Evaluer le bénéfice clinique dans les essais randomisés en utilisant les comparaisons par paire généralisées incluant des données de survie

Julien Péron

UNIVERSITE CLAUDE BERNARD - LYON 1

Evaluer le bénéfice clinique dans les essais randomisés en utilisant les comparaisons par paire généralisées incluant des données de survie

THESE

Présentée et publiquement soutenue

Le 30 octobre 2015

Par Julien PERON

Né le 30 Juillet 1985 à SURESNES (092)

Pour obtenir le grade de DOCTEUR de L'UNIVERSITE CLAUDE BERNARD - LYON 1

SPECIALITE : Biostatistiques

Directeur de thèse : Pascal ROY

Ecole doctorale E2M2 : Evolution Ecosystèmes Microbiologie Modélisation
Laboratoire de Biométrie et Biologie Evolutive - CNRS UMR 5558

Membres du Jury de la Thèse :

Mme Sylvie CHEVRET, Rapporteure (PU-PH)
Mr Raphael PORCHER, Rapporteur (MCU-PH)
Mr Marc BUYSE, Examinateur (Professeur associé, Hasselt University in Belgium)
Mr Gilles FREYER, Examinateur (PU-PH)
Mr François-Noël GILLY, Examinateur (PU-PH)
Mr Jacques BENICHOU, invité (PU-PH)
Mr Pascal ROY, Directeur de thèse (PU-PH)

UNIVERSITE CLAUDE BERNARD - LYON 1

Evaluer le bénéfice clinique dans les essais randomisés en utilisant les comparaisons par paire généralisées incluant des données de survie

THESE

Présentée et publiquement soutenue

Le 30 octobre 2015

Par Julien PERON

Né le 30 Juillet 1985 à SURESNES (092)

Pour obtenir le grade de DOCTEUR de L'UNIVERSITE CLAUDE BERNARD - LYON 1

SPECIALITE : Biostatistiques

Directeur de thèse : Pascal ROY

Ecole doctorale E2M2 : Evolution Ecosystèmes Microbiologie Modélisation
Laboratoire de Biométrie et Biologie Evolutive - CNRS UMR 5558

Membres du Jury de la Thèse :

Mme Sylvie CHEVRET, Rapporteure (PU-PH)
Mr Raphael PORCHER, Rapporteur (MCU-PH)
Mr Marc BUYSE, Examinateur (Professeur associé, Hasselt University in Belgium)
Mr Gilles FREYER, Examinateur (PU-PH)
Mr François-Noël GILLY, Examinateur (PU-PH)
Mr Jacques BENICHOU, invité (PU-PH)
Mr Pascal ROY, Directeur de thèse (PU-PH)

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est - Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud - Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme Caroline FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur Georges TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. Jean-Claude PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Remerciements

Je remercie Pascal Roy pour m'avoir accueilli au sein de son laboratoire, pour avoir dirigé cette thèse avec bienveillance et enthousiasme, et pour ses conseils originaux et avisés.

Je remercie avec chaleur Marc Buyse, pour avoir accepté de me confier le développement des comparaisons par paire, pour s'être rendu aussi disponible tout au long de ce travail, pour ses qualités scientifiques et humaines.

Je remercie Gilles Freyer, pour m'avoir orienté tout initialement vers le monde des biostatistiques, et pour son soutien sans faille tout le long de mon internat. Merci également pour cette incroyable ouverture d'esprit qui fait qu'un oncologue peut devenir statisticien et travailler avec des anthropologues sans aucune barrière.

Je remercie les rapporteurs de cette thèse, Sylvie Chevret et Raphaël Porcher pour la qualité de leurs rapports. Vos critiques et commentaires m'ont beaucoup apporté. Ils m'ont permis d'améliorer ce manuscrit, et m'ont donné de nombreux éclairages utiles que je garderais en mémoire pour mes travaux futurs.

Je remercie François-Noël Gilly, et Jacques Benichou pour avoir accepté de juger ce travail.

Je remercie l'ensemble du service de biostatistique, secrétaires, ingénieurs, chercheurs, médecins pour avoir su intégrer l'oncologue que je suis au sein de cette formidable équipe. C'est grâce à chacun d'entre vous que cet univers de travail est si riche, à la fois au niveau scientifique et humain. La diversité de nos parcours est la source de la qualité de nos échanges. Merci évidemment à Stéphanie, Coraline et Brice, qui m'ont patiemment supporté au quotidien. Votre mérite est immense. Brice, tes compétences de programmeur m'ont ébloui. Merci pour ton efficacité et la qualité de nos échanges.

Je remercie l'ensemble du service d'oncologie médicale. Vous m'avez toujours soutenu dans mes projets. De nouvelles aventures s'annoncent pour nous. La cohésion et l'amitié qui nous unissent sont un atout précieux. Merci tout particulièrement à Benoit, Denis, Olivier et Olivia, la dream-team lyonnaise de la revue systématique.

Enfin je remercie les Hospices Civils de Lyon, le comité médaille d'or, et la fondation Nuovo-Soldati pour le financement de cette thèse.

Je remercie ma famille et mes amis pour m'avoir soutenu dans cet objectif. Mes études touchent à leur fin, incroyable ! Je sais que vous serez à mes côtés pour m'accompagner dans ce bouleversement majeur ! J'ai la chance d'être entouré d'une famille formidable et d'amis sur lesquels je peux compter. Je profite de cette thèse pour vous remercier de toutes les marques d'affection, de soutien, ou juste de camaraderie qui m'accompagnent chaque jour depuis toujours.

A ma femme Noura. Tu sais tout l'amour que j'ai pour toi. C'est grâce à ton soutien à toute épreuve, grâce à tes mots rassurants, et au bonheur que tu m'apportes chaque jour que j'ai pu me lancer sereinement dans ce travail de thèse. C'est aussi toi qui m'aide à définir mes priorités, et n'en doute pas, tu seras toujours ma priorité numéro 1.

Table des matières

TABLE DES FIGURES (HORS ARTICLES)

Listes des tableaux (hors articles)

Notations

n	Taille de la population dans le bras traitement
m	Taille de la population dans le bras contrôle
τ_l	Seuil de bénéfice minimal cliniquement significatif du critère de jugement analysé en priorité l.
x_i et y_j	Valeur du critère de jugement *X* pour le patient i issu du groupe *T* et du critère de jugement *Y* pour le patient j issu du groupe *C*. En cas de critère de type temps jusqu'à événement, x_i et y_j sont les temps jusqu'à observation.
x_i^0 et y_j^0	Temps jusqu'à événement pour le patient i issu du groupe *T* et pour le patient j issu du groupe *C*.
u_i et v_j	Temps jusqu'à censure pour le patient i issu du groupe *T* et pour le patient j issu du groupe *C*.
γ_i et ε_j	Indicatrice d'événement pour le patient i issu du groupe *T* et pour le patient j issu du groupe *C*.
$p_{ij}(l)$	Score de propension au succès pour le critère de jugement analysé en priorité l, et la paire de patients formé du patient i issu du groupe *T* et du patient j issu du groupe *C*.
$\delta(l)$	Propension au succès associé au critère de jugement analysé en priorité l
$\Delta(l)$	Propension au succès associé au critère de jugement analysé en priorité l
$S_T(t) = \mathbb{P}\left[x_i^0 \geq t\right]$	Fonction de survie des patients issus du groupe T
$S_C(t) = \mathbb{P}\left[y_j^0 \geq t\right]$	Fonction de survie des patients issus du groupe C
f	Proportion de paires informatives

Abréviations utilisées

BRAM : *Benefit-Risk Assessment Model*

CONSORT : *Consolidated Standard of Reporting Trials*

CRP : Critère de jugement rapporté par les patients

EMA : *European Medicines Agency*

ECR : essai contrôlé randomisé

EORTC: *European Organization for Research and Treatment of Cancer*

FOLFIRINOX : Association de 5-fluorouracile, d'oxaliplatine, de leucovorine et d'irinotecan

Groupe T : Groupe des patients recevant le traitement expérimental

Groupe C : Groupe des patients recevant le traitement contrôle

NNH : Nombre de sujets à traiter pour observer un événement indésirable

NNT : Nombre de sujets à traiter pour observer un succès

OTU : *Overall Treatment Utility*

Q-TWiST : *quality-adjusted Time Without Symptoms of disease progression or Toxicity of treatment*

Introduction

Un essai clinique est défini comme toute étude systématique d'un médicament ou d'une intervention de santé chez l'homme, qu'il s'agisse de volontaires malades ou sains. Les essais contrôlés randomisés (ECR) ont pour but de comparer l'efficacité d'une intervention de santé expérimentale à une intervention servant de contrôle. Dans un ECR, les patients sont aléatoirement répartis (randomisation) parmi les groupes correspondants à chaque intervention thérapeutique testée. Lorsque l'intervention de santé est un nouveau traitement, on parle alors d'essais thérapeutiques randomisés. La randomisation a pour but de rendre les deux groupes de patients comparables en tous points, en dehors du traitement administré. Les essais thérapeutiques randomisés permettent l'évaluation de l'efficacité et de la balance bénéfice-risque des traitements. Les médicaments en développement peuvent obtenir une autorisation de mise sur le marché lorsque ces évaluations sont favorables. Les essais thérapeutiques randomisés sont donc souvent réalisés sur un grand nombre de malades, et leur méthodologie doit être particulièrement rigoureuse. La méthodologie des essais thérapeutiques repose sur la démarche hypothético-déductive. L'investigateur doit définir a priori un critère de jugement principal et un plan d'analyse statistique, qui serviront à définir le nombre de sujets à inclure en fonction de l'ampleur du bénéfice attendu, de la puissance et du risque de première espèce souhaités.

Il est fréquent que plusieurs critères de jugement soient nécessaires pour évaluer l'effet clinique d'un traitement. Prenons l'exemple d'un essai randomisé réalisé en oncologie médicale évaluant un nouveau traitement. Si la survie globale est souvent choisie comme critère de jugement principal, l'évaluation globale de l'effet thérapeutique repose également sur les événements indésirables du traitement, sur l'évolution de la tumeur, et souvent sur la qualité de vie des patients. Ces critères sont alors analysés de façon secondaire. De plus leur importance dans l'évaluation globale de l'effet du traitement dépend de l'ampleur de l'effet du traitement sur le critère de jugement principal et sur les critères de jugement secondaire. La double nécessité de prédéfinir le plan d'analyses statistiques sur un critère de jugement robuste et de prendre en compte l'ensemble des critères de jugement relatifs à l'état clinique des patients a parfois conduit à des conclusions contradictoires. En effet les résultats d'un essai sont d'abord analysés en suivant le plan d'analyse statistique, et le niveau de significativité statistique de l'essai est alors évalué. Puis une analyse du niveau de

significativité clinique est réalisée, souvent de façon informelle, prenant en compte les critères de jugement secondaires et l'ampleur des effets thérapeutiques.

La première partie de cette thèse porte sur les méthodes actuellement utilisées pour rapporter les événements indésirables dans les essais randomisés de phase III en oncologie médicale. Une revue systématique de la littérature a été conduite dans ce sens. La rédaction des manuscrits rapportant des essais contrôlés randomisés est considérée comme de bonne qualité si les informations importantes à la compréhension du plan d'étude, de sa conduite et de l'analyse des données sont rapportées de manière exhaustive. Le manuscrit doit contenir toutes les informations permettant d'évaluer la validité interne et la validité externe des résultats d'un essai. Le **premier objectif** de cette thèse était d'évaluer la qualité du rapport des événements indésirables dans les manuscrits rapportant les essais de phase III en oncologie, ainsi que l'homogénéité des méthodes utilisées pour rapporter ces événements indésirables. En effet des méthodes de rapport homogènes et correctement décrites sont le prérequis indispensable pour une évaluation non biaisée de la toxicité du traitement par le lecteur d'un manuscrit. Cette évaluation de la toxicité est en effet souvent utilisée pour évaluer subjectivement la balance bénéfice-risque des nouveaux traitements.

La seconde partie de cette thèse porte sur les méthodes utilisées pour analyser et rapporter les critères de jugement rapportés par les patients. Ces critères de jugement sont ceux directement rapportés par les patients, sans interprétation des réponses des patients par leurs médecins ou toute autre personne. Ces critères de jugement rapportés par les patients (CRPs) peuvent varier en complexité, allant d'une question à un item jusqu'à des instruments multidimensionnels comme les mesures de la qualité de vie des patients. Les CRPs permettent de mesurer le ressenti des patients par rapport à un traitement évalué dans un essai clinique, et sont donc un reflet à la fois de l'efficacité et de la toxicité des traitements. Ils peuvent apporter la preuve la plus directe qu'un traitement améliore le bien-être des patients, les symptômes liés à la maladie, ou les symptômes liés à la toxicité du traitement. Le **second objectif** de cette thèse était de réaliser une revue systématique de l'utilisation actuelle des critères de jugement rapportés par les patients en oncologie médicale, et d'évaluer la qualité de rédaction des manuscrits rapportant les essais contrôlés randomisés par rapport à ces CRPs.

Afin de réconcilier l'évaluation du niveau de significativité statistique et du niveau de significativité clinique, des méthodes ont été proposées permettant d'analyser simultanément l'ensemble des critères de jugement pertinents pour évaluer le bénéfice clinique d'un

traitement. Dans la troisième partie de cette thèse, nous développerons la méthode des comparaisons par paire généralisées. Les principes de cette méthode, proposée par le Pr Marc Buyse en 2010, seront rappelés [1]. Lorsqu'au moins un critère de jugement de type temps jusqu'à événement est inclus dans l'analyse globale de l'effet d'un traitement, la procédure standard des comparaisons par paire généralisées ne prend pas en comptes les temps jusqu'à censure pour estimer la « propension au succès » du traitement expérimental, traduction de l'anglais « chance of a better outcome ». Le **troisième objectif** de cette thèse est de montrer comment la prise en compte des temps jusqu'à censure, en utilisant l'estimateur des fonctions de survie de Kaplan-Meier [2], permet de réaliser une estimation non biaisée de la propension au succès, et d'augmenter les performances du test de l'hypothèse nulle.

Les autorités administratives d'enregistrement des nouveaux traitements, américaine (US Food and Drug Administration) comme européenne (European Medicines Agency), ont souligné l'importance de réaliser une évaluation transparente et rigoureuse de la balance bénéfice-risque des nouveaux traitements [3]–[5]. Dans la quatrième partie de cette thèse, les principales méthodes existantes permettant d'évaluer la balance bénéfice-risque des nouveaux traitements seront rappelées. Le **quatrième objectif** de cette thèse est de montrer comment une évaluation pertinente et standardisée de la balance bénéfice-risque des nouveaux traitements peut être réalisée en utilisant la méthode des comparaisons par paire généralisées, en prenant l'exemple des nouvelles thérapies systémiques développées pour traiter l'adénocarcinome du pancréas métastatique.

