Impact des critères de jugement sur l'optimisation de la détermination du nombre de sujet nécessaires pour qualifier un bénéfice clinique dans des essais cliniques en cancérologie

Alhousseiny Pam

UNIVERSITE BOURGOGNE FRANCHE-COMTE

ECOLE DOCTORALE « ENVIRONNEMENTS-SANTE »

Année 2017

THESE

Pour obtenir le grade de

Docteur de l'université Bourgogne Franche-Comté

Spécialité : Biostatistiques

Présentée et soutenue publiquement

Le 19/12/2017

Par

Alhousseiny PAM

Thèse dirigée par : **Pr Franck BONNETAIN**

Co-encadrée par : **Dr Amélie ANOTA**

Impact des critères de jugement sur l'optimisation de la détermination du nombre de sujet nécessaires des essais cliniques en cancérologie devant qualifier un bénéfice clinique.

JURY

Dr Amélie ANOTA, PhD, Université de Franche-Comté (co-encadrante de thèse)

Pr Franck BONNETAIN, PU-PH, Université de Franche-Comté (directeur de thèse)

Dr Caroline BASCOUL MOLLEVI, PhD/HDR, Institut du Cancer de Montpellier (examinateur)

Dr Carine BELLERA, PhD/HDR, Université de Bordeaux (examinateur)

Dr Emmanuel CHAMOREY, PhD/HDR, Université de Nice-Sophia Antipolis (rapporteur)

Pr Pascal HAMMEL, PU-PH, Hôpital Beaujon (rapporteur)

A la mémoire de mon directeur de thèse: Pr Franck BONNETAIN

Remerciements

Cette thèse est en grande partie à la mémoire de mon directeur de thèse le Professeur Franck Bonnetain. Tu nous as quittés trop tôt dans une situation très douloureuse. Je ne pourrai jamais te remercier assez de m'avoir donné une opportunité de faire cette thèse. Je me suis donné tous les moyens pour que cette thèse aboutisse pour ta mémoire.

Un grand merci à Amélie ma co-encadrante pour toutes les heures que tu as passé sur ce projet de thèse depuis le début jusqu'à la fin. Merci pour tous les encouragements, ta disponibilité, sans toi ça aurait été plus compliqué.

Je remercie le Professeur Olivier Adotevi de l'université de Franche-Comté de m'avoir encouragé et motivé pour que cette thèse aboutisse dans les meilleures conditions après la disparition de Franck.

Je remercie le Professeur Pascal HAMMEL, chef du service d'Oncologie Digestive, Hôpital Beaujon et Coordonnateur du DES d'Hépato-Gastroentérologie d'Ile-de-France, et Monsieur Emmanuel CHAMOREY, Praticien Spécialiste des CLCC d'avoir accepté d'être rapporteur de ce travail de thèse. C'est un grand honneur pour moi.

J'adresse ma gratitude à Madame Caroline MOLLEVI Chercheur à l'Institut du Cancer de Montpellier et Madame Carine BELLERA, chercheur à l'Institut Bergonié de Bordeaux d'être membre du jury de ma thèse.

Je remercie toute l'équipe de l'unité de méthodologie et de qualité de vie en cancérologie, CHRU de Besançon pour tous les bons moments passés ensemble.

Je remercie toute ma famille, mes proches et mes amis pour le soutien pendant toutes ces années de thèse.

Table des matières

Abréviations 7

Résumé 8

Abstract 10

Productions scientifiques 11

I. Introduction 14

II. Design des essais cliniques de phase III 17

III. Les critères de jugement en cancérologie 22

A. Les critères centrés sur le patient 24

1. La survie globale 24

2. La qualité de vie relative à la santé 25

B. Les critères centrés sur la tumeur 30

1. Les critères composites intermédiaires 30

2. Les critères de substitution 32

IV. La validation des critères de substitutions 33

A. Approche par essai 33

1. Critère de validation de Prentice 33

2. Freedman : Proportion d'effet du traitement (PE) expliqué par le critère de substitution 38

3. Buyse et Molenberghs : Effet relatif et association ajustée 39

B. Approche par méta-analyse 40

1. Régression linéaire pondérée 41

2. Modélisation jointe 44

C. Classification de Fleming: hiérarchisation des critères de substitution 48

V. Les critères conjoints 48

VI. Le calcul du nombre de sujets nécessaires 49

A. Calcul du nombre d'événements à observer 52

B. Calcul du nombre de patients à inclure 55

VII. Objectives 56

VIII. Les travaux réalisés 59

A. Surrogate endpoints for overall survival in pancreatic cancer trials: an individual patient meta-analysis. (Jama oncology). 59

B. La qualité de vie relative à la santé comme co-critère de jugement principal dans les essais cliniques randomisés en oncologie 80

C. Calcul du nombre de Sujets nécessaires en phase III incluant deux critères conjoints en cancérologie : Package R Coprimary. 95

IX. Discussion 136

X. Conclusion 141

XI. Les références 143

Abréviations

AMM : Autorisation de Mise sur le Marché.

ASCO : American Society of Clinical Oncology.

CONSORT : Consolidated Standards of Reporting Trials

DATECAN : Definition for the Assessment of Time-to-event Endpoints in CANcer trials.

DL : Dose Limitante.

ECR : Essais Cliniques Randomisés.

EORTC : European Organization for Research and Treatment of Cancer

FACTG : Functional Assessment of Cancer Therapy General.

FDA : Food and Drug Administration (FDA) .

FLIC : Functional Living Index-Cancer.

OMS : Organisation Mondiale de la Santé

PFS : Progression Free Survival

PROs : Patient-Reported Outcomes

QdV : Qualité de Vie relative à la santé.

RECIST : Response Evaluation Criteria In Solid Tumours

RSCL : Rotterdam Symptom Checklist

SSP : Survie Sans Progression

SG : Survie Globale

TJD: Temps Jusqu'à Détérioration

TTP : Time To Progression

WHO : World Health Organization

WHOQOL : The World Health Organization Quality of Life

Résumé

La survie globale (SG) est considérée comme le critère de jugement principal de référence et le plus pertinent et objectif dans les essais cliniques en oncologie. Les critères de survie composites, tels que la survie sans progression, sont couramment utilisés dans les essais de phase III comme critère de substitution de la SG. Leur développement est fortement influencé par la nécessité de réduire la durée des essais cliniques, avec une réduction du coût et du nombre de sujets nécessaires. Cependant, ces critères sont souvent mal définis, et leurs définitions sont très variables entre les essais. De plus, leur capacité de substitution à la SG, n'a pas toujours été rigoureusement évaluée. Le projet DATECAN-1 a permis de proposer des recommandations pour la définition et donc l'homogénéisation entre essais cliniques randomisés (ECR) de ces critères.

De plus, la majorité des essais cliniques de phase III intègrent désormais la qualité de vie relative à la santé (QdV) comme critère de jugement afin d'investiguer le bénéfice clinique de nouvelles stratégies thérapeutiques pour le patient. Une alternative est de considérer un co-critère de jugement principal : un critère tumoral tel que la survie sans progression et la QdV afin de s'assurer du bénéfice clinique pour le patient. Lors de la conception d'un essai clinique avec des critères de jugements principaux multiples, il est essentiel de déterminer la taille de l'échantillon appropriée pour pouvoir indiquer la signification statistique de tous les co-critères de jugements principaux tout en préservant la puissance globale puisque l'erreur de type I augmente avec le nombre de co-critères de jugements principaux.

Le premier objectif de ma thèse était d'étudier l'influence des définitions des critères de survie issus du consensus du DATECAN-1 sur les résultats et les conclusions des essais. Le second objectif était d'étudier les propriétés des critères de substitution à la survie globale. Le dernier objectif était de proposer un design du calcul du nombre de sujets nécessaires pour

une étude clinique de phase III avec des co-critères de jugement de type temps jusqu'à événement. L'objectif final de mon projet de thèse était de développer un package R pour le calcul du nombre de sujets nécessaires avec les co-critères de jugement principal et d'étudier les critères de substitutions à la SG pour le cancer du pancréas.

Mes résultats ont permis de confirmé la variabilité des critères de survie utilisés dans les essais cliniques randomisés dans le cancer du pancréas. Beaucoup d'événements nécessaires à la définition des critères de survie recommandés par le consensus DATECAN-1 sont manquants ou mal définis. Il est primordial et nécessaire d'appliquer les critères issus des recommandations de DATECAN-1 sur les nouveaux essais cliniques randomisés pour uniformiser les essais et faire correctement les comparaisons entre les résultats publiés.

Pour le cancer du pancréas en situation avancée, nous avons montré que les critères tels que la survie sans progression, le temps jusqu'à la progression et le temps jusqu'à détérioration du statut OMS sont potentiellement des critères substitutifs à la SG.

Nos travaux ont permis de proposer un design alternatif en associant un critère de jugement intermédiaire composite comme SSP avec la QdV en tant que co-critères de jugement principaux. La combinaison de plusieurs critères principaux permet d'augmenter la puissance statistique de l'étude. Cependant, le calcul du nombre de sujets nécessaires avec des critères conjoints est nécessaire et s'effectue en deux étapes pour détecter les effets sur chaque critère de jugement.

Le package R que nous avons développé permet de déterminer le nombre de sujet nécessaires avec un ou deux critères de jugement de type temps jusqu'à événement. Il permet aussi de vérifier la cohérence du design d'un essai clinique comparant deux groupes de traitement. Il est disponible en libre accès sur le site CRAN pour une large utilisation afin d'améliorer la palification des essais cliniques avec des critères de survie de temps jusqu'à événement.

Abstract

Overall survival (OS) is the gold standard endpoint in phase III cancer clinical trials but with the increasing number of effective salvage treatments available in many types of cancer, more inclusion of patients and much longer follow-up are necessary to observe the number of deaths required to achieve significant statistical power and then to demonstrate that treatments may improve OS. Consequently the cost of clinical trials increases. Thus, progression-free survival (PFS), which is assessed earlier, is frequently used as primary endpoint but suffers from important limitations especially heterogeneity and a lack of validation as surrogate of OS. In this context, health-related quality of life (HRQOL) which is considered by the FDA as an endpoint assessing direct clinical benefit could be an outcome to judge efficacy of a treatment, particulary in advanced cancer. HRQOL could be combined with PFS as co-primary endpoint to ensure a clinically benefit for patients. However researches must be pursued to define clearly the methodology and decision rules of such design of trials. One of the major problem in such trials is how to determine sample size.

The objectives of my thesis project are:

1) To study the impact of the definitions of time-to-event endpoint from the DATECAN-1 consensus on the results and conclusions of the trials published in pancreatic cancer.

2) To study the properties of the potential surrogate to the overall survival.

3) To propose a design for the determination of sample size necessary for a phase III clinical study with co-primary time-to-event such as progression-free survival and time to quality of life deterioration.

The final objective of my thesis project is to develop an R package for the calculation of the number of subjects needed with the co-primary time-to-event.

Productions scientifiques

Publications scientifiques:

- Fiteni, F., **Pam, A.,** Anota, A., Vernerey, D., Paget-Bailly, S., Westeel, V., & Bonnetain, F. (2015). Health-related quality-of-life as co-primary endpoint in randomized clinical trials in oncology. *Expert review of anticancer therapy*, *15*(8), 885-891.
- **A. Pam**, A. Anota, C. Mollevi, T. Filleron, F. Bonnnetain

  Sample size determination in oncology phase III clinical trials with two primary time-to-event endpoints: co-primary R package. Soumission sur Computer Methods and Programs in Biomedicine.

Soumission du package :

- **A. PAM.** Sample Size Calculation for two Primary Time-to-Event Endpoints in Clinical Trials. Publication du package « coprimary » sur le site de CRAN (https://cran.r-project.org/) 14-12-2016.

Articles à soumettre:

A. Pam, A. Anota, F. Fiteni, L. Collette[2], C. Louvet[3], B. Baron, L. Bedenne, E. Briasoulis, B. Chauffert, T. L. Dahan, M.P. Ducreux, P. Fumoleau, G. Hans, K. Haustermans, J. Jeekel, F. Levi , M.P. Lutz, J.F. Seitz, J. Taieb, I. Trouilloud, J.L. Van-Laethem, D.J.T Wagener, S. Gourgou, C. Bellara, F. Bonnetain : Surrogate endpoints for overall survival in pancreatic cancer trials: an individual patient meta-analysis. (Jama oncology).

Communications affichées:

- **Dans un congrès international**

- 17th ESMO WGI (World congress Gastrointestinal) Barcelona 2015 - PD-002:

A. Pam, D. Vernerey, B. Chauffert, C. Louvet, F. Levi, H. Galjaard, S. Gourgou, C. Bellera, A. Anota and F. Bonnetain- Impact of the definition of time to event endpoint on randomized clinical trials results in oncology (DATECAN-2): an analysis of 9 pancreatic cancer trials. (Poster discussion)

- 8th ECCO - 40th ESMO European Cancer Congress Vienna 2015:

A. Pam, D. Vernerey, B. Chauffert, C. Louvet, F. Levi, H. Galjaard, S. Gourgou, C. Bellera, A. Anota and F. Bonnetain : Impact of the definition of time to event endpoint on randomized clinical trials results in oncology (DATECAN-2): an analysis of 9 pancreatic cancer trials.

- **Dans un congrès national**

- **Congrès EPICLIN 11/ 24è journées des statistiques des CLCC 2017, Saint Etienne**

  A. Pam, A. Anota, D. Vernerey, L. Collette, S. Gourgou, C. Bellera, F. Bonnetain – Analyse de 9 essais cliniques randomizes du cancer de pancréas.

- **Congrès EPICLIN 11/ 24è journées des statistiques des CLCC 2017, Saint Etienne**

  A. Pam, A. Anota, C. Mollevi, T. Filleron, F. Bonnnetain

  Calcul du nombre de Sujets nécessaires en phase III incluant deux critères conjoints en cancérologie : Package R Coprimary.

- **Congrès EPICLIN 10/ 23è journées des statistiques des CLCC 2016, Strasbourg:**

  A. Pam, A. Anota, D. Vernerey, L. Collette, C. Louvet, ... & F. Bonnetain - Impact of the definition of time to event endpoint on randomized clinical trials results in oncology (DATECAN-2): an analysis of 9 pancreatic cancer trials

- **Congrès EPICLIN 9/ 22è journées des statistiques des CLCC 2015, Montpellier**:

  A. Pam, A. Anota, F. Fiteni, S. Dabakuyo-Yonli, S. Gourgou, C. Bascoul-Mollevi, F. Bonnetain. Calcul du nombre de sujets nécessaires en phase III incluant deux critères conjoints en cancérologie.

I. Introduction

En cancérologie, le but recherché par la mise en place d'un nouveau traitement est la guérison, l'allongement de la durée de vie, ou la réduction des morbidités sans altération du pronostic vital. C'est pourquoi les essais cliniques de phases III cherchent à mesurer de tels bénéfices en utilisant des critères de jugement, c'est-à-dire qui mesurent un réel bénéfice ressenti par le patient.

Les critères de jugement sont des mesures cliniques et biologiques dans l'évaluation et le développement des thérapeutiques. Ils peuvent être classés en deux catégories : les "critères de jugement centrés sur le patient" comprenant la survie globale (SG) et la qualité de vie relative à la santé (QdV) et les "critères de jugement centrés sur la tumeur" comme la survie sans progression (SSP).

Les critères de substitution sont des critères de jugement centrés sur la tumeur dont l'objectif est de se substituer aux critères de jugement centrés sur le patient, notamment la SG. Le choix du critère de jugement principal dans les essais cliniques en oncologie est un problème majeur.

La survie globale (SG) est considérée comme le critère de jugement principal de référence; il s'agit du critère de jugement le plus pertinent et objectif dans les essais cliniques en oncologie. La « US Food and Drug Administration » (FDA) considère la SG comme étant une mesure unanimement admise du bénéfice direct du traitement et le critère de jugement privilégié pour les essais de phase III (Rock et al.; Peppercorn et al.).

Cependant, avec l'augmentation du nombre de traitements efficaces disponibles pour une majorité de cancers, un nombre plus important de patients à inclure et un suivi plus long est nécessaire afin d'avoir une puissance statistique suffisante pour pouvoir mettre en évidence une amélioration de la SG. De ce fait les critères de jugement composites, tels que la survie sans progression, sont couramment utilisés dans les essais de phase III comme critère de substitution de la SG.

Leur développement est fortement influencé par la nécessité de réduire la durée des essais cliniques, avec une réduction du coût et du nombre de sujets nécessaires. Cependant, ces critères sont souvent mal définis, et leurs définitions sont très variables entre les essais, rendant difficile la comparaison des résultats entre les essais (Mathoulin-Pelissier et al., « Survival End Point Reporting in Randomized Cancer Clinical Trials »). De plus, leur capacité de substitution à la SG, c'est-à-dire la capacité à prédire un bénéfice sur la SG à partir des résultats de l'essai sur le critère d'évaluation, n'a pas toujours été rigoureusement évaluée.

Dans ce contexte, le projet international DATECAN-1(Definition for the Assessment of Time-to-event Endpoints in CANcer trials) a permis de proposer des recommandations pour la définition et donc l'homogénéisation entre essais cliniques randomisés (ECR) de ces critères (Bonnetain, Bonsing, et al.). De plus, la majorité des essais cliniques de phase III intègrent désormais la qualité de vie relative à la santé (QdV) comme critère de jugement afin de s'assurer du bénéfice clinique de nouvelles stratégies thérapeutiques pour le patient (Carolyn Cook Gotay et al.).

Une alternative serait de considérer un co-critère de jugement principal : un critère tumoral tel que la survie sans progression et la QdV afin de s'assurer du bénéfice clinique pour le patient. Bien que la QdV soit reconnue comme second critère de jugement principal par l'ASCO (« American Society of Clinical Oncology ») et la FDA en l'absence d'effet sur la survie globale(Beitz et al.), elle est encore peu prise en compte comme co-critère de jugement principal dans les essais (Fiteni et al.). L'évaluation, l'analyse et l'interprétation des résultats de QdV demeurent complexes, et les résultats restent encore peu pris en compte par les cliniciens du fait de son caractère subjectif et dynamique.

Lors de la conception d'un essai clinique avec des critères de jugements principaux multiples, il est essentiel de déterminer la taille de l'échantillon appropriée pour pouvoir indiquer la signification statistique de tous les co-critères de jugements principaux tout en préservant la puissance globale puisque l'erreur de type I augmente avec le nombre de co-critères considérés.

Plusieurs méthodes ont été développées pour l'ajustement du taux d'erreur de type I. Elles utilisent généralement un fractionnement du taux d''erreur de type I appliqué à chacune des hypothèses testées. Toutes ces méthodes sont investiguées dans ce projet.

Dans ce contexte, une première partie de ma thèse de sciences était dédiée à l'étude de l'impact de définitions des critères de DATECAN-1 sur les résultats et les conclusions des essais cliniques randomisés (ECR) publiés dans le cancer du pancréas. Une validation puis une hiérarchisation de ces critères a ensuite été proposée en fonction de leurs propriétés de substitution à la SG, pour la localisation du pancréas en situation adjuvante et métastatique.

Une deuxième partie de mon travail de thèse consistait à élaborer une méthodologie du calcul du nombre de sujets nécessaires avec des critères de survie conjoints dans une étude clinique de phase III. Enfin, l'objectif final était de développer un package R pour le calcul du nombre de sujets nécessaires avec les co-critères de jugement principal de temps jusqu'à événement.

II. Design des essais cliniques de phase III

L'utilisation chez l'homme d'une nouvelle thérapeutique nécessite actuellement des essais conduits avec une méthodologie rigoureuse, constituant l'une des branches de recherche clinique et de l'épidémiologie. La méthodologie fait largement appel aux méthodes statistiques.

L'expérimentation sur l'homme ou étude clinique pose de multiples problèmes éthiques. Les déclarations d'Helsinki (1964) (World Medical Association) et de Tokyo (1975) par l'association médicale mondiale (AMM), sous l'égide de l'Organisation Mondiale de la Santé (OMS), ont défini les précautions à prendre et conduit une législation précise, en France en particulier.

Une étude (ou un essai) clinique est un protocole d'investigation visant à mettre en évidence ou non l'effet d'un traitement sur la guérison d'une maladie ou d'un traumatisme. Lors de la mise en place des essais cliniques en oncologie, plusieurs questions se posent tant au niveau scientifique qu'au niveau éthique.

Différentes approches ont été proposées pour améliorer la conduite des essais cliniques dans le but de réduire le nombre de patients à traiter et d'optimiser l'information obtenue durant ces essais. Le développement d'un médicament peut être considéré comme un processus

d'accumulation d'informations qui se termine lorsque l'information rassemblée est résumée et présentée aux agences d'autorisation de mise sur le marché. Ce processus comprend différentes étapes alternant apprentissage et confirmation.

Pour chaque étape du développement d'un médicament, le protocole doit spécifier les objectifs primaires et secondaires, les critères d'inclusion, le plan de traitement tout au long de l'étape, ainsi que les méthodes d'analyse de données. L'enjeu financier d'un essai clinique est en général très important : le résultat d'une telle expérimentation détermine si une nouvelle molécule peut être commercialisée (obtention de l'Autorisation de Mise sur le Marché), et de déterminer son degré de remboursement par la sécurité sociale le cas échéant. C'est la raison principale pour laquelle les protocoles sont extrêmement stricts.

Un essai clinique passe par différentes phases (cf. Figure 1).

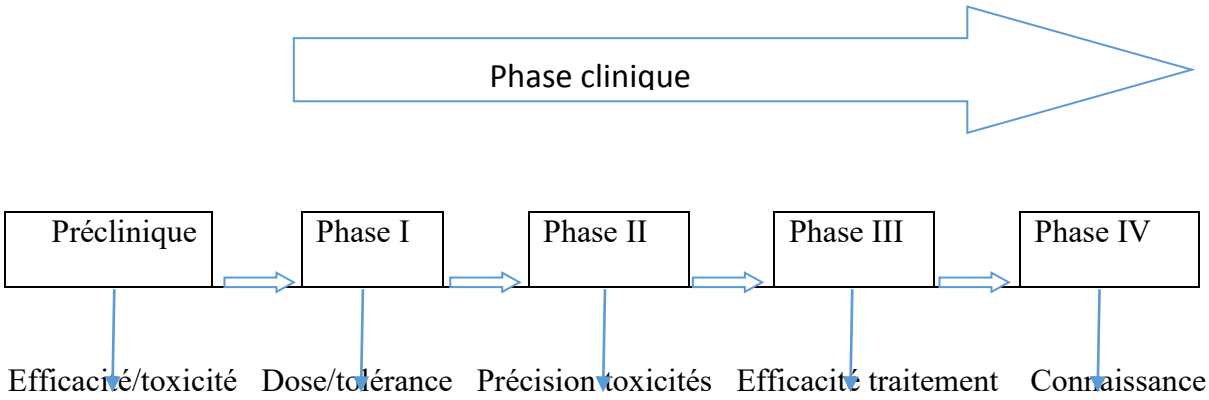


Figure 1 : Etapes des essais cliniques.

- **La phase préclinique**:

- Vérification des conditions de sécurité (études de toxicologie).

- Cette phase a pour but de déterminer à partir de quelle dose le produit est toxique chez l'animal, et quels sont les organes qui en souffrent en premier.
- Cette étude doit être pratiquée chez au moins 2 espèces de mammifères, avec 2 voies d'administration.
- L'objectif est de déterminer la dose qui tue 0%, 50% et 100% des animaux étudiés (DL 0, DL 50, DL 100), ainsi que les causes de mort.
- La toxicité est définie par kilo.

-Les études de toxicité en administration réitérée.

- D'abord, l'objectif est de définir les doses/modes d'administration (posologies) tolérées lors d'un usage prolongé. Cette étude est réalisée sur les mêmes espèces que précédemment.
- Ensuite, la voie d'administration est la même que celle que l'on a prévu pour l'usage thérapeutique chez l'homme.
- Enfin, tous les animaux sont sacrifiés à la fin de l'étude afin d'analyser leurs organes (étude histologique).

-Les études de reproduction.

- L'objectif est d'étudier l'impact du produit sur la fertilité.
- On commence d'abord par une étude de fertilité avec administration du produit avant accouplement ;
- Ensuite on procède à une étude sur les embryons de femelles en gestation, ayant reçu le produit ;

- Enfin une étude sur des éventuels effets sur la périnatalité (en particulier la compatibilité avec l'allaitement) est réalisée.

-Études de mutagénèse : l'objectif est de déterminer ou non si l'administration du médicament a une incidence sur la modification du code génétique.

- Étude de cancérogénèse.
- Études pharmacocinétiques : l'objectif est de déterminer les conditions d'absorption, de diffusion et d'élimination du produit (Pazdur, « Endpoints for Assessing Drug Activity in Clinical Trials »).

- **<u>La phase clinique</u>:**

- **Étude de phase I**
  - L'Objectif est de déterminer les conditions de tolérance humaine au produit, et connaître la posologie entraînant les premiers effets indésirables et les premiers effets curatifs souhaités.
  - L'étude est menée sur des volontaires sains (si la toxicité du médicament est limitée).
  - La première dose est égale à 1/100 de la dose maximale tolérée chez l'animal. Les doses augmentent progressivement.
  - Les premiers effets indésirables sont étudiés et détaillés.
  - En cancérologie, cette phase est réalisée chez des malades pour lesquels toutes les thérapeutiques mises en œuvre par le passé ont échouées.

- **Étude de phase II**
  - L'objectif est de déterminer les conditions optimales de prescription.
  - L'étude est menée sur des volontaires sains ou malades.
  - On cherche à définir le mode d'administration optimal.
  - On réalise une analyse statistique d'un modèle dose/effet (régression normale ou logistique). On détermine ainsi la dose optimale qui sera administrée en phase III

- **Étude de phase III**
  - L'objectif est de déterminer l'efficacité et la tolérance du médicament.
  - Chaque essai est réalisé selon un protocole précis. Il est nécessaire de spécifier une unique hypothèse principale à tester. Des hypothèses secondaires peuvent aussi être faites, mais la conclusion de l'étude portera principalement sur la validation ou infirmation de l'hypothèse principale.
  - L'effet thérapeutique est apprécié sur un groupe de patients souffrant de la maladie à traiter, et recevant le produit étudié.
  - L'effet thérapeutique est mesuré sur un paramètre précis, et étudié de façon comparative par rapport à un placébo ou un traitement de référence.
  - Un calcul préliminaire du nombre de patients à inclure dans l'étude doit être réalisé afin d'avoir une puissance statistique suffisante pour conclure.
  - Le nombre de patients à inclure varie en fonction des risques de premières et secondes espèces spécifiées par le statisticien.

- **Étude de phase IV**
    - Elles sont conduites après que le médicament ait reçu une autorisation de mise sur le marché.
    - Elles ont pour but d'étudier l'efficacité et la tolérance du médicament lors d'une utilisation prolongée.
    - L'objectif est de mettre en évidence d'éventuels effets secondaires rares.
    - Ce genre de surveillance est effectué par le centre de pharmacovigilance du laboratoire, mais aussi par les médecins, qui en sont tenus par la loi.

III. Les critères de jugement en cancérologie

Les critères de jugement sont des mesures cliniques et biologiques dans l'évaluation et le développement des thérapeutiques. Le choix d'un critère de jugement est une étape capitale et difficile qui conditionne l'intérêt des résultats.

La mesure des critères peut se faire de façon :

- Objective: dosages, mensurations, photographies, radiographies,
- Subjective: appréciation de la douleur, pathologie psychiatrique.

Le nombre de critères doit être aussi faible que possible.

La comparaison de deux groupes dans un essai clinique de phase III passe par le choix de critères de jugement, toujours discutables, mesurant l'efficacité et la tolérance du traitement. Un critère de jugement doit (Velentgas et al.) :

- être pertinent cliniquement;
- faire l'objet d'un consensus de la communauté médicale ;
- être disponible chez tous les sujets;
- être acceptable pour les patients (pénibilité);

- être de mesure simple;
- être reproductible (fiable);
- être sensible (apte à détecter de petites différences) ;
- être objectif (dosage biologique fourni par un automate par exemple) mais beaucoup de critères sont subjectifs (qualité de vie relative à la santé, douleur) ou impliquent une interprétation (échographie)... ;
- être direct (traduisant directement le phénomène biologique ou médical étudié, c'est à dire traduisant l'activité thérapeutique);
- être unique. Cependant, rares sont les pathologies se résumant à un seul critère. Il est alors nécessaire de recourir à plusieurs critères de jugement qu'on hiérarchise (principal et secondaires).

Le critère de jugement principal peut être très divers, allant d'un marqueur biologique à la qualité de vie relative à la santé, en passant par le comportement (activité physique hebdomadaire, observance thérapeutique) et un indicateur médico-économique (recours aux soins, consommation de produits de santé, hospitalisations en urgence, nombre de jours d'arrêt de travail…).

Un critère de jugement doit être pertinent par rapport à la maladie, au traitement et à l'objectif de l'étude. L'objectif de l'étude peut être clinique, c'est-à-dire qui évalue l'intérêt immédiat du malade. L'objectif peut être aussi explicatif ou scientifique basé sur la maladie. En cancérologie, les critères de jugement peuvent être classés en deux catégories : les « critères de jugement centrés sur le patient » comprenant la survie globale (SG) et la qualité de vie relative à la santé (QdV) et les « critères de jugement centrés sur la tumeur » comme la survie sans progression (Bonnetain, Bonsing, et al.; Fiteni et al.).

A. Les critères centrés sur le patient

Les critères de jugement centrés sur le patient sont des caractéristiques qui sont le reflet de ce que le patient perçoit, d'une fonction ou d'une survie. Ils sont utilisés pour mesurer l'efficacité d'un traitement et reflètent un bénéfice clinique direct pour le patient.

1. La survie globale

La survie globale est définie comme l'intervalle de temps entre la date de randomisation et la date du décès toutes causes confondues. Les patients vivants ou perdus de vue en cours d'étude sont censurés à la date de dernière nouvelle. C'est le critère de jugement de référence et le plus pertinent pour évaluer l'efficacité d'un traitement dans les essais cliniques en cancérologie. La US Food and Drug Administration estime que la SG doit être un critère de jugement principal pour les études cliniques de phase III (Blumenthal et al.).

La survie globale est un critère objectif, facile à recueillir, mais qui nécessite un suivi particulier afin de ne pas avoir trop de perdus de vue. La durée du suivi nécessaire sera d'autant plus longue que les événements attendus (décès) surviennent de façon tardive au cours de l'évolution de la maladie.

La survie spécifique au cancer est souvent utilisée comme critère de jugement chez les personnes âgées. Elle est définie comme l'intervalle de temps entre la date de randomisation ou le début du traitement et la date du décès lié au cancer qu'il soit lié à la progression, à la tumeur primitive, à un second cancer, ou au protocole de traitement. Les décès non liés au cancer, les patients perdus de vue et les patients encore en vie à la date de point sont censurés. Ce critère est composite et prend en compte le décès spécifique au cancer, il est différent de la survie globale.

Contrairement aux autres critères de survie qui peuvent être utilisés, la survie globale ne pose pas de problème de définition ni de problème d'interprétation clinique. La survie globale est généralement utilisée comme le critère de jugement principal pour les cancers à mauvais pronostic en situation adjuvante car c'est avant tout ce bénéfice clinique que l'on va chercher à améliorer. Cependant, le contexte actuel des essais de stratégie des traitements efficaces, il est nécessaire d'inclure un nombre plus important de patient et un suivi plus long afin d'observer un nombre d'événements requis et ainsi attendre une puissance statistique suffisante et démontrer une amélioration de la SG. De ce fait, nous pouvons faire face à un risque important de perdus de vue. L'ensemble de ces raisons a rendu nécessaire la proposition de nouveaux critères de jugement évaluant, temporellement, d'une façon plus précise l'efficacité d'un traitement.

