate is Non-inferior to Efavirenz/Emtricitabine/Tenofovir Disoproxil Fumarate in Treatment-naive Adults With Human Immunodeficiency Virus-1 Infection: Week 48 Results of the DRIVE-AHEAD Trial. Clin. Infect. Dis. 68, 535–544. https://doi.org/10.1093/cid/ciy540

Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents.Department of Health and Human Service. [WWW Document], n.d. URL https://aidsinfo.nih.gov/contentfiles/lvguidelines/adultandadolescentgl.pdf (accessed 5.2.17).

Papageorgiou, S.N., Tsiranidou, E., Antonoglou, G.N., Deschner, J., Jäger, A., 2015. Choice of effect measure for meta-analyses of dichotomous outcomes influenced the identified heterogeneity and direction of small-study effects. Journal of Clinical Epidemiology 68, 534–541. https://doi.org/10.1016/j.jclinepi.2015.01.004

Piaggio, G., Elbourne, D.R., Altman, D.G., Pocock, S.J., Evans, S.J.W., CONSORT Group, for the, 2006. Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement. JAMA 295, 1152. https://doi.org/10.1001/jama.295.10.1152

Raffi, F., Babiker, A.G., Richert, L., Molina, J.-M., George, E.C., Antinori, A., Arribas, J.R., Grarup, J., Hudson, F., Schwimmer, C., 2014. Ritonavir-boosted darunavir combined with raltegravir or tenofovir–emtricitabine in antiretroviral-naive adults infected with HIV-1: 96 week results from the NEAT001/ANRS143 randomised non-inferiority trial. The Lancet 384, 1942–1951.

Riddler, S.A., Haubrich, R., DiRienzo, A.G., Peeples, L., Powderly, W.G., Klingman, K.L., Garren, K.W., George, T., Rooney, J.F., Brizz, B., Lalloo, U.G., Murphy, R.L., Swindells, S., Havlir, D., Mellors, J.W., 2008. Class-Sparing Regimens for Initial Treatment of HIV-1 Infection. N Engl J Med 358. https://doi.org/10.1056/NEJMoa074609

Rothman, K.J., Greenland, S., Lash, T.L., 2008. Modern Epidemiology. Lippincott Williams & Wilkins.

Royston, P., Parmar, M.K., 2013. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology 13. https://doi.org/10.1186/1471-2288-13-152

Royston, P., Parmar, M.K.B., 2011. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Statistics in Medicine 30, 2409–2421. https://doi.org/10.1002/sim.4274

Sax, P.E., Tierney, C., Collier, A.C., Fischl, M.A., Mollan, K., Peeples, L., Godfrey, C., Jahed, N.C., Myers, L., Katzenstein, D., Farajallah, A., Rooney, J.F., Ha, B., Woodward, W.C., Koletar, S.L., Johnson, V.A., Geiseler, P.J., Daar, E.S., 2009. Abacavir–Lamivudine versus Tenofovir–Emtricitabine for Initial HIV-1 Therapy. New England Journal of Medicine 361, 2230–2240. https://doi.org/10.1056/NEJMoa0906768

Saluja, R., Everest, L., Cheng, S., Cheung, M., Chan, K.K.W., 2019. Assessment of Whether the American Society of Clinical Oncology's Value Framework and the European Society for Medical Oncology's Magnitude of Clinical Benefit Scale Measure Absolute or Relative Clinical Survival Benefit: A Meta-analysis of Randomized Clinical Trials. JAMA Oncol. https://doi.org/10.1001/jamaoncol.2019.0818

Seruga, B., Pond, G.R., Hertz, P.C., Amir, E., Ocana, A., Tannock, I.F., 2012. Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. Annals of Oncology 23, 2977–2982. https://doi.org/10.1093/annonc/mds174

Schechtman, E., 2002. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat--which of these should we use? Value Health 5, 431–436. https://doi.org/10.1046/J.1524-4733.2002.55150.x