Dans la cinquième partie, la méthode des comparaisons par paires est proposée comme alternative aux critères de jugement composites. Les critères de jugement centrés sur le patient reflètent ce qu'il perçoit de son bien-être ou de sa survie, et permettent ainsi une évaluation directe des bénéfices cliniques d'une intervention thérapeutique [6]. De manière générale, en oncologie, ces critères comprennent la qualité de vie relative à la santé et surtout la survie globale, historiquement considérée comme le critère de référence et le plus convaincant en termes d'efficacité. Néanmoins, du fait de l'augmentation du nombre de traitements efficaces et de l'amélioration des soins de support, il devient de plus en plus difficile de démontrer un allongement statistiquement significatif de la survie globale. En effet l'effet du traitement peut être dilué par les traitements reçus ultérieurement, notamment lorsque les patients du bras contrôle reçoivent le traitement expérimental après une première progression. Pour contourner cette difficulté, il est possible de s'orienter vers des critères intermédiaires composites comme la survie sans progression. Il s'agit du temps entre la date

de randomisation et le décès ou la progression tumorale, selon celui qui survient en premier. Les critères composites sont souvent utilisés pour analyser de façon combinée la fréquence de plusieurs types d'événement. Ces critères permettent d'augmenter le nombre d'événement inclus dans le critère de jugement, et ainsi d'augmenter la puissance du test de la différence en survie sans événement. Les critères composites sont mis en défaut lorsque l'effet d'un traitement est hétérogène sur les différents types d'événement inclus dans le critère composite, et lorsque l'importance clinique de ces types d'événement est inégale [7]. Par exemple une différence observée en survie sans progression ne garantit pas une différence en termes de survie globale. Lorsque les décès sont des événements minoritaires, il est même possible d'observer une amélioration de la survie sans progression en faveur d'un traitement malgré une augmentation de la fréquence des décès. Le **cinquième objectif** de cette thèse est de montrer que les comparaisons par paire généralisées peuvent être utilisées pour analyser conjointement deux critères de jugement sur lesquels un bénéfice thérapeutique est attendu. La démonstration reposera sur une évaluation de la puissance de l'analyse combinée, sur la comparaison de la souplesse de la méthode par rapport aux critères composites, et sur la façon d'interpréter les résultats.

Cette thèse a été conduite au sein de l'équipe Biostatistiques-Santé, équipe du Laboratoire de Biométrie et Biologie Evolutive - UMR CNRS 5558. Cette équipe a la particularité d'être adossée au Service de Biostatistique des Hospices Civils de Lyon, structure de soutien en biostatistique, opérant pour les équipes hospitalières mais également pour différentes institutions publiques ou privées. Il s'agit donc d'un environnement associant une dimension de recherche en méthodologie statistique à une dimension de biostatistique appliquée à la recherche clinique. Ce travail a été réalisé sous la direction du Pr Pascal ROY, et en collaboration étroite avec le Pr Marc BUYSE, qui est à l'origine de la méthode des comparaisons par paire généralisées [1].

Chapitre I

I. Les événements indésirables dans les rapports d'essais contrôlés randomisés

I.1. La qualité du rapport des événements indésirables : revue systématique de la littérature

Un des objectifs principaux de cette thèse est de proposer une méthode d'analyse conjointe des bénéfices et des effets secondaires des traitements dans les essais contrôlés randomisés. Le prérequis de ce type d'analyse est de disposer de variables évaluant de façon non biaisée les effets secondaires des traitements. Il peut s'avérer intéressant de comparer la toxicité de traitements évalués dans des essais thérapeutiques différents, par exemple à l'occasion d'une méta-analyse. Les variables mesurant la toxicité des traitements doivent donc être standardisées, ou au moins parfaitement définies. Les événements indésirables sont les événements délétères survenant lors de la conduite d'un essai thérapeutique. Le lien de causalité entre les événements observés et les traitements administrés est souvent difficile à affirmer. Dans les ECRs, la fréquence de ces événements indésirables peut être comparée dans les deux groupes malgré l'absence de certitude sur la causalité de ces événements.

Une revue systématique des manuscrits rapportant des essais contrôlés randomisés publiés entre 2007 et 2011 et évaluant des traitements systémiques anticancéreux a été réalisée. Dans cette revue systématique, la qualité du rapport des événements indésirables a été évaluée en fonction de l'adhérence aux critères de qualité de rapport des événements indésirables dans les essais randomisés définis par le groupe CONSORT (*Consolidated Standard of Reporting Trials*). L'hétérogénéité dans la définition des variables mesurant les effets secondaires des traitements a également été évaluée. Les résultats de cette étude ont été rapportés dans le *Journal of Clinical Oncology* [8]. Les messages principaux de cet article sont rappelés en fin de sous-chapitre, après la présentation du manuscrit en format édité par le *Journal of Clinical Oncology*.

VOLUME 31 · NUMBER 31 · NOVEMBER 1 2013

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Adherence to CONSORT Adverse Event Reporting Guidelines in Randomized Clinical Trials Evaluating Systemic Cancer Therapy: A Systematic Review

Julien Péron, Denis Maillet, Hui K. Gan, Eric X. Chen, and Benoit You

Julien Péron, Denis Maillet, and Benoit You, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, Pierre-Bénite; Julien Péron, Hospices Civils de Lyon; Julien Péron and Benoit You, Université de Lyon, Lyon; Julien Péron, Centre National de la Recherche Scientifique Unité Mixte de Recherche 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Villeurbanne; Benoit You, EMR UCBL/HCL 3738, Faculté de Médecine Lyon-Sud, Oullins, France; Hui K. Gan, Joint Austin-Ludwig Oncology Unit, Austin Hospital, Melbourne, Victoria, Australia; Eric X. Chen, Princess Margaret Hospital, University Health Network; and Eric X. Chen, University of Toronto, Toronto, Ontario, Canada.

Published online ahead of print at www.jco.org on September 23, 2013.

J.P. is the recipient of a grant from the Nuovo-Soldati Research Foundation.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Benoit You, MD, PhD, Service d'Oncologie Médicale, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, F-69310, 165, Chemin du Grand Revoyet 69495 Pierre-Bénite, France; e-mail: benoit.you@chu-lyon.fr.

0732-183X/13/3131w-3957w/$20.00

DOI: 10.1200/JCO.2013.49.3981

ABSTRACT

Purpose
The Consolidated Standards of Reporting Trials (CONSORT) guidance was extended in 2004 to provide a set of 10 specific and comprehensive guidelines regarding adverse event (AE) reporting in randomized clinical trials (RCTs). Limited data exist regarding adherence to these guidelines in publications of oncology RCTs.

Methods
All phase III RCTs published between 2007 and 2011 were reviewed using a 16-point AE reporting quality score (AERQS) based on the 2004 CONSORT extension. Multivariable linear regression was used to identify features associated with improved reporting quality.

Results
A total of 325 RCTs were reviewed. The mean AERQS was 10.1 on a 16-point scale. The most common items that were poorly reported were the methodology of AE collection (adequately reported in only 10% of studies), the description of AE characteristics leading to withdrawals (15%), and whether AEs are attributed to trial interventions (38%). Even when reported, the methods of AE collection and analysis were highly heterogeneous. The multivariable regression model revealed that industry funding, intercontinental trials, and trials in the metastatic setting were predictors of higher AERQS. The quality of AE reporting did not improve significantly over time and was not better among articles published in journals with a high impact factor.

Conclusion
Our findings show that some methodologic aspects of AE collection and analysis were poorly reported. Given the importance of AEs in evaluating new treatments, authors should be encouraged to adhere to the 2004 CONSORT guidelines regarding AE reporting.

J Clin Oncol 31:3957-3963. © 2013 by American Society of Clinical Oncology

INTRODUCTION

Randomized clinical trials (RCTs) are considered to be the gold standard in assessing medical interventions. The primary outcome in RCTs is usually a measure of the response or survival, whereas adverse outcomes are usually assessed as secondary outcomes. However, both the benefits and adverse effects are important to understand the benefit-risk balance of an intervention. Although there are several well-known limitations to toxicity assessment in RCTs (eg, underpowered to detect small but important toxicity differences[1]; emphasis of short-term rather than long-term toxicities), these are further exacerbated by inadequacies in their reporting in publications of RCTs. It is well known that inconsistencies in the methods of adverse event (AE) collection, analysis, and reporting may affect the number of reported AEs and thus the perceived toxicity of an intervention.[2-4] Thus, it is important that RCT reports provide sufficient and appropriate details regarding AEs.

The Consolidated Standards of Reporting Trials (CONSORT) statement provides guidance to authors regarding essential items that should be included in RCT reports.[5] Since its publication, the quality of RCT reporting has shown improvement over time.[6,7] The original CONSORT statement did not provide recommendations for AE reporting.[5] In 2001, the first update was published, adding a single item on AE reporting.[8] The guidance was further extended in 2004 to provide a set of 10 specific recommendations on AE reporting (the 2004 CONSORT extension).[4]

Adherence to these guidelines are particularly relevant for RCTs evaluating systemic cancer therapies because oncology drugs have lower therapeutic indices compared with drugs in other therapeutic

areas.[9] Moreover, the frequent use of multiagent and multimodality regimens substantially increases the risk of toxicity.[10] Because there are few data regarding the adequacy of toxicity reporting in oncology RCTs,[11] this review was performed to evaluate the quality of AE reporting. In addition, we investigated article characteristics associated with better quality in AE reporting.

METHODS

Trial Selection

We searched MEDLINE via PubMed (http://www.pubmed.gov) to identify all publications of RCTs assessing systemic therapies for solid tumors published at least 3 years after the publication of the 2004 CONSORT extension (between January 2007 and December 2011) in the following 10 English-language journals where oncology RCTs are frequently published: *Annals of Oncology*; *British Journal of Cancer*; *Breast Cancer Research and Treatment*; *Cancer*; *European Journal of Cancer*; *Journal of Clinical Oncology*; *Journal of the National Cancer Institute*; *The Lancet*; *The Lancet Oncology*; and *The New England Journal of Medicine*. The search was performed in March 2012, using the terms "randomized" and "cancer" as keywords and "English" plus "clinical trials" or "randomized controlled trial" as limits. Exclusion criteria were as follows: pediatric studies; hematologic trials; phase I, II, or IV trials; meta-analyses, overviews, and publications using pooled data from two or more trials; and secondary reports on previously published trials.[12-14]

Development of a Quantitative Scoring System for Quality of AE Reporting

Similar to previous studies where overall quality-of-reporting scores were used,[11,15-17] an AE reporting quality score (AERQS) based on the 2004 CONSORT extension was defined by three of the authors (J.P., D.M., and B.Y.). The score was based on 16 items derived from the 10 recommendations (Table 1). These items were chosen because they all refer to objectively mea-

Table 1. Quality of Harms Reporting Rating Using Items From the 2004 Extended CONSORT Statement (n = 325)

2004 CONSORT Recommendation No.	Descriptor of the CONSORT Criteria	Descriptor of the Reporting Quality Criteria	Trials in Which Item Was Adequately Reported	
			No.	%
1	If the study collected data on harms and benefits, the title or abstract should so state.	1. Adverse events (AEs) mentioned in the title or abstract	291	90
		AEs mentioned in the title	6	2
		AEs mentioned in the abstract	291	90
2	If the trial addresses both harms and benefits, the introduction should so state.	2. Information on AEs mentioned in introduction	213	66
3	List addressed AEs with definitions for each (with attention, when relevant, to grading, expected *v* unexpected events, reference to standardized and validated definitions, and description of new definitions).	3a. If article mentioned use of a validated instrument to report AE severity	243	75
		3b. If article mentioned definition of AE	141	43
4	Clarify how harms-related information was collected (mode of data collection, timing, attribution methods, intensity of ascertainment, and harms-related monitoring and stopping rules, if pertinent).	4a. Description of how harms data were collected (eg, diaries, phone interviews, or face-to-face interviews)	34	10
		4b. Description of when AE data were collected	171	53
		4c. Whether or not AEs were attributed to trial drugs	125	38
5	Describe plans for presenting and analyzing information on harms (including coding, handling of recurrent events, specification of timing issues, handling of continuous measures, and any statistical analyses).	5. Description of methods for presenting and/or analyzing AEs	157	48
6	Describe for each arm the participant withdrawals that are a result of harms and their experiences with the allocated treatment.	6a. If article reported the number of withdrawals caused by AEs in each arms	228	70
		6b. Description of AEs leading to withdrawals	49	15
		6c. If article reported the number of deaths caused by AEs in each arms	232	71
		Description of AEs leading to death	131	40
7	Provide the denominators for analyses on harms.	7a. If article provided denominators for AEs	280	86
		7b. If article provided definitions used for analysis set	260	80
		ITT	37	11
		Per protocol	208	64
		Safety data available	15	5
		Unclear	65	20
8	Present the absolute risk per arm and per AE type, grade, and seriousness, and present appropriate metrics for recurrent events, continuous variables, and scale variables, whenever pertinent.	8a. Results presented separately for each arm	301	93
		8b. Separate reporting of severe AEs (eg, grade > 2 or serious AEs)	296	91
9	Describe any subgroup analyses and exploratory analyses for harms.	—		
10	Provide a balanced discussion of benefits and harms with emphasis on study limitations, generalizability, and other sources of information on harms.	10. If the discussion was balanced with regard to efficacy and AEs	274	84

Abbreviations: AE, adverse event; ITT, intent to treat.

surable, different, and important aspects of AE reporting. Each item was scored as 1 if it was adequately reported or 0 if it was not clearly reported or not reported at all; each item was weighted with equal importance. For those recommendations with several subcomponents, a score was provided for each subcomponent. The ninth recommendation of the 2004 CONSORT extension was excluded because subgroup analysis for AEs was rarely performed.

The scoring system was piloted on 20 randomly selected trials (320 items) by two investigators (J.P. and D.M.) who were blinded to each other's results. Among these 320 items, 16 discrepancies were identified; however, all were successfully resolved by consensus. On the basis of this finding, a standardized data extraction form was used by two authors (J.P. and D.M.) to capture the remaining data in this review. This included the following guidelines to ensure homogenous data extraction for those recommendations potentially open to interpretation: AEs were considered adequately defined by the authors (item 3b) if relevant AEs were formally defined or if AEs were collected according to a commonly accepted standard (such as National Cancer Institute Common Toxicity Criteria or WHO criteria); an adequate reporting of "how harms data were collected" (item 4a) required at least a description of the collection circumstances (eg, during periodic physical examination or phone interviews or using diaries); and for the requirement that there be a separate reporting of severe AEs (item 8b), reporting was adequate if the frequencies of grade 3 and 4 AEs were provided separately or in aggregate.

In addition, data were also captured regarding the presentation of AEs in the results, specifically whether tables and figures were used, whether a clear attribution to trial intervention was stated, whether only a selection of toxicity data was presented, whether commonly used toxicity grading systems were used, and whether statistical comparisons of toxicity between treatment arms were reported. Attribution of AEs was considered unclear if the authors did not specifically state that they were all possibly related, probably related, and/or definitely related to trial intervention.

Definition of Trial Characteristics

Trials were considered as industry funded if an RCT received any form of industry funding with the exception of studies where only drug(s) was provided but no funding. Trials were considered intercontinental when patients from more than one continent were included. A positive trial was defined as one in which the experimental arm was deemed to be superior to the standard arm in superiority trials by trial investigators, not inferior in noninferiority trials, or alike in equivalence trials. A negative study was defined as one in

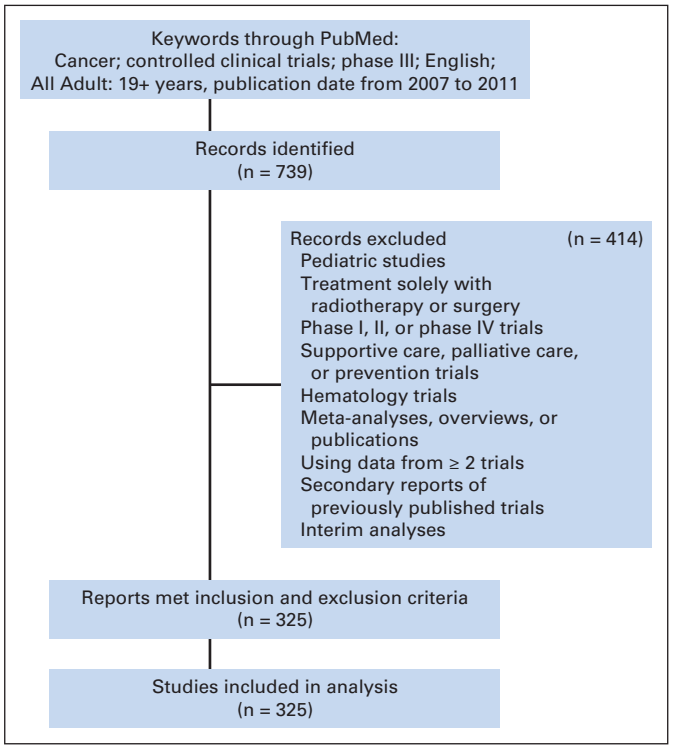


Fig 1. Selection of randomized clinical trials in the systematic review.