2. La qualité de vie relative à la santé

La santé a été définie par Organisation Mondiale de la Santé (OMS) en 1948, comme étant "un état de bien-être physique, mental et social complet, et pas simplement l'absence de maladie" (WHO, 1948)(« WHO | Constitution of WHO »).

En 1993, l'OMS définie la qualité de vie (QdV) relative à la santé comme « la perception qu'a un individu de sa place dans l'existence, dans le contexte de la culture et du système de valeurs dans lequel il vit, en relation avec ses objectifs, ses attentes, ses normes et ses inquiétudes. Il s'agit donc d'un large champ conceptuel, englobant de manière complexe la santé physique de la personne, son état psychologique, son niveau d'indépendance, ses relations sociales, ses croyances personnelles et sa relation avec les spécificités de son environnement » (WHOQOL, 1993) (Sartorius).

La qualité de vie relative à la santé (QdV) découle de cette définition, et est donc un concept subjectif, dynamique, et multidimensionnel incorporant au moins trois domaines: les fonctionnements physique, psychologique et social (Bonnetain). Dans un contexte médical, le terme de QdV a un sens un peu plus spécifique, et considère principalement l'appréciation du patient sur le vécu de son traitement et de sa maladie, même si certaines conséquences indirectes (chômage, difficultés financières,. . . etc) sont parfois prises en compte. Le terme de « qualité de vie relative à la santé » (QdV) est alors préféré.

La QdV entre dans le champ des « Patient-Reported Outcomes » (PROs) (Carolyn C. Gotay et al.), comme la fatigue, la douleur, la satisfaction des soins, etc , et sont des résultats rapportés par les sujets eux-mêmes. Elle est considérée comme second critère de jugement principal par l' « American Society of Clinicial Oncology » et la « Food and Drug Administration» en l'absence d'effet sur la SG (Beitz et al.; Rock et al.).

La QdV constitue donc un critère de jugement alternatif pertinent et disponible pour s'assurer de l'intérêt du traitement pour le patient et le système de santé. Dans le cas de certains cancers (notamment le cancer du sein en situation adjuvante, c'est-à-dire après une chirurgie ou une radiothérapie), les nouveaux traitements ne permettent pas d'espérer des différences significatives sur la SG.

Des critères de substitution intermédiaires sont de plus en plus préconisés afin de réduire la durée des études et/ou de limiter le nombre de patients à recruter. Dès lors, peu de critères de substitution étant validés, la QdV pourrait être considérée comme critère de jugement principal ou co-critère principal avec un critère tumoral (Moinpour et al.).
Certains cliniciens ont également suggéré que la QdV pouvait se substituer à la SG

comme critère principal en situation avancée (Methy et al.). Des freins conceptuels (définition, évaluation subjective par le patient) et méthodologiques (sens clinique d'un changement, données longitudinales, nature multidimensionnelle, données manquantes, etc) limitent l'évaluation de la QdV et l'utilisation des résultats en pratique courante.

De nombreux questionnaires de QdV spécifiques du cancer sont disponibles et validés : tels que le questionnaire EORTC QLQ-C30 développé par l'organisation européenne de recherche et traitement contre le Cancer (European Organization for Research and Treatment of Cancer (EORTC), le Functional Assessment of Cancer Therapy-General (FACT-G), le Rotterdam Symptom Checklist (RSCL) et le Functional Living Index-Cancer (FLIC).

Le questionnaire QLQ-C30 est le plus utilisé dans les essais cliniques de phase III en cancérologie (Fayers et al.). Il est composé de 30 items permettant de mesurer 15 échelles de QdV : 5 échelles fonctionnelles, 3 échelles de symptômes, 6 items uniques (symptômes ou problèmes) et 1 échelle de QdV/santé globale. Des modules spécifiques pour la plupart des localisations cancéreuses peuvent être ajoutés (Bergman et al.). L'EORTC recommande au minimum 3 mesures au cours de l'essai : avant, pendant et après le traitement (*Glossary | EORTC*). Une évaluation plus intensive de la QdV est cependant encouragée pour permettre la mise en évidence d'une différence cliniquement intéressante (Bonnetain, Fiteni, et al.).

La QdV est considérée comme un critère de jugement qui évalue un bénéfice clinique pour le patient. Par conséquent la QdV apparaît comme candidat en tant que critère de jugement principal des essais cliniques en cancérologie avec deux avantages : une durée d'évaluation plus courte que celui de SG et une évaluation du bénéfice clinique pour le patient.

Il conviendrait de remédier aux difficultés d'analyse de la QdV dues aux types de données générées et à son aspect multidimensionnel afin d'assurer une standardisation des analyses (longitudinales) qui permettra une comparabilité entre les résultats de différents essais (Bottomley et al.).

Les données manquantes sont un problème majeur. Dans les études longitudinales, il peut manquer certaines données intermittentes car le patient ne s'est pas rendu à sa consultation ou n'a pas rempli certains questionnaires. Par conséquent, certains biais dus aux données manquantes informatives (i.e. non aléatoires) peuvent impactés les résultats de QdV et donc diminuer la confiance dns les résultats.

Les techniques statistiques robustes de gestion des données manquantes ont été proposées comme l'imputation multiple, et le score de propension (Anota, Mouillet, et al.). Pour les données manquantes aléatoires, les patterns mixture modèles pour les données manquantes non aléatoires.

Une des voies de recherche est le développement du critère de jugement "temps jusqu'à détérioration d'un score de QdV" pouvant être défini comme l'intervalle de temps entre la date d'inclusion dans l'étude et la détérioration du score de QdV (Bonnetain). Plusieurs définitions de la détérioration peuvent être données par exemple l'intervalle de temps entre la date d'inclusion et l'apparition d'une première détérioration du score de QdV qui soit cliniquement pertinente par rapport au score de QdV à l'inclusion (Hamidou et al.). Cette définition a été recommandée en situation adjuvante (Anota, Hamidou, et al.).

En situation avancée ou métastatique, il est préférable d'étudier une détérioration définitive de la QdV. Ainsi, le temps jusqu'à détérioration de la QdV a été définie en situation avancée comme l'intervalle de temps depuis la date d'inclusion et l'apparition d'une première détérioration cliniquement pertinente par rapport au score à l'inclusion, sans amélioration ultérieure cliniquement pertinente par rapport au score à l'inclusion, en incluant ou non le décès comme évènement (Bonnetain, Dahan, et al.; Anota, Hamidou, et al.).

Des guides de bonne pratique devraient être proposés comme des critères RECIST (Response Evaluation Criteria In Solid Tumours) mais appliqués à la QdV (Anota, Hamidou, et al.).

Pour les essais cliniques, des recommandations ont été proposées sur les éléments à rapporter dans les publications sous la forme de critères CONSORT (« Consolidated Standards of Reporting Trials ») (Schulz et al, 2010) (Schulz et al.). Ces recommandations sont suivies à travers le monde et sont demandées dans la plupart des revues lorsqu'il s'agit de publier les résultats d'un essai clinique.

Des PROs CONSORT (Calvert et al.) ont été publiés mais ils ne traitent pas ou peu de l'analyse statistique de la QdV. Par conséquent, des recommandations sur l'analyse statistique de la QdV devraient être développées. Quelques recommandations ont été données sur les définitions de temps jusqu'à détérioration (TJD) à appliquer selon les situations thérapeutiques par Anota et al. (Anota, Hamidou, et al.). Ainsi, en situation adjuvante, le TJD simple en tant qu'état transitoire semble la définition la plus adaptée puisque le patient a de grandes chances de guérir de son cancer.

En revanche, en situation avancée ou métastatique, il parait plus pertinent d'étudier le TJD définitif, intégrant ou non le décès dans la définition de l'évènement, selon la définition de Bonnetain et al (Bonnetain, Dahan, et al.).

Un projet a été initié par l'EORTC afin de proposer une standardisation de l'analyse de la QdV. Cette standardisation facilitera la comparaison entre les essais cliniques et permettra à long terme une meilleure considération des résultats de QdV.

B. Les critères centrés sur la tumeur

Les critères de jugement centrés sur la tumeur sont des marqueurs biologiques/tumoraux. La plupart de ces critères sont des réponses pharmacologiques à une intervention thérapeutique. Par exemple, la réponse tumorale selon RECIST (Eisenhauer et al.), le taux de cellules tumorales circulantes, la survie sans maladie et la survie sans progression sont des critères de jugement centrés sur la tumeur. Néanmoins, ils ne reflètent pas un bénéfice direct pour le patient. Les conclusions des essais cliniques qui utilisent ce genre type de critères sont remises en question quant au bénéfice réel, les agences de régulation demandent souvent une analyse complémentaire de la QdV si celle-ci n'a pas été réalisée. Leur validation comme critère de substitution à des critères centrés sur le patient est un préalable nécessaire au préalable.

1. Les critères composites intermédiaires

Les critères de jugement centrés sur la tumeur sont généralement des critères composites, ils sont basés sur l'évaluation de la tumeur et combinent différents évènements comme la progression locale et à distance, la récidive locale et à distance, la survenue d'un second cancer, le décès ou une toxicité sévère (Fleming et Powers). Les plus fréquemment utilisés et

cités sont la survie sans maladie, la survie sans récidive, la survie sans progression, le temps jusqu'à progression, ou la survie spécifique.

La variété des critères de jugement composites de type temps jusqu'à évènement et la variabilité de leurs définitions, notamment de la survie sans progression et de la survie sans maladie, sont reconnues comme un problème méthodologique majeur et les recherches doivent être poursuivies pour améliorer leur fiabilité et leur reproductibilité (Mathoulin-Pelissier et al., « Survival End Point Reporting in Randomized Cancer Clinical Trials »; Kemp et Prasad).

Les critères composites tels que la survie sans progression sont de plus en plus utilisés, non seulement en phase II mais également comme critère de substitution à la place de la SG dans les phases III. Pourtant, malgré l'utilisation répandue de ces critères, ils sont souvent mal définis, et lorsqu'ils le sont leurs définitions peuvent être très variables entre essais cliniques randomisés (ECR). Cette variabilité peut avoir un impact important sur la puissance statistique de l'étude, l'estimation de l'effet des traitements, et ainsi sur les conclusions des ECR. De plus, cela rend difficile la comparaison des résultats entre les essais.

Dans ce contexte, le projet Definition for the Assessment of Time-to-event Endpoints in CANcer trials (DATECAN-1) a été développé pour élaborer des définitions standardisées des critères de jugement de type temps jusqu'à évènement en utilisant une méthode de consensus formalisée (Bonnetain, Bonsing, et al.; Bellera et al.; Gourgou-Bourgade et al.; Kramar et al.)(Moinpour et al.). Ceci est la première étape avant l'utilisation de tels critères composites et l'évaluation de leur capacité de substitution à la SG (projet DATECAN-2).

2. Les critères de substitution

Un critère de substitution est un critère intermédiaire mesurable de façon fiable et reproductible (imagerie, biologie, …), et qui permet de prédire l'effet du traitement sur le critère clinique. Ainsi en cancérologie, on cherche à substituer un critère principal dont l'événement est trop long à observer ou pas assez spécifique.

La SG est le critère de référence mais avec l'augmentation du nombre de traitements efficaces disponibles dans beaucoup de cancers, il est nécessaire d'inclure un nombre plus important de patients et de les suivre plus longtemps ce qui augmente le coût des essais cliniques. Ainsi les critères de jugement centrés sur la tumeur qui pourraient être évalués plus précocement sont utilisés comme critères de substitution de la SG. Ils sont de plus en plus étudiés mais la plupart d'entre eux n'ont pas de définitions standardisées limitant la comparaison des résultats de différents essais cliniques et ne sont pas validés comme critères de substitution.

Un critère de substitution idéal doit posséder les caractéristiques suivantes :

- Il doit être fiable, reproductible, facilement disponible et facilement mesurable.
- Il doit exister une relation entre le critère de substitution et la maladie (le critère clinique).
- Il doit être un véritable prédicteur de la maladie (ou du risque de maladie) et non pas une simple co-variable associée à la survenue de la maladie.
- Il doit être sensible.
- Il doit être spécifique : un résultat « négatif » doit prédire avec une quasi-certitude l'absence de résultat clinique.

La relation entre le critère de substitution et la maladie doit être biologiquement plausible. Il doit y avoir un seuil précis de séparation entre les valeurs normales et anormales. Les changements dans le critère de substitution doivent rapidement et précisément refléter la

réponse au traitement. En particulier, les taux doivent être normalisés dans les états de guérison ou de rémission.

IV. La validation des critères de substitutions

De nombreuses techniques statistiques ont été développées ces dernières années dans le but de valider les critères intermédiaires comme critère de substitution (Weintraub et al.). Avant qu'un critère de substitution puisse remplacer un critère de référence, il doit être formellement valide. Il existe différentes approches pour la validation des critères de substitution.

Malgré le fait que la survie sans progression et le temps jusqu'à progression aient été souvent utilisés et acceptés comme critère de substitution à la SG en cancérologie, la méthodologie pour établir la substitution est toujours en cours d'évolution. Aucun consensus n'existe sur les méthodes à utiliser pour identifier des critères de substitution. Cependant, deux grands types d'approches se dégagent : celles basées sur un essai clinique et la méthode méta-analytique.

A. Approche par essai

1. Critère de validation de Prentice

Dans un article publié en 1989, Prentice et al. proposent une définition des critères de substitution (Prentice). Ils expliquent qu'un critère de substitution devrait « apporter une information claire de l'effet différentiel du traitement sur le vrai critère de jugement ».

Prentice défini un critère de substitution comme : « A response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint »(Prentice), soit « une variable réponse pour laquelle un test d'hypothèse nulle de l'absence de relation au traitement

est également un test valide de l'hypothèse nulle correspondante pour le critère final (critère de référence) ».

Supposons que la variable X corresponde au bras de traitement, T corresponde au critère final, et S corresponde au critère de substitution.

Selon la définition de Prentice, on peut l'écrire mathématiquement comme :

$$f(X|S) = f(S) \Leftrightarrow f(T|S) = f(T) \quad (1)$$

Avec $f(X)$ la distribution de probabilité de X et $f(X|S)$ la distribution conditionnelle.

4 critères opérationnels ont été proposés pour vérifier que le triplet (X, T, S) satisfait cette définition :

1)	$f(S\|X) \neq f(S)$	Effet du traitement statistiquement significatif sur S
2)	$f(T\|X) \neq f(T)$	Effet du traitement statistiquement significatif sur T
3)	$f(T\|S) \neq f(T)$	Effet pronostique de S sur T statistiquement significatif
4)	$f(T\|S,X) = f(T\|S)$	Effet du traitement non statistiquement significatif sur T après ajustement sur S, i.e. S capte complétement l'effet du traitement sur T

Cependant, la principale limite de cette méthode est la démonstration de la dernière condition, puisque le rejet de l'hypothèse nulle peut être simplement dû à un manque de puissance.

Pour simplifier les notations, nous considérons les critères T et S comme étant des critères binaires (T ou S=0 si absence d'événement et T ou S=1 si présence d'événement).

Par la modélisation mathématique:

Si on note $P(T = 1|X)$ l'estimateur de $P(T = 1|X = x)$, on peut écrire alors :

$$P(T = 1|X) = \int P(T = 1|X, S)f(S|X)ds \quad (1.1)$$

Avec $f(S|X)$ la notation pour $f(S = s|X = x)$. $f(S)$ est la distribution de probabilité de la variable aléatoire S.

La condition (4) équivaut à la formule suivante:

$$P(T = 1|S, X) = P(T = 1|S). \quad (1.2)$$

Si S et X sont indépendants, on a alors :

$$f(S|X) = f(S) \quad (1.3)$$

D'après les trois équations (1.1), (1.2) et (1.3), on a:

$$P(T = 1|X) = \int P(T = 1|X, S)f(S|X)ds \quad \text{(étape 1.1)}$$

$$= \int P(T = 1|S)f(S|X)ds \quad \text{(étape 1.2)}$$

$$= \int P(T = 1|S)f(S)ds \quad \text{(étape 1.3)}$$

Donc :

$$P(T = 1|X) = P(T = 1) \quad (1.4)$$

Si l'équation (1.2) est vérifiée : $(1.3) \Rightarrow (1.4)$ i.e il n'y a pas d'effet traitement sur S, donc pas d'effet traitement sur T.

Soit : non (1.4) $\Rightarrow$ non (1.3) i.e. il y a un effet traitement sur T, ce qui implique qu'il y a un effet traitement sur S.

On pose :

$$P(T = 1|S) \neq P(T = 1) \quad (1.5)$$

Cela signifie que T et S sont corrélés (non indépendants).

Si (1.5) n'est pas vérifié, i.e $P(T = 1|S) = P(T = 1)$ alors l'étape (1.3) peut être remplacée par :

$P(T = 1|X) = \int P(T = 1)f(S|X)dS = P(T = 1)$ sans la condition (1.3).

La condition (1.5) est donc nécessaire à la relation : non (1.3) ⇒ non (1.4), i.e effet traitement sur S ⇒ effet traitement sur T.

Nous pouvons en déduire que:

$f(S|X) \neq f(S) \Rightarrow \int P(T = 1)f(S|X)dS \neq \int P(T = 1|S)f(S)ds$

On a alors non (1.3) ⇒ (1.4).

Buyse et al. (Buyse et al.) ont discuté et détaillé les fondements théoriques des critères de Prentice. Ils ont démontré que la condition (4) est nécessaire pour les critères binaires, mais pas dans le cas général. Dans le cas des critères de survie, les critères peuvent n'être ni nécessaires ni suffisants pour être validé comme des critères de substitutions. On peut démontrer que si la condition (4) est vérifiée, alors un effet de traitement significatif sur le critère final implique un effet de traitement significatif sur le critère de substitution, mais la réciproque n'est pas systématique vraie.

Limites :

- Un critère de substitution ne peut être validé que s'il y a effectivement un effet du traitement sur S et T, ce qui n'est pas toujours le cas.
- La condition (3) précise simplement qu'il existe un lien entre T et S, ce qui devrait être le cas de tout candidat.
- La dernière condition (4) est le plus importante : une fois ajusté au critère de substitution, l'effet du traitement sur T n'est plus significatif.

Ces critères sont critiquables. Premièrement, ils reposent sur des tests d'hypothèses qui dépendent des erreurs de type I et II. Cela implique qu'un manque de puissance peut conduire à une fausse conclusion.

Une autre faiblesse concerne le dernier critère, qui revient à prouver la validité de l'hypothèse nulle. Alors qu'un test d'hypothèse peut seulement accepter ou rejeter l'hypothèse nulle. Tout ceci a conduit à considérer une approche par estimation et prédiction plutôt que par des tests d'hypothèse pour la validation des critères de substitution.

Exemple :

Considérons un critère de substitution S=SSP (survie sans progression) dépendant du temps, la survie globale T=SG et X (X=0 ou 1) le bras de traitement.

On vérifie les 4 conditions de Prentice :

1). L'effet significatif du traitement par le test du log-rank sur les courbes de Kaplan-Meyer obtenues avec le critère de substitution S. Critère vérifié à 5% de risque d'erreur si la p-value < 0.05.

2). L'effet significatif du traitement par le test du log-rank sur les courbes de Kaplan-Meyer obtenues avec le critère final T. Critère vérifié à 5% d'erreur si la p-value < 0.05.

3). On calcule la corrélation entre S et T.

4). On ajuste aux données le modèle de Cox $P(T = t|S, X) = \mu_{T|X,S}(t)\exp(\beta_S X + \gamma_X S)$. Le critère est considéré comme validé si $\beta_S = 0$, i.e si la p-value du test d'hypothèse nulle « $\beta_S = 0$ » est grande (à définir, par exemple si p-value > 0.1).

2. Freedman : Proportion d'effet du traitement (PE) expliqué par le critère de substitution

Une approche plus directe a été développée par Freedman (Freedman et al.), qui consiste à estimer la proportion d'effet de traitement (PE) expliqué par le critère de substitution. Un estimateur naturel est donné par :

$$PE = 1 - \left(\frac{\beta_S}{\beta}\right)$$

Où β_S est l'estimation de l'effet du traitement sur T (issu d'un modèle de régression, modèle de Cox, ou autre) après ajustement sur le critère de substitution et β l'estimation de l'effet du traitement sur T non ajusté.

La proportion d'effet de traitement (PE) est grande si β_S est petit devant β. Le critère de Prentice c4 nécessite que $\beta_S = 0$, soit que $PE = 1$.

On peut fixer des limites pour conclure si S explique suffisamment l'effet du traitement, telles que la borne inférieure de l'IC à 95% de cet estimateur doit être supérieure à 0.5 ou 0.75, de telle sorte que le critère de substitution explique au moins 50% ou 75% de l'effet de traitement.

Par la modélisation mathématique:

1. On estime β à l'aide d'un modèle adapté (ex: modèle de Cox pour des données de survie)
2. On estime β_S à l'aide du modèle précédent en incluant le critère de substitution
3. On calcule $\widehat{PE}$

La proportion d'effet de traitement (PE) est un ratio de deux paramètres, il est donc possible de calculer son intervalle de confiance grâce au théorème de Fieller ou à la delta-méthode (Cox).

Exemple : $T = OS, S = PFS$ avec 2 bras de traitement $(X = 0\ ou\ 1)$

On ajuste les modèles semi-paramétriques de Cox ci-dessous afin d'estimer β et β_S puis PE :

1. $P(T = t|X) = \mu_{T|X}(t)\exp(\beta X) \quad \Rightarrow \quad \hat{\beta}$
2. $P(T = t|S, X) = \mu_{T|X,S}(t)\exp(\beta_S X + \gamma_X S) \quad \Rightarrow \quad \hat{\beta}_S$
3. $\widehat{\mathrm{PE}} = 1 - \left(\frac{\hat{\beta}}{\hat{\beta}_S}\right)$

Cependant, certaines limites ont été mises en évidence :

- L'intervalle de confiance est généralement trop large pour apporter une quelconque information, sauf lorsque le paramètre β est très significatif (> 4 x l'écart-type), ce qui arrive rarement puisque cela implique un effet de traitement très important, et donc facilement décelable sur T.
- la proportion d'effet de traitement (PE) n'est pas une proportion : il peut être supérieur à 1 si β_S et β sont de signe opposé, c'est-à-dire si l'ajustement sur S modifie le sens de l'effet de X sur T. L la proportion d'effet de traitement peut aussi être négatif dans certains cas. Il est possible de construire des exemples où la PE peut prendre n'importe quelle valeur prédéfinie.

3. Buyse et Molenberghs : Effet relatif et association ajustée

Buyse et Molenberghs (Molenberghs et al.) ont développé deux approches pouvant remplacer le PE :

- L'effet relatif (RE) $RE = \frac{\beta}{\alpha}$
- L'association ajustée γ_X, l'effet de S sur T après ajustement sur X.

On peut en fait montrer que $E = \frac{\gamma_X}{RE}$, i.e PE n'est en fait que le ratio de l'association ajustée et de l'effet relatif. Il est cependant préférable de garder ces deux valeurs séparées.

Par la modélisation mathématique:

On estime β et α en ajustant les modèles adaptés (exemple : modèles semi-paramétriques de Cox dans le cas de données de survie). On peut alors estimer RE par $\widehat{RE} = \frac{\hat{\beta}}{\hat{\alpha}}$

On estime γ_X en ajustant le modèle adapté.

On estime RE et PE (cf paragraphe précédent), puis on estime ρ_X par $\widehat{\rho_X} = \widehat{PE} \times \widehat{RE}$

Exemple : $T = SG, S = SSP$ avec deux bras de traitement $(X = 0 \; ou \; 1)$

On ajuste les modèles semi-paramétriques de Cox ci-dessous afin d'estimer RE et γ_X:

1. $P(T = t|X) = \mu_{T|X}(t)\exp(\beta X) \quad \Rightarrow \quad \hat{\beta}$
2. $P(T = s|X) = \mu_{S|X}(t)\exp(\alpha X) \quad \Rightarrow \quad \hat{\alpha}$
3. $\widehat{RE} = \frac{\hat{\beta}}{\hat{\alpha}}$
4. $P(T = t|S, X) = \mu_{T|X,S}(t)\exp(\beta_S X + \gamma_X S) \quad \Rightarrow \quad \hat{\gamma}_X$

B. Approche par méta-analyse

La méta-analyse permet de démontrer une relation forte entre un critère intermédiaire et un critère objectif. Un essai unique ne pas apporter une réponse statistiquement solide pour la validation d'un critère de substitution. Il est essentiel et convaincant d'évaluer une association forte entre deux critères sur plusieurs essais. Il est possible de choisir une autre unité que l'essai, par exemple le centre, le pays, etc. Le choix de l'unité dépend du nombre des unités permettant d'estimer un effet de traitement et la précision des estimateurs des différents paramètres liés à la taille des unités. Cette approche utilise généralement les notions de prédiction et d'association pour évaluer les relations entre

les critères de jugement. Les deux notions essentielles sont : l'association au niveau individuel et l'association au niveau de l'étude.

1. Régression linéaire pondérée

La méthode de la régression linéaire pondérée peut s'appliquer pour étudier les critères de substitutions au niveau individuel ou bien au niveau de l'essai (Paoletti et Bonnetain) (www.unitheque.com).

On note $\mathbf{j = 1, \dots, J}$ les essais, n_j la taille de l'essai j, $\mathbf{x = 1, \dots, X}$ les traitements, $\mathbf{i = 1, \dots, } n_J$ les patients pour chaque essai.

On se place dans le cas de données de survie, avec S et T des taux de survie.

Association au niveau individuel :

La régression linéaire pondérée par la taille de chacun des essais permet de mesurer le coefficient de corrélation entre un critère de substitution S et un critère final T. Cela peut être modélisé par :

$$S_{j,x} = \gamma_0 + \gamma_1 * \omega_j * T_{j,x} + \varepsilon_{j,x} \qquad (2.1)$$

Avec ω_j la pondération pour l'essai j dépendant de n_j et $\varepsilon_{j,X} \sim N(0,1)$ terme d'erreur pour le bras de traitement X de l'essai j. La variabilité de T expliquée par S est quantifiée par le coefficient R^2.

Association au niveau de l'étude :

On ne modélise plus la relation entre les taux pour chaque bras de traitement, mais la relation entre effets du traitement sur S et T. Dans le cas de données de type survie, on peut donc estimer les coefficients hazard ratios (calculés à l'aide d'un modèle semi-paramétrique de Cox) et exprimer leur relation à l'aide d'un modèle linéaire pondéré :

$$HR_j^T = \gamma_0 + \gamma_1 * \omega_j * HR_j^{TS} + \varepsilon_j \qquad (2.2)$$

Avec ω_j pondération pour l'essai j (dépend de j) et $\varepsilon_j \sim N(0,1)$ terme d'erreur pour l'essai j.

De même, la variabilité de T expliquée par S est quantifiée par le coefficient R^2.

Une mesure de concordance peut être calculée pour estimer la similitude des conclusions du test du log-rank obtenues avec S et T (coefficient Kappa de Cohen) (Bland et Altman; McHugh).

Plus le R^2 est grand, meilleur est le critère de substitution : $R^2 = 1$ signifie que les effets du traitement sur S et T sont multiples l'un de l'autre (à une constante près), soit un critère de substitution parfait.

Un paramètre γ_1 significativement différent de 1 indique une atténuation ou une augmentation de l'effet du traitement sur le critère de substitution S par rapport à l'effet du traitement sur le critère final T.

Par la modélisation mathématique:

La démarche est la même dans le cas des associations au niveau individuel et de l'essai :

1. On estime les $S_{j,x}$ (resp HR_j^S) et $T_{j,x}$ (resp HR_j^T) pour chaque essai ;
2. On ajuste le modèle (2.1) (resp (2.2)) aux données pour estimer γ_1 et γ_0.

Exemple : Notons $T = SG, S = SSP$ avec deux bras de traitement $(X = 0 \text{ ou } 1)$

On ajuste les modèles semi-paramétriques de Cox ci-dessous afin d'estimer RE et γ_X:

1. $P(T_j = t|X) = \mu_{T_j|X}(t)\exp(\beta_j X) \quad \Rightarrow \quad \hat{\beta}_j = \widehat{HR}_j^T$
2. $P(S_j = s|X) = \mu_{S_j|X}(t)\exp(\alpha_j X) \quad \Rightarrow \quad \hat{\alpha}_j = \widehat{HR}_j^S$

Puis on ajuste le modèle de régression linéaire suivant pour estimer γ_1 et R^2 :

3. $\widehat{HR}_j^T = \gamma_0 + \gamma_1 * \omega_j * \widehat{HR}_j^S + \varepsilon_j \; \forall \, j = 1, \ldots, J \qquad \Rightarrow \qquad \hat{\gamma}_1$ et R^2

Considérons un nouvel essai $j = 0$ de taille n_0 pour lequel on connait $\widehat{HR}_0^S$. Après estimation des paramètres, on a donc $\hat{\gamma}_1$ et $\hat{\gamma}_0$ connus.

L'effet du traitement sur le critère HR_0^T final pourra alors être estimé à l'aide de l'équation :

$$\widehat{HR}_0^T = \hat{\gamma}_0 + \hat{\gamma}_1 * \omega_j * \widehat{HR}_0^S$$

Les avantages de cette méthode sont nombreux :

- Il s'agit d'une approche simple et intuitive ;
- Elle permet une représentation graphique et des interprétations simples ;
- Identification des observations extrêmes ;
- C'est une approche cohérente avec la planification des essais (taux à un temps donné) ;
- Les estimations et prédiction sont simples.

Les limites de cette méthode sont également importantes:

- cette méthode réduit les données d'un bras de traitement à une unique observation par essai ;
- Elle ne permet pas d'inclusion directe de variables ;
- Elle ne prend pas en compte l'erreur d'estimation des HR ce qui implique que la précision des prédictions et du R^2 peut-être surestimée, en particulier si les études sont de taille modeste ;
- Le fait de calculer la survie à un temps donné peut limiter l'information disponible car selon les études l'ensemble du suivi des patients n'est pas forcément pris en compte ;
- En ce qui concerne la mesure de concordance, cette approche dépend fortement de la puissance des comparaisons avec chacun des critères. Ainsi, si le critère de substitution est plus sensible que le critère final, un effet du traitement mesuré ne sera pas forcément retrouvé avec le critère principal à cause d'un manque de puissance.

2. Modélisation jointe

i. Critères gaussiens

Deux approches stratégiques peuvent être suivies pour modéliser de manière conjointe les deux critères d'évaluation, lorsque ceux-ci sont gaussiens (Buyse et al 2000) (Buyse et al.).

Dans la première, on considère tout d'abord un modèle linéaire à effets fixes :

$$S_{ji}|X_{ji} = \mu_{Sj} + \alpha_j X_{ji} + \varepsilon_{Sji} \qquad (3.1)$$

$$T_{ji}|X_{ji} = \mu_{Tj} + \beta_j X_{ji} + \varepsilon_{Tji}$$

Avec :

- μ_{Sj} et μ_{Tj} interceptes propres à chaque essai
- α_j et β_j effets du traitement X sur les critères spécifiques à chaque essai
- ε_{Sji} et ε_{Tji} termes d'erreur corrélés gaussiens, supposés centrés
- $\Delta t \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ . & \sigma_{TT} \end{pmatrix}$ matrice de variance-covariance des résidus.