Smith, K.Y., Patel, P., Fine, D., Bellos, N., Sloan, L., Lackey, P., Kumar, P.N., Sutherland-phillips, D.H., Vavro, C., Yau, L., Wannamaker, P., Shaefer, M.S., 2009. Randomized, double-blind, placebo-matched, multicenter trial of abacavir/lamivudine or

tenofovir/emtricitabine with lopinavir/ritonavir for initial Hiv treatment. Aids 23, 1547–1556. https://doi.org/10.1097/QAD.0b013e32832cbcc2
Spruance, S.L., Reid, J.E., Grace, M., Samore, M., 2004. Hazard Ratio in Clinical Trials. Antimicrob Agents Chemother 48, 2787–2792. https://doi.org/10.1128/AAC.48.8.2787-2792.2004
SSRMST: vignettes/vignette-ssrmst.Rmd [WWW Document], n.d. URL https://rdrr.io/cran/SSRMST/f/vignettes/vignette-ssrmst.Rmd (accessed 4.4.19).
Thigpen, M.C., Kebaabetswe, P.M., Paxton, L.A., Smith, D.K., Rose, C.E., Segolodi, T.M., Henderson, F.L., Pathak, S.R., Soud, F.A., Chillag, K.L., Mutanhaurwa, R., Chirwa, L.I., Kasonde, M., Abebe, D., Buliva, E., Gvetadze, R.J., Johnson, S., Sukalac, T., Thomas, V.T., Hart, C., Johnson, J.A., Malotte, C.K., Hendrix, C.W., Brooks, J.T., 2012. Antiretroviral Preexposure Prophylaxis for Heterosexual HIV Transmission in Botswana. N. Engl. J. Med. 367, 423–434. https://doi.org/10.1056/NEJMoa1110711
Thompson, M.A., Aberg, J.A., 2010. Antiretroviral treatment of adult HIV infection: 2010 Recommendations of the international AIDS Society-USA Panel.
Trinquart, L., Bill-Axelson, A., Rider, J.R., 2019. Restricted Mean Survival Times to Improve Communication of Evidence from Cancer Randomized Trials and Observational Studies. European Urology 76, 137–139. https://doi.org/10.1016/j.eururo.2019.04.002
Trinquart, L., Jacot, J., Conner, S.C., Porcher, R., 2016. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. Journal of Clinical Oncology 34, 1813–1819. https://doi.org/10.1200/JCO.2015.64.2488
Tripepi, G., Jager, K.J., Dekker, F.W., Wanner, C., Zoccali, C., 2007. Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat.' Kidney international 72, 789–791.
Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., Wei, L.-J., 2014. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol 32, 2380–2385. https://doi.org/10.1200/JCO.2014.55.2208
Uno, H., Wittes, J., Fu, H., Solomon, S.D., Claggett, B., Tian, L., Cai, T., Pfeffer, M.A., Evans, S.R., Wei, L.-J., 2015. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. Annals of Internal Medicine 163, 127. https://doi.org/10.7326/M14-1741
Van Damme, L., Corneli, A., Ahmed, K., Agot, K., Lombaard, J., Kapiga, S., Malahleha, M., Owino, F., Manongi, R., Onyango, J., Temu, L., Monedi, M.C., Mak'Oketch, P., Makanda, M., Reblin, I., Makatu, S.E., Saylor, L., Kiernan, H., Kirkendale, S., Wong, C., Grant, R., Kashuba, A., Nanda, K., Mandala, J., Fransen, K., Deese, J., Crucitti, T., Mastro, T.D., Taylor, D., 2012. Preexposure Prophylaxis for HIV Infection among African Women. N. Engl. J. Med. 367, 411–422. https://doi.org/10.1056/NEJMoa1202614
Venter, W.D.F., Kambugu, A., Chersich, M.F., Becker, S., Hill, A., Arulappan, N., Moorhouse, M., Majam, M., Akpomiemie, G., Sokhela, S., Poongulali, S., Feldman, C., Duncombe, C., Ripin, D.H.B., Vos, A., Kumarasamy, N., 2019. Efficacy and Safety of Tenofovir Disoproxil Fumarate Versus Low-Dose Stavudine Over 96 Weeks: A Multicountry Randomized, Noninferiority Trial. J. Acquir. Immune Defic. Syndr. 80, 224–233. https://doi.org/10.1097/QAI.0000000000001908
Walter, S.., 2000. Choice of effect measure for epidemiological data. Journal of Clinical Epidemiology 53, 931–939. https://doi.org/10.1016/S0895-4356(00)00210-9