Table 2. Trial Characteristics

Characteristic	Studies (N = 325) No.	%
Year of publication		
2007	70	22
2008	75	23
2009	52	16
2010	55	17
2011	73	23
Tumor site		
Lung	72	22
Breast	74	23
Urinary system	34	11
Colon/rectum	48	15
Others	97	30
Sources of trial funding		
No industry funding	101	31
Funded by industry	198	61
Unknown	26	8
Journal		
Journal of Clinical Oncology	151	47
Annals of Oncology	34	11
The New England Journal of Medicine	32	10
The Lancet	25	8
European Journal of Cancer	22	7
Other journals	61	19
Journal impact factor		
< 10	85	26
10-20	183	56
> 20	57	18
Region in which RCT was led		
International	92	28
North America	60	18
Europe	141	43
Others	32	10
Type of investigational therapy		
Cytotoxic chemotherapy	178	55
Hormonal therapy	27	8
Molecular targeted therapy	103	32
Immunotherapy	12	4
Other	5	2
Cancer stage		
Adjuvant and/or neoadjuvant	93	29
Metastatic	232	71
Sample size, No. of patients		
Median	491	
Interquartile range	270-795	
Results of the primary outcome		
Positive	131	40
Negative	194	60
Toxicity profile conclusions*		
Equivalent	104	32
Investigational arm more toxic	136	42
Control arm more toxic	34	11
No conclusion	51	16

Abbreviation: RCT, randomized controlled trial.
*Conclusion of the trial's authors.

which the experimental arm was deemed not superior, inferior, or not equivalent to the standard arm. The authors' assessment of the overall toxicity profile of the experimental arm was based on conclusions in the abstract or in the discussion section of articles. The toxicity of experimental arms was categorized as more toxic, less toxic, equivalent, or unknown according to authors' conclusions. In particular, experimental arms that had toxicities qualitatively different from standard arms but comparable in intensity were considered equivalent.

Statistical Analysis

The AERQS was the sum of the number of items that were adequately reported (Table 1) and expressed as an integer between 0 and 16. AERQS scores were summarized using descriptive statistics such as mean, CI, and range. Single-item frequencies were compared between categories using χ^2 tests.

Univariate and multivariable linear regression analyses were used to identify factors associated with higher AERQS. The following trial characteristics were investigated: year of publication, tumor site, presence of an industry partner, journal impact factor (IF), geographic region, type of investigational therapy, cancer stage, sample size, results of the primary outcome, and conclusion of authors regarding the toxicity profile of the investigational arm. The multivariable model included all of the previously mentioned covariates. No covariate selection was performed because it was deemed desirable to include as many factors associated with reporting quality as possible. Covariates were considered statistically associated with AERQS if the associated $P < .05$.

It was also hypothesized that articles from the same journal might have AERQSs that were more closely correlated to each other than articles from different journals. Therefore, mixed-effects models were used as a supportive regression analysis with the incorporation of publishing journal in the model as a random effect. Results were similar to the linear model without the assumption of correlation. Therefore, for simplicity, only the results of the linear model are reported here. Statistical analyses were performed using R software (http://www.R-project.org/). All statistical tests were two-sided.

Table 3. Results of Regression Analyses of Factors Predicting 2004 AERQS (0 to 16 scale)

			Linear Regression					
	AERQS		Univariate Analysis			Multivariate Analysis		
Study Characteristic	Mean	SE	Estimate*	SE	*P*	Estimate*	SE	*P*
Year of publication, continuous	—		0.32	0.10	.0012	0.17	0.09	.065
Tumor site								
Lung	9.88	2.55	Reference		.289	Reference		.025
Breast	10.16	2.72	0.29	0.43		1.17	0.41	
Urinary system	9.85	2.38	−0.02	0.55		0.74	0.52	
Colon/rectum	10.79	2.30	0.91	0.49		1.25	0.44	
Others	9.56	2.84	−0.02	0.41		0.60	0.37	
Sources of trial funding								
No industry funding	9.20	2.74	Reference		< .001	Reference		< .001
Funded by industry	10.68	2.44	1.48	0.31		1.14	0.31	
Unknown	8.81	2.23	−0.39	0.55		−0.24	0.53	
Journal impact factor								
< 10	9.54	2.76	Reference		.065	Reference		.71
10-20	10.17	2.48	0.63	0.34		0.21	0.32	
> 20	10.52	2.84	0.99	0.45		0.38	0.47	
Region in which RCT was led								
Intercontinental	11.21	2.18	Reference		< .001	Reference		.002
North America	9.8	2.28	−1.42	0.42		−0.29	0.41	
Europe	9.27	2.84	−1.95	0.34		−0.91	0.35	
Other	10.78	2.11	−0.44	0.52		0.65	0.52	
Type of investigational therapy								
Cytotoxic chemotherapy	9.99	2.51	Reference		< .001	Reference		.067
Hormonal therapy	8.81	3.57	−1.18	0.53		−1.21	0.51	
Molecular targeted therapy	10.74	2.19	0.74	0.32		−0.37	0.32	
Immunotherapy	9.25	2.86	−0.74	0.77		−0.57	0.71	
Other	7.60	4.56	−2.39	1.16		−1.92	1.06	
Cancer stage								
Metastatic disease	10.43	2.39	Reference		< .001	Reference		.001
Neoadjuvant/adjuvant	9.16	2.98	−1.27	0.32		−1.00	0.31	
Sample size/100, continuous	—		0.02	0.02	.28	0.01	0.02	.54
Results of the primary outcome								
Negative	9.76	2.64	Reference		.0091	Reference		.76
Positive	10.53	2.57	0.77	0.30		0.09	0.28	
Conclusion of safety outcome for authors								
Equivalent	10.5	2.29	Reference		< .001	Reference		< .001
Investigational arm more toxic	10.42	2.23	−0.08	0.32		−0.20	0.31	
Investigational arm less toxic	10.71	1.85	0.21	0.49		0.34	0.46	
No conclusion	7.82	3.51	−2.68	0.42		−2.18	0.40	

Abbreviation: AERQS, adverse event reporting quality score; RCT, randomized controlled trial.
*Scale range of 0 to 16. The estimates shown indicate the incremental benefit observed compared with the reference level. Any positive value indicates benefit compared with reference, whereas any negative value indicates detriment compared with reference.

RESULTS

Characteristics of Selected RCTs

From the 739 trials initially screened, a total of 325 RCTs were included in this analysis (Fig 1; Appendix, online only). Characteristics of these RCTs are listed in Table 2. The number of published RCTs per year was stable, with 28% of trials being intercontinental studies. Journals with higher IF were most likely to publish intercontinental RCTs (IF > 20: 42%; IF between 20 and 40: 30%; IF < 20: 15%; P = .001). Conversely, trials led outside of Europe and North America were less frequently published in journals with a high IF. Sixty-seven percent of articles were published in three journals (*Annals of Oncology, Journal of Clinical Oncology,* and *The New England Journal of Medicine*; Table 2). Most trials were at least partially industry funded (n = 198, 61%). Forty percent of the trials were positive based on the stated primary outcome, and this frequency was stable over time. Trials published in high-IF journals were more frequently positive (75% in journals with IF > 20). The number of trials investigating molecular targeted therapies increased progressively from 12 (17%) in 2007 to 32 (44%) in 2011.

Rating of Overall Quality Score

The mean AERQS for all items was 10.1 on a 16-point scale (range, 0 to 16; 95% CI, 9.8 to 10.4), with 22 publications (7%) having an AERQS ≤ 5. Only two trials were found with a score of 16. Items pertaining to methods of AE data collection and analysis (items 3 to 5) were poorly reported. Specifically, how AEs were collected was adequately reported in only 10% of publications, AEs leading to withdrawals were adequately described in 15%, the attribution of AEs to trial interventions was discussed in 38%, AEs were clearly defined in 43%, the description of methods for presenting and/or analyzing AEs was present in 48%, and when AEs were collected was stated in 53%. Among these poorly reported items, none showed a statistically significant improvement over time using the Cochran-Armitage test for trend. The definition of AEs was adequately described more frequently by RCTs with at least partial industry funding (n = 105 of 198 RCTs, 53%) than by those with exclusive government or academic funding (n = 31 of 101 RCTs, 31%; P = .028).

Factors Associated With Reporting Quality

The multivariable regression model subsequently revealed that presence of industrial funding (P < .001), intercontinental trials (P = .002), cancer stage (P = .001), tumor site (P = .025) and authors' conclusion on safety outcome (P < .001) were independent predictors of higher AERQS. The estimated effects of such variables on AERQS were adjusted by the year of publication, the journal IF, the type of investigational therapy, the trial sample sizes, and the results of the primary outcome. Articles of RCTs with industrial funding had an AERQS on average 1.14 points higher than those without industrial funding. Publications of intercontinental-led trials had an AERQS that was 0.29 point higher than trials led in North America, 0.91 point higher than trials led in Europe, but 0.65 point lower than trials led in other continents. The AERQS of trials in the metastatic setting was higher than those in the adjuvant and neoadjuvant settings by a mean of 1.00 point. Finally, when authors did not conclude about the relative toxicity profile of the experimental arm, the mean AERQS of these publications was lower by 2.18 points (Table 3).

Table 4. Presentation of AEs in the Results Section of Articles

Presentation of AEs	No. of Trials	%
Mode of presentation		
Text only	21	6
Table and text	296	91
Figure and text	4	1
Figure, table, and text	4	1
Attribution of AEs to trial drugs		
Yes	88	27
No	37	11
Unclear	200	62
Selection of AEs reported*		
Severe AEs	95	29
Frequent AEs	66	20
Difference in frequency between trial arms	22	7
AEs selected by the investigator	18	6
Unclear	132	41
Statistical comparison of AE rates		
Yes	167	51
No	158	49
Scale used to report AE severity		
CTCAE	246	76
WHO	25	8
Other	11	3
No scale or unknown	43	13

Abbreviations: AEs, adverse events; CTCAE, Common Terminology Criteria for Adverse Events.
*Restriction rules could be combined.

Methods of AE Reporting

Most articles used tables to report AEs (Table 4). However, the length of these tables was highly variable. For simplicity and space limitations, reporting was often restricted to severe AEs (n = 95, 29%) and/or to frequent AEs (n = 66, 20%). Less commonly, authors reported only AEs with different frequencies between trial arms (n = 22, 7%) or a subset of AEs considered of primary interest (n = 18, 6%). For 132 articles (41%), the reason for selective AE inclusion was unclear. Among the 125 articles properly describing the attribution process, 88 (27%) reported only AEs attributed by investigators to trial treatments, whereas 37 (11%) reported all AEs whether or not they were related to trial treatments. The method of analysis of AEs was also heterogeneous, with 51% comparing frequencies of AEs through statistical tests and 49% providing descriptions only (Table 4).

DISCUSSION

A careful balance between efficacy and toxicity is of primary importance in medical interventions, especially in developing new cancer therapy. Concerns have been raised previously that anticancer drugs have toxicities that might outweigh their benefits.[9] Both the US Food and Drug Administration[18] and the European Medicines Agency[19] have stressed the importance of a more structured and transparent approach to the benefit-risk assessment in the evaluation of new therapies. Because toxicity assessment should ideally be performed using all available and high-quality data, standardized reporting is essential.[20] This is the first systematic review of the quality of AE reporting in RCTs evaluating systemic cancer therapies.

We used a scoring system derived from the AE-reporting CONSORT guidelines.[4] Previously, similar evaluations of AE reporting quality have been performed in other medical specialties.[11,21-30] Most of these studies did not use any quality score[21-27] or proposed various quality scores based on the CONSORT guidelines.[11,28-30] Because these scores were not specific to the needs of RCTs evaluating systemic cancer therapy and have not been widely accepted as standard tools, we defined an oncology-specific 16-item score in which each item was considered to have the same weight. Factors associated with higher scores were investigated. A 1-point difference in mean score between two groups was considered meaningful, because it would be equivalent to one group failing to report on one more toxicity requirement than the studies in the other group.

Overall, most articles included in our review had several deficiencies in reporting AEs. Items pertaining to methods of AE collection and analysis (items 3 to 5) were poorly reported. This finding may result from a perception that the methodology of AE collection and analysis is implicit and homogeneous. However, where it was possible to identify these aspects, we noted a high degree of heterogeneity among trials. For example, the population chosen for AE analysis, the process of attributing AEs, the criteria used to decide which AEs are reported, and the realization or not of statistical comparison of AE rates varied across trials (Tables 1 and 4). Rates of withdrawal and death related to toxicity were reported in 70% of trials, but the type of AEs responsible of withdrawal or death was insufficiently detailed in 15% and 40% of articles, respectively. Failure to report clinically relevant AEs limits the ability to make a meaningful risk-benefit assessment. It is interesting that articles in which toxicity data were better reported (higher AERQS score) were also significantly more likely to have an overall conclusion regarding treatment toxicity. One potential explanation for this finding is that the better quality of data in these RCTs allows a more definitive interpretation, although it is also possible that both merely reflect a greater awareness of authors on the important of AE reporting.

Although the reporting of RCTs has improved in general,[6,14] there has been no parallel improvement in AE reporting over a similar period of time. This finding corroborates the results of previous studies in nononcology trials.[11,28] This divergence may be related to the slower uptake or awareness of the 2004 CONSORT extension compared with the general CONSORT statement. Only two of the 10 journals included in this review cited this extension in their recommendations for authors (last accessed December 2012), although all of them referred to the general CONSORT statement. According to the Web of Science,[31] the 2004 CONSORT extension has been cited only 314 times, compared with 718 times for the more recent 2010 CONSORT statement and 3,376 times for the 2001 CONSORT statement.

A number of factors were associated with improved quality of AE reporting, including the source of trial funding. The higher AERQS observed among industry-funded RCTs in this review and others[11,21,28-30,32] might be explained by a higher quality of the toxicity data collection, an employment of medical writers, and/or an anticipated stricter scrutiny of the oncology community for safety data arising from industry-funded trials. The AERQS was higher for cytotoxic chemotherapy trials compared with trials investigating other types of drugs. The higher rate of safety concerns and the more standardized toxicity profiles of cytotoxic chemotherapies might explain this finding.[33] However, the marked increase in the number of trials investigating noncytotoxic therapies and the increased frequency of chronic but lower grade toxicities with these agents make this a pertinent observation for future trials. The lowest quality of AE reporting is found in adjuvant/neoadjuvant trials, and this observation is difficult to explain. Most of these trials were designed to demonstrate a benefit from the experimental arm in potentially curable disease, possibly reducing authors' interest for safety concerns.

Although being potentially subject to publication biases related to the limited number of assessed journals, the analysis was certainly representative of reporting quality in oncology because the most significant RCT articles seem to be published in a few leading journals. The reason for poor reporting of AEs is difficult to ascertain. One obvious reason is the desire to minimize AEs. However, AEs were actually better reported in RCTs with industrial funding, which is counter to concerns raised previously.[34] Poor AE reporting may also result from the assumption that toxicity outcomes are less important to readers than efficacy results or from space limitations posed by journals. The space devoted to AEs in articles reporting RCTs is usually small.[11,22,25] However, we believe that improvement in AE reporting would not necessarily mean longer articles. Use of standard AE definitions with appropriate references in Methods sections and of toxicity outcomes in tables may limit the need for more descriptive wording.