Puis on suppose que :

$$\begin{pmatrix} \mu_{Sj} \\ \mu_{Tj} \\ \alpha_j \\ \beta_j \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Sj} \\ m_{Tj} \\ a_j \\ b_j \end{pmatrix} \qquad (3.2)$$

Où $\begin{pmatrix} m_{Sj} \\ m_{Tj} \\ a_j \\ b_j \end{pmatrix}$ est supposé suivre une loi normale centrée de matrice de variance-covariance :

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ . & d_{TT} & d_{Ta} & d_{Tb} \\ . & . & d_{aa} & d_{ab} \\ . & . & . & d_{bb} \end{pmatrix}$$

La seconde approche combine en une seule étape effets fixes et aléatoires à l'aide du modèle :

$$S_{ji}|X_{ji} = \mu_S + m_{Sj} + \alpha X_{ji} + a_j X_{ji} + \varepsilon_{Sji} \quad (3.3)$$

$$T_{ji}|X_{ji} = \mu_T + m_{Tj} + \beta X_{ji} + b_j X_{ji} + \varepsilon_{Tji}$$

Les avantages et inconvénients des deux approches ont été discuté par de nombreux auteurs (Thompson et Pocock, 1991 ; Fleiss, 1993 ; Tompson, 1993 ; Senn, 1998) (Thompson et Pocock; Fleiss; Thompson; Senn). Il apparait cependant que les conclusions des deux méthodes ne divergent que dans le cas de situations pathologique ou bien lorsque le nombre d'unités (essai, centre, …) disponible est très petit.

Association au niveau individuel :

L'association entre T et S après ajustement sur l'effet du traitement, est exprimée par (Buyse et al. 2000) :

$$R^2_{indiv} = R^2_{\varepsilon_{TJ}|\varepsilon_{SJ}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}} \quad (3.4)$$

Association au niveau de l'étude :

On s'intéresse cette fois à l'effet de X sur T sachant l'effet de X sur S. On peut quantifier cette association par un autre coefficient de détermination, notée R^2_{essai} (Buyse et al. 2000)

$$R^2_{essai} = R^2_{b_J|a_J} = \frac{d^2_{ab}}{d_{aa}d_{bb}} \quad (3.5)$$

On aura $R^2_{essai} = 1$ si les effets traitement sont des multiples l'un de l'autre, soit dans le cas d'un critère de substitution parfait.

Par la modélisation mathématique:

Approche à 2 étapes

On ajuste tout d'abord le modèle linéaire mixte (3.1) afin d'estimer les paramètres du vecteur $\begin{pmatrix} \mu_{Sj} \\ \mu_{Tj} \\ \alpha_j \\ \beta_j \end{pmatrix}$ et $\Delta = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ . & \sigma_{TT} \end{pmatrix}$.

Puis, on estime $\begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix}$ et $D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ . & d_{TT} & d_{Ta} & d_{Tb} \\ . & . & d_{aa} & d_{ab} \\ . & . & . & d_{bb} \end{pmatrix}$ en ajustant le modèle (3.2).

Il est alors facile de calculer R^2_{indiv} et R^2_{essai}.

On cherche ici à estimer, pour un nouvel essai j=0, l'effet du traitement sur le critère final T sachant l'effet du traitement sur le critère de substitution S, soit $\beta + b_0 | m_{S_0}, a_0$ que l'on peut estimer intuitivement par $E(\beta + b_0 | m_{S_0}, a_0)$.

1. On estime tout d'abord a_0 et μ_{S_0} en ajustant le modèle suivant aux données du nouvel essai :

   $$S_{0i} = \mu_{S0} + a_0 X_{0i} + \varepsilon_{S0i}$$

2. On remarque que $(\beta + b_0 | m_{S_0}, a_0)$ suit une loi normale de moyenne et variance :

   $$E\left(\beta + b_0 \middle| m_{S_0}, a_0\right) = R^2_{essai} = R^2_{b_j | m_{sj}, a_j} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}$$

   Et

$$Var(\beta + b_0 | m_{S_0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}$$

On a donc :

$$\beta + b_0 \mid m_{S0}, a_0 = E(\beta + b_0 \mid m_{S0}, a_0) = \hat{\beta} + \begin{pmatrix} \hat{d}_{Sb} \\ \hat{d}_{ab} \end{pmatrix}^T \begin{pmatrix} \hat{d}_{SS} & \hat{d}_{Sa} \\ \hat{d}_{Sa} & \hat{d}_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\mu}_{S0} - \hat{\mu}_S \\ \hat{\alpha}_0 - \hat{\alpha} \end{pmatrix}$$

Il est possible de calculer un intervalle de confiance autour de l'estimation en appliquant la delta-méthode. Intuitivement un critère de substitution « parfait pour le niveau de l'essai » sera associé à une variance nulle.

Afin d'évaluer la qualité du critère de substitution au niveau de l'essai, on peut donc utiliser la mesure suivante :

$$R^2_{essai} = R^2_{b_j \mid m_{Sj}, a_j} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}$$

Cette mesure s'estime en remplaçons les valeurs du membre de droite par leurs estimations. Dans le cas d'un critère de substitution parfait au niveau de l'essai, on aura R^2_{essai}=1.

ii. Critère de survie

Pour étendre l'approche proposée par Buyse et al. (Buyse et al.) au cas où le critère final et le substitut sont des critères de survie, Burzykowski et al. (2001)(Burzykowski et al., « Validation of Surrogate End Points in Multiple Randomized Clinical Trials with Failure Time End Points ») proposent de remplacer la première étape du modèle par un modèle de copula ; Plus précisément, ils supposent que la distribution du couple (S_{ji}, T_{ji}) puisse s'écrire :

$$F(S,t) = P(S_{ji} \geq S_i, T_{ji} \geq t) = C_\theta \left\{ F_{Sji}(S \mid X_{ji}), F_{Tji}(t \mid X_{ji}) \right\}, s, t \geq 0$$

C. Classification de Fleming: hiérarchisation des critères de substitution

Une hiérarchisation des critères de substitution a été proposée par Fleming en 2012 afin de choisir un meilleur critère de substitution au critère final (Fleming et Powers).

Cette classification est basée sur quatre niveaux :

- Niveau 1: Critère de jugement mesurant un réel bénéfice clinique pour le patient (survie globale et qualité de vie relative à la santé).
- Niveau 2: Un critère de substitution validé, selon l'approche méta-analytique (de tels critères sont rares)
- Niveau 3: Un critère de substitution non validé, mais qui a néanmoins raisonnablement des chances de prédire le bénéfice clinique souhaité, sur des considérations statistiques et cliniques.
- Niveau 4: Une mesure biologique corrélée au critère clinique.

V. Les critères conjoints

Les critères de jugement conjoints (ou multiples) sont une association entre deux ou plusieurs critères de jugement clinique pour former un unique critère sur lequel portera l'analyse statistique. On utilise cette astuce pour augmenter la probabilité de mettre en évidence une différence statistiquement significative entre les médicaments comparés, tout particulièrement lorsque l'essai est de taille réduite ou lorsque les médicaments, d'action voisine, s'avèrent difficiles à départager(Chuang-Stein et al.).

On associe donc dans ces critères dits "conjoints" des situations hétérogènes comme la QdV et la SSP ou la survie sans maladie (SSM, en anglais Disease-Free Survival (DFS)). Ces critères sont différents, ils ne revêtent pas la même importance pour le clinicien ou le patient.

En définitive, il faut proscrire dans l'évaluation des médicaments l'utilisation des critères de jugement conjoints pour utiliser des critères cliniques pertinents (c'est-à-dire importants pour

le patient et le clinicien) et surtout validés (les critères conjoints ne sont généralement pas validés, c'est-à-dire vérifiés).

VI. Le calcul du nombre de sujets nécessaires

L'estimation du nombre de sujets nécessaires (NSN) d'un essai clinique randomisé utilisant un critère de jugement de temps jusqu'à évènement dépend de plusieurs paramètres (Gail) :

- le type d'essai (supériorité, non infériorité ou équivalence)
- La puissance statistique 1-β souhaitée
- Le taux de survie dans chaque bras de traitement ou le rapport des risques (HR, Hasard ratio), soit la différence attendue
- La proportion de patients à inclure dans le bras expérimental
- Le risque d'erreur de première espèce α
- Le nombre d'analyse intermédiaire souhaité
- La durée de la période d'inclusion des patients
- La durée de suivi selon un suivi fixe pour tous les patients ou des suivis variables
- La durée totale de l'étude
- Le taux des perdus de vue

Pour mettre en évidence une différence qui existe réellement ou conclure à son absence, il faut que l'étude ait une puissance statistique suffisante, déterminée pour l'essentiel par le nombre de sujets inclus. Dans le cas particulier des comparaisons de courbes de survie, c'est le nombre d'évènements ou de décès totaux au moment de l'analyse qui donne sa puissance à l'étude.

La puissance statistique d'un essai mesure son aptitude à mettre en évidence l'effet d'un traitement si celui-ci existe. La puissance est égale à 1-β où β est le risque de deuxième espèce. La puissance est donc la probabilité d'obtenir un vrai résultat positif (mettre en évidence l'efficacité d'un traitement).

Une puissance statistique suffisante est nécessaire pour montrer qu'il existe effectivement une différence entre 2 groupes. Plus la différence entre les deux groupes est petite, plus il faudra de puissance statistique pour montrer que les deux groupes sont différents. La question essentielle du nombre de sujets nécessaires est de garantir à une étude une puissance $1-\beta$ donnée afin de mettre en évidence une différence minimale entre plusieurs groupes expérimentaux. Pour notre part, nous nous concentrerons sur le cas de deux groupes au plus.

Le test du log-rank permet de comparer plusieurs courbes de survie calculées avec la méthode de Kaplan-Meier. L'hypothèse nulle (H0) à tester est celle de l'égalité des fonctions de survie de p échantillons. On se limite dans un premier temps à la comparaison de deux groupes d'individus (échantillons) qui suivent respectivement les traitements A et B. Le principe est simple (Mantel, 1966) (Mantel) : si pour un jour donné, le groupe A et le groupe B comptent le même nombre de patients, alors il devrait y avoir en moyenne le même nombre de décès dans le groupe A et dans le groupe B, à moins que le traitement A soit beaucoup plus efficace que le traitement B, ou inversement.

Plus généralement, si les traitements sont similaires la proportion d'évènements au jour considéré devrait être la même au sein des groupes A et B. Nous voulons donc tester :

H_0 : Les proportions de décès parmi les sujets à risque à un temps t_i quelconque sont identiques entre les groupes A et B.

H_1 : Les proportions de décès à un temps t_i quelconque sont différentes entre les groupes A et B.

Un test d'hypothèse est un procédé d'inférence qui a pour but de fournir une règle de décision à une problématique scientifique. Il permet, sur la base de résultats d'échantillons aléatoires, de faire un choix entre deux hypothèses statistiques relatives à une population. Cette généralisation comporte des risques et des limites, que contrôle la démarche statistique.

Les inférences multiples sont présentes dans de nombreux essais cliniques et soulèvent des problèmes d'analyse et d'interprétation si elles ne sont pas correctement prises en compte(*A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. - PubMed - NCBI*; *Multiple Co-primary Endpoints: Medical and Statistical Solutions: A Report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of AmericaTherapeutic Innovation & Regulatory Science - Walter Offen, Christy Chuang-Stein, Alex Dmitrienko, Gary Littman, Jeff Maca, Laura Meyerson, Robb Muirhead, Paul Stryszak, Alex Baddy, Kun Chen, Kati Copley-Merriman, Willard Dere, Sam Givens, David Hall, David Henry, Joseph D. Jackson, Alok Krishen, Thomas Liu, Steve Ryder, A. J. Sankoh, Julia Wang, Chyon-Hwa Yeh, 2007*; Neuhäuser; Chuang-Stein et al.; Senn). Il est essentiel de déterminer la taille de l'échantillon appropriée pour pouvoir indiquer la signification statistique de tous les co-critères de jugements principaux tout en préservant la puissance globale puisque l'erreur de type I augmente avec le nombre de co-critères de jugements principaux(Abel et al.).

L'écueil à éviter est de considérer à tort l'existence d'une différence statistiquement significative. Pour ce faire, on a recours à des méthodes statistiques adéquates. Leur utilisation n'est toutefois pas systématique et dépend fortement des objectifs de l'essai.

Répondre à de multiples questions au cours d'un même essai clinique est un problème bien connu des chercheurs qui s'investissent dans la recherche clinique. Or, si multiplier les analyses est une tentation légitime, le risque d'aboutir dans ce contexte à des conclusions faussement positives est réel et affecte la crédibilité des résultats.

Le contrôle de ces erreurs s'effectue par un ajustement du taux d'erreur de type I. De nombreuses méthodes ont été développées dans cet objectif (« Sample Size Calculations in Clinical Research, Second Edition »). Elles utilisent habituellement un fractionnement du risque d'erreur de type I appliqué à chacune des hypothèses(Dmitrienko et al.; Pocock).
La puissance statistique d'un essai clinique avec un critère de jugement de temps jusqu'à événement dépend essentiellement du nombre d'événements à observer. Le calcul du nombre de sujets nécessaires s'effectue en deux étapes :

A. Calcul du nombre d'événements à observer

L'essai de phase III est mis en œuvre pour répondre à une question précise. L'objectif peut alors être de trois ordres:

- montrer la supériorité du nouveau traitement par rapport au traitement standard;
- montrer l'équivalence du nouveau traitement par rapport au traitement de référence;
- montrer la non-infériorité du nouveau traitement par rapport au traitement de référence.

Chacun de ces trois types d'essais a une méthodologie qui lui est propre, et un essai conçu pour montrer la supériorité du nouveau traitement par rapport au traitement de référence ne peut finalement conclure à une non-infériorité, et inversement.

Etude de supériorité ou efficacité :

Cela concerne la majorité des essais. Les hypothèses sont définies comme suit :

- H_0 : hypothèse nulle, à rejeter: le nouveau traitement est inférieur ou égal au traitement de référence;
- H_1 : hypothèse alternative: le nouveau traitement est supérieur au traitement de référence.

Lorsque l'hypothèse H_0 ne peut pas être rejetée, l'essai est non concluant. L'absence de supériorité ne permet pas de conclure à une équivalence ou à une non-infériorité.

La formule pour déterminer le nombre d'événements requis pour évaluer la supériorité d'un traitement en utilisant le test de log-rank avec un niveau de signification α et de l'erreur de type II 1-β selon (Schoenfeld) est donnée par:

$$e = \frac{\left[Z_{1-\alpha/s} + Z_{1-\beta}\right]^2}{P_E(1-P_E)\left[\ln(h_a)\right]^2}$$

Avec s = 1 (resp s = 2) si le test est unilatéral (et bilatéral) et $Z_{1-\alpha}$ est la valeur critique de (1 - α). P_E la proportion de patients randomisés dans le bras expérimental (avec $0 < P_E < 1$). h_a le hazard ratio sous l'hypothèse alternative.

Essai d'équivalence ou de non-infériorité :

Ce type d'essai est de plus en plus fréquent. Il peut servir à comparer deux formes galéniques, deux rythmes d'administration, ou encore deux molécules appartenant à une même classe thérapeutique. On peut également ne pas être intéressé par une efficacité supérieure mais par

une meilleure tolérance, une plus grande facilité d'utilisation ou un coût plus faible.

La méthodologie statistique mise en œuvre nécessite de déterminer a priori une zone d'équivalence. La valeur seuil de cette zone correspond à la perte d'efficacité maximale à laquelle on consent avec le nouveau traitement. Le choix de ce seuil est donc crucial dans la mise en œuvre d'un tel essai. Le nombre de patients à inclure est d'autant plus grand que cette zone d'équivalence est réduite. La comparaison est ensuite basée sur le test d'hypothèse:

✓ En situation bilatérale: équivalence

- H_0 : les traitements ne sont pas équivalents;
- H_1 : les traitements sont équivalents.

Le nombre d'événements requis est défini par :

$$e = \frac{\left[Z_{1-\alpha} + Z_{1-\beta/2}\right]^2}{P_E(1-P_E)\left[\ln(h_0)\right]^2}$$

✓ En situation unilatérale: non-infériorité

- H_0 : le nouveau traitement est inférieur au traitement de référence;
- H_1 : le nouveau traitement est égal ou supérieur au traitement de référence.

Le nombre d'événements requis est défini par :

$$e = \frac{\left[Z_{1-\alpha} + Z_{1-\beta}\right]^2}{P_E(1-P_E)\left[\ln(h_0)\right]^2}$$

En pratique, l'utilisation d'un test bilatéral n'a que peu d'intérêt puisqu'on court le risque de rejeter l'hypothèse d'équivalence si le nouveau traitement est supérieur au traitement de référence. Le plus souvent, c'est donc la non-infériorité qui intéresse le clinicien, c'est-

à-dire que le nouveau traitement soit au moins aussi efficace que le traitement de référence. La méthodologie statistique mise en œuvre dans ce type d'essai consiste à construire un intervalle de confiance autour du risque relatif et à comparer la borne supérieure de cet intervalle avec la limite de non-infériorité choisie.

B. Calcul du nombre de patients à inclure

Le nombre de patients à inclure dans l'essai pour observer le nombre d'événements requis *e* dépend de la probabilité ψ d'observer le premier événement d'intérêt avec N = *e* / ψ. Les paramètres de cette probabilité sont la durée d'inclusion, le suivi, le risque d'échec dans chaque bras, la durée de l'étude et le taux de perdus de vus. Cette probabilité est définie par :

$$\psi(t,\lambda,\gamma) = P_E \times \psi_E(t,\lambda_E,\gamma_E) + (1 - P_E) \times \psi_C(t,\lambda_C,\gamma_C)$$

VII. Objectives

La survie globale (SG) est considérée comme le critère de jugement principal de référence et le plus pertinent et objectif dans les essais cliniques en oncologie. Cependant, avec l'augmentation du nombre de traitements efficaces disponibles pour une majorité de cancers, un nombre plus important de patients à inclure et un suivi plus long est nécessaire afin d'avoir suffisamment de puissance statistique pour pouvoir mettre en évidence une amélioration de la SG.

De ce fait les critères de survie composites, tels que la survie sans progression, sont couramment utilisés dans les essais de phase III comme critère de substitution de la SG. Leur développement est fortement influencé par la nécessité de réduire la durée des essais cliniques, avec une réduction du coût et du nombre de sujets nécessaires.

Cependant, ces critères sont souvent mal définis, et leurs définitions sont très variables entre les essais, rendant difficile la comparaison entre les essais. Ainsi le consensus du DATECAN-1 a été développé pour standardiser les définitions des critères de jugement utilisés en cancérologie.

Dans ce contexte, le premier objectif de ma thèse était d'évaluer la qualité des définitions des critères utilisés dans essais cliniques en cancérologie. Ensuite, j'ai étudié l'impact de ces définitions sur les résultats et conclusions des essais. Pour répondre à ces deux objectifs, 9 essais cliniques randomisés (ECRs) pour la localisation pancréas ont été sélectionnés, dont un de phase II adjuvant, quatre de phase II métastatique, un de phase III adjuvant et trois de phase III métastatique.

Le second objectif de la thèse était d'étudier les propriétés des critères de jugement intermédiaire de survie en tant que critères de substitution pour la survie globale (SG). La capacité à prédire un bénéfice sur la SG à partir des résultats de l'essai sur le critère d'évaluation, n'a pas toujours été rigoureusement évaluée. La validation des critères de substitution a été faite selon une méta-analyse. La régression linéaire pondérée, a permis d'estimer conjointement le niveau d'association entre un critère final et un critère de substitution par un coefficient R^2.

Bien que la survie globale est toujours considérée comme un « gold standard » pour les critères de jugements principaux, la plupart des essais cliniques intègrent désormais la qualité de vie relative à la santé (QdV) comme critère de jugement afin d'investiguer le bénéfice clinique de nouvelles stratégies thérapeutiques pour le patient.

La majorité des essais cliniques de phase III intègrent désormais la QdV comme critère de jugement afin d'investiguer le bénéfice clinique de nouvelles stratégies thérapeutiques pour le patient. Une alternative serait de considérer un co-critère de jugement principal : un critère tumoral tel que la survie sans progression et la QdV afin de s'assurer du bénéfice clinique pour le patient. Bien que la QdV soit reconnue comme second critère de jugement principal par l'ASCO (American Society of Clinical Oncology) et la FDA (Food and Drug Administration), elle est encore peu prise en compte comme critère de jugement principal ou co-critère de jugement principal dans les essais cliniques. L'évaluation, l'analyse et l'interprétation des résultats de QdV demeurent complexes, et les résultats restent encore peu pris en compte par les cliniciens du fait de son caractère subjectif et dynamique.

Lors de la conception d'un essai clinique avec des critères de jugements principaux multiples, il est essentiel de déterminer la taille de l'échantillon appropriée pour pouvoir indiquer la signification statistique de tous les co-critères de jugements principaux tout en préservant la puissance globale puisque l'erreur de type I augmente avec le nombre de co-critères de jugements principaux. Plusieurs méthodes ont été développées pour l'ajustement du taux d'erreur de type I. Elles utilisent généralement un fractionnement du taux d'erreur de type I appliqué à chacune des hypothèses testées.

Toutes ces méthodes ont donc été investiguées dans mon projet de thèse. Afin de faciliter l'utilisation des co-critères de jugement principal et en particulier la QdV, il était essentiel de proposer un outil statistique pour déterminer facilement le nombre de sujets nécessaires pour ce type de design.

L'objectif final de mon projet de thèse était donc de développer un package R pour le calcul du nombre de sujets nécessaires avec les co-critères de jugement principal et d'étudier les critères de substitutions à la SG pour le cancer du pancréas.

VIII. Les travaux réalisés

A. Surrogate endpoints for overall survival in pancreatic cancer trials: an individual patient meta-analysis. (Jama oncology).

Méta-analyse de 9 essais cliniques randomisés du cancer de pancréas pour évaluer les critères de substitutions.

Résumé :

Dans les essais cliniques randomisés (ECR) en cancérologie, les progrès thérapeutiques ont conduit à l'utilisation de critères de jugement tels que la survie sans progression (SSP) en tant que critère de substitution de la survie globale (SG). Ces critères sont souvent mal définis, et lorsqu'ils le sont, ces définitions peuvent être très variables. Cette variabilité peut impacter les résultats et les conclusions. De plus, leur capacité de substitution à la SG, n'a pas toujours été rigoureusement évaluée. Le projet DATACAN-1 a fourni des recommandations pour les définitions des critères de jugement dans les ECR en cancérologie. Le premier objectif de DATECAN-2 est d'étudier l'impact des définitions des critères de survie sur les résultats à partir de 9 ECRs portant sur le cancer du pancréas métastatique. Le second objectif est d'étudier les propriétés des critères de substitution à la SG.

Chaque ECR a été analysé pour vérifier la cohérence entre les résultats publiés et les résultats de la base reçue. Ensuite les définitions des critères de survie de l'étude ont été comparées avec les définitions des critères issus de recommandations de DATECAN-1 pour évaluer la qualité des définitions des critères de l'étude. Enfin les résultats retrouvés de l'étude sont comparés avec les résultats issus des critères de DATECAN-1 pour vérifier s'il y a un impact sur les conclusions. Les comparaisons sont faites sur les définitions des critères, le nombre d'événement, la médiane de survie, le coefficient hazard ratio de l'effet traitement et la p-

value du test utilisé dans l'article (tel que le Log-rang). Six études en situation métastatique ont été rassemblées pour évaluer les capacités de substitutions des critères de survie à la SG par la régression linéaire pondérée.

Parmi les 9 bases analysées, 7 bases ont été publiées, 1 base est sans publication et 1 base avec des données non valides. Beaucoup d'évènements nécessaires aux définitions des critères de DATECAN-1 sont manquants ou mal définis, surtout le second cancer, le type de progression et les causes de décès ne sont toujours pas bien spécifiées. Malgré l'absence de plusieurs événements nécessaires à la définition des critères issus de recommandation de DATECAN-1, certains critères ont été reconstruits partiellement pour évaluer leurs capacités de substitution à la SG. Les coefficients de corrélations entre OS et les critères temps jusqu'à détérioration du statut OMS, SSP, temps jusqu'à la progression sont égaux respectivement à 0.7858, 7.858 et 0.7852. Et aussi OS est fortement associé au niveau de l'essai avec les trois critères de substitutions. Les valeurs de R^2 des modèles sont égales respectivement à 0.98, 0.96, 0.92.

Les résultats trouvés ne sont pas suffisants pour évaluer l'impact des définitions des critères de survie sur les résultats et les conclusions. Il est primordiale de définir les critères selon les recommandations issues de DATECAN-1 dans les nouveaux essais afin d'uniformiser les définitions et de faciliter la comparaison des résultats entre les essais cliniques. Les critères tels que SSP, temps jusqu'à la progression, temps jusqu'à détérioration du statut OMS sont potentiellement des substitutifs à la SG avec des R^2 proche de 1.

Surrogate endpoints for overall survival in advanced pancreatic cancer trials: an individual patient meta-analysis

A. PAM[1], A. ANOTA[1], D. VERNEREY[1], L. COLLETTE[2], C. LOUVET[3], B. BARON[2], L. BEDENNE[4], E. BRIASOULIS[5], B. CHAUFFERT[6], T. L. DAHAN[7], M.P. DUCREUX[8], P. FUMOLEAU[9], G. HANS[10], K. HAUSTERMANS[11], J. JEEKEL[10], F. LEVI[12], M.P. LUTZ[13], J.F. SEITZ[7], J. TAIEB[14], I. TROUILLOUD[14], J.L. VAN-LAETHEM[15], D.J.T WAGENER[16], S. GOURGOU[17], C. BELLARA[18], F. BONNETAIN[1]

[1]Methodology and Quality of Life Unit in Oncology, INSERM UMR 1098, University Hospital of Besançon, France;
[2]EORTC Data Center, Brussels, Belgium;
[3]Department of Oncology, Institute Mutualiste Montsouris, Paris;
[4]University Hospital, Dijon, France;
[5]Medical Oncology Department, School of Medicine, University of Ioannina, Greece;
[6]CHU Amiens Picardie - Site Sud (Amiens) Oncologie
[7]Hospitals public assistance of Marseille, Timone Hospital, University of Mediterranean, Marseille, France;
[8]Gastrointestinal Unit at the Institut Gustave Roussy, Villejuif, France;
[9]Department of oncology, Anticancer Centre G.F. Leclerc, Dijon;
[10]University Hospital Rotterdam Dijikzigt, Rotterdam, the Netherlands;
[11]U.Z. Gasthuisberg, Leuven, Belgium;
[12]Paul Brousse Hospital & INSERM, Villejuif, France;
[13]University of Ulm, Ulm, Deutshes;
[14]Department of Gastroenterology and Digestive Oncology, Georges Pompidou European Hospital, Paris, France;
[15]Medico-sugical Department of Gastroenerology, Erasme University Hospital, Brussels;
[16]UMC Nijmegen, Nijmegen, the Netherlands;
[17]Institut Du Cancer de Montpellier, Comprehensive Cancer Centre, and Data Center for Cancer Clinical Trials, CTD-INCa, Montpellier, France;
[18]Clinical and Epidemiological Research Unit, Institut Bergonie, Comprehensive Cancer Centre, Bordeaux, France

IMPORTANCE: In Randomized clinical trials (RCTs), time-to-event endpoints such as progression-free survival (PFS) are frequently used as potential surrogate of overall survival (OS). Such endpoints required to be precisely defined and in a standard way and surrogacy for OS should be assessed for internal validation.

OBJECTIVES: The main objective was to study the impact of the definitions proposed in DATECAN-1 on the results and conclusions on RCTs in pancreatic cancer. The potential surrogate for OS of some key time to event endpoints was then assessed.

Data Sources: Nine RCTs analyzed between 1999 and 2014 for pancreatic cancer were selected.

Data selection: We have selected pancreatic cancers phase II and III trials published and assessing at least one composite time to-event endpoint from DATECAN-1 recommendations as well as OS.

Data Extraction and Synthesis: Each study was analyzed to check the consistency between the results published and the results of the database received. Then the definitions of the time-to-event endpoint from the study selected and DATECAN-1 were compared to evaluate the quality of the definitions of time-to-event endpoint of the study and impact of the definitions on the results. After pooled studies, weighted linear regression was used to estimate trial-level association (surrogacy R^2) between treatment effects on the potential surrogate and OS. **RESULTS:** Among the 9 databases analyzed, two were excluded for unpublished results and due to availability of updated database only, respectively. Despite the absence of several events necessary to the definition of time to event endpoint from the DATECAN-1 recommendations, some endpoints were partially reconstructed to assess their capacity to substitute of OS in metastatic setting. PFS and time to performance status deterioration (TPSD) was most strongly correlated with OS (r_{pfs}=r_{tpsd}=0.7858), flowed by time to

progression (TTP) (r_{ttp}=0.7852). TPSD, PFS and TTP were validated as surrogate for OS with R^2_{tpsd}=0.98, R^2_{pfs}=0.96 and R^2_{ttp}=0.92.

CONCLUSION AND RELEVANCE: Our results show that RCT failed to collect precisely each type of event that contribute in the definition of composite time-to-event analysis. Thus we could not apply all DATECAN-1 definitions. Despite the absence of several events necessary to define the endpoints recommended by DATECAN-1, TTP, PFS and TTP are potentially surrogate for OS in metastatic pancreatic cancer. To be more rigorous and to allow standardization, we suggest reporting of the precise localization of progression and/or recurrence in future RCTs. Moreover, all surrogate endpoints could be classify based on their substitution capabilities to OS in the future.

1. INTRODUCTION

In oncology clinical trial, overall survival (OS) benefit is the gold standard for the approval of new anticancer by the regulatory agencies as the FDA(1,2). The need to reduce long-term follow-up, sample size and cost of clinical trials led to use some intermediate endpoints for OS to assess earlier efficacy or futility of treatments. Intermediate endpoints such as progression-free survival (PFS), disease-free survival (DFS) are often used as primary endpoints because they can be assessed earlier and most of the time, these endpoints are composite because they combine different events such as local and distant progression, local and distant recurrence, development of metachronous cancer, death or severe toxicity(3,4). Nevertheless, composite endpoints suffer from important limitations specially the variability of their definitions which is recognized as a major methodological problem(5–9). In this context, the Definition for the Assessment of Time-to-event Endpoints in CANcer trials (DATECAN-1) project has been developed to provide recommendations and to standardize definitions of time-to-event composite endpoints for each specific diseases and at each

specific stage by use of a formal consensus methodology. It has already led to guidelines for time-to-event endpoints definitions in sarcomas and gastro-intestinal stromal tumors, pancreatic cancer, renal cancer and breast cancer trials(10–14). Each endpoint has been defined by the date of origin, the events of interest and the rules of censorship. Nevertheless, these intermediate endpoints have not been systematically validated as surrogate endpoints for OS(15). The DATECAN-2 project has been developed to assess the surrogacy on OS of the time-to-event endpoints as defined in the DATECAN-1 project. As a part of the DATECAN-2 project, our study assessed the surrogacy on OS of time-to-event endpoints as defined in the DATECAN-1 project in pancreatic cancer trials. To validate surrogate endpoints, Buyse and colleagues developed a method based on individual-data meta-analysis, which assesses "individual-level" surrogacy and "trial-level" surrogacy, which is considered as the gold standard. Based on the results of previously completed trials and using individual data, this approach jointly estimates: the correlation between the candidate surrogate and the final endpoints and the correlation between the treatment effect on the candidate surrogate and its effect on OS.

2. METHOD

Search Strategy and Study Selection

We included 6 RCT. In all trials, we compared the definitions of the endpoints in the trials with the DATECAN-1 guideline. Then, we assessed the surrogacy of the different endpoints defined with the DATCAN-1 guideline on OS.