WHO | Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection [WWW Document], n.d. . WHO. URL http://www.who.int/hiv/pub/guidelines/arv2013/download/en/ (accessed 1.14.19).

Zhao, L., Tian, L., Uno, H., Solomon, S.D., Pfeffer, M.A., Schindler, J.S., Wei, L.J., 2012. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. Clinical Trials 9, 570–577. https://doi.org/10.1177/1740774512455464

ANNEXES

Algorithme pour créer un ensemble de données brutes avec le logiciel DigizeIt à partir d'une courbe de Kaplan-Meier.

```
#Algorithm to create a raw dataset from DigizeIt readings from a Kaplan-Meier curve

library("MASS")
library("splines")
library("survival")

###FUNCTION INPUTS
path<-"C:\\PHD\\algorithm\\reliability exercice\\"
digisurvfile<-"data initials study2 figA arm1 time1.txt"                #Input survival times from graph
reading
nriskfile<-"nrisk study2 figA arm1 time1.txt"                           #Input reported number at risk
KMdatafile<-"KMdata study2 figA arm1 time1 ne.txt"                      #Output file events and cens
KMdataIPDfile<-"KMdataIPD study2 figA arm1 time1 ne.txt"        #Output file for IPD
tot.events<-"NA"                                                #tot.events = total no. of events reported. If not re-
ported, then tot.events="NA"
arm.id<-1 #arm indicator
###END FUNCTION INPUTS

#Read in survival times read by digizeit
digizeit<- read.table(paste(path,digisurvfile,sep=""),header=TRUE)
t.S<-digizeit[,2]
S<-digizeit[,3]

#Read in published numbers at risk, n.risk, at time, t.risk, lower and upper

# indexes for time interval
pub.risk<-read.table(paste(path,nriskfile,sep=""),header=TRUE)
t.risk<-pub.risk[,2]
lower<-pub.risk[,3]
upper<-pub.risk[,4]
n.risk<-pub.risk[,5]
n.int<-length(n.risk)
n.t<- upper[n.int]


#Initialise vectors
arm<-rep(arm.id,n.risk[1])
n.censor<- rep(0,(n.int-1))
n.hat<-rep(n.risk[1]+1,n.t)
cen<-rep(0,n.t)
d<-rep(0,n.t)
KM.hat<-rep(1,n.t)
last.i<-rep(1,n.int)
sumdL<-0

if (n.int > 1){
 #Time intervals 1,...,(n.int-1)
  for (i in 1:(n.int-1)){
    #First approximation of no. censored on interval i
        n.censor[i]<- round(n.risk[i]*S[lower[i+1]]/S[lower[i]]- n.risk[i+1])
       #Adjust tot. no. censored until n.hat = n.risk at start of interval (i+1)
         while((n.hat[lower[i+1]]>n.risk[i+1])||((n.hat[lower[i+1]]<n.risk[i+1])&&(n.censor[i]>0))){
                    if (n.censor[i]<=0){
```

```
                        cen[lower[i]:upper[i]]<-0
                        n.censor[i]<-0
                }
                if (n.censor[i]>0){
                        cen.t<-rep(0,n.censor[i])
                        for (j in 1:n.censor[i]){
                                cen.t[j]<- t.S[lower[i]] +
                j*(t.S[lower[(i+1)]]-t.S[lower[i]])/(n.censor[i]+1)
        }
                #Distribute censored observations evenly over time. Find no. censored on each
time interval.
                cen[lower[i]:upper[i]]<-hist(cen.t,breaks=t.S[lower[i]:lower[(i+1)]],
                plot=F)$counts
        }

        #Find no. events and no. at risk on each interval to agree with K-M estimates read from