In conclusion, our findings show that methodologic aspects of AE collection and analysis were often poorly reported. The quality of AE reporting has not significantly improved over time, possibly as a result of the lower profile of the 2004 CONSORT extension statement. Adequate reporting of AEs is essential to adequately evaluate the risk-benefit ratio of experimental treatments. Hence, greater use of guidelines in the 2004 CONSORT extensions should be encouraged.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: Julien Péron, Denis Maillet, Hui K. Gan, Eric X. Chen, Benoit You
Collection and assembly of data: Julien Péron, Denis Maillet
Data analysis and interpretation: Julien Péron, Hui K. Gan, Eric X. Chen, Benoit You
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Moher D, Dulberg CS, Wells GA: Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 272:122-124, 1994
2. Tugwell P, Judd MG, Fries JF, et al: Powering our way to the elusive side effect: A composite outcome "basket" of predefined designated endpoints in each organ system should be included in all controlled trials. J Clin Epidemiol 58:785-790, 2005
3. Edwards JE, McQuay HJ, Moore RA, et al: Reporting of adverse effects in clinical trials should be improved: Lessons from acute postoperative pain. J Pain Symptom Manage 18:427-437, 1999
4. Ioannidis JP, Evans SJ, Gøtzsche PC, et al: Better reporting of harms in randomized trials: An extension of the CONSORT statement. Ann Intern Med 141:781-788, 2004
5. Begg C, Cho M, Eastwood S, et al: Improving the quality of reporting of randomized controlled trials: The CONSORT statement. JAMA 276:637-639, 1996

6. Moher D, Jones A, Lepage L: Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. JAMA 285:1992-1995, 2001

7. Hopewell S, Dutton S, Yu LM, et al: The quality of reports of randomised trials in 2000 and 2006: Comparative study of articles indexed in PubMed. BMJ 340:c723, 2010

8. Moher D, Schulz KF, Altman DG: The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357:1191-1194, 2001

9. Niraula S, Seruga B, Ocana A, et al: The price we pay for progress: A meta-analysis of harms of newly approved anticancer drugs. J Clin Oncol 30: 3012-3019, 2012

10. Socinski MA, Stinchcombe TE, Moore DT, et al: Incorporating bevacizumab and erlotinib in the combined-modality treatment of stage III non–small-cell lung cancer: Results of a phase I/II trial. J Clin Oncol 30:3953-3959, 2012

11. Haidich AB, Birtsou C, Dardavessis T, et al: The quality of safety reporting in trials is still suboptimal: Survey of major general medical journals. J Clin Epidemiol 64:124-135, 2011

12. You B, Gan HK, Pond G, et al: Consistency in the analysis and reporting of primary end points in oncology randomized controlled trials from registration to publication: A systematic review. J Clin Oncol 30:210-216, 2012

13. Gan HK, You B, Pond GR, et al: Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. J Natl Cancer Inst 104:590-598, 2012

14. Péron J, Pond GR, Gan HK, et al: Quality of reporting of modern randomized controlled trials in medical oncology: A systematic review. J Natl Cancer Inst 104:982-989, 2012

15. Toulmonde M, Bellera C, Mathoulin-Pelissier S, et al: Quality of randomized controlled trials reporting in the treatment of sarcomas. J Clin Oncol 29:1204-1209, 2011

16. Rios LP, Odueyungbo A, Moitri MO, et al: Quality of reporting of randomized controlled trials in general endocrinology literature. J Clin Endocrinol Metab 93:3810-3816, 2008

17. Kober T, Trelle S, Engert A: Reporting of randomized controlled trials in Hodgkin lymphoma in biomedical journals. J Natl Cancer Inst 98:620-625, 2006

18. US Food and Drug Administration: Guidance for industry and Food and Drug Administration staff: Factors to consider when making benefit-risk determinations in medical device premarket approvals and de novo classifications. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm267829.htm

19. Committee for Medicinal Products for Human Use: Reflection paper on benefit-risk assessment methods in the context of the evaluation of marketing authorisation applications of medicinal products for human use. http://www.emea.europa.eu/docs/en_GB/document_libr

20. Scharf O, Colevas AD: Adverse event reporting in publications compared with sponsor database for cancer clinical trials. J Clin Oncol 24:3933-3938, 2006

21. Turner L-A, Singh K, Garritty C, et al: An evaluation of the completeness of safety reporting in reports of complementary and alternative medicine trials. BMC Complement Altern Med 11:67, 2011

22. Ioannidis JP, Lau J: Completeness of safety reporting in randomized trials: An evaluation of 7 medical areas. JAMA 285:437-443, 2001

23. Chowers MY, Gottesman BS, Leibovici L, et al: Reporting of adverse events in randomized controlled trials of highly active antiretroviral therapy: Systematic review. J Antimicrob Chemother 64: 239-250, 2009

24. Pitrou I, Boutron I, Ahmad N, et al: Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med 169:1756-1761, 2009

25. Papanikolaou PN, Churchill R, Wahlbeck K, et al: Safety reporting in randomized trials of mental health interventions. Am J Psychiatry 161:1692-1697, 2004

26. Ethgen M, Boutron I, Steg PG, et al: Reporting of harm in randomized controlled trials evaluating stents for percutaneous coronary intervention. Trials 10:29, 2009

27. Ethgen M, Boutron I, Baron G, et al: Reporting of harm in randomized, controlled trials of nonpharmacologic treatment for rheumatic disease. Ann Intern Med 143:20-25, 2005

28. Shukralla AA, Tudur-Smith C, Powell GA, et al: Reporting of adverse events in randomised controlled trials of antiepileptic drugs using the CONSORT criteria for reporting harms. Epilepsy Res 97:20-29, 2011

29. Breau RH, Gaboury I, Scales CD Jr, et al: Reporting of harm in randomized controlled trials published in the urological literature. J Urol 183: 1693-1697, 2010

30. Smith SM, Chang RD, Pereira A, et al: Adherence to CONSORT harms-reporting recommendations in publications of recent analgesic clinical trials: An ACTTION systematic review. Pain 153:2415-2421, 2012

31. Web of Knowledge: Web of Science. http://www.webofknowledge.com

32. Bernal-Delgado E, Fisher ES: Abstracts in high profile journals often fail to report harm. BMC Med Res Methodol 8:14, 2008

33. Mansi L, Thiery-Vuillemin A, Nguyen T, et al: Safety profile of new anticancer drugs. Expert Opin Drug Saf 9:301-317, 2010

34. Djulbegovic B, Lacevic M, Cantor A, et al: The uncertainty principle and industry-sponsored research. Lancet 356:635-638, 2000

Un des messages principaux de cet article est que certains items méthodologiques, notamment ceux relatifs aux méthodes de recueil et d'analyse des événements indésirables, n'étaient souvent pas décrits. Par exemple, la description du support utilisé par les investigateurs pour recueillir les événements indésirables, la fréquence de ce recueil, et la sélection ou non des événements indésirables rapportés en fonction de leur relation avec le traitement estimée par les investigateurs n'étaient décrites que dans 10%, 53% et 38% des manuscrits respectivement. Les manuscrits rapportant des essais ayant un financement industriel, des essais incluant des patients sur plusieurs continents, et des essais réalisés dans un contexte de maladie tumorale métastatique avaient globalement une meilleure qualité de rapport des événements indésirables.

Un autre message principal de cet article était que la les variables utilisées pour rapporter les événements indésirables étaient hétérogènes, leur définition précise n'étant toutefois pas claire dans la majorité des manuscrits.

Cette revue systématique a également permis d'identifier que certains événements cliniques graves liés aux événements indésirables n'étaient fréquemment pas rapportés. Les événements cliniques graves liés aux événements indésirables étaient définis comme les décès toxiques, les arrêts de traitements, ou les modifications de dose liés à des événements indésirables. De plus lorsque la fréquence de ces événements cliniques graves liés aux événements indésirables était rapportée, la nature des événements indésirables était souvent non rapportée. Par exemple, le nombre de décès toxiques était rapporté dans 71% des manuscrits, et la nature des événements indésirables responsables de ces décès toxiques n'était décrite que dans 40% des manuscrits. Les événements cliniques graves liés aux événements indésirables semblent être une information importante pour évaluer la tolérance et donc la faisabilité d'un traitement, ainsi que sa balance bénéfice-risque. Une étude complémentaire centrée sur cette problématique a donc été réalisée.

I.2. La description des événements cliniques graves liés aux événements indésirables, comparaison d'une revue systématique de la littérature à l'avis des membres d'une société européenne de recherche clinique

L'étude présentée dans cette sous-section a été réalisée du fait de l'observation que les événements cliniques graves liés aux événements indésirables, tels que les décès toxiques, les arrêts de traitements, ou les modifications de dose, étaient souvent mal ou non rapportés dans les articles rapportant des essais contrôlés randomisés en oncologie médicale. Les résultats de la revue systématique présentée dans le sous-chapitre I.1 ont été utilisés pour définir la fréquence actuelle d'utilisation de ces événements cliniques graves dans les manuscrits rapportant des ECR. Une enquête d'opinion a été réalisée auprès des membres de l'*EORTC* (*European Organization for Research and Treatment of Cancer* - eortc.eu), afin d'évaluer l'opinion de professionnels engagés dans la recherche clinique en oncologie sur l'importance de rapporter les événements cliniques graves liés aux événements indésirables, ainsi que les événements indésirables de grade élevé. Les messages principaux de cet article sont rappelés en fin de sous-chapitre, après la présentation du manuscrit qui a été accepté pour publication dans le journal *Annals of Oncology* (In press).

The reporting of critical adverse events in oncology phase III trials: a comparison of the current status versus the expectations of the EORTC members

Denis Maillet[1]; Jean-Yves Blay [2,3]; Benoit You[1, 4]; Awatif Rachdi[1], Hui K. Gan[5]; Julien Péron[1,6]

1. Medical oncology department, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, F-69310, Pierre-Bénite, France;
2. Medical oncology department, Centre Léon Bérard, Lyon, France.
3. EORTC, 83 Avenue Mounier, Bruxelles, Belgium
4. EMR UCBL/HCL 3738, Faculté de Médecine Lyon-Sud, F-69600, Oullins, France; Université de Lyon, F-69622, Lyon, France.
5. Joint Austin-Ludwig Oncology Unit, Austin Hospital, Melbourne, Victoria, Australia
6. Hospices Civils de Lyon, Biostatistics unit, F-69003, Lyon, France ; CNRS UMR 5558, Biometry and Evolutionary Biology laboratory, Health and biostatistics team, Villeurbanne, France ; Université de Lyon, F-69622, Lyon, France.

Corresponding Author Information:

Dr Julien Péron

Medical oncology department, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, F-69310. 165, chemin du grand revoyet 69495 Pierre-Bénite, France

Tel: +33.(0)4.78.86.43.18

Fax: +33.(0)4.78.86.43.56

Email: julien.peron@chu-lyon.fr

Running head: Critical adverse events reporting in medical oncology trials.

Conflict of interest: The authors have no conflict of interest to disclose

Funding: None

ABSTRACT

Purpose

Determination of drug safety and tolerability is usually based on the frequency and nature of critical adverse events (AEs) rather than the frequency of all-grade toxicities. We assessed the reporting of critical-AEs in medical oncology randomized controlled trials (RCTs) and compared that to the expectations of the EORTC membership.

Methods

RCTs reports published between 2007 and 2011 were reviewed regarding the reporting of critical-AEs outcomes. Critical-AEs comprised severe-AEs (SAEs i.e. grade 3/4 AEs); lethal-AEs (LAEs); and AEs resulting in study withdrawal or in dose reduction. Study characteristics associated with better reporting of critical-AEs were investigated. In parallel, a survey was conducted among EORTC members to determine their expectations on critical-AEs reporting.

Results

The frequency of SAEs was reported in most cases (96%), but reporting thresholds were infrequently described (17%). LAEs frequency and nature were correctly reported in 161 (50%) of manuscripts; AEs leading to study withdrawal in 61 manuscripts (19%); and AEs leading to dose reduction in 43 manuscripts (13%). In contrast most EORTC members expected a comprehensive reporting of LEAs (96% of EORTC members), AEs leading to study withdrawal (86%) and AEs leading to dose reduction (70%). In multivariate analysis, LAEs frequencies were poorly reported in European trials (p=0.004). Frequencies of AEs leading to withdrawals were more frequently reported in trials funded by industry (p=0,005) and also in trials including patients with breast or urological cancers (p=0,006).

Conclusion

These findings suggest that current practice of critical-AEs reporting remains highly variable and sometimes inadequate in oncology RCTs reports. Current standards for safety reporting in randomized trials should be revised to highlight the importance of critical-AEs reporting.

INTRODUCTION

The primary aim of anticancer therapies is to improve patient survival. However, the toxicities of these therapies and their impact on patient quality of life are equally important. The precise knowledge of treatment-related adverse events (AEs), as well as the resulting impact on quality of life often influences physicians' choice of an anticancer regimen for an individual patient. The main source of information about treatment-related AEs are from the publications of randomized controlled trials (RCTs) (1). To optimize the reporting of RCTs data, the Consolidated Standards of Reporting Trials (CONSORT) guidelines provide a checklist of essential items that should always be reported (2). In 2004, the CONSORT guidelines were extended to include 10 recommendations for toxicity reporting (3). However toxicity reporting in RCTs remains suboptimal in both oncology and non-oncology trials (4, 5, 6, 7, 8, 9, 10, 11, 12).

Although a description of all emergent toxicities is important and relevant, determination of overall drug toxicity and tolerability are usually based on critical-AEs such as Severe-AEs (SAEs), Lethal-AEs (LAEs), or those resulting in study withdrawal or dose reduction. Accurate reporting of these data is therefore essential.

In this study, we systematically reviewed the reporting of critical-AEs in all oncology RCTs reports published between 2007 and 2011. RCTs characteristics associated with a better reporting of critical-AEs were also investigated. Members of the EORTC (European Organization for Research and Treatment of Cancer - eortc.eu) network were invited to an online survey regarding their expectations on critical-AEs reporting in phase III reports. The results of this survey were compared with the current status of critical-AEs reporting.

METHODS

Study Selection

We identified using MEDLINE via PubMed (http://www.pubmed.gov) all English publications of RCTs assessing systemic anti-cancer therapies published between January 2007 and December 2011. We selected 10 major journals in which oncology RCTs are frequently published: *Annals of Oncology*; *British Journal of Cancer*; *Breast Cancer Research and Treatment*; *Cancer*; *European Journal of Cancer*; *Journal of Clinical Oncology*; *Journal of the National Cancer Institute*; *Lancet*; *Lancet Oncology*; and *New England Journal of Medicine*.

Exclusion criteria were: pediatric studies; treatment with radiotherapy or surgery only; phase I, II, or IV trials; supportive care, palliative care or prevention trials; meta-analyses, overviews, or publications using pooled data from two or more trials; and secondary reports of previously published trials (13, 14, 15).

Data extraction and quality assessment

All manuscripts were reviewed independently by two investigators (DM and JP) for eligibility. If eligible, data were independently extracted by both invetigators. Discrepancies were resolved by consensus. For the purposes of data extraction, SAEs were grade 3/4 AEs as defined by recognized toxicity grading scales such as the CTCAE scales. The reporting of SAEs, LAEs, AEs leading to study withdrawal frequencies, and AEs leading to dose reductions was assessed for each manuscript. For all-grade AEs and SAEs, the invetigators noted whether a reporting threshold (i.e. the frequency below which an adverse event was deemed sufficiently infrequent that it would not be explicitly reported in the study) was explicitly stated.