Statistical Analysis

Time to event curves were estimated according to the Kaplan-Meier method and compared by treatment arm using log-rank tests, except for non-comparative phase II clinical trials. Univariate Cox regression models were used to estimate the HR with its 95% confidence interval (CI) as indicated in the published paper, except for FFCD/ SFRO 2000-01 for which the 99% CI was reported.

The rank correlation coefficient between distributions of the candidate surrogate endpoints and overall survival at the individual level was assessed with a bivariate survival model that takes censoring into account (15–17). The trial-level correlations between treatment effects (log hazard ratios) on the candidate surrogate endpoints and overall survival were quantified through a linear regression model, weighted by trial size. The squared correlation coefficients or coefficients of determination "i.e, ρ^2 at the individual level and R^2 at the trial level" were calculated to investigate the amount of variation explained by the surrogate. The candidate surrogate endpoints were deemed acceptable only if both correlation coefficients were close to 1. We classified squared correlation values higher than 0.9 as excellent, higher than 0.75 as very good, higher than 0.5 as good, higher than 0.25 as moderate, and equal to or lower than 0.25 as poor. Statistical analyses were done with SAS (version 9.4).

3. RESULTS

Nine RCTs published between 1999 and 2014 for pancreatic cancer were selected, including one adjuvant Phase II trial (EORTC 40013), four metastatic phase II trial (FIRGEM, EORTC 40924, EORTC 40984 and EORTC 16994P)), one adjuvant phase III trial (EORTC 40891), three metastatic phase III trial (FFCD 0301, FFCD/ SFRO 2000-01 and EORTC 05962) (Table 1). Among them, seven RCTs were finally retained for analysis since one was not published and the other due to availability of updated database only. Among the seven RCTs analyzed, OS was considered as the primary endpoint for two phase III trial. Regarding the five phase II clinical trials, the survival endpoint (OS, PFS, TTP, and DFS) was a secondary objective. In all tables, the empty boxes for the HR and p-value results in the article concern non-comparative Phase II clinical trials.

Comparison between the definition from the Article and DATECAN-1 (Table 2)

Regarding FIRGEM phase II trial, PFS was defined as the time interval between the date of randomization and the date of first local progression or regional progression or metastatic progression or the date of death from any cause; tumor response was evaluated according to RECIST criteria. In FFCD 0301 and FFCD/ SFRO 2000-01 and EORTC 40984 trials, PFS was defined as the time interval between the date of randomization and the first disease progression or death (all causes). Patients alive without progression were censored. In the database, only the presence of a progression was declared without specifying the type of progression (local progression, regional progression, distant progression, second cancer, etc.). Tumor response was evaluated according to RECIST criteria in FFCD 0301 and FFCD / SFRO 2000-01 and was evaluated and graded using WHO criteria (WHO Handbook for Reporting Results of Cancer Treatment, 1979) in EORTC 40984. In DATECAN-1, PFS was defined as the time interval between the day of reference in the study (date of randomization,

date of diagnosis, etc…) and the date of local or regional progression or metastases progression or occurrence of distant metastases (including liver or non-liver metastases) or occurrence of a second pancreatic cancer or death (all causes), whichever occurs first.

In EORTC 40013, DFS was defined as the time interval from random assignment to disease recurrence or death (all causes), whichever came first. Recurrence was not specifying by local recurrence or regional recurrence or second cancer. In DATECAN-1, DFS was defined as time interval between the day of reference in the study (date of randomization, date of diagnosis, etc.) and the date of local relapse/recurrence or regional relapse/recurrence or occurrence of distant metastases (liver or non-liver) or appearance of second pancreatic cancer or death (all causes), whichever occurs first.

In EORTC 40924, the duration of response (confused with time to progression) and time to treatment failure were calculated from the randomization date to disease progression. In EORTC 16994p, the same endpoint was measured from the date a first objective response and was documented until the first sign of progression and was assessed by an imaging/ radiological method. The type of progression/ relapse was not specifying, only the progression/ relapse (yes or no) was documented. In DATECAN-1, TTP was defined as the time interval between the day of reference in the study and the date of local or regional progression or metastases progression or occurrence of distant metastases (including liver or non-liver metastases), whichever occurs first.

In all study trials, OS was defined as the interval between the date of randomization and the date of death (all causes).

Comparison between the results from the Article and Database received (Table 3)

Regarding OS results from the Article and Database received, the numbers of events were identical, except for EORTC 40984 in arm B with 42 events reported in the article against 43 for the database received. The median times were also identical, except for the EORTC

16994p with a median time for all patients of 5.3 [3.9-7.1] months in the article against 5.5 [4–7.1] months the database received. All HR were the same between published articles and corresponding databases. The p-value results between the articles and the databases are similar for both phase III trial FFCD 0301 and FFCD/ SFRO 2000-01.

For the comparison between PFS results from the Article and Database received, the numbers of events were the same between the article and database, except for EORTC 40984 for which the results was not presented in the paper. The median times and HR were identical between the article and the database. The p-value result between article and database was identical for FFCD/ SFRO 2000-01 and different for FFCD 0301 with 0.67 for the article against 0.58 for database.

For the comparison between TTP results from the Article and Database, the number of event is identical between the article and database for EORTC 40924 and the result was not presented for EORTC 16994p. The median time for EORTC 16994p was 1.4 months for the entire group in article against 1.8 [1.4–2.7] months for the database.

For the comparison between DFS results from the Article and Database, the number of event and the median time are identical between the article and database in EORTC 40013.

Potential surrogate for OS (figure 1)

In total, 582 patients from six clinical trials with metastatic setting (202 from FFCD 0301, 119 from FFCD/ SFRO 2000-01, 96 from EORTC 40984, 58 from Firgem, 34 from EORTC 16994, 33 from 40924) were grouped for the surrogacy analyses. The correlation coefficient between OS and surrogate endpoint was statistically significant. We found the strongest correlation between OS and the three endpoints TPSD, TTP, PFS with r_{tpsd}=0.7858, r_{pfs}=0.7858 and r_{ttp}=0.7852. The association at the trial level between log HR_OS and log HR_TPSD, log HR_PFS and log HR_TTP was strong, with a coefficient of determination,

R^2, adjusted , The results of weighted linear regression models of HR_{OS}, HR_{PFS}, HR_{TPSD} and HR_{TTP}, indicating a significant regression between OS and their potential surrogates.

4. Discussion

To our knowledge, this is the first individual patient data meta-analyses of the surrogacy for OS of intermediate endpoints defined according to guidelines in advanced pancreatic cancer. In 2016, Hamada et al.(18) performed an aggregate data meta-analysis of 50 trials in advanced pancreatic cancer and concluded of the validity of the surrogacy of PFS for OS. Nevertheless, Buyse and colleagues have developed a method based on individual data meta-analysis which is considered as the gold standard and currently, there is no comparison of methods based on IPD meta-analysis and on aggregate data meta-analysis.

Our assessment of a large sample of data for patients with metastatic pancreatic cancer provides high-level evidence that PFS, TTP and TPSD are valid surrogate endpoints for OS in studies of metastatic chemotherapy in patients with pancreatic cancer. Those endpoints could, therefore, be considered for use as primary endpoints for chemotherapy trials to avoid the biases on the assessment of OS due to cross-over and subsequent lines of treatment after the first-line.

Nevertheless, regardless of evidence for the validity of intermediate endpoints as a surrogate for survival, the magnitude of benefit is very important for the approval of new drugs. The European Medicines Agency has accepted PFS as primary endpoints in trials where there is a "large effect on progression-free survival, a long expected survival after progression, or a clearly favorable safety profile."

Moreover, a major difficulty for the validation of surrogate endpoints arises from the fact that they must be validated with respect to a specific therapeutic class, for a specific disease, at a

specific stage. The surrogacy of PFS, TTP and TPSD for OS might be different in trials with targeted therapy or immunotherapy.

Therefore, OS remains the gold standard. Furthermore, health-related quality of life is an important endpoint which measures a direct clinical benefit for the patient especially in advanced pancreatic cancer where the treatment is palliative.

5. Conclusion

Our results showed that time-to-event endpoints were defined and evaluated differently between articles published. The definitions of endpoints are confused with other, for example the duration of response and time to progression. Indeed, local progression and distant progression are considered as the same, while there is a significant difference between them.

The use of composite endpoints in clinical trials is problematic. Components are often unreasonably combined, inconstantly defined, and inadequately reported. Unfortunately, few RCTs provide a good rationale for the choice of components to associate.

When the criteria are not defined identically between clinical trials, it is difficult to compare the results between studies. Especially when the criteria is not validated by any international guidelines for use time to event endpoint definition.

The DATECAN-1 project provided guidelines for the definition of time-to-event endpoints in pancreatic cancer. The aim of the DATECAN-2 project is to assess the surrogacy on OS of the endpoints defined in DATECAN-1 project.

In conclusion, using a standardized definition of intermediate endpoints, we demonstrated the surrogacy of PFS, TTP and TPSD in advanced pancreatic cancer. Nevertheless, OS and HRQoL are the most valid endpoints to demonstrate a clinical benefit for the patients.

INDEX

Table 1: Studies selected for analysis

Ref	Clinical trial	Cancer	Phase	Treatment	Stage	N (arm A + arm B)	Endpoints
(19)	FIRGEM	Pancreas	II	Chemotherapy	Metastatic	58 (49 + 49)	OS, PFS
(20)	FFCD 0301	Pancreas	III	Chemotherapy	Metastatic	202 (102 + 100)	OS, PFS
(21)	FFCD/ SFRO	Pancreas	III	Chemotherapy	Metastatic	119 (59 + 60)	OS, PFS
(22)	EORTC 40013	Pancreas	II	Chemotherapy	Adjuvant	90 (45 + 45)	OS, DFS
(23)	EORTC 40924	Pancreas	II	Chemotherapy	Metastatic	33 (18 + 15)	OS, TTP
(24)	EORTC 40984	Pancreas	II	Chemotherapy	Metastatic	96 (49 + 47)	OS, PFS
(25)	EORTC 16994p	Pancreas	II	Chemotherapy	Metastatic	34 (16 + 18)	OS, TTP
	EORTC 40891	Pancreas	III	Chemotherapy	Adjuvant	Unavailable data	
	EORTC 05962	Pancreas	III	Chemotherapy	Metastatic	Not published	

Abbreviations: Ref: references; **OS**: overall survival; **PFS**: Progression-Free survival; **DFS**: disease-free survival; **TTP**: time to progression.

Table 2: Comparison between the definition from the Article and DATECAN reconstituted

	PFS				TTP		DFS
Events	**FIRGEM**	**FFCD 0301**	**FFCD/ SFRO**	**EORTC 40984**	**EORTC 40924**	**EORTC 16994p**	**EORTC 40013**
Definition of Article							
Progression/ relapse	progression locoreginal or metastatic	Progression YES or NO	Progression YES or NO	Progression YES or NO	progression/ relapse	Progression YES or NO	Disease recurrence
Death	Death (all causes)	death (all causes)	death (all causes)	death (all causes)			death (all causes)
DATECAN reconstituted							

Local relapse/recurrence							Y
Local progression	Y	N	N	N	N	N	
Regional relapse/recurrence							missing
Regional progression	Y	N	N	N	N	N	
Appearance/occurrence of distant metastases							Y
Progression of metastases/distant progression	Y	N	N	N	N	N	
Appearance/occurrence of liver metastases	missing	missing	missing	missing	missing	missing	missing
Appearance/occurrence of non-liver metastases	missing	missing	missing	missing	missing	missing	missing
Second pancreatic cancer	missing	missing	missing	missing			missing
Death related to primary cancer	Y	Y	Y	Y	Y	Y	Y
Death related to second cancer	missing	missing	missing	missing			missing
Death related to protocol treatment	Y	Y	Y	Y			Y
Other cause of death	Y	Y	Y	Y			Y
Unknown cause of death	Y	missing	Y	Y			Y

N and **missing** indicates: the event was not clearly specified and defined in the trial analyzed , respectively. **Y**: means the event is clearly defined in the article and usable in DATECAN-1 definition.

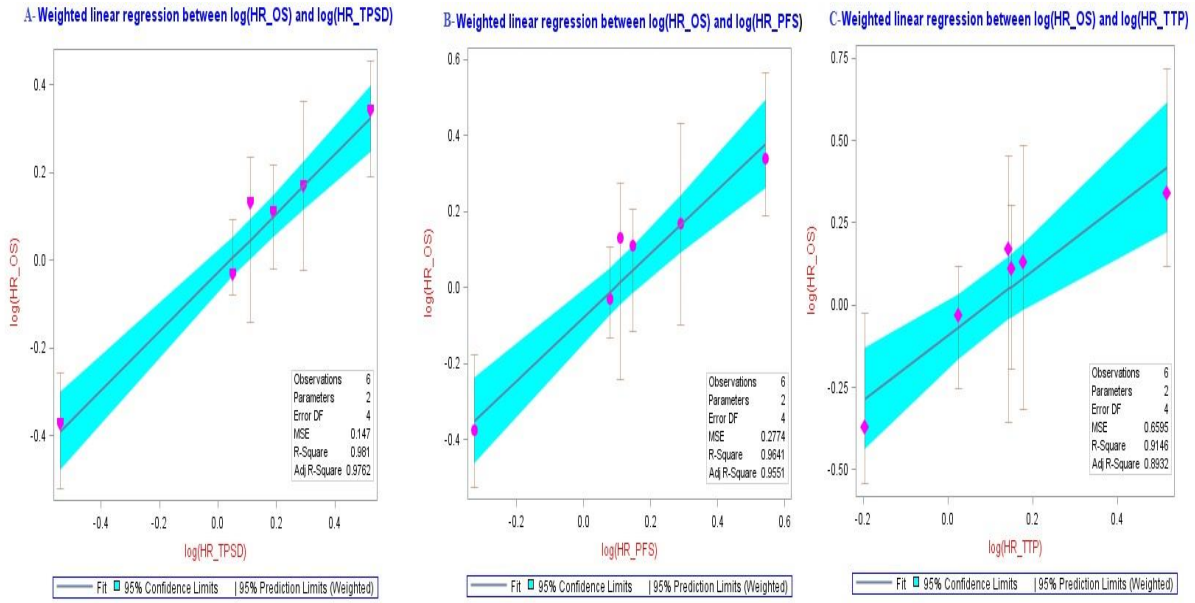
Table 3: comparison between the results from the Article and Database received

			Event		Median (months)		HR (CI)		p_value	
	Group	N	Article	Database	Article	Database	Article	Database	Article	Database
OS										
FFCD 0301	arm A	102	94	94	6.7 [5.4-8.6]	6.7 [5.4-8.6]	ref	ref	0.83	0.83
	arm B	100	98	98	8.03 [5.9-9.8]	8.04 [5.9-9.9]	0.97 [0.73 - 1.29]	0.97 [0.73 - 1.29]		
EORTC 40984	arm A	49	46	46	7.4 [5.6-11.0]	7.4 [5.5-10.2]]				
	arm B	47	42	43	7.1 [4.8-8.7]	7.1 [4.6-8.7]				
FIRGEM	arm A	49	37	37	11 [5.8-13.6]	10.7 [5.4-13.6]	0.71 [0.46 - 1.10]	0.71 [0.46 - 1.10]		
	arm B	49	44	44	8.2 [5.3-9.2]	8.2 [5.3-9.2]	ref	ref		
EORTC 40013	arm A	45	25	25	24.3 [20.5-.]	24.3 [18 - .]				
	arm B	45	26	26	24.4 [21.5-.]	24.4 [19.5-.]				
EORTC 40924	arm A	18	15	15	6.5	6.5 [2.3-9.0]				
	arm B	15	14	14	5	5 [1.1-11.1]				
FFCD/ SFRO 2000-01	arm A	59	54	54	8.6 [7.1-11.4]	8.6 [7.1-11.4]	ref	ref	0.057	0.057
	arm B	60	52	52	13 [8.7-18.1]	13 [8.7-18.1]	0.69 [0.41 - 1.14]	0.69 [0.41 - 1.14]		
EORTC 16994p	arm A	16		14	5.3 [3.9-7.1]	5.5 [4.0–7.1]				
	arm B	18		17						

PFS										
FFCD 0301	arm A	102	99	99	3.4 [2.4-4.4]	3.4 [2.4-4.4]	ref	ref	0.67	0.58
	arm B	100	100	100	3.5 [2.4-4.1]	3.4 [2.4-4.1]	1.06 [0.8 - 1.4]	1.08 [0.8 - 1.4]		
EORTC 40984	arm A	49		46	3.9 [3.0-4.7]	4.1 [2.8-5.6]				
	arm B	47		45	2.8 [2.6-4.6]	2.9 [2.6-4.8]				
FIRGEM	arm A	49	42	42	5 [3.8-8.8]	5 [3.8-8.8]	0.59 [0.38 - 0.90]	0.58 [0.38 - 0.89]		
	arm B	49	47	47	3.4 [2.6-5.2]	3.6 [2.6-5.2]	ref	ref		
FFCD/ SFRO 2000-01	arm A	59	57	57	6 [4.0-8.0]	6 [4.0-8.0]	ref	ref	0.088	0.089
	arm B	60	57	57	6.7 [4.5-11.0]	6.7 [4.5-11.0]	0.72 [0.44 - 1.18]	0.72 [0.44 - 1.18]		
TTP										
EORTC 40924	arm A	18	16	16	5	5[1.4 - 6.1]				
	arm B	15	14	14	3	3 [0.9 –3.9]				
EORTC 16994p	arm A	16		16	1.4	1.8 [1.4– 2.7]				
	arm B	18		18						
DFS										
EORTC 40013	arm A	45	34	34	11.8 [10.1-19.3]	11.8 [10-16.8]				
	arm B	45	37	37	10.9 [8.3-16.0]	10.9 [7.9-15.1]				

Empty box: missing in the article; **ref**: selected as the reference class.

Figure 1: Correlation between treatment effects on intermediate endpoint and OS



References

1. Pazdur R. Endpoints for Assessing Drug Activity in Clinical Trials. The Oncologist. 1 avr 2008;13(Supplement 2):19‑21.

2. Fetting J, Anderson P, Ball H, Benear J, Benjamin K, Bennett C, et al. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol. févr 1996;14(2):671‑9.

3. End Points and United States Food and Drug Administration Approval of Oncology Drugs: Journal of Clinical Oncology: Vol 21, No 7 [Internet]. [cité 13 nov 2017]. Disponible sur: http://ascopubs.org/doi/abs/10.1200/JCO.2003.08.072

4. Chibaudel B, Bonnetain F, Shi Q, Buyse M, Tournigand C, Sargent DJ, et al. Alternative End Points to Evaluate a Therapeutic Strategy in Advanced Colorectal Cancer: Evaluation of Progression-Free Survival, Duration of Disease Control, and Time to Failure of Strategy—An Aide et Recherche en Cancérologie Digestive Group Study. J Clin Oncol. 1 nov 2011;29(31):4199‑204.

5. Proposal for Standardized Definitions for Efficacy End Points in Adjuvant Breast Cancer Trials: The STEEP System: Journal of Clinical Oncology: Vol 25, No 15 [Internet]. [cité 13 nov 2017]. Disponible sur: http://ascopubs.org/doi/abs/10.1200/jco.2006.10.3523

6. Punt CJA, Buyse M, Köhne C-H, Hohenberger P, Labianca R, Schmoll HJ, et al. Endpoints in Adjuvant Treatment Trials: A Systematic Review of the Literature in Colon Cancer and Proposed Definitions for Future Trials. JNCI J Natl Cancer Inst. 4 juill 2007;99(13):998‑1003.

7. Cheson BD. The International Harmonization Project for Response Criteria in Lymphoma Clinical Trials. Hematol Oncol Clin North Am. 1 oct 2007;21(5):841‑54.

8. Juweid ME, Wiseman GA, Vose JM, Ritchie JM, Menda Y, Wooldridge JE, et al. Response Assessment of Aggressive Non-Hodgkin's Lymphoma by Integrated International Workshop Criteria and Fluorine-18–Fluorodeoxyglucose Positron Emission Tomography. J Clin Oncol. 20 juill 2005;23(21):4652‑61.

9. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival End Point Reporting in Randomized Cancer Clinical Trials: A Review of Major Journals. J Clin Oncol. 1 août 2008;26(22):3721‑6.

10. Chibaudel B, Bonnetain F, Tournigand C, de Larauze MH, de Gramont A, Laurent-Puig P, et al. STRATEGIC-1: A multiple-lines, randomized, open-label GERCOR phase III study in patients with unresectable wild-type RAS metastatic colorectal cancer. BMC Cancer. 4 juill 2015;15:496.

11. Bonnetain F, Bonsing B, Conroy T, Dousseau A, Glimelius B, Haustermans K, et al. Guidelines for time-to-event end-point definitions in trials for pancreatic cancer.

Results of the DATECAN initiative (Definition for the Assessment of Time-to-event End-points in CANcer trials). Eur J Cancer. 1 nov 2014;50(17):2983‑93.

12. Bellera CA, Penel N, Ouali M, Bonvalot S, Casali PG, Nielsen OS, et al. Guidelines for time-to-event end point definitions in sarcomas and gastrointestinal stromal tumors (GIST) trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials). Ann Oncol. 1 mai 2015;26(5):865‑72.

13. Kramar A, Negrier S, Sylvester R, Joniau S, Mulders P, Powles T, et al. Guidelines for the definition of time-to-event end points in renal cell cancer clinical trials: results of the DATECAN project. Ann Oncol. 1 déc 2015;26(12):2392‑8.

14. Gourgou-Bourgade S, Cameron D, Poortmans P, Asselain B, Azria D, Cardoso F, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials). Ann Oncol. 1 mai 2015;26(5):873‑9.

15. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics. 1 mars 2000;1(1):49‑67.

16. Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, Alonso A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Control Clin Trials. 1 déc 2002;23(6):607‑25.

17. Gilbert PB, Gabriel EE, Huang Y, Chan ISF. Surrogate Endpoint Evaluation: Principal Stratification Criteria and the Prentice Definition. J Causal Inference. 1 sept 2015;3(2):157‑75.

18. Hamada T, Nakai Y, Isayama H, Yasunaga H, Matsui H, Takahara N, et al. Progression-free survival as a surrogate for overall survival in first-line chemotherapy for advanced pancreatic cancer. Eur J Cancer Oxf Engl 1990. sept 2016;65:11‑20.

19. Trouilloud I, Dupont-Gossard A-C, Malka D, Artru P, Gauthier M, Lecomte T, et al. Fixed-dose rate gemcitabine alone or alternating with FOLFIRI.3 (irinotecan, leucovorin and fluorouracil) in the first-line treatment of patients with metastatic pancreatic adenocarcinoma: An AGEO randomised phase II study (FIRGEM). Eur J Cancer. déc 2014;50(18):3116‑24.

20. Dahan L, Bonnetain F, Ychou M, Mitry E, Gasmi M, Raoul J-L, et al. Combination 5-fluorouracil, folinic acid and cisplatin (LV5FU2-CDDP) followed by gemcitabine or the reverse sequence in metastatic pancreatic cancer: final results of a randomised strategic phase III trial (FFCD 0301). Gut. 1 nov 2010;59(11):1527‑34.

21. Chauffert B, Mornex F, Bonnetain F, Rougier P, Mariette C, Bouché O, et al. Phase III trial comparing intensive induction chemoradiotherapy (60 Gy, infusional 5-FU and intermittent cisplatin) followed by maintenance gemcitabine with gemcitabine alone for locally advanced unresectable pancreatic cancer. Definitive results of the 2000–01 FFCD/SFRO study. Ann Oncol. 1 sept 2008;19(9):1592‑9.

22. Laethem J-LV, Hammel P, Mornex F, Azria D, Tienhoven GV, Vergauwe P, et al. Adjuvant Gemcitabine Alone Versus Gemcitabine-Based Chemoradiotherapy After Curative Resection for Pancreatic Cancer: A Randomized EORTC-40013-22012/FFCD-9203/GERCOR Phase II Study. J Clin Oncol. 10 oct 2010;28(29):4450‑6.

23. Wagener DJT, Wils JA, Kok TC, Planting A, Couvreur ML, Baron B. Results of a randomised phase II study of cisplatin plus 5-fluorouracil versus cisplatin plus 5-

fluorouracil with α-interferon in metastatic pancreatic cancer: an EORTC gastrointestinal tract cancer group trial. Eur J Cancer. mars 2002;38(5):648‑53.

24. Lutz MP, Cutsem EV, Wagener T, Laethem J-LV, Vanhoefer U, Wils JA, et al. Docetaxel Plus Gemcitabine or Docetaxel Plus Cisplatin in Advanced Pancreatic Carcinoma: Randomized Phase II Study 40984 of the European Organisation for Research and Treatment of Cancer Gastrointestinal Group. J Clin Oncol. 20 déc 2005;23(36):9250‑6.

25. Briasoulis E, Pavlidis N, Terret C, Bauer J, Fiedler W, Schöffski P, et al. Glufosfamide administered using a 1-hour infusion given as first-line treatment for advanced pancreatic cancer. A phase II trial of the EORTC-new drug development group. Eur J Cancer. nov 2003;39(16):2334‑40.

B. La qualité de vie relative à la santé comme co-critère de jugement principal dans les essais cliniques randomisés en oncologie

Résumé

La SG est considéré comme critère de référence par la FDA et l'ASCO mais avec l'augmentation du nombre de traitements efficaces disponibles pour une majorité de localisations cancéreuses et de situation thérapeutique, il est nécessaire d'inclure un nombre plus important de patients et de les suivre plus longtemps ce qui augmente le coût des essais cliniques. Ainsi, les critères de jugement dit intermédiaires centrés sur la tumeur qui sont évalués plus précocement et considérés comme critères substitutifs de la survie globale sont de plus en plus utilisés. Cependant, la plupart d'entre eux ne sont pas validés comme critères de substitution selon la méthodologie rigoureuse, ce qui ne permet pas de s'assurer du bénéfice clinique pour le patient. Ainsi la QdV constitue un critère de jugement pertinent pour évaluer un bénéfice clinique direct pour le patient. Un design alternatif serait d'associer un critère de jugement intermédiaire composite avec la QdV en tant que co-critères de jugement principaux. Dans cet article nous discutons les problématiques méthodologiques soulevées par un tel design : les règles de décision, les procédures de contrôle de l'erreur de type I ainsi que le calcul du nombre de sujets nécessaires.

EXPERT REVIEWS

Health-related quality-of-life as co-primary endpoint in randomized clinical trials in oncology

Expert Rev. Anticancer Ther. Early online, 1–7 (2015)

Fre!de!ric Fiteni*[1,2], Alhousseiny Pam[1], Ame!lie Anota[1,3,4], Dewi Vernerey[1], Sophie Paget-Bailly[1], Virginie Westeel[5] and Franck Bonnetain[1,3,4,6]

[1]University Hospital of Besanc¸on, Methodology and Quality of Life in Oncology Unit, Besanc¸on, France
[2]Department of Medical Oncology, University Hospital of Besanc¸on, Besanc¸on, France

Overall survival (OS) has been considered as the most relevant primary endpoint but trials using OS often require large numbers of patients and long-term follow-up. Therefore composite endpoints, which are assessed earlier, are frequently used as primary endpoint but suffer from important limitations specially a lack of validation as surrogate of OS. Therefore, Health-related quality of life (HRQoL) could be considered as an outcome to judge efficacy of a treatment. An alternative approach would be to combine HRQoL with composite endpoints as co-primary endpoint to ensure a clinical benefit for patients of a new therapy. The decision rules of such design, the procedure to control the Type I error and the determination of sample size remain questions to debate. Here, we discusses HRQoL as co-primary endpoints in randomized clinical trials in oncology and provide some solutions to promote such design.

KEYWORDS: clinical trial . composite endpoint . co-primary endpoint . endpoint . health-related quality of life . methodology

Expert Review of Anticancer Therapy

[3]EA 3181 University of Franche-Comte!, Besanc¸on, France
[4]The French National Platform Quality of Life and Cancer, Besanc¸on, France
[5]Chest Disease Department, University Hospital of Besanc¸on, Besanc¸on, France [6]EORTC QOL Group, Brussels, Belgium *Author for correspondence: Tel.: + 33 3 81 21 88 97
fredericfiteni@gmail.com

Endpoints refer to clinical and biological measurements that assess the efficacy of therapeutic strategies. As the American Society of Clinical Oncology stated, active treatment in cancer is generally undertaken with the goal of provid-ing improved quantity and/or quality of patient survival [1]. Cancer randomized clinical trials are conducted to obtain clinical evidence on the safety and efficacy of new interventions. The selection of an appropriately valid primary endpoint is an important aspect of clinical trial design to achieve this objective. Endpoints in cancer clinical trials can be classified into two main categories: patient-centered clinical endpoints, including overall survival (OS) and health-related quality of life (HRQoL), and tumor-centered clinical endpoints, such as progression-free survival (PFS) [2,3]. The US FDA considers OS benefit as the foundation for the approval of new anticancer drugs in the USA. Nevertheless, the increasing number of effective salvage treatments available in many types of cancer (i.e. subsequent lines of treatments) has resulted in the need for a larger number of patients to be included and/ or the need of a more prolonged observation period to attain sufficient events that can achieve planned statistical power; this increases the cost of clinical trials and requires a longer duration to obtain results [4,5]. Consequently, tumour-centered clinical endpoints such as PFS are often used as primary endpoints because they are assessed earlier (i.e., interme-diate endpoints). These endpoints are fre-quently composite endpoints. However, there is a lack of consistency in their definitions, and they are not systematically validated as surrogate endpoints for OS [4]. In this context, HRQoL could constitute an alternative end-point, which ensures earlier assessment of direct clinical benefit for the patient [5] and is recognized as a component endpoint for can-cer therapy approvals by the American Society of Clinical Oncology and the FDA. An extract from the Lancet Handbook of essential con-cepts in clinical research says that "Researchers should restrict the number of primary end-points tested. They should specify a priori the primary endpoint or endpoints in their proto-col" but a new methodological approach would be to combined HRQoL with other endpoints, for example PFS as co-primary

endpoints, especially when OS is difficult to assess, to ensure the analysis of the clinical benefit of a new therapeutic. This paper discusses HRQoL as co-primary endpoints in random-ized clinical trials in oncology.

Composite endpoints

'Tumor-centered endpoints' are, most of the time, composite endpoints based on tumor assessment and combine different events such as local and distant progression, local and distant recurrence, development of metachronous cancer, death or severe toxicity. Composite endpoints combine multiple events (called components) into a single endpoint. An event is said to occur if any one of the prospectively defined components of the composite occurs. In the absence of at least one component considered in the composite endpoint, the patient is censored at the time of the last follow-up. For example, in the metastatic setting, PFS combined two components: tumor progression and death for any causes.

These types of endpoints are frequently used as primary outcome in oncologic clinical trials for several reasons. First, they can increase statistical power. A higher event rate than with OS is observed, so clinical trials need fewer numbers of patients to achieve required power [7]. Then they are assessed earlier than OS. The fewer number of subjects and the smaller duration of trials contribute to an economic benefit of less costly trials [8,9]. Then the use of a composite endpoint could lead to information preservation and thus reduction of bias due to information censoring [7]. For instance, PFS include death with a nonfatal event (Definition for the Assessment of Time-to-event Endpoints in CANcer trials) [15] was developed to obtain standardized consensus definitions for multiple cancer sites based on a formal consensus process and has already led to guidelines for time-to-event endpoint definitions in sarcomas and gastrointestinal stromal tumors and pancreatic cancers trials [16,17].