curves
        n.hat[lower[i]]<-n.risk[i]
last<-last.i[i]
        for (k in lower[i]:upper[i]){
                if (i==1 & k==lower[i]){
                        d[k]<-0
                        KM.hat[k]<-1
         }
         else {
         d[k]<-round(n.hat[k]*(1-(S[k]/KM.hat[last])))
                KM.hat[k]<-KM.hat[last]*(1-(d[k]/n.hat[k]))
         }
         n.hat[k+1]<-n.hat[k]-d[k]-cen[k]
    if (d[k] != 0) last<-k
        }
        n.censor[i]<- n.censor[i]+(n.hat[lower[i+1]]-n.risk[i+1])
    }
    if (n.hat[lower[i+1]]<n.risk[i+1]) n.risk[i+1]<-n.hat[lower[i+1]]
    last.i[(i+1)]<-last
    }
  }
#Find no. events and no. at risk on each interval to agree with K-M estimates read from curves
n.hat[lower[n.int]]<-n.risk[n.int]
last<-last.i[n.int]
for (k in lower[n.int]:upper[n.int]){
                if(KM.hat[last] !=0){
                d[k]<-round(n.hat[k]*(1-(S[k]/KM.hat[last])))} else {d[k]<-0}
        KM.hat[k]<-KM.hat[last]*(1-(d[k]/n.hat[k]))
        n.hat[k+1]<-n.hat[k]-d[k]-cen[k]
        #No. at risk cannot be negative
        if (n.hat[k+1] < 0) {
                n.hat[k+1]<-0
                cen[k]<-n.hat[k] - d[k]
        }
        if (d[k] != 0) last<-k
  }

#If total no. of events reported, adjust no. censored so that total no. of events agrees.
if (tot.events != "NA"){
        if (n.int>1){
        sumdL<-sum(d[1:upper[(n.int-1)]])
         #If total no. events already too big, then set events and censoring = 0 on all further
time intervals
         if (sumdL >= tot.events){
```

```
                        d[lower[n.int]:upper[n.int]]<- rep(0,(upper[n.int]-lower[n.int]+1))
                        cen[lower[n.int]:(upper[n.int]-1)]<- rep(0,(upper[n.int]-lower[n.int]))
                         n.hat[(lower[n.int]+1):(upper[n.int]+1)]<- rep(n.risk[n.int],(upper[n.int]+1-
lower[n.int]))
                }
                        }

                #Otherwise adjust no. censored to give correct total no. events
                if ((sumdL < tot.events)|| (n.int==1)){
                                sumd<-sum(d[1:upper[n.int]])
                                while ((sumd > tot.events)||(sumd< tot.events)&&(n.censor[n.int]>0))){
                                        n.censor[n.int]<- n.censor[n.int] + (sumd - tot.events)
                                        if (n.censor[n.int]<=0){
                                                cen[lower[n.int]:(upper[n.int]-1)]<-0
                                                n.censor[n.int]<-0
                                         }
                                         if (n.censor[n.int]>0){
                                                 cen.t<-rep(0,n.censor[n.int])
                                                 for (j in 1:n.censor[n.int]){
                                                                 cen.t[j]<- t.S[lower[n.int]] +
                                             j*(t.S[upper[n.int]]-t.S[lower[n.int]])/(n.censor[n.int]+1)
                                                                 }