The description of the nature/symptomatology of critical-AEs was also collected. A critical-AEs outcome was considered to be correctly described when both frequency and nature/symptomatology of the critical-AEs were reported in the manuscript.

Analysis

We explored whether critical-AEs reporting was influenced by: funding characteristics (solely or partially sponsored by industry); geographic regions; type of investigational therapy; year of publication; journal impact factor; the result of primary outcomes (positive or negative study); the treatment line (adjuvant or metastatic); and tumor type. Univariate and multivariable logistic regression analyses were used to identify factors associated with a correct description of critical-AEs outcomes (especially LAEs, AEs leading to study withdrawal, and AEs leading to dose reductions).

Questionnaire to EORTC members regarding AE reporting

The EORTC membership was invited to an online survey of 11 questions regarding their expectations on AEs reporting (Table 3). We divided the survey in two distinct sections. The first section (5 questions) was related to the reporting of SAEs (i.e. grade 3/4 AEs). The second section was related to the reporting of critical-AEs outcomes (AEs leading to death, study withdrawal or dose reduction). This survey was approved by the EORTC executives (Denis Lacombe) and validated by the EORTC board in July 2013. The respondents had no access to the preliminary results of the survey before the end of the questionnaire. After six months and one reminder, we closed the survey and centralized all responses for analysis.

RESULTS

Characteristics of selected RCTs

From the 739 trials initially screened, a total of 325 RCTs were included in this analysis (Figure 1). Characteristics of selected RCTs are presented in Table 1. The number of RCTs published in the 10 selected journals was broadly stable between 2007 and 2011. The most commonly enrolled tumor types were lung cancers (n=72, 22% of RCTs); breast cancers (n=74, 23%), colon/rectum cancers (n=48, 15%) and urinary tract system cancers (n=34, 11%). Most commonly, the investigational agents were cytotoxic chemotherapy in 55% of the RCTs and molecular targeted therapy in 32% of the RCTs. The frequency of targeted agents increased from 17% in 2007 to 44% of the RCTs in 2011 (p<0,0001). Forty percent of the trials were positive based on the stated primary endpoint measure and this frequency was stable over time. Seventy-one percent of treatments were tested for metastatic disease. The majority of RCTs were sponsored by industry (61%). Seventy-four percent of articles were published in journals with an impact factor higher than 10. Trials published in high impact-factor journals were more frequently positive (75% in journals with impact factor> 20). Sixty seven percent of articles were published in three journals *(Annals of Oncology, Journal of Clinical Oncology, and New England Journal of Medicine)*. The majority of articles (43%) were from European trials, with 18% from North America trials and 28% being intercontinental. However, journals with higher IF were most likely to publish intercontinental RCTs (IF > 20: 42%; IF between 20 and 40: 30%; IF < 20: 15%, p-value = .001). Conversely, trials performed outside Europe and North America were less frequently published in journals with and IF>20.

Description of all grade and grade 3/4 AE reporting in phase III trials manuscripts

Among the 325 RCT reports, the majority (96%) provided data about SAEs, and 61% reported data about all-grade AEs. The reporting thresholds were explicitly reported in 60 manuscripts (18%) for all grades AEs and in 56 manuscripts (17%) for grade severe AEs. The median values of reporting thresholds for all-grade and severe AEs were 10% (range 0% to 25%) and 3% (range 0% to 20%) respectively. In nine manuscripts (3%), the frequency thresholds were different for all grade AEs and grade 3/4 AEs. Moreover in 88 RCTs (27%), authors clearly specified that the reported AEs were at least possibly related to the trials drugs whilst in 37 RCTs (11%) the reported AEs were not necessarily related to the trials drugs. Finally in 200 RCTs (62%), it was not clear whether or not the reported AEs were those related to the trials drugs.

Reporting of LEAs, AEs leading to study withdrawal or dose reduction in phase III trials manuscripts

A correct description of LAEs, AEs resulting in study withdrawal or dose reduction was infrequent (Table 2). Only 50% of the manuscripts correctly described the frequency and nature of LAEs encountered in the study. Another 23% reported the frequency of LAEs but did not provide information on nature/symptomatology of LAEs, whilst 27% provided no information on LAEs. A multivariate analysis revealed that the conduction of a trial in Europe was an independent factor associated with a less frequent reporting of LAEs frequency (p-value = 0.004) (Table 3). There was no factor predicting a correct description of LAEs nature/symptomatology (Table A in appendix).

Frequency and nature of AEs leading to study withdrawal were adequately reported in only 19% of manuscripts (Table 2). Another 55% reported the frequency of these AEs but without an adequate description of their natures. A total of 26% did not report on AEs leading to study withdrawal at all. A higher frequency of AEs leading to withdrawal was independently associated with industrial funding (p-value = 0.0051) and with breast or urological cancer types (p-value = 0.0061, Table B in appendix). There was no factor predicting a correct description of AEs leading to drug discontinuation nature (Table C in appendix).

Only 10% of the manuscripts correctly described AEs leading to dose reductions (Table 2). Another 26% reported the frequency of AEs leading to dose reduction but without any description on their nature. Finally, 61% of the manuscripts did not report the number of dose reduction due to AEs. The dose reduction frequencies were independently more frequently reported in studies testing a cytotoxic chemotherapy (p-value = 0.0006) (Table D, appendix). Studies testing a cytotoxic chemotherapy (p-value = 0.028) was also predictive of a correct description of AEs leading to dose reduction nature (Table E in appendix).

Questionnaire within EORTC membership about critical-AEs reporting

Between September 2013 and February 2014, an online survey was sent to the EORTC membership regarding their expectations about AEs reporting. The survey was sent to a total of 3323 EORTC members from 304 different institutions. In March 2014, we closed the survey and collected a total of 210 responses from 156 different institutions (51%) in 29 different countries (Figure 2). Among the participant of this survey, 142 were oncologist / radiotherapist / organ specialists M.D (68%), 55 were technical specialist M.D (26%) and 13 were non M.D (7%). Adequate reporting of SAEs (Grades 3/4) was important to a majority

with 172 (82%) EORTC members expecting a table dedicated to grade 3/4 AEs to be provided in addition to the description of all grade AEs (table 4). Only 58 (27%) EORTC members expected a focus only on severe AEs considered by investigators to be at least possibly related to the treatment. The reporting threshold for grade 3/4 AEs should not exceed 0%, 3%, 5%, 10% and 20% for 80 (38%), 44 (21%), 59 (28%), 20 (10%) and 7 (3%) EORTC members respectively. The reporting threshold for all-grade AEs should not be below than 0%, 3%, 5%, 10% and 20% for 41 (20%), 36 (17%), 73 (35%), 45 (21%), 15 (7%) EORTC members respectively.

Similarly, the vast majority of EORTC investigators felt that it was important to systematically specify the LEAs frequency in each phase III reports (99%) and also to describe the nature of LAEs (96%). A majority of EORTC members also felt that the frequency of AEs leading to study withdrawal should always be reported (98%) and that the nature of these AEs should always be described (86%). Finally, AEs leading to dose reduction frequency and nature should always be reported for 88% and 70% of the EORTC members respectively. Thus, a very substantial gap exists between the expectations of the EORTC members for critical-AEs reporting and the current status of critical-AEs reporting (Figure 3).

DISCUSSION

The reporting of AEs in RCTs offers an unique opportunity to assess and compare the tolerance of treatments. It represents a key source of information for therapeutic decision and also significantly influences new drug development. In the present study, we assessed the reporting of critical-AEs in oncology phase III reports published between 2007 and 2011. These results confirm that despite the publication of the 2004 CONSORT extension, reporting of toxicity remains suboptimal. LAEs were clearly described in only 50% of reports; AEs leading to study withdrawal in 19% of reports; and AEs leading to dose reduction was provided in 13% of reports. Given these results, one may wonder how clinicians can assess the real benefit/toxicity ratio of novel anticancer therapies.

These results are even more startling when contrasted with the expectations of clinical researchers in a large clinical research network such as the EORTC. It is clear that a very substantial gap exists between the expectations of the EORTC members for critical-AEs reporting and the current status of critical-AEs reporting.

Only 6.3% of the total number of EORTC members responded to our survey, such rates of response are in line with many similar survey-based studies reported response rates (17, 18, 19). However, we received responses from more than 50% of the different institutions affiliated to EORTC and this from 29 different countries. The responses are likely to be similar among investigators working daily in a same center, so we believe the outcomes to our survey are probably a good reflection of what many trialists think. Moreover it is possible that the wording of the survey (words like “always”) could biased the results and explained in parts the discrepancy between what EORTC investigators say they want and what they actually do. But when we wrote this survey, we were seeking a clear cut opinion of EORTC investigators concerning their expectations on critical-AEs reporting precisely to highlight the

discrepancies between the reality and the wishes of EORTC investigators. This gap might be partly explained by the constraints of RCT reports redaction, like the limited number of words allowed in a manuscript, the willingness to highlight the results concerning the primary endpoint and not necessarily the treatment toxicities and also the difficulty to report clearly large and complex data such as AE data. Moreover it might be difficult for investigators to determine whether a critical-AEs such as LAEs is related to a toxicity of the study treatment or to another cause.

Improvement in reporting of critical-AEs is clearly needed. This will require to strike a balance between a comprehensive reporting of AEs and the risk of over-reporting. Certain areas of AEs reporting are routinely under-reported and it presents opportunities for rapid improvements. The reporting LAES, AEs leading to withdrawal or to dose reduction should be reported because these data reflect the tolerance of study treatment in a short manner. One could also debate whether all grade 3/4 AEs should be reported rather than those which exceed a particular reporting threshold, as rare events may still be relevant if severe or life-threatening. Based on the expectations of EORTC members and on the outcomes of the present study, some areas might be delineated, in order to improve quality of AE reporting :

- Frequencies of critical-AEs outcomes (especially LAEs, AEs leading to study withdrawal or dose reduction) should be clearly reported.

- Nature/Symptomatology of all critical-AEs should be clearly described in phase III reports.

- A table dedicated to grade 3/4 AEs, separate from the table with all grade AEs, should be considered, so clinicians could easily discriminate routine-AEs and severe-AEs.

- Two different reporting thresholds should be used for all grade AEs and grade 3/4-AEs .

- The frequency threshold for grade 3/4 AEs reporting should be as low as possible and should not exceed 5% as suggested by EORTC members. This strategy otherwise introduce a bias in reporting as many severe events are rare given the limited sample size in most trials.

- The most frequent all-grade AEs should be reported in order to distinguish the AEs prone to disturb patient's quality of life. The percentage above which all-grade AEs are reported should not be less than 10% as suggested by a majority of EORTC members.

RCTs are not the sole source of data on AEs. Toxicity data can also come from non-randomized studies (20); as data from national and international pharmacovigilance organizations and post-marketing surveillance. Meta-analyses of AEs are an interesting method to obtain a relevant insights on the tolerance and the relative risk of anticancer therapies (21). However, few examples of meta-analysis have helped to detail the toxicity profile of drugs widely used in practice (22,23). The standardization of the methods for AEs reporting might help to perform meta-analysis of AE data in the future.

CONCLUSION

These findings suggest that critical-AEs data are not comprehensively reported in oncology phase III reports, in stark contrast with the expectations of the EORTC membership. The emergence of new anticancer therapies undoubtedly provides news hope for patient care,

but proper estimations of the toxicity profiles of these new drugs are needed. Our data provides a framework of key areas of AE reporting that could be improved.

ACKNOWLEDGEMENT

Dr Julien Péron is the recipient of a grant from the Nuovo-Soldati research foundation

JY Blay : is supported by INCA-DHOS- LYRIC 4664, DevweCAn ANR-10-LABX-0061, Eurosarc FP7-278472, NetSARC, INTERSARC.

Thanks to Denis Lacombe for his participation and help to this study.

BIBLIOGRAPHY

1. Ioannidis JPA, Lau J. The impact of high-risk patients on the results of clinical trials. Journal of Clinical Epidemiology. 1997 Oct;50(10):1089–98.
2. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. Ann Intern Med. 2001 Apr 17;134(8):663–94.
3. Ioannidis JP, Evans SJ, Gøtzsche PC, O'neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Annals of internal medicine. 2004;141(10):781–8.
4. Haidich A-B, Birtsou C, Dardavessis T, Tirodimos I, Arvanitidou M. The quality of safety reporting in trials is still suboptimal: survey of major general medical journals. Journal of clinical epidemiology. 2011;64(2):124–35.
5. Pitrou I, Boutron I, Ahmad N, Ravaud P. REporting of safety results in published reports of randomized controlled trials. Arch Intern Med. 2009 Oct 26;169(19):1756–61.
6. Vries TW de, Roon EN van. Low quality of reporting adverse drug reactions in paediatric randomised controlled trials. Arch Dis Child. 2010 Dec 1;95(12):1023–6.
7. Shukralla AA, Tudur-Smith C, Powell GA, Williamson PR, Marson AG. Reporting of adverse events in randomised controlled trials of antiepileptic drugs using the CONSORT criteria for reporting harms. Epilepsy research. 2011;97(1):20–9.
8. Turner L-A, Singh K, Garritty C, Tsertsvadze A, Manheimer E, Wieland LS, et al. An evaluation of the completeness of safety reporting in reports of complementary and alternative medicine trials. BMC complementary and alternative medicine. 2011;11(1):67.
9. Smith SM, Chang RD, Pereira A, Shah N, Gilron I, Katz NP, et al. Adherence to CONSORT harms-reporting recommendations in publications of recent analgesic clinical trials: An ACTTION systematic review. PAIN. 2012 Dec;153(12):2415–21.
10. Breau RH, Gaboury I, Scales Jr CD, Fesperman SF, Watterson JD, Dahm P. Reporting of harm in randomized controlled trials published in the urological literature. The Journal of urology. 2010;183(5):1693–7.
11. Péron J, Maillet D, Gan HK, Chen EX, You B. Adherence to CONSORT Adverse Event Reporting Guidelines in Randomized Clinical Trials Evaluating Systemic Cancer Therapy: A Systematic Review. JCO. 2013 Nov 1;31(31):3957–63.
12. Sivendran S, Latif A, McBride RB, Stensland KD, Wisnivesky J, Haines L, et al. Adverse event reporting in cancer clinical trial publications. J. Clin. Oncol. 2014 Jan 10;32(2):83–9.

13. You B, Gan HK, Pond G, Chen EX. Consistency in the analysis and reporting of primary end points in oncology randomized controlled trials from registration to publication: a systematic review. J. Clin. Oncol. 2012 Jan 10;30(2):210–6.

14. Gan HK, You B, Pond GR, Chen EX. Assumptions of Expected Benefits in Randomized Phase III Trials Evaluating Systemic Treatments for Cancer. JNCI J Natl Cancer Inst. 2012 Apr 18;104(8):590–8.

15. Péron J, Pond GR, Gan HK, Chen EX, Almufti R, Maillet D, et al. Quality of reporting of modern randomized controlled trials in medical oncology: a systematic review. J. Natl. Cancer Inst. 2012 Jul 3;104(13):982–9.

16. Giro C, Berger B, Bölke E, Ciernik IF, Duprez F, Locati L, et al. High rate of severe radiation dermatitis during radiation therapy with concurrent cetuximab in head and neck cancer: results of a survey in EORTC institutes. Radiother Oncol. 2009 Feb;90(2):166–71.

17. Warren GW, Marshall JR, Cummings KM, Toll BA, Gritz ER, Hutson A, et al. Addressing tobacco use in patients with cancer: a survey of American Society of Clinical Oncology members. J Oncol Pract. 2013 Sep;9(5):258–62.