- Most of composite endpoints have not been validated as a surrogate endpoint for OS. Nevertheless, PFS is frequently used as the primary outcome in other types of cancer, and many trials have demonstrated improved PFS with no improvement of OS. Currently, in cancer clinical trials, PFS was validated as a surrogate for OS only in 5-fluorouracil-based chemotherapy in advanced colorectal cancer [18], in T3/T4 rectal cancer [19].
- If an important component is not substantially modified by the effects of treatment, then less rather than more statistical power to detect effects may be the result [7].
- Although the composite endpoint as a whole may appear to be affected by the treatment, the benefit may be different between the components. This is the case in rectal cancer, and neoadjuvant chemo-radiotherapy has a higher effect on local recurrence than in metastatic recurrence [19].
- Competing risks interfere with the ability to observe events of primary interest in relation to the therapy being tested. Examples of competing risks can be second malignancies, treatment-related adverse effects. Such events interfere with analysis of effects on disease-specific events, which are the events directly affected by cancer therapies.

(tumor progression). In the case of a cancer with a poor prognosis, drop-out of the patient before the planned end of the study could correspond to patient's health deterioration. The censoring in this context seems to be informative; therefore, the inclusion of a nonfatal event reduces this bias. Moreover, the use of PFS may be appropriate in particular disease settings and has been accepted by regulatory authorities on many occasions [10,11].

Nevertheless, composite endpoints suffer from important limitations:

- There is heterogeneity of the definitions of composite endpoints like PFS, disease-free survival (DFS), time-to-treatment failure and so on. Consequently the definition of the same 'endpoint' is variable in different studies of the same disease [12]. Birgisson et al. [13] demonstrated that the inclusion of a second primary cancer other than the incident colo-rectal cancer as an event in the definition of DFS significantly impacted the results. The estimated DFS rate for patients with stage I–III disease was 62% after 5 years if this event was not counted as an event, compared with 58% if it was. The difference was larger for stage II (68 versus 60%) than for stage III (49 vs 47%). Another example is the PETACC 03 randomized study [14] where results were either significant or nonsignificant depending on whether or not second primary tumors were accounted for in the DFS definition. In the context of heterogeneity of definitions of survival endpoints, the international DATECAN initiative

HRQoL as co-primary endpoint

HRQoL constitutes an alternative endpoint that allows earlier assessment of direct clinical benefit for the patient. Over the last decade, a tremendous growth in the incorporation of HRQoL in cancer clinical trials has been observed. The FDA considers HRQoL as an endpoint that assesses direct clinical benefit for the patient [20,21]. In assessing new therapies and approaches, recent authors have first published the results of efficacy and elsewhere the analysis of HRQoL; for example, in breast cancer: BOLERO-2 [22,23], EMILIA [24,25], CLEOPA-TRA [26,27] trials, in lung cancer: LUX-lung 3 [28,29], OPTI-MAL [30,31], AVAPERL [32,33] trials, in ovarian cancer: AURELIA [34,35], GOG 0218 [36,37] trials, in pancreatic cancer: PRODIGE 4/ACCORD 11 [38,39] trial. This indicates that HRQoL becomes a major consideration in treatment choice and influences adoption of new treatments. HRQoL could be implemented as composite primary endpoint (in combination with efficacy endpoints). Nevertheless, this approach might lead to problems of composite endpoints. This could develop new endpoints with heterogeneity between trials. The benefit may be different between the components; therefore the clinical meaningful of the endpoint would be difficult to interpret. Competing risks could interfere with the ability to observe events of primary interest in relation to the therapy being tested. It would be difficult for clinicians to interpret the

doi: 10.1586/14737140.2015.1047768

treatment effect. For example, in the AVAglio trial [42] that investigated the impact of adding bevacizumab to the standard therapy of newly diagnosed glioblastomas, deterioration-free survival was a composite endpoints defined as the time to: a ‡10-point deterioration (regarded as clinically meaningful) from baseline without a subsequent ‡10-point improvement from baseline; or progression or death. An exploratory sensitiv-ity analysis was conducted without progression as an event. In the prespecified primary analysis, deterioration-free survival was significantly longer among patients in the bevacizumab group than among those in the placebo group for all five prespecified dimensions and all 21 nonprespecified dimensions, while in the sensitivity analysis, deterioration-free survival was significantly longer for only three of the five prespecified dimensions and 9 of 21 nonprespecified dimensions.

Another approach could be to combine intermediate end-point, with HRQoL as co-primary endpoints, especially when treatments are associated with significant toxicity and surrogacy of other traditional efficacy clinical endpoints has not been demonstrated. A trial with co-primary endpoints is a trial with two or more primary efficacy endpoints. These endpoints can be composite endpoints. For example, Hussain et al. [41] com-pared intermittent and continuous androgen deprivation in metastatic prostate cancer, the co-primary objectives were to assess whether intermittent therapy was noninferior to continu-ous therapy with respect to survival, and whether HRQoL dif-fered between the groups. C Von Plessen et al. compared a policy of three versus six courses of platinum-based combina-tion chemotherapy with regard to Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire QLQ-C30 and the Functional Assessment of Cancer Therapy-General (FACT-G) are the most widely used cancer-specific instrument in Phase III clini-cal trials [44,45]. A challenge would be to promote, through can-cer site and treatment modalities, guidelines for selecting the best questionnaires allowing for direct comparison of results across trials. As an example, two recent Phase III clinical trials, OPTIMAL and LUX-Lung 3, focused on patients with advanced non-small cell lung cancer with EGFR mutations and investigated an inhibitor of EGFR (erlotinib for OPTIMAL trial and afatinib for LUX-Lung 3 trial respectively) versus che-motherapy [28,30]. HRQoL was assessed by the FACT-L ques-tionnaire in OPTIMAL trial while EORTC QLQ-C30 and LC13 module were used in LUX-Lung 3 trial. These two clini-cal trials could hardly be compared since EORTC and FACT questionnaires for lung cancer do not contain the same dimen-sions in regard to impact of lung cancer on HRQoL.

The decision rules of trials with HRQoL as co-primary end-point remains a question to debate. If an intermediate endpoint and HRQoL are co-primary endpoints, we can presume that the investigator might want to give equal importance to the endpoints. Therefore, a decision rule could be that the interme-diate composite endpoint must be positive (e.g., PFS) and at least one dimension of HRQoL must be positive without dete-rioration of the other. Another approach of decision rules could be that statistical significance is needed for the intermediate composite endpoint and all HRQoL dimensions. This proce-dure is called intersection-union test. The global null hypothe-sis can be expressed as the

effects on HRQoL and sur-vival in patients with stage IIIB or IV non-small cell lung cancers [42]. Nevertheless, this approach is recent and methodo-logical research must be pursued to clearly define decision-making rules to promote such design.

First, two main statistical methods have been developed to longitudinally analyze HRQoL: the linear mixed model for repeated measure (LMM) and the time to HRQoL deteriora-tion (TTD), and at this time, no guidelines for the longitudi-nal analysis have been proposed, which compromise comparison between trials [6]. For example, two recent Phase III trials investigating the impact of adding bevacizumab to the standard therapy of newly diagnosed glioblastomas have applied two different approaches (LMM and TTD) to analyze longitudinal HRQoL data [40,43]. The results are divergent and compromise the conclusion about clinical value of adding beva-cizumab since OS was not improved. Therefore, the analysis of HRQoL requires improved standardization in the future. Hence, LMM and TTD could systematically be applied or data could be available for researches, in order to make alterna-tive analyses allowing the comparisons between trials.

Then, the treatment effect on HRQoL might be confounded by subsequent treatment. Nevertheless, most of the time, HRQoL is measured until progression; therefore, in this case, HRQoL is not confounded by subsequent treatment.

Then, in oncology, many self-completion HRQoL question-naires have been developed and validated, and The European union of all single hypotheses regarding the individual endpoints; this union is tested versus the global alternative that can be expressed as the intersection of all single alternative hypotheses. Moreover, we know that there is general agreement concerning the multidimensional concept of HRQoL taking into account levels of physical, men-tal, social and patient satisfaction with treatment. Therefore, the improvement of only one dimension of HRQoL reaches the problem of the holistic sense of HRQoL. The inclusion of an overall HRQoL dimension that captures everything and the inclusion of more than one dimension of HRQoL might be systematically discussed. The choice of the co-primary end-points and dimensions of HRQoL must be discussed between clinicians and methodologists, and the decision rules must be clearly described in the protocol.

In confirmatory clinical trials with multiple endpoints, the use of multiple test procedures is mandatory and CONSORT Statement [46,47] recommends a multiplicity adjustment in case of multiple testing. However, no procedure to control the Type I error was implemented in the trials that analyzed HRQoL as secondary endpoint. In biomedical clinical research, it is well known that testing multiple hypotheses without any adjustment may increase the probability of erroneously rejecting at least one true null hypothesis. In an intersection-union test procedure, as all individual null hypotheses have to be rejected to claim significance, no multiplicity adjustment, that is, no

doi: 10.1586/14737140.2015.1047768

special method to consider the multiplicity of tests, is needed. Nevertheless, the Type II error rate inflates, that is, there is an increased probability of not rejecting a false null hypothesis [48]. This inflation must be taken into account for sample size determination. If the decision rules are not an intersection-union test procedure, the Type I error rate increases substantially with the total number of outcomes analyzed. Many procedures have been developed to control the Type I error at 0.05. Two main approaches can be distinguished: single-step procedures and stepwise procedures [48]. In a single-step procedure, such each single test is tested without reference to any other. The simplest approach to controlling such risk is the Bonferroni correction. In this procedure, the significance level a is split equally among the k hypotheses, and each is tested at level a/k. The individual null hypothesis H(i) is rejected if p-value < a/k, and the global null hypothesis H is rejected if min p-value < a/k. The Bonferroni procedure is conservative and consequently leads to a diminution of power if the number of hypotheses k is large. As an illustration of this approach, Hussain et al. [41] used such design, the co-primary objectives were to assess whether intermittent therapy was noninferior to continuous therapy with respect to survival, and whether HRQoL differed between the groups. The five targeted HRQoL dimensions (impotence or erectile dysfunction, libido, vitality, mental health and physical functioning) were clearly specified in the protocol and the p-value of 0.01 was chosen for HRQoL to control the overall Type I error rate at 0.05. In stepwise proce-dures, such as testing in a hierarchical order, the decision on already tested hypotheses influences whether subsequent co-primary endpoints, we can presume that the investigator might want to give equal importance to these two endpoints. The weighted procedure might be rather discussed for the dif-ferent dimensions of HRQoL related to their clinical impor-tance. Moreover, these procedures lead to a reduction of power especially when the number of tests increases. Therefore, the greater the number of co-primary endpoints, the more stringent the alpha level becomes for each endpoint after adjustment and the sample size required increases. The determination of sample size, taking into account the adjustment of the type I error, is an urgent statistical challenge. One approach could be to repeat the sample size estimates for each endpoint and then select the largest number as the sample size required to answer all the questions of interest.

Expert commentary

The variety of composite time-to-event tumor-centered endpoints and the variability of their definitions are recognized as a major methodological problem and efforts must be pursued to improve their reliability and reproducibility. Moreover, most of composite endpoints have not been validated as a surrogate endpoint for OS and when evaluating a new class of therapy, trials must be performed to validate a surrogate endpoint; therefore the clinical benefit of these trials using such endpoints remains uncertain.

Consequently, HRQoL constitutes an alternative endpoint that allows earlier assessment of direct clinical benefit for the patient. A new approach is to combine intermediate endpoint,

tests are conducted. The procedure is carried out sequentially where H(i) is tested at level a as long as all previous hypotheses H(1),. . .H(i–1) are rejected; testing stops with the first non-significant result. Holm developed a sequential procedure that improved upon the Bonferroni method. The univariate p-values are classified in increasing order such that P(1) £ P(2) £ . . . £ P(k) with corresponding hypotheses H(1), H(2),. . .H(k). Then, the smallest p-value, P(1), is compared against the most conservative a-level, a/k. If P(1) £ a/k holds, reject H(1) and con-tinue stepwise, with each successive test conducted at progressively higher significance levels, a/(k-1), a/(k-2),. . .,a/l. Otherwise, fail to reject H(1), and stop the procedure. The acceptance of H(i) implies acceptance of H(j), for all j > i. It is possible to assign different weights to the endpoints according to their importance within the Holm procedure. This method provides a substantial increase in power than the Bonferroni method, but a major problem of the sequential procedure is that once a hypothesis is not rejected, no further testing is per-mitted. The Bonferroni technique is the most classical but is conservative. This conservatism and the accompanied lack of power when the endpoints are correlated is the major drawback of the Bonferroni procedure. Holm's weighted sequential Bon-ferroni procedure gives the option to weight the individual hypotheses and proportionally allocate the p-level to account for the relative 'importance' of the individual endpoint [48]. Nevertheless, if an intermediate endpoint and HRQoL are

with HRQoL as co-primary endpoints, especially when treatments are associated with significant toxicity, and when surrogacy of other traditional efficacy clinical endpoints has not been demonstrated. Implementation of HRQoL as co-primary can provide a useful measure of the impact of intervention in a way that it is relevant to clinicians and patients especially when OS is difficult to measure.

Five-year view

In oncologic clinical trials, the Definition for the Assessment of Time-to-event Endpoints in CANcer trials (DATECAN) project has been developed to provide recommendations and to stan-dardize definitions of time-to-event composite endpoints for each specific diseases and at each specific stage by use of a formal consensus methodology [15]. This is the preliminary step before assessing their surrogate capabilities (DATECAN 2 project).

HRQoL as co-primary endpoint is a recent approach, and methodological research must be pursued to promote such design. First, the measurement, analysis and reporting of HRQoL requires improved standardization in the future. Then, software packages which calculate the sample size in Phase III cancer clinical trials with HRQOL as co-primary endpoints should be developed where the investigators could specify the treatment effect for each co-primary and the procedure to adjust the type I error to obtain the sample size. Then, algo-rithms to clearly define decision-making rules of trials with HRQoL as co-primary endpoint must be developed.

doi: 10.1586/14737140.2015.1047768

Financial & competing interests disclosure
The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Key issues

. Composite endpoints combine multiple events (called components) into a single endpoint.

. There is heterogeneity of the definitions of composite endpoints like PFS.

. Most of composite endpoints have not been validated as a surrogate endpoint for OS. Nevertheless, PFS is frequently used as the primary outcome in other types of cancer and many trials have demonstrated improved PFS with no improvement of OS.

. A trial with co-primary endpoints is a trial with two or more primary efficacy endpoints.

. Over the last decade, a tremendous growth in the incorporation of HRQoL in cancer clinical trials has been observed. The US FDA considers HRQoL as an endpoint that assesses direct clinical benefit for the patient.

. Implementation of HRQoL as co-primary can provide a useful measure of the impact of intervention in a way that it is relevant to clinicians and patients.

. The decision rules of trials with HRQoL as co-primary endpoint remains a question to debate. If an intermediate endpoint and HRQoL are co-primary endpoints, we can presume that the investigator might want to give equal importance to the endpoints. Therefore, a decision rule could be that the intermediate composite endpoint must be positive (e.g., PFS) and at least one dimension of HRQoL must be positive without deterioration of the other. The decision rules must be clearly described in the protocol.

. The procedure to control the type I error must be clearly reported and the determination of sample size, taking into account the adjustment of the type I error, is mandatory.

References

for quality of life to achieve standardization? Qual Life Res 2015;24(1):5-18

trials: a review of major journals. J Clin Oncol 2008;26(22):3721-6

· of interest

·· of considerable interest

1. Peppercorn JM, Smith TJ, Helft PR, et al. American society of clinical oncology statement: toward individualized care for patients with advanced cancer. J Clin Oncol 2011;29(6):755-60
2. Booth CM, Ohorodnyk P, Eisenhauer EA. Call for clarity in the reporting of benefit associated with anticancer therapies. J Clin Oncol 2009;27(33):e213-14
3. Ohorodnyk P, Eisenhauer EA, Booth CM. Clinical benefit in oncology trials: is this a patient-centred or tumour-centred end-point? Eur J Cancer 2009;45(13):2249-52

.. Description of how clinical benefit has been used in oncology trials.

4. Fiteni F, Westeel V, Pivot X, et al. Endpoints in cancer clinical trials. J Visc Surg 2014;151(1):17-22
5. Saad ED, Buyse M. Overall survival: patient outcome, therapeutic objective, clinical trial end point, or public health measure? J Clin Oncol 2012;30(15):1750-4
6. Anota A, Hamidou Z, Paget-Bailly S, et al. Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST
7. Sankoh AJ, Li H, D'Agostino RB. Use of composite endpoints in clinical trials. Stat Med 2014;33(27):4709-14
8. Freemantle N, Calvert M. Composite and surrogate outcomes in randomised controlled trials. BMJ 2007;334(7597): 756-7
9. Freemantle N, Calvert M, Wood J, et al. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 2003;289(19):2554-9
10. Blumenthal GM, Scher NS, Cortazar P, et al. First FDA approval of dual anti-HER2 regimen: pertuzumab in combination with trastuzumab and docetaxel for HER2-positive metastatic breast cancer. Clin Cancer Res 2013;19(18): 4911-16
11. Thornton K, Kim G, Maher VE, et al. Vandetanib for the treatment of symptomatic or progressive medullary thyroid cancer in patients with unresectable locally advanced or metastatic disease: U.S. Food and Drug Administration drug approval summary. Clin Cancer Res 2012; 18(14):3722-30
12. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, et al. Survival end point reporting in randomized cancer clinical

.. Evaluation of the reporting of survival end points in cancer randomized clinical trials. A majority of articles failed to provide a complete reporting of survival end points, thus adding another source of uncontrolled variability.

13. Birgisson H, Wallin U, Holmberg L, et al. Survival endpoints in colorectal cancer and the effect of second primary other cancer on disease free survival. BMC Cancer 2011;11:438
14. Van Cutsem E, Labianca R, Bodoky G, et al. Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. J Clin Oncol 2009;27(19): 3117-25
15. Bellera CA, Pulido M, Gourgou S, et al. Protocol of the Definition for the Assessment of Time-to-event Endpoints in CANcer trials (DATECAN) project: formal consensus method for the development of guidelines for standardised time-to-event endpoints' definitions in cancer clinical trials. Eur J Cancer 2013;49(4):769-81
16. Bellera CA, Penel N, Ouali M, et al. Guidelines for time-to-event endpoint definitions in sarcomas and gastro-intestinal stromal tumors (GIST) trials. Results of the

doi: 10.1586/14737140.2015.1047768

DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials). Ann Oncol 2015;26(5): 865-72

17. Bonnetain F, Bonsing B, Conroy T, et al. Guidelines for time-to-event end-point definitions in trials for pancreatic cancer. Results of the DATECAN initiative (Definition for the Assessment of Time-to-event End-points in CANcer trials). Eur J Cancer 2014;50(17):2983-93
18. Buyse M, Burzykowski T, Carroll K, et al. Progression free survival is a surrogate for survival in advanced colorectal cancer. J Clin Oncol 2007;25(33):5218-24
19. Bonnetain F, Bosset JF, Gerard JP, et al. What is the clinical benefit of preoperative chemoradiotherapy with 5FU/leucovorin for T3-4 rectal cancer in a pooled analysis of EORTC 22921 and FFCD 9203 trials: surrogacy in question? Eur J Cancer 2012; 48(12):1781-90
20. Beitz J, Gnecco C, Justice R. Quality-of-life end points in cancer clinical trials: the U.S. Food and Drug Administration perspective. J Natl Cancer 1996(20):7-9

The Oncologic Drugs Advisory Committee (US FDA) has recommended that beneficial effects on quality of life and/or survival be the basis for approval of new

randomized phase 3 study of trastuzumab emtansine (T-DM1) versus capecitabine and lapatinib in human epidermal growth factor receptor 2-positive locally advanced or metastatic breast cancer. Cancer 2014; 120(5):642-51
26. Baselga J, Cortes J, Kim SB, et al. Pertuzumab plus trastuzumab plus docetaxel for metastatic breast cancer. N Engl J Med 2012;366(2):109-19
27. Cortes J, Baselga J, Im YH, et al. Health-related quality-of-life assessment in CLEOPATRA, a phase III study combining pertuzumab with trastuzumab and docetaxel in metastatic breast cancer. Ann Oncol 2013;24(10):2630-5
28. Wu YL, Zhou C, Hu CP, et al. Afatinib versus cisplatin plus gemcitabine for first-line treatment of Asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-Lung 6): an open-label, randomised phase 3 trial. Lancet Oncol 2014;15(2):213-22
29. Yang JC, Hirsh V, Schuler M, et al. Symptom control and quality of life in LUX-Lung 3: a phase III study of afatinib or cisplatin/pemetrexed in patients with advanced lung adenocarcinoma with EGFR mutations. J Clin Oncol 2013;31(27): 3342-50

therapy in AVAPERL (MO22089). J Thorac Oncol 2013;8(11):1409-16
34. Pujade-Lauraine E, Hilpert F, Weber B, et al. Bevacizumab combined with chemotherapy for platinum-resistant recurrent ovarian cancer: The AURELIA open-label randomized phase III trial. J Clin Oncol 2014;32(13):1302-8
35. Stockler MR, Hilpert F, Friedlander M, et al. Patient-reported outcome results from the open-label phase III AURELIA trial evaluating bevacizumab-containing therapy for platinum resistant ovarian cancer. J Clin Oncol 2014;32(13):1309-16
36. Burger RA, Brady MF, Bookman MA, et al. Incorporation of bevacizumab in the primary treatment of ovarian cancer. N Engl J Med 2011;365(26):2473-83
37. Djaïrv T, Metcalfe C, Avery KN, et al. Prognostic value of changes in health-related quality of life scores during curative treatment for esophagogastric cancer. J Clin Oncol 2010;28(10):1666-70
38. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl J Med 2011;364(19):1817-25
39. Gourgou-Bourgade S, Bascoul-Mollevi C, Desseigne F, et al. Impact of FOLFIRINOX compared with gemcitabine on quality of life

anticancer drugs.

The FDA welcomes the opportunity to explore with investigators the use of quality of life instruments in the design of cancer clinical trials.

21. Johnson JR, Temple R. Food and drug administration requirements for approval of new anticancer drugs. Cancer Treat Rep 1985;69(10):1155-9
22. Piccart M, Hortobagyi GN, Campone M, et al. Everolimus plus exemestane for hormone-receptor-positive, human epidermal growth factor receptor-2-negative advanced breast cancer: overall survival results from BOLERO-2†. Ann Oncol 2014;25(12):2357-62
23. Burris HA 3rd, Lebrun F, Rugo HS, et al. Health-related quality of life of patients with advanced breast cancer treated with everolimus plus exemestane versus placebo plus exemestane in the phase 3, randomized, controlled, BOLERO-2 trial. Cancer 2013; 119(10):1908-15
24. Verma S, Miles D, Gianni L, et al. Trastuzumab emtansine for HER2-positive advanced breast cancer. N Engl J Med 2012;367(19):1783-91
25. Welslau M, Dielras V, Sohn JH, et al. Patient-reported outcomes from EMILIA, a
30. Zhou C, Wu YL, Chen G, et al. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. Lancet Oncol 2011;12(8):735-42
31. Chen G, Feng J, Zhou C, et al. Quality of life (QoL) analyses from OPTIMAL (CTONG-0802), a phase III, randomised, open-label study of first-line erlotinib versus chemotherapy in patients with advanced EGFR mutation-positive non-small-cell lung cancer (NSCLC). Ann Oncol 2013;24(6): 1615-22
32. Barlesi F, Scherpereel A, Rittmeyer A, et al. Randomized phase III trial of maintenance bevacizumab with or without pemetrexed after first-line induction with bevacizumab, cisplatin, and pemetrexed in advanced nonsquamous non-small-cell lung cancer: AVAPERL (MO22089). J Clin Oncol 2013;31(24):3004-11
33. Rittmeyer A, Gorbunova V, Vikstro¨m A, et al. Health-related quality of life in patients with advanced nonsquamous non-small-cell lung cancer receiving bevacizumab or bevacizumab-plus-pemetrexed maintenance in patients with metastatic pancreatic cancer: results from the PRODIGE 4/ACCORD 11 randomized trial. J Clin Oncol 2013;31(1):23-9
40. Chinot OL, Wick W, Mason W, et al. Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. N Engl J Med 2014;370(8):709-22
41. Hussain M, Tangen CM, Berry DL, et al. Intermittent versus continuous androgen deprivation in prostate cancer. N Engl J Med 2013;368(14):1314-25
42. Von Plessen C, Bergman B, Andresen O, et al. Palliative chemotherapy beyond three courses conveys no survival or consistent quality-of-life benefits in advanced non-small-cell lung cancer. Br J Cancer 2006;95(8):966-73
43. Gilbert MR, Dignam JJ, Armstrong TS, et al. A randomized trial of bevacizumab for newly diagnosed glioblastoma. N Engl J Med 2014;370(8):699-708
44. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85(5):365-76

doi: 10.1586/14737140.2015.1047768

45. Cella D, Eton DT, Fairclough DL, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol 2002;55(3):285-95

46. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134(8): 663-94

47. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg Lond Engl 2012;10(1):28-55

48. Neuhs¨user M. How to deal with multiple endpoints in clinical trials. Fundam Clin Pharmacol 2006;20(6):515-23

- This article discusses statistical methods that can be applied to control the rate of false-positive conclusions at an acceptable level for trials with multiple endpoints.

doi: 10.1586/14737140.2015.1047768

C. Calcul du nombre de Sujets nécessaires en phase III incluant deux critères conjoints en cancérologie : Package R Coprimary.

Résumé :

Le package R «coprimary» permet de planifier des essais de supériorité et non-infériorité ou équivalence avec des analyses intermédiaires (efficacité et/ou futilité). Il contient 6 fonctions pour calculer le NSN avec un ou deux critères de jugement de type temps jusqu'à événement. La fonction **nsurvival** permet de calculer le NSN avec un seul critère de jugement et aussi intègre le nombre de dimensions considérées pour la qualité de vie relative à la santé (QdV). La fonction **ncoprimary** permet de calculer le NSN avec deux critères de jugement principaux. Les deux critères peuvent être un critère composite comme SSP ou la QdV. Ces deux fonctions font appel aux quatre autres fonctions :

- La fonction **datacheck** permet de vérifier la cohérence du design ;
- La fonction **nbevent** calcul le nombre d'événements nécessaires pour le calcul du NSN ;
- La fonction **probafix** estime la probabilité d'évènement lorsque le suivi est fixe ;
- La fonction **probanofix** estime la probabilité d'événement lorsque le suivi est non fixe.

Les critères conjoints permettent de détourner l'attention d'un critère d'évaluation très important en clinique, mais pour lequel il n'y a pas de différence statistiquement significative. En combinant deux critères de survie on augmente nos chances d'obtenir une différence statistiquement significative. Lorsque les critères conjoints prennent en compte à la fois des bénéfices pour le clinicien et le patient, ils contribuent à évaluer la balance bénéfices risques (avantage ou désavantage global). Cependant, lorsqu'ils ne comportent que des critères d'efficacité, il faut au moins exiger que ces critères d'efficacité soient tous cliniques et d'importance à peu près voisine pour le patient. Le package **coprimary** est le premier package R qui permet le calcul du NSN avec co-critère conjoint de temps jusqu'à événement.

Il est disponible pour tous les utilisateurs du logiciel R en libre accès sur le site du CRAN (https://CRAN.R-project.org/package=coprimary)

Sample size determination in oncology phase III clinical trials with two primary time-to-event endpoints: *coprimary* R package

Article soumis à *computer methods and programs in biomedicines*

Alhousseiny PAM[1], Amélie ANOTA[1,4], Caroline MOLLEVI[2,4], Thomas FILLERON[3,4], Franck BONNNETAIN[1,4]

[1]Methodology and Quality of Life in oncology Unit, INSERM UMR 1098, University Hospital of Besançon, Besançon, France.

[2]Unité de Biométrie, Institut du Cancer Montpellier, France.

[3]Institut Claudius Régaud, IUCT-O, Toulouse, France.

[4]French National Platform Quality of Life and Cancer, France.

Correspoding author

Adresse e-mail: alhousseiny.pam@gmail.com (Alhousseiny PAM)

Tel : +33 (0)6 60 50 32 84

ABSTRACT

Background and Objective: In oncology, an important challenge in the conception of clinical trials is the sample size calculation according to the main objective(s) and the ability to manage multiple co-primary endpoints. Overall survival (OS) is considered as the most relevant primary endpoint for demonstrating clinical benefit for cancer drugs. However, multiple co-primary endpoints is an alternative to assess the efficacy of treatment in phase III clinical trial.

Methods: The statistical power of the study using time-to-event endpoints depends on the number of events observed during the trial and not on the number of patients included. The "*coprimary"* R package computes the required number of patients for two time-to-event endpoints as co-primary endpoint.

Results: This R package contains six functions to check the consistency of the design and to compute the sample size with one or two time to event endpoints. The endpoints can be OS, progression free survival (PFS) or the time to health-related quality of life (HRQoL) score deterioration as example. The *"nsurvival"* function computes the sample size for one time-to-event endpoint. The "*ncoprimary"* computes the sample size for two primary time to-event endpoints. The "*DataCheck"* checks the consistency between the arguments specified. The "*nbevent"* computes the number of events required to determine the number of patients to include. The "*probafix"* estimates the probability of event when the follow-up is fixed. The "*probanofix"* estimates the probability of event when the follow-up is not fixed.

Conclusions: The "*coprimary*" package can be accessed and download on the CRAN website (https://CRAN.R-project.org/package=coprimary). The *coprimary* R package depends on: *stats*, *gsDesign*, *digest*, *plyr*, and *proto* R packages.

Keywords

Sample size calculation, co-primary endpoints, clinical trial design, multiple testing

INTRODUCTION

Co-primary endpoints are more and more used in phase III cancer clinical trials. Use of co-primary endpoints creates challenges in the evaluation of the sample size determination during the trial design. Every clinical trial should be planned. This plan should include the objective(s) of the trial, the sample size with scientific justification, the statistical methods used and the assumptions.

Most commonly, a single composite endpoint (i.e. multiple events all treated as one endpoint) is selected as the primary endpoint. It is used as the basis for the trial design including sample size determination, as well as for interim monitoring and final analyses. Composite endpoints are overused and often improperly used, because a composite endpoint lumps together several outcomes. It is key to select appropriate endpoints before a trial begins. One alternative could be to use multiple primary endpoint as co-primary endpoint to assess the efficacy of treatment in clinical trials[1–4]. Clinical trials with time-to-event endpoint as co-primary endpoints are common in many areas such as oncology, cardiovascular disease, etc. For time to event endpoints, the statistical power depends on the total number of events rather than on total sample size. The determination of sample size is fundamental and critical elements in the design of phase III clinical trials [5,6]. If the sample size is too small, important effects may not be detected. If it is too large, it represents a waste of resources and unethically puts more participants at risk than necessary. Typical problems that arise during the study such as drop-outs, and uncertainty in the control arm event rates are difficult to enter completely without the use of specific design and simulation tools. However, a sufficient number of patients must be entered into the trial and followed for a sufficient length of time in order to observe the required number of events. When utilizing multiple primary endpoints, sample size determination is designed with the aim to detect the effects on all endpoints (referred as "multiple co-primary endpoints").

This article introduces an R package, named "*coprimary*", created to compute the sample size

in clinical trials with one or two time-to-event endpoints. In section 2, the methods are presented. In Section 3, the function and arguments with one time-to-event endpoint used are described and the methods are illustrated with examples. In section 4, the function and arguments with two co-primary time-to-event endpoints used are described and the methods are illustrated with examples.

METHODOLOGY

1.1 Required sample size to compare the log-hazard rates

The logrank test is one of the most popular tests to compare two survival distributions. It is easy to apply and is usually more powerful than an analysis based simply on proportions. This section allows to compute the sample size with two co-primary endpoints.