                                           cen[lower[n.int]:(upper[n.int]-1)]<-
hist(cen.t,breaks=t.S[lower[n.int]:upper[n.int]],
                                            plot=F)$counts
                                           }
                                           n.hat[lower[n.int]]<-n.risk[n.int]
                                           last<-last.i[n.int]
                                           for (k in lower[n.int]:upper[n.int]){
                                                        d[k]<-round(n.hat[k]*(1-(S[k]/KM.hat[last])))
                                                         KM.hat[k]<-KM.hat[last]*(1-(d[k]/n.hat[k]))
                                                         if (k != upper[n.int]){
                                                                 n.hat[k+1]<-n.hat[k]-d[k]-cen[k]
                                                                 #No. at risk cannot be negative
                                                                 if (n.hat[k+1] < 0) {
                                                                         n.hat[k+1]<-0
                                                                         cen[k]<-n.hat[k] - d[k]
                                                                  }
                                                          }
                                                          if (d[k] != 0) last<-k
                                                 }
                                        sumd<- sum(d[1:upper[n.int]])
                                        }
                 }
}
write.table(matrix(c(t.S,n.hat[1:n.t],d,cen),ncol=4,byrow=F),paste(path,KMdatafile,sep=""),sep="\t")

### Now form IPD ###

#Initialise vectors
t.IPD<-rep(t.S[n.t],n.risk[1])
event.IPD<-rep(0,n.risk[1])
#Write event time and event indicator (=1) for each event, as separate row in t.IPD and event.IPD
k=1
for (j in 1:n.t){
        if(d[j]!=0){
                t.IPD[k:(k+d[j]-1)]<- rep(t.S[j],d[j])
                event.IPD[k:(k+d[j]-1)]<- rep(1,d[j])
                k<-k+d[j]
         }
```

```
}
#Write censor time and event indicator (=0) for each censor, as separate row in t.IPD and event.IPD
for (j in 1:(n.t-1)){
        if(cen[j]!=0){
                t.IPD[k:(k+cen[j]-1)]<- rep(((t.S[j]+t.S[j+1])/2),cen[j])
                event.IPD[k:(k+cen[j]-1)]<- rep(0,cen[j])
                k<-k+cen[j]
            }
}

#Output IPD
IPD<-matrix(c(t.IPD,event.IPD,arm),ncol=3,byrow=F)
write.table(IPD,paste(path,KMdataIPDfile,sep=""),sep="\t")


#Find Kaplan-Meier estimates
IPD<-as.data.frame(IPD)
KM.est<-survfit(Surv(IPD[,1],IPD[,2])~1,data=IPD,type="kaplan-meier",)
KM.est
summary(KM.est)
#Find Cox hazard ratio
data0<- read.table("C:\\PHD\\algorithm\\reliability exercice\\KMdataIPD study2 figA arm0 time2
no.txt",header=TRUE)
data1<- read.table("C:\\PHD\\algorithm\\reliability exercice\\KMdataIPD study2 figA arm1 time2
no.txt",header=TRUE)
data<-merge(data0,data1,all=T)
cox<-coxph(formula = Surv(data[,1], data[,2]) ~ data[,3], data=data)
summary(cox)
```