18. Lockhart AC, Brose MS, Kim ES, Johnson DH, Peppercorn JM, Michels DL, et al. Physician and stakeholder perceptions of conflict of interest policies in oncology. J. Clin. Oncol. 2013 May 1;31(13):1677–82.

19. Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. CMAJ. 2006 Feb 28;174(5):635–41.

20. Hernandez AV, Walker E, Ioannidis JPA, Kattan MW. Challenges in meta-analysis of randomized clinical trials for rare harmful cardiovascular events: the case of rosiglitazone. Am. Heart J. 2008 Jul;156(1):23–30.

21. Cortes J, Calvo V, Ramírez-Merino N, O'Shaughnessy J, Brufsky A, Robert N, et al. Adverse events risk associated with bevacizumab addition to breast cancer chemotherapy: a meta-analysis. Ann Oncol. 2012 May 1;23(5):1130–7.

22. Huang H, Zheng Y, Zhu J, Zhang J, Chen H, Chen X. An updated meta-analysis of fatal adverse events caused by bevacizumab therapy in cancer patients. PLoS ONE. 2014;9(3):e89960.

Figure 1: Se lection of randomized clinical trials in the systematic review.

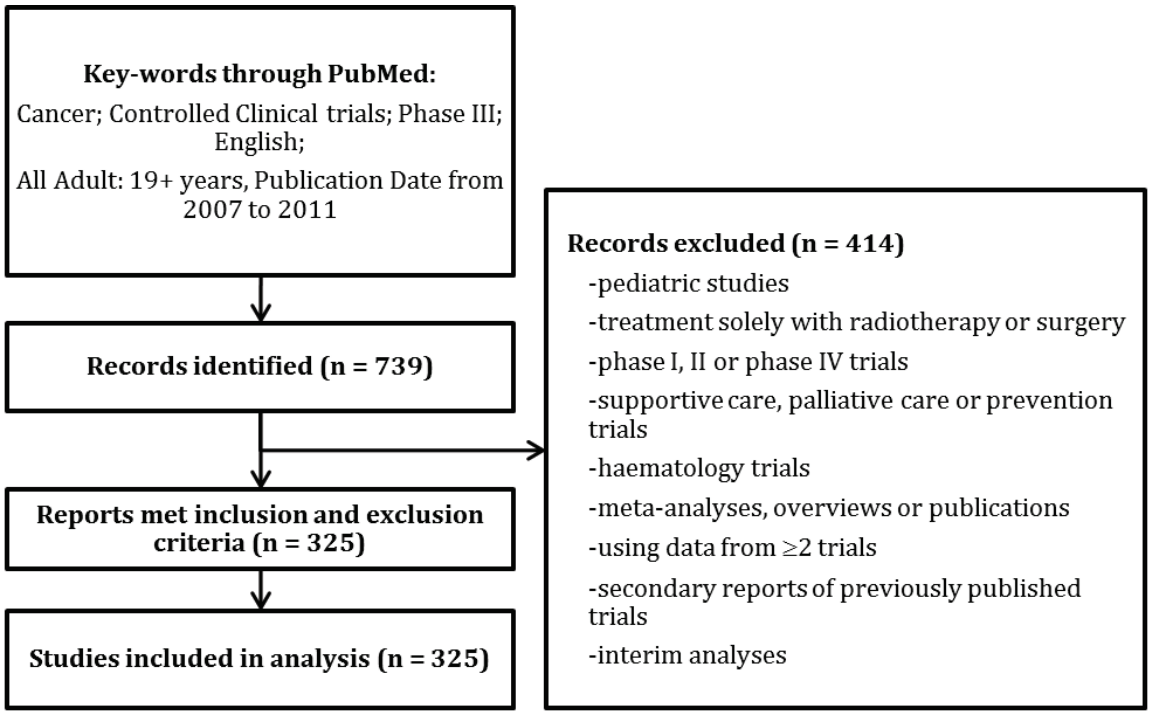


Figure 2: Map of the EORTC investigators responders to survey concerning the Critical AEs reporting

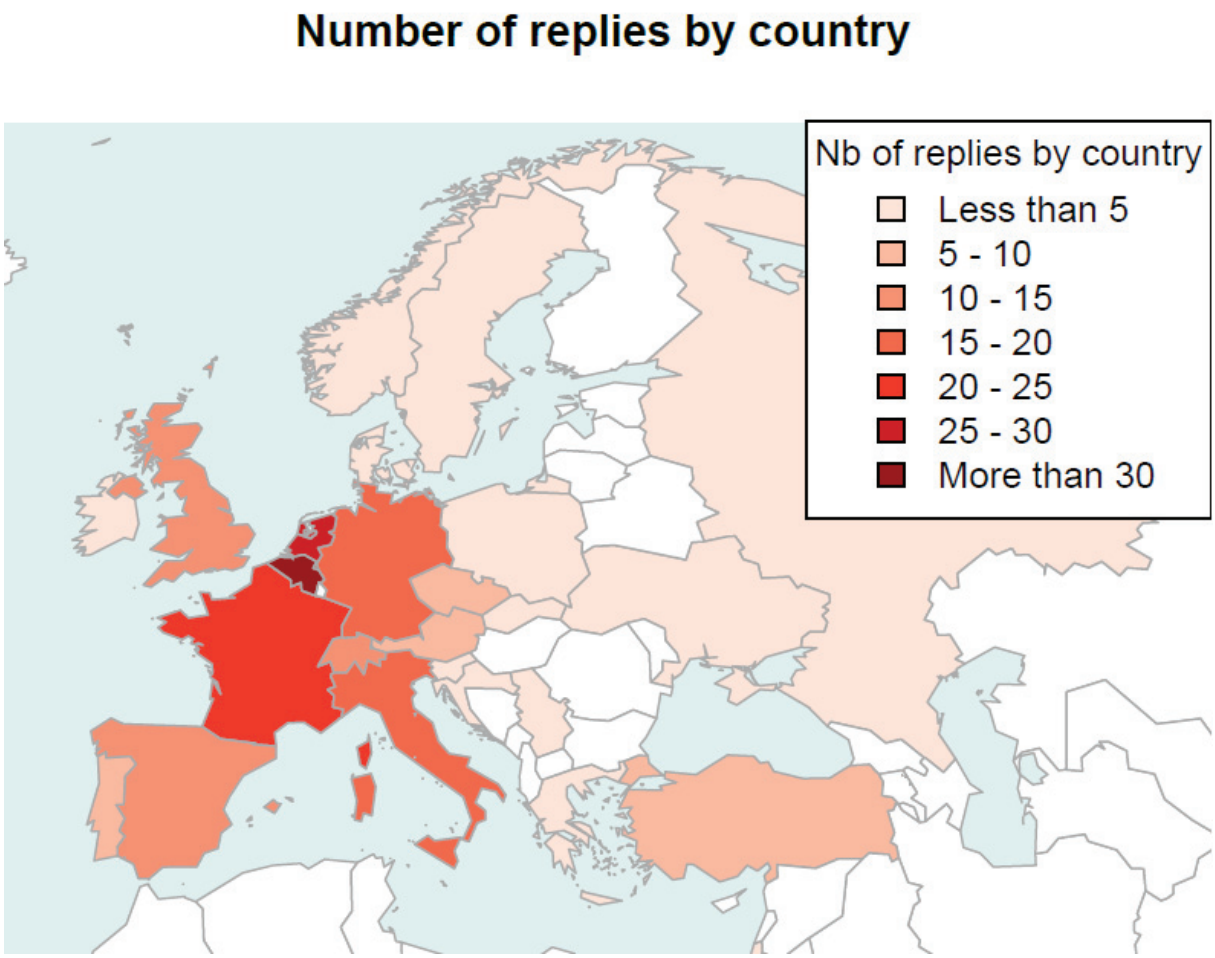
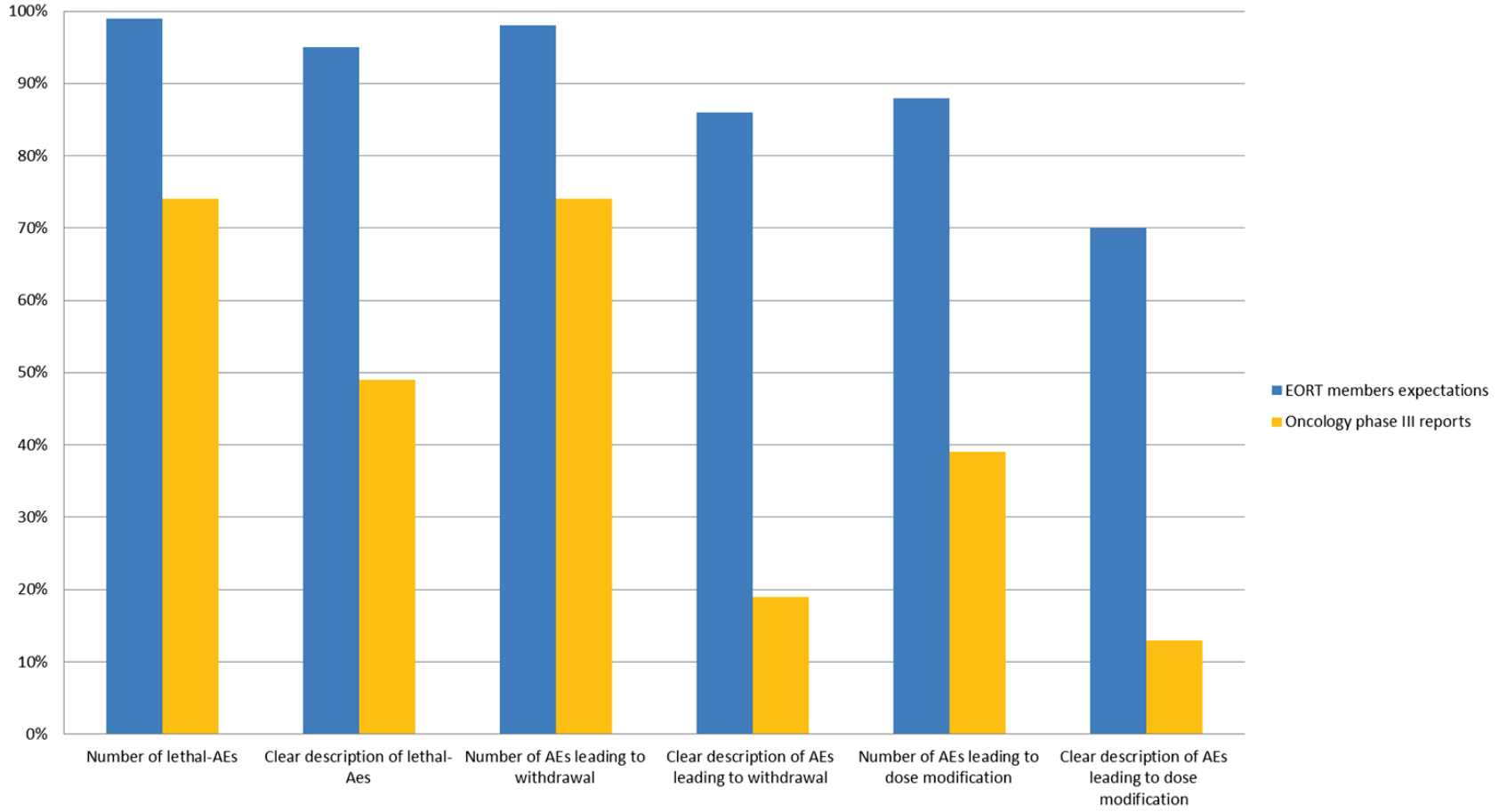


Figure 3: Comparison between EORTC expectation and observed reporting of critical-AEs related outcomes reporting in oncology phase III reports.



Abbreviations: sAEs, severe adverse events; AEs, adverse events

1

Table 1. Trial characteristics (*n* = 325)			
Study characteristics		Studies	
		n	%
Year of publication	2007	71	22
	2008	75	23
	2009	51	16
	2010	55	17
	2011	73	23
Tumor site	Lung	72	22
	Breast	74	23
	Urinary System	34	11
	Colon/Rectum	48	15
	Others	97	30
Sources of trial funding	Government/Foundation	101	31
	Funded by industry	198	61
	Funding not reported	26	8
Journal	Journal of Clinical Oncology	151	47
	Annals of Oncology	34	11
	New England Journal of Medicine	32	10
	Lancet	25	8
	European Journal of Cancer	22	7
	Other journals	61	19
Journal impact factor	<10	85	26
	10-20	183	56
	>20	57	18
Regions in which RCTs was led	Intercontinental	92	28
	North America	60	18
	Europe	141	43
	Others	32	10
Investigational therapy	Cytotoxic chemotherapy	178	55
	Hormonal therapy	27	8

	Molecular targeted therapy	103	32
	Immunotherapy	12	4
	Other	5	2
Cancer stage	Adjuvant and/or neoadjuvant	93	29
	Metastatic	232	71
Sample size	Median		492
	Interquartile range		270-803
Results of the primary outcome	Positive	131	40
	Negative	194	60
Toxicity profile conclusions*	Equivalent	104	32
	Investigational arm more toxic	136	42
	Control arm more toxic	34	11
	No Conclusion	51	16

1

2 * Conclusion of the trials authors

3 RCT = Randomized controlled trial

Table 2 Reporting of AEs-related outcomes		
Major endpoint	**Phase III reports**	
	N = 325	***%***
Lethal adverse events (LAEs)		
Number reported	237	73
• No LAE	31	10
• Presence of LAEs without any description of the LAEs	76	23
• Presence of LAEs with adequate description of the LAEs	130	40
Number not reported	88	27
LAEs correctly reported (Number reported and adequate description of the LEAS OR no LAE)	161	50
AEs relating to study withdrawal		
Number reported	240	74
• No AE leading to study withdrawal	12	4
• Presence of AE- related study withdrawal without description of the AEs	179	55
• Presence of AE- related study withdrawal with adequate description of the AEs	49	15
Number not reported	85	26
AEs relating to study withdrawal correctly reported (Number reported and adequate description of the AEs OR no withdrawal)	61	19
AEs relating to dose reduction		
Number reported	128	39
• No AE leading to dose reduction	12	4
• Presence of AE related to dose reduction without description of the AEs	85	26
• Presence of AE related to dose reduction with adequate		

description of the AEs.	31	9
Number not reported	197	61
AEs relating to dose reduction correctly reported (Number reported and adequate description of the AEs OR no dose reduction)	43	13

1

2

Table 3. Results of regression analyses of factors predicting the reporting of LAEs frequency						
		Logistic Regression				
		Univariate analysis			Multivariate analysis	
Study characteristics		Unadjusted OR (95% CI)	*p*-value		Adjusted OR (95% CI)	*p*-value
Year of publication	Continuous	1.09 (0.92-1.29)	0.32		1.02 (0. 84-1.25)	0.82
Tumor site	Lung	REF	0.069		REF	0.11
	Breast	0.50 (0.24-1.03)			0.48 (0.20-1.12)	
	Urinary System	0.79 (0.31-2.04)			1.08 (0.35-3.35)	
	Colon/Rectum	1.67 (0.63-4.43)			1.66 (0.57-4.82)	
	Others	0.67 (0.33-1.36)			0.75 (0.33-1.67)	
Presence of an industry partner	No industry funding	REF	0.078		REF	0.40
	Funded by industry	1.65 (0.97-2.81)			1.46 (0.76-2.79)	
	Unknown	0.78 (0.32-1.90)			0.86 (0.30-2.48)	
Journal impact factor	<10	REF	0.68		REF	0.54
	10-20	1.23 (0.70-2.18)			0.70 (0.35-1.38)	
	>20	2.35 (0.63-2.89)			0.64 (0.23-1.74)	
Region in which RCT was led	Intercontinental	REF	**0.00028**		REF	**0.0048**
	North America	0.65 (0.28-1.50)			1.05 (0.41-2.71)	
	Europe	0.28 (0.14-0.54)			0.35 (0.16-0.78)	
	Others	0.78 (0.27-2.23)			0.92 (0.27-3.11)	
Type of investigational therapy	Cytotoxic chemotherapy	REF	**0.018**		REF	0.24
	Hormonal therapy	0.68 (0.29-1.59)			0.83 (0.31-2.23)	
	Molecular targeted therapy	1.90 (1.04-3.47)			1.18 (0.58-2.42)	
	Immunotherapy	0.40 (0.12-1.30)			0.26 (0.07-1.01)	
	Other	0.27 (0.04-1.64)			0.30 (0.04-2.23)	