Consider a randomized clinical trial designed to compare two treatments with a total of N patients. We denote by p_E the proportion of patients randomized to the experimental arm (with $0 < p_E < 1$). Let $n_E = p_E N$ patients be assigned to the experimental group (E) and $n_C = (1 - p_E)N$ participants to the control group (C). Two time- to-event endpoints (T1, T2) are to be evaluated as primary endpoints of analysis. Thus, we have n_E paired time-to-event endpoints $(E_{T1i}, E_{T2i})(i = 1, \ldots, n_E)$ for the experimental group and n_C paired time-to-event endpoints (C_{T1j}, C_{T2j}) $(j = 1, \ldots, n_C)$ for the control group. Assume that the time-to-event endpoints (E_{T1i}, E_{T2i}) and (C_{T1j}, C_{T2j}) follow the exponential distribution with constant hazard rates $\lambda_{Ek}(t) = \lambda_{Ek}$ and $\lambda_{Ck}(t) = \lambda_{Ck}$ for all $t > 0$, $k = 1, 2$, respectively. In addition, the proportion of survivors after t years is given by $S_{Ek}(t) = \exp(-\lambda_{Ek}t)$ for the experimental group and $S_{Ck}(t) = \exp(-\lambda_{Ck}t)$ for the control group. Furthermore, assume that the two time-to-event endpoints within individual for the E and C are correlated with ρ_E and ρ_C, that is, $\rho_E = \text{corr}[E_{T1i}, E_{T2i}]$ and $\rho_C = \text{corr}[C_{T1j}, C_{T2j}]$, respectively, but that observations from different individuals are independent.

Suppose patients entered the trial uniformly with accrual duration of length T_a and followed by duration of length T_f. The follow-up period resulting in a trial is $T = T_a + T_f$.

Let $S_{C1}(t)$, $S_{C2}(t)$ and $S_{E1}(t)$, $S_{E2}(t)$ denote the survival estimates at time t in the control and experimental arms for T1 and T2 respectively and h_1, h_2 the respective hazard ratio. The hazard ratios between the two groups with a proportional hazards model for survival times can be estimated by $h_1 = \ln[S_{E1}(t)] / \ln[S_{C1}(t)]$ and $h_2 = \ln[S_{E2}(t)] / \ln[S_{C2}(t)]$. The hazard ratios are constant $h_1 = \lambda_{E1} / \lambda_{C1}$, $h_2 = \lambda_{E2} / \lambda_{C2}$ when the survival distribution is exponential, with λ_{E1}, λ_{E2} and λ_{C1}, λ_{C2} the hazard rates in the experimental and control group for T1 and T2.

The type I error α1 and α2 correspond to the probability to reject the null hypothesis when the null hypothesis is true for T1 and T2, respectively. The type II error β represents the probability to not reject the null hypothesis when the alternative hypothesis is true.

The sample size determination in a clinical trial with two co-primary endpoints is computed in two steps for each endpoint. In the first step, the number of events required is computed. In the second step, the number of patients that we have to include is computed.

1.1.1 The required number of events

Superiority trial

Superiority trials are designed to demonstrate that one treatment is more effective than another. This type of design is often used to test the effectiveness of a treatment compared to the standard treatment. The statistical analysis will consist of a two-sided test of the null hypothesis of equality of the hazard function,

H_{01}: $\lambda_{E1}=\lambda_{C1}$ against H_{11}: $\lambda_{E1} \neq \lambda_{C1}$
H_{02}: $\lambda_{E2}=\lambda_{C2}$ against H_{12}: $\lambda_{E2} \neq \lambda_{C2}$

which is equivalent to test in terms of the hazard ratio:

H_{01}: $h_1 = \lambda_{E1}/\lambda_{C1} = 1$ against H_{11}: $h_1 \neq 1$
H_{02}: $h_2 = \lambda_{E2}/\lambda_{C2} = 1$ against H_{12}: $h_2 \neq 1$

or in terms of the log-hazard ratio

H_{01}: $\ln(h_1) = 0$ against H_{11}: $\ln(h_1) \neq 0$
H_{02}: $\ln(h_2) = 0$ against H_{12}: $\ln(h_2) \neq 0$

For the one sided test, the null hypothesis of equality of the hazard rates is,

H_{01}: $\lambda_{E1}=\lambda_{C1}$ against H_{11}: $\lambda_{E1} < \lambda_{C1}$
H_{02}: $\lambda_{E2}=\lambda_{C2}$ against H_{12}: $\lambda_{E2} < \lambda_{C2}$

The equivalent expression using the hazard ratio is

H_{01}: $h_1 = \lambda_{E1}/\lambda_{C1} = 1$ against H_{11}: $h_1 < 1$
H_{02}: $h_2 = \lambda_{E2}/\lambda_{C2} = 1$ against H_{12}: $h_2 < 1$

or in terms of the log-hazard ratio

H_{01}: $\ln(h_1) = 0$ against H_{11}: $\ln(h_1) < 0$
H_{02}: $\ln(h_2) = 0$ against H_{12}: $\ln(h_2) < 0$

The formula to determine the number of events required to evaluate the superiority of a treatment using the log-rank test with a significance levels α1, α2 and type II error 1-β according to Schoenfeld[7] is given by:

$$e = \max\left(\frac{\left[Z_{1-\alpha_1/s} + Z_{1-\beta}\right]^2}{p_E(1-p_E)\left[\ln(h_{a1})\right]^2}, \frac{\left[Z_{1-\alpha_2/s} + Z_{1-\beta}\right]^2}{p_E(1-p_E)\left[\ln(h_{a2})\right]^2} \right)$$

with s = 1 (resp s = 2) if the test is one sided (resp. two-sided) and $Z_{1-\alpha}$ is the $100(1-\alpha)$th percentile of the standard normal distribution.

Non inferiority trial

Non-inferiority trials are designed to demonstrate that a treatment is at least not appreciably worse than another by a small amount. This amount is known as the non-inferiority margin or

delta δ. This type of design is often employed when comparing a new treatment to an established medical standard of care[8]. Here, we are interested in the rejection of the hypothesis that the experimental treatment is worse than the control treatment by a non-inferiority margin:

H_{01}: $h_1 = \lambda_{E1}/\lambda_{C1} \geq 1+\delta_{01}$ against H_{11}: $h_1 = \lambda_{E1}/\lambda_{C1} < 1+\delta_{01}$
H_{02}: $h_2 = \lambda_{E2}/\lambda_{C2} \geq 1+\delta_{02}$ against H_{12}: $h_2 = \lambda_{E2}/\lambda_{C2} < 1+\delta_{02}$

where δ_{01}, δ_{02} (≥ 0) are the non-inferiority margins used to define that the experimental treatment is not inferior to the control or standard treatment by more than this amount for T1 and T2. The rejection of the null hypothesis means the experimental arm is not less than the control arm. In terms of the log-hazard ratio, this is equivalent to testing the following hypotheses:

H_{01}: $\ln(h_1) \geq \ln(1+\delta_{01}) = \ln(h_{01})$ against the one-sided alternative H_{11}: $\ln(h_1) < \ln(1+\delta_{01}) = \ln(h_{01})$

H_{02}: $\ln(h_2) \geq \ln(1+\delta_{02}) = \ln(h_{02})$ against the one-sided alternative H_{12}: $\ln(h_2) < \ln(1+\delta_{02}) = \ln(h_{02})$

The number of events required is thus:

$$e = \max\left(\frac{\left[Z_{1-\alpha_1} + Z_{1-\beta}\right]^2}{p_E(1-p_E)\left[\ln(h_{01})\right]^2}, \frac{\left[Z_{1-\alpha_2} + Z_{1-\beta}\right]^2}{p_E(1-p_E)\left[\ln(h_{02})\right]^2} \right)$$

Equivalence trial

Equivalence trials are designed to demonstrate that the experimental treatment is therapeutically similar to the treatment control for each endpoint. In terms of log hazard ratio, this corresponds to the following hypothesis:

H_{01}: $|\ln(h_1)| \geq \delta_{01}$ against H_{11} : $|\ln(h_1)| < \delta_{01}$
H_{02}: $|\ln(h_2)| \geq \delta_{02}$ against H_{12} : $|\ln(h_2)| < \delta_{02}$

with δ_{01}, δ_{02} the equivalence margin in terms of log hazard ratio for T1 and T2. This is equivalent to the following hypotheses:

$H_{01} : \ln(h_1) \geq \delta_{01}$ against $H_{11} : \ln(h_1) < \delta_{01}$
$H_{01'} : \ln(h_1) \leq -\delta_{01}$ against $H_{11'} : \ln(h_1) > -\delta_{01}$

$H_{02} : \ln(h_2) \geq \delta_{02}$ against $H_{12} : \ln(h_2) < \delta_{02}$
$H_{02'} : \ln(h_2) \leq -\delta_{02}$ against $H_{12'} : \ln(h_2) > -\delta_{02}$

$$e = \max\left(\frac{\left[Z_{1-\alpha_1} + Z_{1-\beta}\right]^2}{p_E(1-p_E)[\ln(h_{01})]^2}, \frac{\left[Z_{1-\alpha_2} + Z_{1-\beta}\right]^2}{p_E(1-p_E)[\ln(h_{02})]^2} \right)$$

with the equivalence margins $\ln(h_{01}) = \delta_{01}$ and $\ln(h_{02}) = \delta_{02}$.

1.1.2 The required number of patients

The number of patients to observe the number of events required e is dependent on the probability ψ to observe the first event of interest with $N = e/\psi$. The parameters of this probability are the accrual duration, the follow-up, the risk of failure in each arm, the duration of the trial and the rate of drop outs. The probability of weighted sum of the probability to observe a failure in the experimental and control arms (respectively denoted by ψ_E and ψ_C) denoted by $\psi_k(t, \lambda_k, \gamma)$ is given by:

$$\psi_k(t, \lambda_k, \gamma) = p_E \times \psi_E(t, \lambda_{E_k}, \gamma_E) + (1 - p_E) \times \psi_C(t, \lambda_{C_k}, \gamma_C)$$

with $\lambda_k = (\lambda_{E_k}, \lambda_{C_k})$ (resp. $\gamma = (\gamma_E, \gamma_C)$) the hazard rate of the event of interest (resp. drop-out rate) in each arm and k=1,2. There are two types of follow-up: variable and fixed.

No fixed follow-up

When the follow-up is variable (i.e not fixed) in the design, all patients are followed until the end of the study ($T_f = \infty$). When the drop-out rate is null in each arm (i.e. $\gamma = (0,0)$), the formulation of the ψ_E and $_C$ can be estimated as follows [9] :

$$\psi_E(t,\lambda_{E_k},0) = \begin{cases} 1 - \frac{1-\exp(-\lambda_{E_k} t)}{\lambda_{E_k}} & if\ t \le T_a \\ 1 - \left(\frac{\exp(-\lambda_{E_k} t)}{\lambda_{E_k}}\right)\left(\frac{\exp(-\lambda_{E_k} T_a)}{T_a}\right) & if\ t > T_a \end{cases}, \text{ with k=1,2}$$

Fixed follow-up

When the follow-up is fixed, all patients will be followed during a fixed period ($T_f < \infty$). When the drop-out rate is null in each arm (i.e. $\gamma = (0,0)$), the formulation of the ψ_k can be estimated as follows [9]:

$$\psi_E(t,\lambda_{E_k},0) = \begin{cases} \left[1 - \exp(-\lambda_{E_k} T_f)\right]\frac{T'}{t} + \left\{\left(t - T'\right) - \frac{1}{\lambda_{E_k} t}\left[1 - \exp(\lambda_{E_k}(T' - t))\right]\right\} & \text{if } t \le T_a \\ \left[1 - \exp(-\lambda_{E_k} T_f)\right] + \frac{1}{T_a}\left\{(T_a - T') - \frac{\exp(-\lambda_{E_k} t)}{\lambda_{E_k}}\left[\exp(\lambda_{E_k} T_a) - \exp(\lambda_{E_k} T')\right]\right\} & \text{if } t > T_a \end{cases}$$

with k=1, 2

Drop outs

When the drop-out rate is not null in each arm (i.e. $\gamma \neq (0,0)$), the probability to observe an event at time *t* in the experimental arm is thus:

$$\psi_{E_k}(t,\lambda_{E_k},\gamma) = \frac{\lambda_{E_k}}{\lambda_{E_k} + \gamma_E}\psi_k(t,\lambda_{E_k} + \gamma_E,0)$$

with k=1, 2

1.2 Multiple testing in clinical trial

In clinical research, it is well known that testing multiple hypotheses without any adjustment may increase the probability of erroneously rejecting at least one true null hypothesis. Thus, if two endpoints are considered as co-primary endpoints in a phase III cancer clinical trial, the probability of a false positive finding can be substantial if no adjustment on multiple testing is

made. If some interim analyses are planned, the probability of observing a "significant" result may also increase. To illustrate, it has been shown that, with 10 analysis times, the overall type I error rate of 5% increases to about 20%[10]. Thus, several methods for adjustment on multiple testing have been proposed, the simplest is Bonferroni method. We proposed below the more important methods we have implemented in our package.

When two endpoints are used as co-primary endpoints, then two hypothesis are tested for each endpoint, the overall type I error α is manually adjusted in two parties α1 and α2. When the HQoL is considered as primary endpoint with several dimensions, the overall type I error α is adjusted according the number of dimensions by Bonferroni method. The analysis plan should describe ways to determine how the endpoints are tested, including the order of testing and α level applied to each specific test.

1.2.1 The Bonferroni procedure

In the Bonferroni procedure the significance level α is split equally among the k hypotheses, and each is tested at level α/k[11]. The individual null hypothesis H_{0i} is rejected if $p_i < \alpha/k$, and the global null hypothesis H_0 is rejected if $\min p_i < \alpha/k$. However, the Bonferroni procedure is conservative and consequently leads to a diminution of power if the number of hypotheses k is large. Thus, this method is not recommended in case of lots of endpoints considered.

1.2.2 The Fixed sequence procedure (hierarchical)

If the sequence of hypotheses is fixed a priori on the basis of the relative importance of the individual hypotheses or expected treatment effects, one can apply a fixed sequence procedure [12]. The procedure is carried out sequentially where H(i) is tested at level α as long as all previous hypotheses H(1),….H(i-1) are rejected; the test is stopped when the first non-significant result is observed. A major problem of the fixed sequence procedure is that once a hypothesis is not rejected, no further testing is permitted. The advantage of this procedure is to not adjust the type I error rate.

1.3 Interim analyses

For ethical and economic reasons, the technique of sequential testing has been developed to enable the examination of data at a series of intermediate times [13,14]. At any stage in the trial, if the boundary is crossed, the study is stopped and an appropriate conclusion drawn [15,16]. Group sequential designs are based on the principle that there are a total of H analyses: H−1 interim analyses and one final analysis. The number of the interim analyses should be pre-specified in the protocol. Since multiple analyses to test the same hypothesis lead to multiple comparisons, the overall type I error rate increases. Several approaches were thus developed to adjust the type I error rate.

1.3.1 The Pocock procedure

Pocock et al [17] came up with an approach that effectively compared the p-value against an adjusted alpha level where the adjusted alpha level was constant across each interim evaluation. The Pocock boundaries are depended on the type I error α and the number of interim analysis K. For each interim analysis, a constant p-value is used to compare the treatments using the test statistics η_k.

1.3.2 The O'Brien and Fleming procedure

O'Brien and Fleming introduced a group sequential method which is used for interim analysis in clinical trials. The O'Brien-Fleming approach[18] is the most popular group sequential approach because the significance level at the final analysis is near the overall desired significance level. At each interim analysis K the boundary of the α_k type I error increases. The procedure use more conservative stopping boundaries at each interim analysis. These bounds spend little alpha at each interim analysis and lead to boundary values at the final interim analysis that are close to the values fixed by the design.

1.3.3 Alpha spending function

Lan and DeMets extended the group sequential concept to a very flexible method that controls the overall alpha level while allowing for the number and exact timing of the interim analyses to remain unspecified a priori. The stopping boundaries are generated by two functions $\alpha_1(t)$ and $\alpha_2(t)$ that closely resemble the Pocock and O'Brien-Fleming stopping boundaries. The functions $\alpha_1(t)$ and $\alpha_2(t)$ are given by:

$$\alpha_1(t) = \begin{cases} 2 - 2\phi\left(\frac{Z_{\alpha/2}}{\sqrt{t}}\right) & \text{for one - sided test} \\ 4 - 4\phi\left(\frac{Z_{\alpha/4}}{\sqrt{t}}\right) & \text{for two - sided test} \end{cases}$$

and $\alpha_2(t) = \alpha ln[1+(e-1)t]$

PROGRAM DESCRIPTION: the *"nsurvival" function*

The *"nsurvival"* function computes the sample size for one time to event endpoint. This endpoint can be a survival endpoint, such as OS or PFS, or HRQoL considering time to HRQoL score deterioration. For HRQoL, several HRQoL dimensions can be considered and Bonferroni method is used for adjustment type I error. This function is identical to the "*plansurvct.func*" function developed by Filleron et al. [19], but the particularity of our function is that several HRQoL dimensions can be specified in the parameters. Our function requires the following arguments:

```
nsurvival(design,Survhyp,alpha,duraccrual,durstudy,power,pe,
look,fup,dropout,dqol)
```

1.4 ARGUMENTS

The R function ***nsurvival*** has eleven arguments:

Table 1: The parameters of the "*nsurvival"* function according to the type of trial.

	Superiority	**Non inferiority**	**Equivalence**
design	c(1,1) : 1-sided c(1,2) : 2-sided	c(2)	c(3)
survhyp	c(thyp,t,hype,Sc)	c(thyp,t,hype,Sc,hypeA)	c(t,delta,Sc)
alpha	alpha : 1-sided c(alpha.low,alpha.up) : 2-sided	alpha	alpha
duraccrual	duraccrual	duraccrual	duraccrual
durstudy	durstudy	durstudy	durstudy
power(0.80)	power	power	power
pe(0.5)	pe	pe	pe
look(1)	c(1) : one final analysis c(nb,bound,timing) : > 1 bound=c(bound.eff,bound.fut) : 1-sided bound=c(bound.lown,bound.up):2-sided	c(1) : one final analysis c(nb,bound,timing) : > 1 bound=c(bound.eff,bound.fut)	c(1) : one final analysis
fup(0)	c(0) : No fixed follow-up c(1,durfollow) : Fixed follow-up	c(0) : No fixed follow-up c(1,durfollow) : Fixed follow-up	c(0) : No fixed follow-up c(1,durfollow) : Fixed follow-up
dropout(0)	c(0) : No drop out c(1,gammae,gammac)	c(0) : No drop out c(1,gammae,gammac)	c(0) : No drop out c(1,gammae,gammac)
dqol(0)	dqol	dqol	dqol

*the value in parenthesis is the default value

The R function "*nsurvival"* has eleven arguments:

- **design:** Four tests are possible

- c(1,1): one-sided superiority test

- c(1,2): two-sided superiority test

- c(2): non-inferiority test

- c(3): equivalence test

- **Survhyp:** Survival rate or hazard ratio according the test

- c(thyp,t,hype,Sc): one-sided or two-sided superiority test

* thyp=1: comparison between the survival rate in experimental arm (hype, 0 < hype <1) and the survival rate in control arm (Sc, 0 < Sc < 1) at time t.

* thyp=2: comparison between the hazard ratio in experimental arm (hype, 0 < hype <1) and the hazard ratio in control arm (Sc, 0 < Sc < 1) at time t.

- c(thyp,t,hype,Sc,hypeA): non-inferiority test

* thyp=1: comparison between the survival rate in the experimental arm (hype, 0 < hype <1 (respectively hypeA, 0 < hypeA <1) under the null hypothesis (respectively under the alternative hypothesis)) and the survival rate in the control arm (Sc, 0 < Sc < 1) at time t.

* thyp=2: comparison between the hazard ratio in the experimental arm (hype, 0 < hype <1 (respectively hypeA, 0 < hypeA <1)) under the null hypothesis (respectively under the alternative hypothesis) and the hazard ratio in the control arm (Sc, 0 < Sc < 1) at time t.

- c(t,delta,Sc): equivalence test

* delta the log-hazard ratio equivalence margin and Sc the survival rates at time t in the control arm (with 0 < Sc < 1)

- **pe :** Proportion of patients assigned to the experimental arm

- 0 < pe < 1: all tests

- **alpha :** Type I error rate

- 0 < alfa < 1 : for a non-inferiority test, an equivalence test and a one sided superiority test.

- c(alfa.low,alfa.up): for a two-sided superiority test

* alfa.low and alfa.up the lower and upper alpha boundaries (with 0 < alfa.low < 1, 0 < alfa.up < 1, 0 < alfa.low + alfa.up < 1). If there is no interim analysis, the upper and lower alpha boundaries need to be equal (i.e. alfa.low = alfa.up).

- **power = 1-beta:** where beta is the probability of a type II error, the chance of missing a clinically significant difference.

- 0 < power < 1: all tests

- **duraccrual:** Accrual duration, expressed in t time units

- **durstudy:** Study duration, expressed in time units

- **look:** Specifies the number of looks at the data and in case of planned interim analyses, the type of bound, and their timing:

- c(1): if there is only one final analysis, all tests.

- c(nb,bound,timing) : at least one interim analysis for superiority test or non-inferiority test

* nb: the number of planned analysis

* timing: timing of interim analyses. The values should satisfy the follow-ing relation: 0<timing[1]<timing[2]<· · ·<timing[Last Interim]<1.

* bound=c(bound.eff,bound.fut) : one-sided superiority test or non-inferiority. bound.eff and bound.fut correspond to the type of boundaries used for efficacy (i.e. reject H0) and futility (i.e. reject H1). The following values can be used for the parameter bound.fut: 0: No futility monitoring, 1: Lan deMets O'Brien Fleming, 2: Lan deMets Pocock.

* bound= c(bound.lown,bound.up): two-sided superiority test. bound.low and bound.up correspond the type of lower and upper boundaries used (1: Lan deMets O'Brien Fleming, 2: Lan deMets Pocock, 3: O'Brien Fleming, 4: Pocock). If alpha boundaries are asymmetric (i.e. alfa.low not equal to alfa.up) only Lan deMets spending function can be used.

* for an equivalence trial, no interim analysis is planned with this function.

- **fup:** follow-up information

- c(0) : no fixed follow-up, all patients are followed until the end of study, all tests.

- c(1,durfollow) : fixed follow-up, all patients have a fixed duration of follow-up, all tests.

* durfollow: duration of follow-up.

- **dropout:** drop out information

- c(0): no drop out, all tests

- c(1,gammae,gammac) : drop out, all tests.

* gammae and gammac correspond the hazard drop-out rates per time units in the experimental and control arms respectively.

- **`dqol:`** number of dimension targeted for the quality of life, all tests.

*Any convenient time unit can be used, such as years, months or days, as long as it is used consistently for all the items (duration accrual, study duration, follow-up. . .).

1.5 APPLICATION: one primary endpoint

Depending on the parameters, the "*nsurvival*" function estimates the number of events to observe and computes the number of subjects to include in clinical study in order to observe these events. The program returns the parameter values, the number of events estimated and finally the number of patients that we have to include in the clinical study.

1.5.1 Superiority trial

- Example 1: two-sided superiority trial with HRQoL as the primary endpoint.

Let consider a randomized superiority clinical trial with HRQoL as the primary endpoint with three targeted dimensions, using the time to HRQoL deterioration approach. The 5-year rate of patients without HRQOL deterioration is estimated to 68% in the control arm and to 77% in experimental arm. We consider a two sided type I error of 5%, an accrual duration of 2 years, a study duration of 6 years, and a follow-up of 5 years for each patient. By default, the statistical power is fixed to 80% (power=0.80), the ratio is 1:1, i.e. the proportion of patients included in the experimental arm is 50% (pe=0.5), no interim analysis is performed (look=1), and no drop out is observed (dropout=0).

```
ns1 <-nsurvival(design=c(1,2),Survhyp=c(1,4,0.77,0.68),
alpha=c(0.02,0.03),duraccrual=2,durstudy=6, fup=c(1,5),dqol=3)
```

- **Result 1:** Results obtained for example 1 using the *"nsurvival"* function

Results	values
Number of targeted HRQoL dimensions	3
Alpha	alpha.low=0.007, alpha.up=0.010
Statistical Power	0.80 (80%)

Ratio in arm E	0.50
Hazad ratio (HR) Arm E vs. Arm C	0.678
Number of events required	265
Total number of subjects to include	838
Number of subjects (Arm E, Arm C)	(419, 419)

The hazard ratio between the experimental arm and the control arm is 0.678=log (0.77) / log (0.68). The number of event estimated is 265 and the number of subjects required in the study is 838 patients. The number of subjects required in each arm is 419 patients.

- **Example 2**: one-sided superiority with progression free survival (PFS) endpoint as the primary endpoint.

Let consider a randomized superiority clinical trial with the PFS as the primary endpoint. The 3-year progression free survival rate is 69% in the control arm and the hazard ratio under alternative hypothesis is 76% with one sided type I error of 5%, accrual duration of 3 years, study duration of 6 years, follow-up of 4 years. By default, the statistical power is fixed to 80% (power=0.80), the ratio is 1:1 i.e. the proportion of patients included in the experimental arm is 50% (pe=0.5), no interim analysis is performed (look=1), no drop out is observed (dropout=0), and no targeted dimensions of HRQoL (dqol=0).

```
ns2 <-nsurvival(design=c(1,1),Survhyp=c(2,3,0.76,0.69),
alpha=0.05,duraccrual=3, durstudy=6,fup=c(1,4))
```

- **Result 2:** Results obtained for example 2 using the *"nsurvival"* function

Results	**values**
Alpha	alpha=0.05
Statistical Power	0.80 (80%)
Ratio in arm E	0.50

Survival rate in control arm	0.69
Hazard ratio (HR) Arm E vs. Arm C	0.76
Survival rate in experimental arm	0.754
Number of events required	328
Total number of subjects to include	968
Number of subjects (Arm E, Arm C)	(484, 484)

The survival rate in the experimental arm is 0.754 computed from the hazard ratio 0.76 and the survival rate in the control arm 0.69. The number of events estimated is 328 and the number of subjects required in the study is 968 patients. The number of patients required in each arm is 484 patients.

1.5.2 Non-inferiority trial

- **Example 3:** non-inferiority trial with the HRQoL as the primary endpoint.

Let consider a randomized non-inferiority trial with the HRQoL as the primary endpoint with three targeted dimensions. The 3-year rate of patients without HRQOL deterioration under null hypothesis and alternative hypothesis is 60% and 70% in the experimental arm, respectively and 70% in control arm with type I error of 5%, accrual duration of 4 years, study duration of 8 years, 2 interim analysis are performed. By default, the statistical power is fixed to 80% (power=0.80), the ratio is 1:1 i.e. the proportion of patients included in the experimental arm is 50% (pe=0.5), all patients are followed until the end of study (fup=0), and no drop out is observed (dropout=0).

```
ns3 <- nsurvival(design=c(2),Survhyp=c(1,5,0.60,0.70,0.70),
alpha=0.05, duraccrual=4,durstudy=8, look=c(3,c(1,1),
c(1/3,2/3)), dqol=3)
```

- **Result 3:** Results obtained for example 3 using the "*nsurvival"* function

Results	**values**
Alpha	alpha=0.05
Statistical Power	0.80 (80%)
Ratio in arm E	0.50

Number analysis planned	3
Survival rate in control arm	0.70
Survival rate in E_{h0} arm	0.60
Survival rate in E_{ha} arm	0.70
Hazard ratio (HR) Arm E_{h0} vs. Arm C	1.43
Hazard ratio (HR) Arm E_{ha} vs. Arm C	1
Number of events required	298
Total number of subjects to include	864
Number of subjects (Arm E, Arm C)	(432, 432)

The hazard ratio under the null hypothesis and alternative hypothesis is 1.43 = log (0.60) / log (0.70) and 1 = log (0.70) / log (0.70), respectively. The number of events estimated is 298 and the number of subjects required in study is 864 patients. The number of patients required in each arm is 432 patients.

1.5.3 Equivalence trial

- **Example 4**: equivalence trial with the disease free survival (DFS) as the primary endpoint

Suppose a randomized equivalence trial with the DFS as the primary endpoint with two targeted dimensions. The 3-year rate of DFS is 65% in control arm and a log hazard ratio equivalence margin is 30%, type I error of 10%, accrual duration of 3 years, study duration of 5 years, and drop out of 10% is observed. By default, the statistical power is fixed to 80% (power=0.80), the ratio is 1:1 i.e. the proportion of patients included in the experimental arm is 50% (pe=0.5), no interim analysis is performed (look=1), all patients are followed until the end of study (fup=0).

```
ns4 <-survival(design=c(3),Survhyp=c(3,0.30,0.65),alpha=0.10,
duraccrual=3, durstudy=5,dropout= c(1,0.05,0.05))
```

- **Result 4:** Results obtained for example 4 using the "*nsurvival"* function

Results	**values**
Alpha	alpha=0.10
Statistical Power	0.80 (80%)
Ratio in arm E	0.50

Number analysis planned	1
Survival rate in control arm	0.65
log hazard ratio equivalence margin	0.30
Log hazard ratio H_0	1.35
Log hazard ratio H_1	0.741
Number of events required	292
Total number of subjects to include	910
Number of subjects (Arm E, Arm C)	(455, 455)

The log hazard ratio under the null hypothesis is 1.35 and the log hazard ratio under the alternative hypothesis is 0.741. The number of events estimated is 292 and the number of subjects required in the study is 910 patients. The number of patients required in each arm is 455 patients.