Vignette pour le package SSRMST: Calcul de la taille de l'échantillon en utilisant la différence de RMST (Restricted Mean Survival Time).

Vignette for SSRMST Package: Sample Size Calculation using Restricted Mean Survival Time"

In SSRMST: Sample Size Calculation using Restricted Mean Survival Time

1. Introduction

The hazard ratio (HR) associated with the Cox proportional hazards model is conventionally used in a randomized clinical trial that employs a right-censored, time-to-event outcome. When the proportional hazards assumption is satisfied, the HR is a valid measure for between-group differences. However, when the assumption is not satisfied, the HR variability over time makes assessment of the between-group differences difficult. In addition, the clinical interpretation of the HR is difficult regardless of the level of adequacy in the proportional hazards assumption. This is because of the absence of an absolute hazard as a reference.

Useful alternatives to the HR based on restricted mean survival times (RMST) have been proposed^[Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, Schrag D, Takeuchi M, Uyama Y, Zhao L, Skali H, Solomon S, Jacobus S, Hughes M, Packer M, Wei LJ. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of clinical Oncology 2014, 32, 2380-2385.]. The RMST is defined as the area under the survival function up to the truncation time, tau. When we consider death as an event, we can interpret the RMST as the tau-year life expectancy. The simple interpretation of the RMST would be, on average the life expectancy with this treatment over the next tau is ~years. RMST-based measures, (e.g., the difference in RMST), can be used as alternatives to the HR.

Because these are model-free and clinically interpretable, we therefore recommend using them in clinical trials. For implementing RMST-based measures, there is a **survRM2 package**, which is a user-friendly tool for RMST estimations. The vignette for the survRM2 package helps one to understand not only how to calculate RMSTs but also the cadre of advantages associated with the use of RMST-based measures.

In addition, the SSRMST package is helpful in designing superiority and non-inferiority studies using difference in RMST. The package calculates the study sample size and power by the method introduced in Appendix 2 of Uno et al. (2015)^[Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, Cai T, Pfeffer MA, Evans SR, Wei LJ. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Non-inferiority Studies. Annals of Internal Medicine 2015, 163, 127-134.]. This vignette thus explains how to properly use the main function ***ssrmst*.** In section 2, an example is provided for calculating sample sizes and powers for both superiority and non-inferiority tests using observed data from a recent clinical trial.

The SSRMST package was made and tested on the R version 3.4.0.

2. Function- ***ssrmst***

2.1) Setting

In this section, function arguments are introduced. These values can be determined by referring to results of previously analyzed clinical trials. Furthermore in this section, the same values presented in Appendix 2 of Uno et al. (2015) are utilized. These values are the observed data from a recent placebo-controlled clinical trial to assess the potential cardiovascular risk of saxagliptin in patients with type-2 diabetes mellitus (The Saxagliptin Study)^[Scirica BM, Bhatt DL, et al. Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus. N Engl J Med. 2013, 369, 1317-26.].

```
ac_rate   = 30
ac_period = 70
```

tot_time = 908

The first argument (ac_rate) is a scalar value of accrual rate: the number of patients per unit time. The second argument (ac_period) is a scalar value of accrual period: the time point at last accrual. The third argument (tot_time) is a scalar value of total study time: the time point at last follow-up. Note that the time units of these values should be same. According to the saxagliptin study, the observed accrual rate, accrual period, and total study time are 30 patients per day, 70 days, and 908 days, respectively.

From the version 0.1.1, the argument (ac_number) is added. ac_number is a scalar value of accrual number: the number of patients needed. Please specify only two paramerts from accrual information (ac_rate, ac_period, and ac_number). By specifying the arguments ac_rate and ac_period, the number of patients per unit time is automatically calculated (accrual type 1). By specifying ac_period and ac_number, the accrual rate is automatically calculated (accrual type 2). By specifying ac_rate and ac_number, the accrual period is automatically calculated (accrual type 3).

Next, the parameters for the Weibull distribution as the survival function are required:

shape0 = 1.05

shape1 = 1.05

scale0 = 8573

scale1 = 8573

The arguments (shape0 and shape1) are scalar values of the shape parameter for the Weibull distribution in the control (arm0) and the treatment (arm1). When the default (shape0 = 1 and shape1 = 1), the Weibull distribution reduces to the Exponential distribution with the scale parameter. The arguments (scale0 and scale1) are scalar values of the scale parameter for the Weibull distribution in the control (arm0) and the treatment (arm1). Note that when the shape parameters in both arms are same (shape0 = shape1), the value of the scale parameter in the

treatment (arm1) needs to be larger than or equal to that in the control (arm0), because the difference of the RMSTs (arm1 minus arm0) is always of interest in this package. In the saxagliptin study, the shape and scale parameters in both arms were 1.05 and 8573, respectively. Next, the derivation of tau is required.

```
tau  = 900
```