Cancer stage	Metastatic disease	REF	0.44	REF	0.64
	Neoadjuvant / Adjuvant	0.81 (0.48-1.38)		1.17 (0.61-2.26)	
Sample Size / 100	Continuous	1.01 (0.98-1.05)	0.36	1.00 (0.97-1.04)	0.79
Results of primary outcome	Negative	REF	0.12	REF	0.13
	Positive	1.17 (0.70-1.95)		0.99 (0.54-1.80)	
Conclusion of safety outcome for authors	Equivalent	0.78 (0.38-1.59)	0.31	0.94 (0.43-2.07)	0.49
	Investigational arm more toxic	REF		REF	
	Investigational arm less toxic	1.44 (0.80-2.57)		1.39 (0.73-2.66)	
	No conclusion	1.38 (0.56-3.38)		1.83 (0.65-5.15)	

1

Table 4: Survey of EORTC members concerning sAEs reporting in phase III reports.			
	Questions	**EORTC members responses (N=210)**	
		YES	**%**
1	All grade 3/4 AEs, whether or not related to the study treatment should be reported	153	73%
2	Only grade 3/4 AEs at least possibly related to the study treatment should be reported	58	27%
3	A table dedicated to grade 3/4 AEs separated from the table with all AEs, should always be added	172	82%
4	The reporting threshold for grade 3/4 AEs should not exceed		
	0%	80	20%
	3%	44	17%
	5%	59	35%
	10%	20	21%
	20%	7	7%
5	The reporting threshold for all-grade AEs should not be below than		
	0%	41	20%
	3%	36	17%
	5%	73	35%
	10%	45	21%
	20%	15	7%
6	The frequency (in %) of LAEs should always be reported	208	99%
7	The frequency (in %) of AEs leading to study withdrawal should always be reported	207	98%
8	The frequency (in %) of AEs leading to dose reduction should always be reported	186	88%
9	When they occurred, the types of AEs leading to death should always be described	201	96%

10	When they occurred, the types of AEs leading to study withdrawal should always be described	181	86%
11	When they occurred, the types of AEs leading to a dose reduction should always be described	148	70%

Abbreviations: AEs, adverse events; LEAs, lethal adverse events; sAEs: severe AEs.

Une première conclusion de cette étude est que la fréquence des événements indésirables de grade élevé était le plus souvent correctement rapportée. Néanmoins, il était souvent difficile de savoir si un seuil de fréquence déclenchait le rapport d'un événement indésirable d'une certaine nature. Quand ce seuil était décrit, il était variable d'une étude à l'autre. Les événements cliniques graves liés aux événements indésirables étaient souvent décrit en terme de fréquence (73% des manuscrits rapportaient la fréquence des décès toxiques ; 74% la fréquence des arrêts de traitement pour toxicité ; et 13% la fréquence des réductions de dose pour toxicité). La nature des événements indésirables responsables de ces événements cliniques graves était moins souvent rapportée (respectivement 50%, 19%, et 13% des manuscrits). Globalement ces chiffres étaient assez bas en comparaison des attentes des membres de l'*EORTC*.

En conclusion de ces deux études, les méthodes utilisées actuellement pour recueillir les événements indésirables, pour sélectionner les données à rapporter, et pour rapporter les données étaient hétérogènes. De plus la qualité de rédaction des manuscrits était souvent insuffisante pour comprendre les méthodes effectivement utilisées. Il n'est donc pas raisonnable d'envisager de réaliser une évaluation non biaisée de la balance bénéfice-risque des essais contrôlés randomisés en oncologie médicale uniquement à partir des données publiées dans les revues scientifiques. Cette conclusion s'applique à l'évaluation de la balance bénéfice-risque d'un essai individuel, ou dans le cadre de méta-analyses. Ce type d'analyse doit donc idéalement être réalisé sur données individuelles.

Chapitre II

II. Les critères de jugement rapportés par les patients dans les rapports d'essais contrôlés randomisés

Les critères de jugement rapportés par les patients (CRPs) permettent de mesurer l'effet d'un traitement en utilisant des variables recueillies auprès des patients eux-mêmes, sans interprétation par leurs médecins ou n'importe quelle autre personne. Ils peuvent varier en complexité, d'une réponse à une question unique jusqu'à des questionnaires contenant de multiples questions et permettant de mesurer de multiples domaines de qualité de vie ou de l'état fonctionnel des patients. Les CRPs permettent d'évaluer à la fois l'efficacité et la toxicité des traitements, puisqu'ils peuvent intégrer tout événement influençant la perception des patients de leur maladie ou de leurs traitements. Ces critères sont donc certainement informatifs pour évaluer l'effet global d'un traitement, en plus des critères classiquement utilisés en oncologie médicale comme la survie globale, la variation en taille de la tumeur, et la survenue d'événements indésirables rapportés par les investigateurs.

Pour pouvoir être intégrés dans une analyse globale de l'effet d'un traitement, les CRPs doivent donc être standardisés, ou au moins parfaitement définis. Une revue systématique des manuscrits rapportant des essais contrôlés randomisés publiés entre 2007 et 2011 et évaluant des traitements systémiques anticancéreux a donc été réalisée. Dans cette revue systématique, l'utilisation d'au moins un CRP a été recherchée. La qualité du rapport des CRPs a été évaluée en fonction de l'adhérence aux critères de qualité de rapport de ce type de variable, définis par le groupe CONSORT. Les méthodes utilisées pour recueillir, analyser et rapporter les CRPs ont également été analysées. Les résultats de cette étude ont été rapportés dans un article publié dans le journal *Annals of Oncology* [9]. Les messages principaux de cet article sont rappelés en fin de sous-chapitre, après la présentation du manuscrit en format édité par le journal *Annals of Oncology*.

Annals of Oncology 26: 231–237, 2015
doi:10.1093/annonc/mdu489
Published online 29 October 2014

Poor patient-reported outcomes reporting according to CONSORT guidelines in randomized clinical trials evaluating systemic cancer therapy

O. Bylicki[1,2], H. K. Gan[3], F. Joly[4,5,6], D. Maillet[7], B. You[7,8,9] & J. Péron[7,10,11]*

[1]Department of Pneumology, Desgenettes Hospital; [2]Department of Medical Oncology, Centre Léon BERARD, University of Lyon, Lyon, France; [3]Joint Austin-Ludwig Oncology Unit, Austin Hospital, Melbourne, Victoria, Australia; [4]INSERM, U1086, Caen; [5]Clinical Research Unit, François Baclesse Center, Caen; [6]Department of Medicine, CHU de Caen, Caen; [7]Department of Medical Oncology, Lyon-Sud Hospital Center, Hospices Civils de Lyon, Pierre-Bénite; [8]EMR UCBL/HCL 3738, Faculté de Médecine Lyon-Sud, Oullins; [9]Department of Medical Oncology, University of Lyon, Lyon; [10]Biostatistics Unit, Hospices Civils de Lyon, Lyon; [11]Biometry and Evolutionary Biology Laboratory, Health and Biostatistics Team, CNRS UMR 5558, Villeurbanne, France

Received 23 June 2014; revised 21 September 2014; accepted 30 September 2014

Background: The Consolidated Standards of Reporting Trials (CONSORT) guidance was extended in 2013 to provide a set of specific recommendations regarding patient-reported outcomes (PROs) reporting in randomized clinical trials (RCTs). There is limited data regarding how well current publications of oncology RCTs report PROs if assessed using these guidelines.

Design: All phase III medical oncology RCTs published between 2007 and 2011 were reviewed according to the 2013 PROs CONSORT recommendations and an 11-point PROs reporting quality score (PRORQS) was defined based on the criteria.

Results: The majority of trials did not report on PROs at all (201 of 325; 62%). Of the remaining 124 trials, the mean PRORQS score was 5.0 on an 11-point scale. The items related to methods of PROs collection and analysis were poorly reported (Description of the prespecified PRO hypothesis: 26% of RCTs; methods for PRO data collection (paper, telephone, electronic, other): 16%; statistical approaches for managing missing data: 37%). The only factor significantly associated with improved PROs reporting was where PROs reporting was the subject of a dedicated secondary manuscript, as was the case in 36 of the 124 (29%) of RCTs.

Conclusion: Despite their clinical relevance, our findings show that some aspects of PROs reporting may greatly be improved, especially critical methodological aspects of PROs collection and analysis. The exceptions were where PROs were described in PROs-specific secondary publication. Use of the 2013 PROs CONSORT extensions should be encouraged and their effects on PROs reporting subsequently reassessed.

Key words: randomized clinical trials, quality of life, patients-reported outcomes, reporting quality

introduction

Randomized clinical trials (RCTs) are considered to be the gold standard in assessing medical interventions. Patient-reported outcomes (PROs) are outcomes reported by the patients themselves, without the interpretation of the patient's responses by a physician or anyone else [1]. PROs measures may vary in complexity, from a single-item question about unique concept, up to multi-item instruments for measuring quality of life, and multiple domains of functional status.

In the field of medical oncology, the primary end points are frequently some measure of patient survival. PROs are complementary to evaluate both benefits and harm of treatments tested in RCTs. They can arguably be considered as important as patient survival, especially since oncology drugs have lower therapeutic indices compared with drugs in other therapeutic areas [2]. PROs data are increasingly used in modern RCTs [3, 4]. They provide the most direct evidence of whether the prescribed treatments actually improve patients' general well-being, tumor-related symptoms as well as treatment side-effects [5]. As there are still many methodological challenges on data collection, appropriate timing of assessment, adequate statistical hypothesis and analysis as well as outcomes reporting and interpretation [6–8], data issued from PROs are not totally accepted. Not surprisingly then, evidence suggests that reporting

Correspondence to: Dr Julien Péron, Department of medical oncology, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, F-69310. 165, chemin du grand revoyet, 69495 Pierre-Bénite, France. Tel: +33-4-78-86-43-18; Fax: +33-4-78-86-43-56; E-mail: julien.peron@chu-lyon.fr

of PROs remains sub-optimal across both oncology and non-oncology RCTs [9–11].

To improve the reporting of PROs in oncology RCTs, a recent extension to the CONSORT statement regarding PROs reporting was published [12]. It included five 'extension' statements to the 2010 CONSORT guidance [13] that each addresses a key reporting item for quality reporting from all RCTs using PROs. In addition, components of the existing 2010 CONSORT guidance that were critically relevant to PROs reporting were expanded by six 'elaboration' statements. This review was carried out to evaluate the quality of PROs reporting in recent oncology RCT reports, according to the 2013 PROs-specific CONSORT extension, thereby establishing the current adequacy of PROs reporting. In addition, we investigated manuscripts' characteristics associated with better quality in PROs reporting.

methods

trial selection

We searched Medline via PubMed (http://www.pubmed.gov) to identify all primary report of RCTs assessing systemic therapies for solid tumors and including at least one PROs published between January 2007 and December 2011 in 10 English language journals where oncology RCTs are frequently published: *Annals of Oncology*; *British Journal of Cancer*; *Breast Cancer Research and Treatment*; *Cancer*; *European Journal of Cancer*; *Journal of Clinical Oncology*; *Journal of the National Cancer Institute*; *Lancet*; *Lancet Oncology*; and *New England Journal of Medicine*. The search was carried out using the terms 'randomized' and 'cancer' as keywords and 'English' plus 'clinical trials' or 'randomized controlled trial' as limits. Exclusion criteria were: lack of PROs reporting; pediatric studies; hematologic trials, phase I, II, or IV trials; meta-analyses, overviews, publications using pooled data from two or more trials; and secondary reports on previously published trials [14, 15]. The presence of at least one PROs (patient-reported HRQL data; patient-reported symptom data; patient-reported satisfaction data) was researched by reviewing of the full manuscripts. We also searched Medline via PubMed for secondary reports of PROs, published secondary or as companion paper to the primary report, using the title and the authors' names of all the article citing the primary report in their references (Figure 1).

definition of trial characteristics

Trials were considered as industry funded if a RCT received any form of industry funding with the exception of those studies where only drug(s) was provided but no funding. Trials were considered intercontinental when patients from more than one continent were included. The journal impact factor (IF) was the mean journals IF between 2007 and 2011. Only the journal IF of the manuscript reporting the main RCT results (primary manuscript) was used for analyses. The type of investigational therapy was the main drug assessed in trials; it could be either tested in combination with other drugs or as a single-agent therapy.

development of a quantitative scoring system for quality of patient-related outcomes reporting

A standardized data extraction form was used by two authors (JP and DM) to capture the data in this review. The two authors examined each article. Where there was a discrepancy in responses to a given item, the authors resolved this by consensus evaluation.

The extraction form included a number of guidelines to ensure homogenous data extraction for those recommendations potentially open to interpretation. PROs relevant domains were considered adequately identified (i.e. item P2b) if one or several domains of primary interest were identified

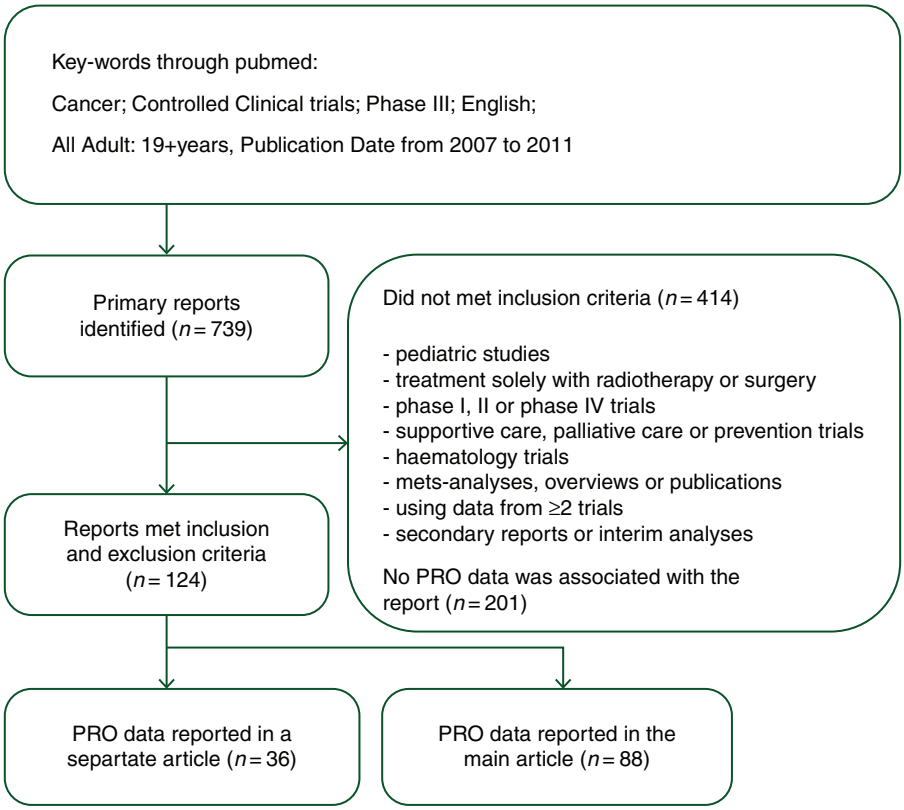


Figure 1. Selection of randomized clinical trials in the systematic review.

by the authors of the reports. The description of the number of PROs outcome data at subsequent time points (i.e. item E13a) was considered adequate if authors provided at least one assessment of the number of PROs data collected in addition to the baseline evaluation. Item E17a was considered correctly reported when the results were reported for all the symptoms or dimensions included in the PROs tool used.