Program description: the "*ncoprimary*" function

The "*ncoprimary"* function computes the sample size for two co-primary endpoints for oncology phase III clinical trials. Both endpoints have to be time to events endpoints. They can be classical survival endpoints such as PFS or OS, or HRQoL considering the time to HRQoL score deterioration method [20]. In case of the use of HRQoL as a co-primary, several HRQoL dimensions can be considered. This function requires the following arguments:

```
ncoprimary(design,Survhyp1,Survhyp2,alpha1,alpha2,duraccrual,d
urstudy,power,pe,look,fup,dropout,dqol)
```

1.6 ARGUMENTS

Table 2: The parameters of the "*ncoprimary"* function according to the type of trial

	Superiority	**Non inferiority**	**Equivalence**

design	c(1,1) : 1-sided c(1,2) : 2-sided	c(2)	c(3)
Survhyp1, Survhyp2	c(thyp,t,hype,Sc)	c(thyp,t,hype,Sc,hypeA)	c(t,delta,Sc)
alpha1, alpha2	alpha : 1-sided c(alpha.low,alpha.up) : 2-sided	alpha	alpha
duraccrual	duraccrual	duraccrual	duraccrual
durstudy	durstudy	durstudy	durstudy
power (0.80)	power	power	power
pe (0.5)	pe	pe	pe
look (1)	c(1) : one final analysis c(nb,bound,timing) : > 1 bound=c(bound.eff,bound.fut) : 1-sided bound=c(bound.lown,bound.up):2-sided	c(1) : one final analysis c(nb,bound,timing) : > 1 bound=c(bound.eff,bound.fut)	c(1) : one final analysis
fup (0)	c(0) : No fixed follow-up c(1,durfollow) : Fixed follow-up	c(0) : No fixed follow-up c(1,durfollow) : Fixed follow-up	c(0) : No fixed follow-up c(1,durfollow) : Fixed follow-up
dropout (0)	c(0) : No drop out c(1,gammae,gammac)	c(0) : No drop out c(1,gammae,gammac)	c(0) : No drop out c(1,gammae,gammac)
dqol (0)	dqol	dqol	dqol

*the value in parenthesis is the default value

The description of the parameters is identical to the description of the parameters of the nsurvival function (section 4.1)

4.2 APPLICATION: two co-primary endpoints

4.2.1 Superiority trial for both endpoints

-**Exemple 5**: two-sided superiority trial with two primary endpoints

Suppose a randomized superiority clinical trial with two primary endpoints. The first endpoint is the HRQoL with two targeted dimensions and the 2-year rate of patients without HRQOL deterioration equals to 35% in the control arm and 55% in experimental arm and two-sided type I error of 2%. The second endpoint is another time to event endpoint, let consider for example the PFS, with the 2-year PFS rate equal to 30% in the control arm and 66% in experimental arm and a two-sided type I error of 3%. This study has accrual duration of 2 years, study duration of 6 years, statistical power of 80%, 3 planned analysis with boundaries

of Lan deMets O'Brien Fleming at times 1/3, 2/3 for interim analysis (look=c(3,c(1,1),c(1/3,2/3))), dropout of 10%. The default values are pe=0.5, fup=0.

```
nc1 <- ncoprimary(design=c(1,2),Survhyp1=c(1,2,0.55,0.35),
Survhyp2=c(1,2,0.66,0.30), alpha1=c(0.01,0.01), alpha2=
c(0.015,0.015),duraccrual=2,durstudy=6, power=0.80,
look=c(3,c(1,1),c(1/3,2/3)),dropout=c(1,0.05,0.05),dqol=3)
```

-**Result 5:** Results obtained for example 5 using the "*ncoprimary"* function

Results	**values**
Number of targeted HRQoL dimensions	3
Alpha	alpha1 =0.02, alpha2=0.03
Statistical power	0.80 (80%)
Ratio of patients included in arm E	0.50
Hazard ratio (HR)	HR1=0.569, HR2=0.345
Number of events required	163
Total number of subjects	210
Number of subjects (arm E, arm C)	(105, 105)

For endpoint 1, the hazard ratio is 0.569. For endpoint 2, the hazard ratio is 0.345. The number of events estimated is 163 and the number of subjects required in the study is 210 patients. The number of patients required in each arm is 105 patients.

```
[1] "#####################################################"
[1] "|   SAMPLE SIZE CALCULATION FOR TWO SURVIVAL ENDPOINTS  |"
[1] "#####################################################"
[1] "+-------------------------------------------------+"
[1] "| SURVIVAL SUPERIORITY TRIAL: TWO-SAMPLE TEST     |"
[1] "+-------------------------------------------------+"
[1] "+-----------------------------------------+"
[1] "|          TEST PARAMETERS                |"
[1] "+-----------------------------------------+"
[1] "  - 1-Sided or 2-Sided Test: 2-Sided"
[1] "  - Significance level alpha1 for endpoint 1: 0.02"
[1] "  - Significance level alpha2 for endpoint 2: 0.03"
[1] "  - Significance level alpha global        : 0.05"
[1] "  - Power: 0.8"
[1] "+-----------------------------------------+"
[1] "|          STUDY PARAMETERS               |"
[1] "+-----------------------------------------+"
[1] "  - Accrual Duration (duraccrual): 2"
[1] "  - Follow-up: No Fixed Follow-up"
[1] "  - Study Duration (durstudy): 6"
[1] "  - Assigned Fraction Experimental Arm: 0.5"
[1] "  - Drop Out: Drop Out with hazard rate (gammae,gammac)=(0.05,0.05)"
[1] "+-----------------------------------------+"
[1] "|          INTERIM ANALYSIS               |"
[1] "+-----------------------------------------+"
[1] "  - Number of Planned Analysis: 3"
[1] "  - Spacing of analysis: 0.333 0.667 1"
[1] "  - Hypothesis to be rejected: Only H0"
[1] "  - Boundary to reject (H0- / H0+): Lan deMets O'Brien Fleming / Lan deMets O'Brien Fleming"
[1] "  - Symmetric Boundary Alpha for endpoint 1: Lower=0.01/ Upper=0.01"
[1] "  - Symmetric Boundary Alpha for endpoint 2: Lower=0.015/ Upper=0.015"
[1] "+-----------------------------------------+"
[1] "|    SURVIVAL PARAMETERS FOR ENDPOINT 1   |"
[1] "+-----------------------------------------+"
[1] "  - time (time) for endpoint 1: 2"
[1] "  - Survival Control (Sc) for endpoint 1: 0.35"
[1] "  - Number of dimension quality of life: 3"
[1] "  - Survival Experimental (Se) for endpoint 1: 0.55"
[1] "  - Hazard Ratio under Alternative Hypothesis (HR) for endpoint 1: 0.569"
[1] "+-----------------------------------------+"
[1] "|    SURVIVAL PARAMETERS FOR ENDPOINT 2   |"
[1] "+-----------------------------------------+"
[1] "  - time (time) for endpoint 2: 2"
[1] "  - Survival Control (Sc) for endpoint 2: 0.3"
[1] "  - Survival Experimental (Se) for endpoint 2: 0.66"
[1] "  - Hazard Ratio under Alternative Hypothesis (HR) for endpoint 2: 0.345"
[1] "+-----------------------------------------+"
[1] "|          SAMPLE SIZE                    |"
[1] "+-----------------------------------------+"
[1] "  - Number of Events: 163"
[1] "  - Total Number of Subjects: 210"
[1] "  - Number of Subjects (Arm Exp., Arm Contr.): (105,105)"
[1] "+ Boundaries "
  Information  Events pvalue reject H0 Boundary reject H0- Boundary reject H0+
1       0.333  54.376            0.000               -5.008                5.008
2       0.667 108.752            0.001               -3.452                3.452
3       1.000 163.129            0.006               -2.758                2.758
  Analysis Time Under H0 Analysis Time Under H1
1                      0                      0
2                      0                      0
3                      0                      0
```

Fig 1 – All results obtained for example 5

4.2.2 Superiority test and non-inferiority test

- **Example 6** : one sided superiority test and non-inferiority test

Suppose a randomized two primary endpoints with one-sided superiority test for the first endpoint and a non-inferiority test for the second endpoint. The first endpoint is OS with the 3-year OS rate equals to 54% in the control arm and 72% in experimental arm and one-sided type I error of 2.5%. The second endpoint is PFS with the 2-year PFS rate equal to 54% in the

control arm and the hazard ratios under null hypothesis and alternative hypothesis are 1.10 and 1.25 in the experimental arm, respectively and one sided type I error of 2.5%. This study has accrual duration of 2 years, study duration of 5 years, statistical power of 80%,The default values are pe=0.5, fup=0, look=0,dropout=0.

```
ns5 <- nsurvival(design=c(1,1),Survhyp=c(1,3,0.72,0.54),
alpha=0.025, duraccrual=2,durstudy=5)
```

```
ns6 <-nsurvival(design=c(2),Survhyp=c(2,2,1.06,0.35,1.45),
alpha=0.025, duraccrual=2,durstudy=5)
```

-**Result 6:** Results obtained for example 5 using the "*nsurvival"* function

Results	**values**
Alpha	Alpha1 =0.025, alpha2=0.025
Statistical power	0.80 (80%)
Ratio of patients included in arm E	0.50
For endpoint 1	
Hazard ratio (HR)	HR=0.533
Number of events required	79
Number of subjects(arm E, arm C)	176 (88, 88)
For endpoint 2	
Survival rate in experimental arm	SeA=0.218
Number of events required	320
Number of subjects(arm E, arm C)	352 (176, 176)
Total number of subjects	352 (176, 176)

For the first endpoint, the number of events required is 79 and the total number of patients required is 176. For the second endpoint, the number of events required is 320 and the total number of patients required is 352. For this study, the total number of patients required is 352.

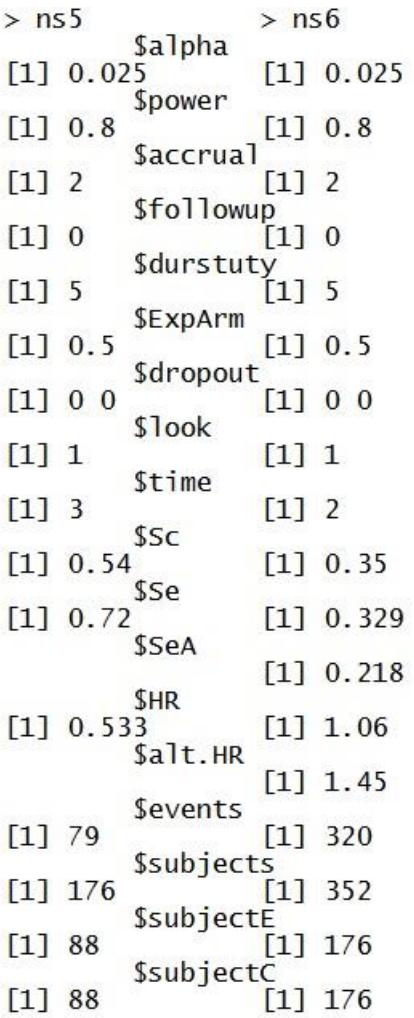


Fig 2 – All results obtained for example 6

Conclusion

The use of co-primary time-to-event endpoints has become common in clinical trials. This multiple endpoints create challenges in the evaluation of statistical power and the calculation of sample size during trial design.

The choice of a primary endpoint may be arduous; the distinction between the primary and some of the secondary endpoints may not be obvious. If the endpoints are not all affected in the same direction, the interpretation of the results may present some difficulties in deciding if

there is an appreciable difference. The use of multiple significance tests is likely to increase the chance of detecting a difference in at least one of the endpoints between two treatments.

In this paper, we outline a simple method for calculating the sample size for randomized clinical trials with two co-primary endpoints when the endpoints are the time-to-event. When designing the trial to detect effects for all of the endpoints, no adjustment is needed to control type I error. The hypothesis associated with each endpoint should be evaluated at the same significance level as is required for all the endpoints. However, type II error will increase as the number of endpoints being evaluated increases. In contrast, when designing the trial to detect an effect for at least one of the endpoints, then an adjustment is needed to control type I error. Different options permit a fixed follow-up for each patient and a potentially different drop-out rate in each arm. It is possible to plan interim analyses using different alpha-spending functions, but it is also possible to compute the timing of the different analyses under the null and alternative hypotheses. All of these options are implemented in the "*coprimary*" package. The "*coprimary*" package can be accessed and download on the CRAN website (https://CRAN.R-project.org/package=coprimary).

References

1. Multiple Co-primary Endpoints: Medical and Statistical Solutions: A Report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of AmericaTherapeutic Innovation & Regulatory Science - Walter Offen, Christy Chuang-Stein, Alex Dmitrienko, Gary Littman, Jeff Maca, Laura Meyerson, Robb Muirhead, Paul Stryszak, Alex Baddy, Kun Chen, Kati Copley-Merriman, Willard Dere, Sam Givens, David Hall, David Henry, Joseph D. Jackson, Alok Krishen, Thomas Liu, Steve Ryder, A. J. Sankoh, Julia

Wang, Chyon-Hwa Yeh, 2007. Available at: http://journals.sagepub.com/doi/abs/10.1177/009286150704100105. (Accessed: 10th August 2017)

2. Neuhäuser, M. How to deal with multiple endpoints in clinical trials. *Fundam. Clin. Pharmacol.* **20,** 515–523 (2006).
3. Chuang-Stein, C., Stryszak, P., Dmitrienko, A. & Offen, W. Challenge of multiple co-primary endpoints: a new approach. *Stat. Med.* **26,** 1181–1192 (2007).
4. Hung, H. M. J. & Wang, S.-J. Some Controversial Multiple Testing Problems in Regulatory Applications. *J. Biopharm. Stat.* **19,** 1–11 (2009).
5. Wiley: Design and Analysis of Clinical Trials: Concepts and Methodologies, 3rd Edition - Shein-Chung Chow, Jen-Pei Liu. Available at: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470887656.html.
6. Sample Size Calculations in Clinical Research, Second Edition. *CRC Press* (2007). Available at: https://www.crcpress.com/Sample-Size-Calculations-in-Clinical-Research-Second-Edition/Chow-Wang-Shao/p/book/9781584889823.
7. Schoenfeld, D. A. Sample-size formula for the proportional-hazards regression model. *Biometrics* **39,** 499–503 (1983).
8. Moffett, P. & Moore, G. The Standard of Care: Legal History and Definitions: the Bad and Good News. *West. J. Emerg. Med.* **12,** 109–112 (2011).
9. Kim, K. & Tsiatis, A. A. Study duration for clinical trials with survival response and early stopping rule. *Biometrics* **46,** 81–92 (1990).
10. Armitage, P., McPherson, C. K. & Rowe, B. C. Repeated Significance Tests on Accumulating Data. *J. R. Stat. Soc. Ser. Gen.* **132,** 235–244 (1969).
11. Burman, C.-F., Sonesson, C. & Guilbaud, O. A recycling framework for the construction of Bonferroni-based multiple tests. *Stat. Med.* **28,** 739–761 (2009).

12. Dmitrienko, A., Wiens, B. L., Tamhane, A. C. & Wang, X. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat. Med.* **26,** 2465–2478 (2007).

13. Todd, S., Whitehead, A., Stallard, N. & Whitehead, J. Interim analyses and sequential designs in phase III studies. *Br. J. Clin. Pharmacol.* **51,** 394–399 (2001).

14. Group Sequential Methods with Applications to Clinical Trials. *CRC Press* (1999). Available at: https://www.crcpress.com/Group-Sequential-Methods-with-Applications-to-Clinical-Trials/Jennison-Turnbull/p/book/9780849303166. (Accessed: 8th November 2016)

15. Todd, S., Whitehead, A., Stallard, N. & Whitehead, J. Interim analyses and sequential designs in phase III studies. *Br. J. Clin. Pharmacol.* **51,** 394–399 (2001).

16. Filleron, T., Gal, J. & Kramar, A. Designing Group Sequential Randomized Clinical Trials with Time to Event End Points Using a R Function. *Comput Methods Prog Biomed* **108,** 113–128 (2012).

17. Pocock, S. J. Interim Analyses for Randomized Clinical Trials: The Group Sequential Approach. *Biometrics* **38,** 153–162 (1982).

18. O'Brien, P. C. & Fleming, T. R. A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556 (1979).

19. Filleron, T., Gal, J. & Kramar, A. Designing group sequential randomized clinical trials with time to event end points using a R function. *Comput. Methods Programs Biomed.* **108,** 113–128 (2012).

20. Anota, A. *et al.* Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? *Qual. Life Res.* **24,** 5–18 (2015).

L'aide du package R « coprimary »

Package 'coprimary'

December 15, 2016

Type Package

Title Sample Size Calculation for Two Primary Time-to-Event Endpoints in Clinical Trials

Version 1.0

Date 2016-12-14

Author Alhousseiny PAM

Maintainer Alhousseiny PAM <alhousseiny.pam@gmail.com>

Description Computes the required number of patients for two time-to-event endpoints as primary endpoint in phase III clinical trial.

License GPL (>= 3.3.2)

URL http://www.umqvc.org/

RoxygenNote 5.0.1

Depends stats, gsDesign, digest, plyr, proto

Collate 'datacheck.R' 'nbevent.R' 'probanofix.R' 'probafix.R' 'ncoprimary.R' 'nsurvival.R'

NeedsCompilation no

Repository CRAN

Date/Publication 2016-12-15 13:52:58

R topics documented:

coprimary-package 2
datacheck 3
nbevent 4
ncoprimary 5
nsurvival 8
probafix 10
probanofix 11

Index 13

coprimary-package Sample size calculation for two primary time-to-event endpoints in phase III clinical trials

Description

The coprimary R package computes the required number of patients for two time-to-event endpoints as primary endpoint. This R package contains six functions to check the consistency of the design and computes the sample size with one or two time-to-event endpoints. Both endpoints can be one time to event endpoint, such as Overall Survival (OS), Progression Free Survival (PFS) or the health health-related quality of life (HRQoL), or two time-to-event endpoints, which could be PFS and time to HRQoL score deterioration as example.

Details

Package:	coprimary
Type:	package
Version:	1.0
Date:	2016-09-20
licence:	GPL(>=3.3.2)

Author(s)

Alhousseiny PAM
Maintainer: <alhousseiny.pam@gmail.com>

References

Legault, Claudine. Analyzing multiple endpoints with a two-stage group sequential design in clini-cal trials. Diss. University of North Carolina at Chapel Hill, 1991.

Chow, Shein-Chung, Hansheng Wang, and Jun Shao. Sample size calculations in clinical research. CRC press, 2007.

Filleron, T., Gal, J., & Kramar, A. (2012). Designing group sequential randomized clinical trials with time to event end points using a R function. Computer methods and programs in biomedicine, 108(1), 113-128.

Examples

```
################################################################################
############ Design superiority: two-sided with two co-primary endpoints ############
################################################################################
## - For endpoint 1: 2 target variables for the health related quality of life with 3-year
## rate without HRQoL deterioration Se=0.75 and Sc=0.67, alpha1=c(0.01,0.01)
```

```
## - For endpoint 2: 4-year survival rates Se=0.86 and Sc=0.80, alpha2=c(0.015,0.015)
## - with accrual duration of 3 years, study duration of 6 years, power=0.90,
## look=c(2,c(1,1),0.5), and default values i.e  pe=0.5, fup=0, dropout=0

nc1 <- ncoprimary(design=c(1,2),Survhyp1=c(1,5,0.75,0.67),Survhyp2=c(1,5,0.86,0.80),
alpha1=c(0.01,0.01),alpha2=c(0.015,0.015),duraccrual=3,durstudy=6,power=0.90,
look=c(2,c(1,1),0.5),dqol=2)


##############################################################################################
############ Design superiority: one-sided with two co-primary endpoints ############
##############################################################################################
## - For endpoint 1: 2-year hazard ratio hype=0.86 and Sc=0.62, alpha1=0.05
## - For endpoint 2: 3-year survival rates hype=0.81 and Sc=0.57, alpha2=0.05
## - with accrual duration of 2 years, study duration of 10 years and default values i.e
## power=0.90, pe=0.5, look=1, fup=0, dropout=0, dqol=0


nc2 <- ncoprimary(design=c(1,1),Survhyp1=c(2,2,0.86,0.62),Survhyp2=c(2,3,0.81,0.57),
alpha1=0.05,alpha2=0.05,duraccrual=2,durstudy=10)

##############################################################################################
###############  Design non-inferiority with one primary endpoint ##################
##############################################################################################
## 5-year rate without HRQoL deterioration are equal under the alternative hypothesis,
## i.e Se=0.60 and Sc=SeA=0.70, with alpha=0.05, accrual duration of 4 years,study duration
## of 8 years, two interim analysis after the occurence 1/3 and 2/3 of the events, 3 target
## variables for the health related quality of life and default values i.e power=0.80, pe=0.5,
## fup=0, dropout=0

ns <- nsurvival(design=c(2),Survhyp=c(1,5,0.60,0.70, 0.70),alpha=0.05,duraccrual=4,
durstudy=8,look=c(3,c(1,1),c(1/3,2/3)), dqol=3)
```

---

datacheck	check the consistency of the parameters

Description

this function check the parameters required to calcul the sample size for nsurvival and ncoprimary functions. datacheck is a simple utility for carrying out parameter checks and reporting on problems or errors.

Usage

```
datacheck(design,Survhyp,pe,alfa,beta,duraccrual,durstudy,look,followup,dropout)
```

Arguments

design	Superiority=c(1,sided)[with sided=1 if 1-sided and 2 if 2-sided]; Non inferior-ity=c(2); Equivalence=c(3)
Survhyp	For Superiority=c(thyp,t,hype,Sc); for Non inferiority=c(thyp,t,hype,Sc,hypeA); for Equivalence=c(t,delta,Sc): parameters at time t if thyp=1 then hype is survival rate in experimental arm under the null hypothesis and hypeA is the survival rate in the experimental arm under the alternative hypothesis; if thyp=2 then hype is the hazard ratio under the null hypothesis and hypeA is the hazard ratio under the alternative hypothesis; Sc is the survival rate in the control arm; delta is the log hazard ratio equivalence margin. When endpoint is HRQoL, the survival rate is replaced by the rate of patients without the HRQoL deterioration.
pe	Proportion (ratio) of patients assigned to the experimental arm (with 0<pe<1)
alfa	Type I error, for Non inferiority, Equivalence and 1-sided superiority, alfa is a vector of length one. For 2-sided superiority, alfa is a vector to length two c(alpha.low, alpha.up).
beta	Probability of a type II error.
duraccrual	Accrual duration, expressed in number of days, months or years
durstudy	Study duration, expressed in number of days, months or years
look	The number interim analyses, c(1) for one final analysis; c(nb, bound, timing) for at least one interim analyses with bound=c(bound.eff,bound.fut):1-sided or bound=c(bound.low,bound.up):2-sided. nb the number of planned looks, bound.eff and bound.fut corresponds to the type of boundaries used for efficacy (i.e. reject H0) and futility (i.e. reject H1). bound.fut=0: No futility monitor-ing, 1: Lan deMets O.Brien Fleming, 2: Lan deMets Pocock. bound.low and bound.up the type of lower and upper boundaries used (1: Lan deMets O.Brien Fleming, 2: Lan deMets Pocock, 3: O.Brien Fleming, 4: Pocock). Default value = 1.
followup	Follow-up information, No fixed:c(0) (follow-up until the end of study); Fixed:c(1, durfollow) with durfollow is the duration of follow-up
dropout	Drop out information, No drop out:c(0); Drop out:c(1,gammae,gammac) with gammae the hazard drop out rates in experimental arm and control arm respec-tively.

Details

the datacheck function performs consistency checks on the arguments

---

nbevent Number of events estimates

Description

To determine the sample size N in clinical trials with time to event endpoint, it is necessary to proceed in two steps. In the first step, the numbers of events that need to be observed (e) are computed. In the second step, we determine the number of patients necessary to observe the number of events required. This function computes the number of event for one-time-to event.

Usage

nbevent(hypsurv,pe,alfa,beta,design)

Arguments

hypsurv	For Superiority=c(Sc,Se); for Non inferiority=c(Sc,Se,SeA); for Equivalence=c(Sc, Se), with Sc is survival rate in the control arm; Se is survival rate in experimen-tal arm; SeA is the survival rate in the experimental arm under the alternative hypothesis.
pe	Proportion (ratio) of patients assigned to the experimental arm (with 0<pe<1).
alfa	Type I error, for Non inferiority, Equivalence and 1-sided superiority, alfa is a vector of length one. For 2-sided superiority, alfa is a vector to length two c(alpha.low, alpha.up).
beta	Probability of a type II error.
design	Superiority=c(1,sided)[with sided=1 if 1-sided and 2 if 2-sided]; Non inferior-ity=c(2); Equivalence=c(1,1)

Details

The nbevent function computes the required number of events to determine the number of patients.

Value

E: Number of events

h: Hazard Ratio under null hypothesis(HR=log(Se)/log(Sc))

h.alt: Hazard Ratio under alternative hypothesis (h.alt=log(SeA)/log(Sc))

References

Chow, S. C., Shao, J., Wang, H. (2003). Sample Size Calculation in Clinical Research. New York: Marcel Dekker.

Schoenfeld. Sample-size formula for the proportional-hazards regression model. Biometrics. 1983 39<499>503.

---

ncoprimary	Sample size calculation in clinical trials with two co-primary time-to-event endpoints

---

Description

ncoprimary() is used to calculate the sample size for phase III clinical trial with two co-primary endpoints to assess the efficacy of treatment between two groups.

Usage

```
ncoprimary(design,Survhyp1,Survhyp2,alpha1,alpha2,duraccrual,durstudy,power,pe,
look,fup,dropout,dqol)
```

Arguments

design	Superiority=c(1,sided)[with sided=1 if 1-sided and 2 if 2-sided]; Non inferior-ity=c(2); Equivalence=c(3)
Survhyp1	For Superiority=c(thyp,t,hype,Sc); for Non inferiority=c(thyp,t,hype,Sc,hypeA); for Equivalence=c(t,delta,Sc): parameters at time t for the first endpoint, if thyp=1 then hype is survival rate or the rate of patients without HRQoL deteri-oration in experimental arm under the null hypothesis and hypeA is the survival rate in the experimental arm under the alternative hypothesis; if thyp=2 then hype is the hazard ratio under the null hypothesis and hypeA is the hazard ratio under the alternative hypothesis; Sc is survival rate in the control arm; delta is the log hazard ratio equivalence margin. When endpoint is HRQoL, the survival rate is replaced by the rate of patients without HRQoL deterioration.
Survhyp2	For Superiority=c(thyp,t,hype,Sc); for Non inferiority=c(thyp,t,hype,Sc,hypeA); for Equivalence=c(t,delta,Sc): parameters at time t for the second endpoint if thyp=1 then hype is survival rate in experimental arm under the null hypothesis and hypeA is the survival rate in the experimental arm under the alternative hy-pothesis; if thyp=2 then hype is the hazard ratio under the null hypothesis and hypeA is the hazard ratio under the alternative hypothesis; Sc is survival rate in the control arm; delta is the log hazard ratio equivalence margin. When endpoint is HRQoL, the survival rate is replaced by the rate of patients without HRQoL deterioration.
alpha1	Type I error assigned to the first endpoint, for Non inferiority, Equivalence and 1-sided superiority is a vector of length one. For 2-sided superiority is a vector to length two c(alpha.low, alpha.up).
alpha2	Type I error assigned to the second endpoint, for Non inferiority, Equivalence and 1-sided superiority is a vector of length one. For 2-sided superiority is a vector to length two c(alpha.low, alpha.up).
duraccrual	Accrual duration, expressed in number of days, months or years
durstudy	Study duration, expressed in number of days, months or years
power	1-Probability of a type II error. Default value = 0.80.
pe	Proportion (ratio) of patients assigned to the experimental arm (with 0<pe<1). Default value = 0.50.
look	

The number of interim analyses, c(1) for one final analysis; c(nb, bound, tim-ing) for at least one interim analyses with bound=c(bound.eff,bound.fut):1-sided or bound=c(bound.low,bound.up):2-sided. nb the number of planned looks, bound.eff and bound.fut corresponds to the type of boundaries used for efficacy (i.e. reject H0) and futility (i.e. reject H1). bound.fut=0: No futility monitor-ing, 1: Lan deMets O.Brien Fleming, 2: Lan deMets Pocock. bound.low and bound.up the type of lower and upper boundaries used (1: Lan deMets O.Brien Fleming, 2: Lan deMets Pocock, 3: O.Brien Fleming, 4: Pocock). Default value = 1.

fup	Follow-up information, No fixed:c(0) (follow-up until the end of study); Fixed:c(1, durfollow) with durfollow is the duration of follow-up. Default value = 0.
dropout	Drop out information, No drop out=c(0); Drop out=c(1,gammae,gammac) with gammae and gammac are the hazard drop out rates in experimental arm and control arm respectively. Default value = 0.
dqol	number of targeted dimensions for the health related quality of life. Default value = 0.

Details

The ncoprimary function computes the sample size for two primary endpoints. Both endpoints can be one time to event endpoint and health related quality of life (HRQoL) or two times to event endpoints.

Value

Event: number of events estimated

Total: number of patients

Ne: number for experimental arm for each endpoint

Nc: number for control arm for each endpoint

HR: Hazard ratio for each endpoint

Examples

```
#############################################################################################
############ Design superiority:one-sided with two co-primary endpoints ############
#############################################################################################
## - For endpoint 1: 3-year survival rates Se=0.75 and Sc=0.65, alpha1=0.02
## - For endpoint 2: 4-year survival rates Se=0.70 and Sc=0.59, alpha2=0.03
## with accrual duration of 2 years, study duration of 4 years and default values i.e
## power=0.80, pe=0.5, look=1, fup=0, dropout=0, dqol=0

nc1 <- ncoprimary(design=c(1,1),Survhyp1=c(1,3,0.75,0.65),Survhyp2=c(1,4,0.70,0.59),
alpha1=0.02,alpha2=0.03,duraccrual=2,durstudy=4)


#############################################################################################
            ############ Design superiority:two-sided with two co-primary endpoints ############
#############################################################################################
## - For endpoint 1: 2 target variables for the health related quality of life with 3-year
## rate without HRQoL deterioration Se=0.75 and Sc=0.67, alpha1=c(0.01,0.01)
## - For endpoint 2: 4-year survival rates Se=0.86 and Sc=0.80, alpha2=c(0.015,0.015)
## with accrual duration of 3 years, study duration of 6 years, power=0.90, look=c(2,c(1,1),0.5),
## and default values i.e  pe=0.5, fup=0, dropout=0

nc2 <- ncoprimary(design=c(1,2),Survhyp1=c(1,5,0.75,0.67),Survhyp2=c(1,5,0.86,0.80),
alpha1=c(0.01,0.01),alpha2=c(0.015,0.015),duraccrual=3,durstudy=6, power=0.90,
look=c(2,c(1,1),0.5),dqol=2)
```

```
##############################################################################################
##############Design non-inferiority with two co-primary endpoints ###############
##############################################################################################
## - For endpoint 1: 3-year survival rates Se=0.75 and Sc=SeA=0.75, alpha1=0.01
## - For endpoint 2: 4-year survival rates Se=0.67 and Sc=SeA=0.80, alpha2=0.04
## with accrual duration of 2 years, study duration of 6 years, power=0.95, pe=0.60 and
## default values i.e look=1, fup=0, dropout=0, dqol=0

nc3 <- ncoprimary(design=c(2),Survhyp1=c(1,4,0.65,0.75,0.75),Survhyp2=c(1,5,0.67,0.80,0.80),
alpha1=0.01,alpha2=0.04,duraccrual=2,durstudy=6,power=0.95,pe=0.60)

##############################################################################################
################       Design superiority with two co-primary endpoints #################
##############################################################################################

## - For endpoint 1: 2-year survival rate Sc=0.65 and log hazard equivalence margin delta=0.15
## and alpha1=0.025
## - For endpoint 2: 1-year survival rate Sc=0.70 and log hazard equivalence margin delta=0.10
## and alpha2=0.025
## with accrual duration of 3 years, study duration of 5 years, drop out hazard rate of 0.025
## per arm and default values i.e power=0.80, pe=0.5, look=1, fup=0, dqol=0

nc4 <- ncoprimary(design=c(3),Survhyp1=c(2,0.15,0.65),Survhyp2=c(1,0.10,0.70),alpha1=0.025,
alpha2=0.025,duraccrual=3,durstudy=5,dropout=c(1,0.025,0.025))
```

---

nsurvival	Sample size calculation in clinical trials with one primary survival endpoint

---

Description

nsurvival() is used to determine the sample size for one time to event endpoint, such as Overall Survival (OS), Progression Free Survival or the health related quality of life (HRQoL). If it is HRQoL, several HRQoL dimension can be considered.