The argument (tau) is a scalar value to specify the truncation time point. It is important to specify an appropriate tau to calculate the RMST, which needs to be long enough to assess what we want to. In case the number of patients at risk is small around the last follow-up time, it would be appropriate to specify tau at a time with a large-enough number of patients at risk. Because the total study time of the saxagliptin study was 908 days, it was appropriate to set tau at 900 days.

Under the setting, 2,000 sets of realizations (default) were generated with a 1:1 treatment allocation (default) for calculating confidence intervals of RMST differences. The argument (ntest) is a scalar value of the number of iterations and (allocation1) is a scalar value of the proportion of patients allocated to arm1. The argument (seed) is a random seed for the sampling. The default is NULL. The confidence intervals of the differences in RMST are then calculated.

2.2) Non-inferiority test

It is conventional to use the HR for designing a non-inferiority test with respect to a time-to-event outcome. As described in Section 1, the HR is not always appropriate as a summary measure for between-group differences. Moreover, the rationale of the conventionally used non-inferiority margin, such as 1.30 in safety studies for diabetes mellitus patients in the United States, has never been fully asserted with indisputable clarity. If the difference in RMST is used for designing a non-inferiority test, a non-inferiority margin implies a clinically acceptable difference in RMST value. This gives us more clinical interpretation.

Next, the derivation of margin is required.

```
margin  = 18
```

The argument (margin) is a scalar value of non-inferiority margin: a clinically acceptable difference in RMST. Note that the time unit should be the same as other values. In the saxagliptin study, non-inferiority margin was determined as 18 days. Here it is assumed that a difference of 18 days out of 900 is clinically acceptable.

Next, the power is calculated, i.e., the chance that the lower bound of 2,000 confidence intervals of the difference in RMST [arm1 minus arm0] falls above a value of minus margin, -18 days. Since this test is one-sided, the nominal type I error level is given as one-sided. In this program, when implementing the default value (one_sided_alpha = 0.025), 95% confidence intervals of the difference in RMST are therefore estimated to calculate the power. The number of expected events at the time point of last follow-up is also calculated. The resultant output of a non-inferiority test is provided as follows. Note that two types of power (separate and pooled) are shown based on separate and pooled variances.

```
library(SSRMST)
ssrmst(ac_rate=ac_rate, ac_period=ac_period, tot_time=tot_time, tau=tau, shape0=shape0,
scale0=scale0, shape1=shape1, scale1=scale1, margin=margin, seed=2017)
```

2.3) Superiority test

When implementing the default value of margin (margin=0), a superiority test is selected. In this case the power (i.e., the chance that the lower bound of 2,000 confidence intervals of the difference in RMST [arm1 minus arm0] falls above 0) is calculated. The resultant output of a superiority test is provided as follows.

```
ssrmst(ac_rate=ac_rate, ac_period=ac_period, tot_time=tot_time, tau=tau, shape0=shape0,
scale0=scale0, shape1=shape1, scale1=scale1, seed=2017)
```

If the resulting chance is found to be lower than 80%, for example, the setting undergoes adjustment to increase the current sample size until an 80% chance is ultimately obtained.

3. Conclusion

The HR, a conventionally used summary measure for between-group differences, is not always appropriate; this is particularly the case when the proportional hazards assumption is not satisfied. In addition, due to the lack of a baseline reference number, the interpretation of the HR can be quite difficult. One of the model-free alternatives to the HR, the differences in RMST, is clinically interpretable. Therefore, it is strongly recommended to use such a measure when designing clinical trials. The SSRMST package will likely be deemed helpful when evaluating non-inferiority or superiority studies via the use of the difference in RMST.