Usage

```
nsurvival(design,Survhyp,alpha,duraccrual,durstudy,power,pe,look,fup,dropout,dqol)
```

Arguments

design	Superiority=c(1,sided)[with sided=1 if 1-sided and 2 if 2-sided]; Non inferior-ity=c(2); Equivalence=c(3)

Survhyp	For Superiority=c(thyp,t,hype,Sc); for Non inferiority=c(thyp,t,hype,Sc,hypeA); for Equivalence=c(t,delta,Sc): parameters at time t if thyp=1 then hype is survival rate in experimental arm under the null hypothesis and hypeA is the survival rate in the experimental arm under the alternative hypothesis; if thyp=2 then hype is the hazard ratio under the null hypothesis and hypeA is the hazard ratio under the alternative hypothesis; Sc is survival rate in the control arm; delta is the log hazard ratio equivalence margin. When endpoint is HRQoL, the survival rate is replaced by the rate of patients without HRQoL deterioration.
alpha	Type I error, for Non inferiority, Equivalence and 1-sided superiority is a vector of length one. For 2-sided superiority is a vector to length two c(alpha.low, alpha.up).
duraccrual	Accrual duration, expressed in number of days, months or years
durstudy	Study duration, expressed in number of days, months or years
power	1- Probability of a type II error. Default value=0.80.
pe	Proportion (ratio) of patients assigned to the experimental arm (with 0<pe<1). Default value = 0.5.
look	The number of interim analyses, c(1) for one final analysis; c(nb, bound, timing) for at least one interim analyses with bound=c(bound.eff,bound.fut):1-sided or bound=c(bound.lown,bound.up):2-sided. nb the number of planned looks, bound.eff and bound.fut corresponds to the type of boundaries used for efficacy (i.e. reject H0) and futility (i.e. reject H1). bound.fut=0: No futility monitor-ing, 1: Lan deMets O.Brien Fleming, 2: Lan deMets Pocock. bound.low and bound.up the type of lower and upper boundaries used (1: Lan deMets O.Brien Fleming, 2: Lan deMets Pocock, 3: O.Brien Fleming, 4: Pocock). Default value = 1.
fup	Follow-up information, No fixed:c(0) (follow-up until the end of study); Fixed:c(1, durfollow) with durfollow is the duration of follow-up. Default value = 0.
dropout	Drop out information, No drop out=c(0); Drop out=c(1,gammae,gammac) with gammae the hazard drop out rates in experimental arm and control arm respec-tively. Default value = 0.
dqol	number of targeted dimensions for the health related quality of life. Default value = 0.

Details

The nsurvival function computes the sample size for one time to event endpoint, such as OS, PFS or HRQoL. HRQoL has become increasingly important in clinical trials over the past two decades.

Value

Event: number of events estimated

Total: number of patients

Ne: number for experimental arm for each endpoint

Nc: number for control arm for each endpoint

HR: Hazard ratio for each endpoint

Examples

```
############################################################
###############Design superiority:one-sided ###############
############################################################
## 7-year survival rates Se=0.57 and Sc=0.53, alpha=0.05, accrual duration of 4 years,
## study duration of 8 years and default values i.e power=0.80, pe=0.5, look=1, fup=0,
## dropout=0, dqol=0

ns1 <- nsurvival(design=c(1,1),Survhyp=c(1,7,0.57,0.53),alpha=0.05,duraccrual=4,durstudy=8)

############################################################
############### Design superiority:two-sided ###############
############################################################
## 5-year rate without HRQoL deterioration Se=0.75 and Sc=0.65, alpha=c(0.04,0.01), accrual
## duration of 2 years, study duration of 6 years, power=0.90, pe=0.55, follow-up 5 years,
## 3 target variables for health related quality of life and default values i.e look=1, dropout=0

ns2 <- nsurvival(design=c(1,2),Survhyp=c(1,5,0.75,0.65),alpha=c(0.04,0.01),duraccrual=2,
durstudy=6,power=0.90,pe=0.55,fup=c(1,5),dqol=3)

############################################################
###############  Design non-inferiority ##################
############################################################
## 5-year survival rates are equal under the alternative hypothesis, i.e Se=0.60 and Sc=SeA=0.70,
## with alpha=0.05, accrual duration of 4 years, study duration of 8 years, two interim analysis
## after the occurence 1/3 and 2/3 of the events and default values i.e power=0.80, pe=0.5, fup=0,
## dropout=0, dqol=0

ns3 <- nsurvival(design=c(2),Survhyp=c(1,5,0.60,0.70, 0.70),alpha=0.05,duraccrual=4,
durstudy=8,look=c(3,c(1,1),c(1/3,2/3)))

############################################################
###############    Design superiority       ################
############################################################
## 3-year rate without HRQoL deterioration Sc=0.80 and log hazard equivalence margin delta=0.1
## with alpha=0.10, accrual duration of 3 years, study duration of 5 years, drop out hazard rate
## of 0.05 per arm, 2 target variables for health related quality of life and default values i.e
## power=0.80, pe=0.5, look=1, fup=0

ns4 <- nsurvival(design=c(3),Survhyp=c(3,0.10,0.80),alpha=0.10,duraccrual=3,durstudy=5,
dropout=c(1,0.05,0.05),dqol=2)
```

---

probafix Probabitility of event when the follow-up is fixed

Description

In a fixed follow-up design, each subject can only be followed during a fixed period (Tf < infinite) and then goes off study. Using similar reasoning to K. Kim and A.A. Tsiatis (1990), it is easy to compute the probability.

Usage

probafix(surv,time,duraccrual,durfollow,limit,gamma)

Arguments

surv	Survival estimates
time	Time estimate
duraccrual	Accrual duration, expressed in t time units
durfollow	Follow-up duration
limit	Time limit to estimate the survival probability
gamma	the probability of observing an event by time t

Details

The probafix function estimates the probability of event when the follow-up is fixed.

Value

probafix: event probability at time limit

References

K. Kim, A.A. Tsiatis, Study duration for clinical trials with survival response and early stopping rule, Biometrics 46 (1990) 81-92

---

probanofix Probabitility of event when the follow-up is no fixed

---

Description

In the design with variable follow-up, each subject is followed until the end of the study (Tf = infinite), i.e. subjects who are enrolled at the beginning of the enrolment phase are followed for a longer time than subjects who are enrolled later. When there are no drop outs (i.e. = (0,0), the probability of failure in each arm can be directly estimated using the formulation proposed by K Kim and A.A. Tsiatis (1990).

Usage

probanofix(surv,time,duraccrual,limit,gamma)

Arguments

surv	Survival estimates
time	Time estimate
duraccrual	Accrual duration, expressed in t time units
limit	Time limit to estimate the survival probability
gamma	the probability of observing an event by time t

Details

The probanofix function estimates the probability of event when the follow-up is no fixed

Value

probanofix: event probability at time limit

References

K. Kim, A.A. Tsiatis, Study duration for clinical trials with survival response and early stopping rule, Biometrics 46 (1990) 81-92

Index

Topic clinical trial
coprimary-package, 2
Topic co-primary
coprimary-package, 2
Topic multiple endpoints
coprimary-package, 2
Topic sample size
coprimary-package, 2

coprimary (coprimary-package), 2
coprimary-package, 2

datacheck, 3

nbevent, 4
ncoprimary, 5
nsurvival, 8

probafix, 10
probanofix, 11

IX. Discussion

La variabilité des définitions des critères de jugement dans les études cliniques

Dans les essais cliniques randomisés (ECR) en cancérologie, les critères de survie sont couramment utilisés à la place de la survie globale dans les essais de phase III. Leur développement est fortement influencé par la nécessité de réduire la durée des essais cliniques, le coût et le nombre de patients à recruter. Alors que ces critères de survie sont fréquemment utilisés en tant que critère de jugement, ils sont souvent mal définis, et lorsqu'ils le sont, ces définitions peuvent être très variables(Mathoulin-Pelissier et al., « Survival End Point Reporting in Randomized Cancer Clinical Trials »). Pour étudier cette variabilité, nous avons analysé 9 essais cliniques randomisés sur le cancer du pancréas en situation adjuvante et métastatique de phase II et phase III.

Nous avons constaté que beaucoup d'évènements nécessaires aux définitions des critères recommandés par le consensus du DATECAN-1 sont manquants ou mal définis, surtout le second cancer, le type de progression et les causes de décès ne sont toujours pas bien spécifiées. Le plus souvent, la progression locale et la progression régionale sont associées dans une même variable progression locorégionale ou une variable progression est indiquée par oui ou non, sans savoir s'il s'agit d'une progression locale, d'une progression régionale, d'une progression métastatique ou d'une progression distante.

Cela pose des problèmes pour reconstituer les définitions des critères de DATECAN-1 étant donné que les types de progressions sont bien différents et les causes de décès bien spécifiées. De ce fait il est difficile de reconstituer totalement les critères recommandés par le consensus DATECAN-1 tels qu'ils sont définis.

Mathoulin et al (Mathoulin-Pelissier et al., « Survival End Point Reporting in Randomized Cancer Clinical Trials ») avaient fait le même constat sur 125 articles avec 267 critères de survie, seulement 33 critères (12%) ont été clairement définis et contiennent les 7 point clés nécessaires à la définition d'un critère de survie. Leurs analyses étaient basées sur tout type de cancer, alors que notre étude concerne uniquement le cancer du pancréas, mais nos résultats restent similaires.

Beaucoup de critères de survie sont manquants et mal défini dans les articles publiés. Il parait donc nécessaire de promouvoir l'utilisation des définitions DATECAN-1 des critères de jugement dans les essais cliniques. Depuis leur publication en 2014, de nombreux essais cliniques nouvellement conçus ont utilisé les critères DATECAN. De plus, déjà de nombreuses citations de ces critères sont référencées dans Pubmed (19 citations) et leur utilisation devrait croitre dans les années futures.

Impact des définitions sur les résultats et les conclusions

En absence de plusieurs événements pour la constitution des critères issus de recommandation sur les bases analysées, les résultats trouvés ne permettent complètement de comparer les différentes définitions d'un même critère. La majorité des définitions ne reposent sur aucune recommandation consensuelle. De même les événements sont mesurés de différentes manières (radiologie, WHO, RECIST, …).

Il est primordial et nécessaire d'appliquer les critères issus des recommandations de DATECAN-1 sur les nouveaux essais cliniques randomisés pour uniformiser les essais et faire correctement les comparaisons entre les résultats publiés. Par la suite, nous pourrons plus

facilement étudier les propriétés des critères de survie intermédiaire en tant que critères de substitution à la survie globale.

La validation des critères de substitution

En cancérologie, le critère principal le plus souvent utilisé dans les essais de phase III est la survie globale ou la survie sans progression (Pazdur, « Endpoints for Assessing Drug Activity in Clinical Trials »). L'utilisation des critères intermédiaires comme critères de substitution ne repose pas en règle générale sur une stratégie argumentée de validation. Cette validation nécessite de déployer des efforts très importants qui doivent être renouvelés pour chaque situation thérapeutique.

Malgré la variabilité de définition des critères et l'absence de plusieurs événements nécessaires à la définition des critères de DATECAN-1, certains critères ont été reconstruits partiellement pour évaluer leurs capacités de substitution à la SG.

Il existe différentes méthodes de validation de critère de substitution par la méta-analyse. Nous avons choisi l'approche par régression linéaire pondérée par la taille des essais. L'association entre le critère de substitution et le critère final a été mesuré par le coefficient R^2.

Nous avons rassemblé six études en situation métastatique avec au total de 582 patients pour évaluer les capacités de substitutions des critères tels que la survie sans progression, le temps jusqu'à la progression et le temps jusqu'à détérioration du statut OMS. Les résultats ont montré que ces critères étaient fortement associés à la survie globale avec des coefficients R^2

égaux à R^2_1=0.98 pour la survie sans progression, R^2_2=0.96 pour le temps jusqu'à détérioration et R^2_3=0.92 pour le temps jusqu'à détérioration du statut OMS.

Il existe d'autres approches pour la validation d'un critère de substitution, comme par exemple l'approche de Burzykowski et les modèles conjoints à fragilités (Burzykowski et al., « Validation of Surrogate End Points in Multiple Randomized Clinical Trials with Failure Time End Points »). Ces approches sont un peu plus complexes mais modélise parfaitement l'association entre les deux critères. Il serait intéressant d'étudier le caractère substitutif de ces critères intermédiaires en utilisant ces différentes approches afin de comparer ces résultats avec ce que nous avons obtenus ici. De plus, les résultats que nous avons obtenus devront être validés dans une validation externe.

Design alternatif avec les co-critères de jugement

Les critères de jugement principal unique ou multiple est le paramètre sur lequel dépend la conclusion de l'essai. Ce critère doit donc être défini très précisément et doit correspondre au critère clinique le plus pertinent vis-à-vis de l'objectif de l'étude.

La combinaison de plusieurs critères principaux permet d'augmenter la puissance statistique de l'étude. Dans ce sens une bonne alternative est de combiner un critère centré sur le patient et un critère centré sur la tumeur par exemple la qualité de vie relative à la santé et la survie sans progression. La qualité de vie relative à la santé est un critère de jugement alternatif pertinent et disponible qui assure l'intérêt du traitement pour le patient et le système de santé.

Calcul du nombre de sujets nécessaires avec des critères conjoints de temps jusqu'à événement

L'utilisation des co-critères de jugement de type temps jusqu'à événement est une alternative pour évaluer l'efficacité des traitements. Cependant, peu de recommandations et de moyen statistique ont été mis en œuvre pour la détermination du nombre de sujets nécessaires pour les co-critères de jugement, ce qui compromet leur utilisation.

Pour déterminer la taille de l'échantillon plusieurs paramètres doivent être pris en compte: la durée des inclusions, la durée de suivi ainsi que le taux de perdus de vues, le taux de survie ou le coefficient du Hasard Ratio sous l'hypothèse alternative. Le calcul du nombre de sujets nécessaires avec deux critères conjoints s'effectue en deux étapes pour détecter les effets sur les deux critères. Pour chaque critère on calcule le nombre d'événement à observer et le nombre de patients à inclure par le test de log rank.

Dans cette optique, nous avons développé un package R nommé « coprimary » qui permet de planifier des essais de supériorité, non-infériorité ou équivalence avec des analyses intermédiaires (efficacité et/ ou futilité). Ce package permet de déterminer le nombre de sujet nécessaires (NSN) avec un ou deux critères de jugement de type temps jusqu'à événement. Il permet aussi vérifier la cohérence du design d'un essai clinique comparant deux lignes de traitement.

La package peut être développé d'avantage pour d'autres types de design. Notamment des designs qui comparent entre plusieurs groupes de population supérieure à deux.

Dans les perspectives, mon grand souhaite est d'implémenter l'ensemble des méthodes d'ajustement de risque d'erreurs pour enrichir davantage le package.

X. Conclusion

Mes travaux de thèse ont permis de mettre en évidence la variabilité des critères de survie utilisés dans les essais cliniques randomisés dans le cancer du pancréas. De plus, j'ai constaté que certaines définitions sont confondues avec d'autres, alors qu'elles sont complétement différentes par exemple entre la durée de la réponse et le temps jusqu'à progression. Certains évènements étaient également parfois regroupés avec d'autres, alors qu'il existe une différence importante entre eux, par exemple entre une progression locale et une progression distante, souvent indifférencié dans les essais que nous avions analysé.

Lorsque les évènements ne sont pas clairement évalués, il est difficile de choisir une définition de survie correspondant aux évènements étudiés. Cela induit à une variabilité des définitions entre les études. Il est impératif d'utiliser des définitions standard comme recommandées par le consensus DATECAN-1 pour se diriger dans le même sens.

Malgré l'absence de plusieurs événements nécessaires à la définition des critères issus de recommandation de DATECAN-1, certains critères ont été reconstruits partiellement pour évaluer leurs capacités de substitution à la SG.

Pour le cancer du pancréas en situation avancée, nous avons montré que les critères tels que la survie sans progression, le temps jusqu'à la progression et le temps jusqu'à détérioration du statut OMS sont potentiellement des critères substitutifs à la SG.

En combinant deux critères de survie on augmente nos chances d'obtenir une différence statistiquement significative. Ainsi la QdV constitue un critère de jugement pour évaluer un bénéfice clinique direct pour le patient. Un design alternatif est d'associer un critère de

jugement intermédiaire composite comme SSP avec la QdV en tant que co-critères de jugement principaux.

Lors de la conception d'un essai clinique avec des co-critères de jugements principaux multiples, il est essentiel d'estimer le nombre de sujets nécessaires pour pouvoir planifier correctement l'essai et ainsi évaluer l'efficacité des traitements. Le package R «coprimary» que j'ai développé lors de mon doctorat permet de planifier des essais de supériorité et non-infériorité ou équivalence avec des analyses intermédiaires (efficacité et/ ou futilité) et aussi permet d'ajuster facilement les risques de premier espèce attribué à chaque critère de jugement. C'est le premier package qui permet le calcul du NSN selon le type de design avec des critères conjoints de temps jusqu'à événement. Il est disponible en libre accès sur le site CRAN pour une large utilisation afin d'améliorer la palification des essais cliniques avec des critères de survie de temps jusqu'à événement.

XI. Les références

A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/22415870. Consulté le 10 août 2017.

Abel, Ulrich R., et al. « Some Issues of Sample Size Calculation for Time-to-Event Endpoints Using the Freedman and Schoenfeld Formulas ». Journal of Biopharmaceutical Statistics, vol. 25, n° 6, novembre 2015, p. 1285‑311. Taylor and Francis+NEJM, doi:10.1080/10543406.2014.1000546.

Anota, Amélie, Guillaume Mouillet, et al. « Sequential FOLFIRI.3 + Gemcitabine Improves Health-Related Quality of Life Deterioration-Free Survival of Patients with Metastatic Pancreatic Adenocarcinoma: A Randomized Phase II Trial ». PloS One, vol. 10, n° 5, 2015, p. e0125350. PubMed, doi:10.1371/journal.pone.0125350.

Anota, Amélie, Zeinab Hamidou, et al. « Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization? » Quality of Life Research, vol. 24, 2015, p. 5‑18. PubMed Central, doi:10.1007/s11136-013-0583-6.

Beitz, J., et al. « Quality-of-Life End Points in Cancer Clinical Trials: The U.S. Food and Drug Administration Perspective ». Journal of the National Cancer Institute. Monographs, n° 20, 1996, p. 7‑9.

Bellera, C. A., et al. « Guidelines for time-to-event end point definitions in sarcomas and gastrointestinal stromal tumors (GIST) trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials) ». Annals of Oncology, vol. 26, n° 5, mai 2015, p. 865‑72. academic.oup.com, doi:10.1093/annonc/mdu360.

Bergman, B., et al. « The EORTC QLQ-LC13: A Modular Supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for Use in Lung Cancer Clinical Trials. EORTC Study Group on Quality of Life ». European Journal of Cancer (Oxford, England: 1990), vol. 30A, n° 5, 1994, p. 635‑42.

Bland, J. Martin, et Douglas G. Altman. « The logrank test ». BMJ : British Medical Journal, vol. 328, n° 7447, mai 2004, p. 1073.

Blumenthal, Gideon M., et al. « Oncology Drug Approvals: Evaluating Endpoints and Evidence in an Era of Breakthrough Therapies ». The Oncologist, vol. 22, n° 7, juillet 2017, p. 762‑67. PubMed Central, doi:10.1634/theoncologist.2017-0152.

Bonnetain, F. « Qualité de vie relative à la santé et critères de jugement en cancérologie. » Cancer Radiother, 2010, p. 515‑18.

Bonnetain, Franck, Bert Bonsing, et al. « Guidelines for time-to-event end-point definitions in trials for pancreatic cancer. Results of the DATECAN initiative (Definition for the Assessment of Time-to-event End-points in CANcer trials) ». European Journal of Cancer, vol. 50, n° 17, novembre 2014, p. 2983‑93. ScienceDirect, doi:10.1016/j.ejca.2014.07.011.

Bonnetain, Franck, Frédéric Fiteni, et al. « Statistical Challenges in the Analysis of Health-Related Quality of Life in Cancer Clinical Trials ». Journal of Clinical Oncology, vol. 34, n° 16, juin 2016, p. 1953‑56. ascopubs.org (Atypon), doi:10.1200/JCO.2014.56.7974.

Bonnetain, Franck, Laetitia Dahan, et al. « Time until Definitive Quality of Life Score Deterioration as a Means of Longitudinal Analysis for Treatment Trials in Patients with Metastatic Pancreatic Adenocarcinoma ». European Journal of Cancer (Oxford, England: 1990), vol. 46, n° 15, octobre 2010, p. 2753‑62. PubMed, doi:10.1016/j.ejca.2010.07.023.

Bottomley, Andrew, et al. « Analysing Data from Patient-Reported Outcome and Quality of Life Endpoints for Cancer Clinical Trials: A Start in Setting International Standards ». The Lancet. Oncology, vol. 17, n° 11, novembre 2016, p. e510‑14. PubMed, doi:10.1016/S1470-2045(16)30510-1.

Burzykowski, Tomasz, et al. « Validation of Surrogate End Points in Multiple Randomized Clinical Trials with Failure Time End Points ». Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 50, nº 4, janvier 2001, p. 405‑22. Wiley Online Library, doi:10.1111/1467-9876.00244.

---. « Validation of Surrogate End Points in Multiple Randomized Clinical Trials with Failure Time End Points ». Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 50, nº 4, janvier 2001, p. 405‑22. Wiley Online Library, doi:10.1111/1467-9876.00244.

Buyse, M., et al. « The Validation of Surrogate Endpoints in Meta-Analyses of Randomized Experiments ». Biostatistics, vol. 1, nº 1, janvier 2000, p. 49‑67. biostatistics.oxfordjournals.org, doi:10.1093/biostatistics/1.1.49.

Calvert, Melanie, et al. « Reporting of Patient-Reported Outcomes in Randomized Trials: The CONSORT PRO Extension ». JAMA, vol. 309, nº 8, février 2013, p. 814‑22. PubMed, doi:10.1001/jama.2013.879.

Chuang-Stein, Christy, et al. « Challenge of Multiple Co-Primary Endpoints: A New Approach ». Statistics in Medicine, vol. 26, nº 6, mars 2007, p. 1181‑92. Wiley Online Library, doi:10.1002/sim.2604.

Cox, Christopher. « Fieller's Theorem, the Likelihood and the Delta Method ». Biometrics, vol. 46, nº 3, 1990, p. 709‑18. JSTOR, doi:10.2307/2532090.

Dmitrienko, Alex, et al. « Tree-Structured Gatekeeping Tests in Clinical Trials with Hierarchically Ordered Multiple Objectives ». Statistics in Medicine, vol. 26, nº 12, mai 2007, p. 2465‑78. PubMed, doi:10.1002/sim.2716.

Eisenhauer, E. A., et al. « New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1.1) ». European Journal of Cancer (Oxford, England: 1990), vol. 45, nº 2, janvier 2009, p. 228‑47. PubMed, doi:10.1016/j.ejca.2008.10.026.

Fayers, PM, et al. EORTC QLQ-C30 scoring manual. Vol. 11, 2001.

Fiteni, Frédéric, et al. « Health-Related Quality-of-Life as Co-Primary Endpoint in Randomized Clinical Trials in Oncology ». Expert Review of Anticancer Therapy, vol. 15, nº 8, 2015, p. 885‑91. PubMed, doi:10.1586/14737140.2015.1047768.

Fleiss, J. L. « The Statistical Basis of Meta-Analysis ». Statistical Methods in Medical Research, vol. 2, nº 2, 1993, p. 121‑45. PubMed, doi:10.1177/096228029300200202.

Fleming, Thomas R., et John H. Powers. « Biomarkers and Surrogate Endpoints In Clinical Trials ». Statistics in medicine, vol. 31, nº 25, novembre 2012, p. 2973‑84. PubMed Central, doi:10.1002/sim.5403.

Freedman, Laurence S., et al. « Statistical Validation of Intermediate Endpoints for Chronic Diseases ». Statistics in Medicine, vol. 11, nº 2, janvier 1992, p. 167‑78. Wiley Online Library, doi:10.1002/sim.4780110204.

Gail, Mitchell H. « Sample Size Estimation When Time-to-Event Is the Primary Endpoint ». Drug Information Journal, vol. 28, nº 3, juillet 1994, p. 865‑77. SAGE Journals, doi:10.1177/009286159402800322.

Glossary | EORTC. http://groups.eortc.be/qol/glossary. Consulté le 18 novembre 2017.

Gotay, Carolyn C., et al. « The Prognostic Significance of Patient-Reported Outcomes in Cancer Clinical Trials ». Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, vol. 26, nº 8, mars 2008, p. 1355‑63. PubMed, doi:10.1200/JCO.2007.13.3439.

Gotay, Carolyn Cook, et al. « Quality-of-Life Assessment in Cancer Treatment Protocols: Research Issues in Protocol Development ». JNCI: Journal of the National Cancer Institute, vol. 84, nº 8, avril 1992, p. 575‑79. academic.oup.com, doi:10.1093/jnci/84.8.575.

Gourgou-Bourgade, S., et al. « Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials) ». Annals of Oncology, vol. 26, nº 5, mai 2015, p. 873‑79. academic.oup.com, doi:10.1093/annonc/mdv106.

Hamidou, Zeinab, et al. « Time to Deterioration in Quality of Life Score as a Modality of Longitudinal Analysis in Patients with Breast Cancer ». The Oncologist, vol. 16, n° 10, octobre 2011, p. 1458‑68. PubMed Central, doi:10.1634/theoncologist.2011-0085.

Kemp, Robert, et Vinay Prasad. « Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? » BMC Medicine, vol. 15, juillet 2017. PubMed Central, doi:10.1186/s12916-017-0902-9.

Kramar, A., et al. « Guidelines for the definition of time-to-event end points in renal cell cancer clinical trials: results of the DATECAN project ». Annals of Oncology, vol. 26, n° 12, décembre 2015, p. 2392‑98. academic.oup.com, doi:10.1093/annonc/mdv380.

Mantel, N. « Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration ». Cancer Chemotherapy Reports, vol. 50, n° 3, mars 1966, p. 163‑70.

Mathoulin-Pelissier, Simone, et al. « Survival End Point Reporting in Randomized Cancer Clinical Trials: A Review of Major Journals ». Journal of Clinical Oncology, vol. 26, n° 22, août 2008, p. 3721‑26. ascopubs.org (Atypon), doi:10.1200/JCO.2007.14.1192.

---. « Survival End Point Reporting in Randomized Cancer Clinical Trials: A Review of Major Journals ». Journal of Clinical Oncology, vol. 26, n° 22, août 2008, p. 3721‑26. ascopubs.org (Atypon), doi:10.1200/JCO.2007.14.1192.

McHugh, Mary L. « Interrater Reliability: The Kappa Statistic ». Biochemia Medica, vol. 22, n° 3, 2012, p. 276‑82.

Methy, N., et al. « Surrogate Endpoints for Overall Survival in Digestive Oncology Trials: Which Candidates? A Questionnaires Survey among Clinicians and Methodologists. » BMC Cancer, vol. 10, 2010, p. 277‑277. europepmc.org, doi:10.1186/1471-2407-10-277.

Moinpour, Carol McMillen, et al. « Quality of Life End Points in Cancer Clinical Trials: Review and Recommendations ». JNCI: Journal of the National Cancer Institute, vol. 81, n° 7, avril 1989, p. 485‑96. academic.oup.com, doi:10.1093/jnci/81.7.485.

Molenberghs, Geert, et al. « Statistical challenges in the evaluation of surrogate endpoints in randomized trials ». Controlled Clinical Trials, vol. 23, nº 6, décembre 2002, p. 607‑25. ScienceDirect, doi:10.1016/S0197-2456(02)00236-2.

Multiple Co-primary Endpoints: Medical and Statistical Solutions: A Report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of AmericaTherapeutic Innovation & Regulatory Science - Walter Offen, Christy Chuang-Stein, Alex Dmitrienko, Gary Littman, Jeff Maca, Laura Meyerson, Robb Muirhead, Paul Stryszak, Alex Baddy, Kun Chen, Kati Copley-Merriman, Willard Dere, Sam Givens, David Hall, David Henry, Joseph D. Jackson, Alok Krishen, Thomas Liu, Steve Ryder, A. J. Sankoh, Julia Wang, Chyon-Hwa Yeh, 2007. http://journals.sagepub.com/doi/abs/10.1177/009286150704100105. Consulté le 10 août 2017.

Neuhäuser, Markus. « How to Deal with Multiple Endpoints in Clinical Trials ». Fundamental & Clinical Pharmacology, vol. 20, nº 6, décembre 2006, p. 515‑23. PubMed, doi:10.1111/j.1472-8206.2006.00437.x.

Pazdur, Richard. « Endpoints for Assessing Drug Activity in Clinical Trials ». The Oncologist, vol. 13, nº Supplement 2, avril 2008, p. 19‑21. theoncologist.alphamedpress.org, doi:10.1634/theoncologist.13-S2-19.

---. « Endpoints for Assessing Drug Activity in Clinical Trials ». The Oncologist, vol. 13 Suppl 2, 2008, p. 19‑21. PubMed, doi:10.1634/theoncologist.13-S2-19.

Peppercorn, Jeffrey M., et al. « American Society of Clinical Oncology Statement: Toward Individualized Care for Patients with Advanced Cancer ». Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, vol. 29, nº 6, février 2011, p. 755‑60. PubMed, doi:10.1200/JCO.2010.33.1744.

Pocock, S. J. « Interim Analyses for Randomized Clinical Trials: The Group Sequential Approach ». Biometrics, vol. 38, nº 1, mars 1982, p. 153‑62.

Prentice, R. L. « Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria ». Statistics in Medicine, vol. 8, nº 4, avril 1989, p. 431‑40.

Rock, Edwin P., et al. « Food and Drug Administration Drug Approval Summary: Sunitinib Malate for the Treatment of Gastrointestinal Stromal Tumor and Advanced Renal Cell Carcinoma ». The Oncologist, vol. 12, nº 1, janvier 2007, p. 107‑13. PubMed, doi:10.1634/theoncologist.12-1-107.

« Sample Size Calculations in Clinical Research, Second Edition ». CRC Press, 22 août 2007, https://www.crcpress.com/Sample-Size-Calculations-in-Clinical-Research-Second-Edition/Chow-Wang-Shao/p/book/9781584889823.

Sartorius, Norman. « A WHO Method for the Assessment of Health-Related Quality of Life (WHOQOL) ». Quality of Life Assessment: Key Issues in the 1990s, Springer, Dordrecht, 1993, p. 201‑07. link.springer.com, doi:10.1007/978-94-011-2988-6_11.

Schoenfeld, D. A. « Sample-Size Formula for the Proportional-Hazards Regression Model ». Biometrics, vol. 39, nº 2, juin 1983, p. 499‑503.

Schulz, Kenneth F., et al. « CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials ». BMJ (Clinical Research Ed.), vol. 340, mars 2010, p. c332.

Senn, Stephen. « Some Controversies in Planning and Analysing Multi-Centre Trials ». Statistics in Medicine, vol. 17, nº 15‑16, août 1998, p. 1753‑65. Wiley Online Library, doi:10.1002/(SICI)1097-0258(19980815/30)17:15/16<1753::AID-SIM977>3.0.CO;2-X.

Thompson, S. G. « Controversies in Meta-Analysis: The Case of the Trials of Serum Cholesterol Reduction ». Statistical Methods in Medical Research, vol. 2, nº 2, 1993, p. 173‑92. PubMed, doi:10.1177/096228029300200205.

Thompson, S. G., et S. J. Pocock. « Can Meta-Analyses Be Trusted? » Lancet (London, England), vol. 338, nº 8775, novembre 1991, p. 1127‑30.

Velentgas, Priscilla, et al. Outcome Definition and Measurement. Agency for Healthcare Research and Quality (US), 2013. www.ncbi.nlm.nih.gov, https://www.ncbi.nlm.nih.gov/books/NBK126186/.

Weintraub, William S., et al. « The perils of surrogate endpoints ». European Heart Journal, vol. 36, nº 33, septembre 2015, p. 2212‑18. PubMed Central, doi:10.1093/eurheartj/ehv164.

« WHO | Constitution of WHO ». WHO, http://www.who.int/about/mission/en/. Consulté le 16 novembre 2017.

World Medical Association. « World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects ». JAMA, vol. 310, nº 20, novembre 2013, p. 2191‑94. PubMed, doi:10.1001/jama.2013.281053.

www.unitheque.com. Méthodes biostatistiques appliquées à la recherche clinique en cancérologie. https://www.unitheque.com/Livre/john_libbey_eurotext/L_innovation_therapeutique_en_cancerologie/Methodes_biostatistiques_appliquees_a_la_recherche_clinique_en_cancerologie-48726.html. Consulté le 18 novembre 2017.