A PROs reporting quality score (PRORQS) based on the 2013 CONSORT extension was defined by two of the authors (OB and JP). The score was based on 11 items derived from the 11 recommendations (either extension or elaboration of the 2010 CONSORT statement) (Table 1). Each item was scored 1 if it was adequately reported or 0 if it was not clearly reported or not reported at all; each item was weighted with equal importance. A 1-point difference in mean score between two groups was considered meaningful, as it would suggest the failure of one group to report on one more item compared with the other group.

In addition, data were also captured regarding the methods of PROs data collection and PROs results reporting. When PROs results were included in the main article, the space (expressed as the percentage of lines devoted to PROs relative to total section size) devoted to PRO in the Methods and Results sections were collected. The type of PROs used was collected (quality-of-life data and/or symptom data). Quality-of-life data were defined as those referencing a multidomain concept representing the patient's general perception of the effect of illness and treatment on physical, psychological, and social aspects of life. Symptom data were defined as any subjective evidence of a disease, health condition, or treatment-related effect that can be noticed and known only by the patient. The main statistical method used to compare PROs between the randomized groups was also collected.

statistical analysis

Categorical trials characteristics among RCTs including at least one PROs and RCTs without PRO were compared using χ^2 tests or fisher's exact tests as appropriate.

The PRORQS was the sum of the number of items that were adequately reported (Table 1) and expressed as an integer between 0 and 11. PRORQS scores were summarized using descriptive statistics such as mean, confidence intervals (CIs) and range. Single-item frequencies were compared between categories using χ^2 tests.

Univariate and multivariable linear regression analyses were used to identify factors associated with higher PRORQS. The following trial characteristics were investigated: year of primary report publication, tumor site, presence of an industrial funding, primary report journal IF, geographic

Table 1. Quality of PROs reporting, rating using items from the 2013 extensions of the CONSORT statement ($N = 124$)

Descriptor of the 2010 CONSORT criteria	Descriptor of the 2013 PRO-specific extension or elaboration PRO-specific extensions are prefaced by the letter P PRO-specific elaborations are prefaced by the letter E	Number of trials in which item was adequately reported, *n* (%)
1b—Structured summary of trial design, methods, results, and conclusions	**P1b** Identification of the PROs in the abstract as a primary or secondary outcome	47 (28)
2a—Scientific background and explanation of rationale	**E2a** Background and rationale for PROs assessment	53 (43)
2b—Specific objectives or hypotheses	**P2b** Identification of the PROs relevant domains	80 (65)
	Statement of the PROs hypothesis	32 (26)
	Statement of the PROs analysis power	12 (10)
6a—Completely defined prespecified primary and secondary outcome measures	**P6a** Evidence of PROs instrument validity	45 (36)
	Reference of the PROs instrument	104 (84)
	Statement of the person completing the PROs	47 (38)
	Methods of data collection (paper, telephone, electronic, other)	20 (16)
12a—Statistical methods used to compare groups	**P12a** Statistical approaches for dealing with missing data are explicitly stated	46 (37)
13a—For each group, the numbers of participants who were randomly assigned, received intended treatment and were analyzed for the primary outcome	**E13a** Description of the number of PROs outcome data at baseline and at subsequent time points	64 (52)
	At baseline	76 (61)
	At subsequent time points	87 (70)
15—A table showing baseline demographic and clinical characteristics for each group	**E15** Including baseline PROs data when collected	49 (40)
16—For each group, number of participants (denominator) included in each analysis	**E16** Required for PROs results	60 (48)
17a—For each primary and secondary outcome, results for each group, the estimated effect size, and its precision	**E17a** For multidimensional PROs results from each domain	54 (43)
20/21—Trial limitations, addressing sources of potential bias and generalizability of the trial findings	**P20/21** PROs-specific limitations and implications for generalizability and clinical practice	43 (35)
22—Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	**E22** PROs data should be interpreted in relation to clinical outcomes including survival data	75 (60)

The 11 items of the 2013 PROs-specific CONSORT extension are in bold. PROs-specific extensions are prefaced by the letter P; PROs-specific elaborations are prefaced by the letter E.
PROs, patient-reported outcomes.

region, type of investigational therapy, cancer stage and presence of a PROs dedicated secondary manuscript. The multivariable model included all the above-mentioned covariates. No covariate selection was carried out because it was deemed desirable to include as many factors associated with reporting quality as possible. Covariates were considered statistically associated with PRORQS if the associated *P*-value was <0.05. When PROs results were included in the main report of RCT, the relationships between the place devoted to PROs, and PRORQS was investigated using univariate linear regression analyses.

It was also hypothesized that primary manuscripts from the same journal might have PRORQSs that were more closely correlated to each other than manuscripts from different journals. Therefore, mixed-effect models were used as a supportive regression analysis with the incorporation of 'primary manuscript journal' in the model as a random effect. Results were similar to the linear model without the assumption of correlation. Therefore, for more simplicity, only the results of the linear model are reported here. Statistical analyses were carried out using R software (http://www.R-project.org/). All statistical tests were two-sided.

results

characteristics of selected randomized, controlled trials

From the 739 trials initially screened, a total of 325 RCTs reporting phase III trials in the field of medical oncology were identified. Data from this dataset have previously been reported regarding the quality of adverse events reporting [14]. Among them, 124 (38%) included at least one PROs and were included in this analysis (Figure 1). The number of published RCTs with PROs per year was stable ($P = 0.29$) (Table 2). The presence of at least one PROs was more frequent among RCTs including patients with metastatic disease than among RCTs including patients in adjuvant or neoadjuvant setting (44% versus 27%, $P = 0.00015$). PROs were also more frequent among RCTs reported in high IF journals (61% for IF >20, 36% for IF between 10 and 20, 27% for IF <10, $P = 0.00013$). The presence of PROs was not statistically associated with the source of funding or the region in which RCTs were led (data not shown).

Table 2. Trial characteristics ($N = 124$)

Study characteristics	Trials	
	n	%
Year of primary report		
2007	27	22
2008	23	19
2009	26	21
2010	21	17
2011	27	22
Tumor site		
Lung	35	28
Breast	17	14
Urinary system	16	13
Colon/rectum	17	14
Others	39	31
Sources of trial funding		
No industry funding	35	28
Funded by industry	83	67
Unknown	6	5
Primary manuscript journal		
Journal of Clinical Oncology	52	42
Annals of Oncology	9	7
New England Journal of Medicine	20	16
Lancet	15	12
Lancet Oncology	11	9
Other journals	17	14
Primary manuscript journal impact factor		
<10	23	19
10–20	66	53
>20	35	28
Regions in which RCTs was led		
International	41	33
North America	16	13
Europe	51	41
Others	16	13
Type of investigational therapy		
Cytotoxic chemotherapy	55	44
Hormonal therapy	16	13
Molecular targeted therapy	47	38
Immunotherapy	3	3
Other	3	3
Cancer stage		
Adjuvant and/or neoadjuvant	20	16
Metastatic	104	84

RCT, randomized clinical trial.

methods of patient-reported outcomes reporting

A secondary PROs-dedicated manuscript was identified for 36 (29%) of the RCTs. The methods and results related to the PROs were included in the main manuscript for the remaining 88 RCTs. When PROs was reported in the main manuscript, the median percentage of the space allocated to the PROs in the Methods section was 16% (interquartile range: 8–22), and the median percentage of the space allocated to the PROs in the results section was 10% (interquartile range: 4–19) (Table 3). Overall, PROs was most frequently a measurement of the patients quality of life ($n = 88$, 71%), a measurement of the patients symptoms ($n = 22$, 18%) or both ($n = 11$, 9%). The instrument used to assess patient quality of life was most often disease-specific ($n = 57$, 58%) or at least cancer-specific ($n = 35$, 35%). Longitudinal models and time to event comparisons were used to compare PROs between groups in, respectively, 26% and 33% of the RCTs, while time point comparisons were used in 34% of the RCTs.

rating of overall quality score

The mean PRORQS for all items was 5.0 on an 11-point scale [range: 0–11, 95% CI of the mean 4.4–5.5], including 17 publications (14%) having a PRORQS ≤1. Four trials (3%) were found with a score of 11. All were reported in a secondary PROs-specific manuscript, and three were pivotal trials with positive results. Ten manuscripts reported correctly all the five extensions of the CONSORT statement (P1b, P2b, P6a, P12a, P20/21). The most frequently reported item included in the 2013 CONSORT extension was the identification of the PROs relevant domains (item P2b,

Table 3. Presentation of PROs results

Study characteristics	Trials	
	N	%
Secondary PROs report		
Yes	36	29
No	88	71
Journal impact factor of PROs-specific manuscript ($n = 36$)		
<10	26	72
10–20	9	25
>20	1	3
Space allocated to PROs in the Methods section ($n = 88$)[a]		
Absolute line number (median, IQR)	19 (8–22)	
Percentage of the Methods section (median, IQR)	16 (7–18)	
Space allocated to PROs in the Results section ($n = 88$)[a] (median, IQR)		
Absolute line number (median, IQR)	10 (4–22)	
Percentage of the Results section (median, IQR)	10 (4–19)	
PROs stated as a primary or secondary end point		
Primary end point	1	1
Secondary end point	121	98
Unclear	2	2
Type of PROs		
HRQL	88	71
Symptom scale	22	18
Both	11	9
Unclear	3	2
Type of HRQL scale ($n = 99$)[b]		
Non-cancer-specific	7	7
Cancer-specific	35	35
Site-specific	57	58
Type of PROs main analysis		
Longitudinal models[c]	32	26
Time point comparison	42	34
Time-to-event comparison	18	15
Descriptive	9	7
Other	4	3
Unclear	19	15

[a]Only manuscripts not dedicated to PROs were included in this analysis.
[b]Only manuscripts reporting HRQL data were included in this analysis.
[c]Other than time-to-event comparison.
PROs, patient-reported outcomes; IQR, interquartile range; HRQL, health-related quality of life.

correctly reported in 65% of the RCTs) (Table 1). A correct description of the prespecified PROs hypothesis was reported in 26% of the reports, and the description of the analysis power in 10% of the reports. The identification of the PROs in the abstract as a primary or secondary outcome (item P1b) was done in 28% of the reports. Also, 16% reported methods for PROs data collection (paper, telephone, electronic, other). Statistical approaches for managing missing data (item P12a) were adequately described in 37% of RCTs. Adequate description of each domain result for multidimensional PROs (item E17a) was found in 43% of manuscripts. PROs-specific limitations (item P20/21) were discussed in 35% of manuscripts. Among the 11 items included in the PRORQS, none was correctly reported by more than 70% of RCT reports.

factors associated with reporting quality

The multivariable regression model subsequently revealed that the presence of a secondary manuscript dedicated to PROs outcome was an independent predictor of higher PRORQS ($P < 0.001$). The estimated effect on PRORQS was adjusted by year of primary report, tumor site, sources of trial funding, primary report IF, region in which RCT was led, type of investigational therapy, and cancer stage. The extent of space dedicated to PROs in the methods and results section when PROs was reported in the main manuscript ($n = 88$), was not associated with an improved PRORQS (univariate linear regression $P = 0.32$). When a secondary PRO-dedicated manuscript was present, the PRORQS increased from a mean of 3.5 to a mean of 9.0, an increase of nearly 5.5 point (Table 4). Reporting in a PROs-specific manuscript is nearly 2.5 times better than if just reported in a primary article.

discussion

Our review showed that, judged against the practice espoused in the recent CONSORT guidelines for PROs reporting, there is scope for significant improvement in PROs reporting in current oncology RCTs. The majority of manuscripts did not report on PROs at all (201 of 325; 62%). When some PROS elements were reported, there was a wide variation in quality of reporting, and several deficiencies were commonly seen. Importantly, a correct description of the prespecified PROs hypothesis was reported in only 26% of the reports, and the description of the analysis power in 10% of the reports. The poor reporting of such important methodological aspects might undermine the confidence of readers in PROs result interpretation.

PRO are direct measures of clinical benefit able to assess any aspect of health status from the patient's perspective and without interpretation by health-care providers or others. One indication of their importance lies in the fact that the Food and Drug Administration (FDA) recently published guidance to inform medical product developers, clinicians, and researchers regarding how the FDA reviews and evaluates existing, modified, or newly created PROs instruments used to support claims in approved medical product labeling [1, 16]. Further recognition of their relevance came from the recent publication of an extension to the CONSORT statement focusing specifically on PROs reporting [12].

The reason for poor reporting of PROs in oncology RCTs is difficult to ascertain. Poor PROs reporting may result from the assumption that PROs outcomes are less important to readers than other end points such as time to death, tumor progression, or other clinician-reported outcomes. The subjective nature of PROs might limit the interpretation of PROs data, arguing for the further development of quality standard for PROs collection, analysis, and reporting. Another obvious reason for poor reporting of PROs might be related to space limitations required by journals, because PROs are essentially used as secondary outcomes in medical oncology. The space devoted to PROs

Table 4. Results of regression analyses of factors predicting 2013 PRORQS (0–11 scale)

Study characteristics	Mean PRORQS (standard error)	Linear regression			
		Univariate analysis		Multivariate analysis	
		Estimate[a] (standard error)	*P*-value	Estimate[a] (standard error)	*P*-value
Year of primary report					
Continuous	–	0.2 (0.2)	0.24	0.2 (0.1)	0.12
Tumor site					
Lung	3.9 (2.6)	Reference	0.025	Reference	0.66
Breast	5.2 (3.4)	1.3 (0.9)		0.2 (0.7)	
Urinary system	6.4 (3.1)	2.4 (0.9)		−0.6 (0.7)	
Colon/rectum	3.9 (3.2)	−0.0 (0.9)		0.2 (0.5)	
Others	5.6 (3.3)	1.7 (0.7)		0.5 (0.8)	
Sources of trial funding					
No industry funding	5.3 (3.2)	Reference	0.52	Reference	0.43
Funded by industry	4.9 (3.2)	−0.5 (0.6)		−0.5 (0.5)	
Unknown	3.8 (1.9)	−1.5 (1.4)		0.3 (1.0)	
Primary manuscript journal impact factor					
<10	4.9 (2.6)	Reference	0.011	Reference	0.63
10–20	4.3 (3.0)	−0.6 (0.7)		−0.5 (0.5)	
>20	6.3 (3.5)	1.4 (0.8)		−0.4 (0.7)	
Region in which RCT was led					
Intercontinental	5.4 (3.3)	Reference	0.66	Reference	0.88
North America	4.4 (2.8)	−1.0 (0.9)		−0.4 (0.7)	
Europe	5.0 (3.2)	−0.4 (0.7)		0.0 (0.5)	
Others	4.5 (3.2)	−0.9 (0.9)		−0.3 (0.7)	
Type of investigational therapy					
Cytotoxic chemotherapy	5.0 (3.1)	Reference	0.86	Reference	0.77
Hormonal therapy	4.9 (3.2)	−0.1 (0.9)		−0.2 (0.5)	
Molecular targeted therapy	4.8 (3.3)	−0.3 (0.6)		−0.5 (0.5)	
Immunotherapy	6.3 (3.8)	1.3 (1.9)		−1.1 (1.3)	
Other	6.3 (3.5)	1.3 (1.9)		0.4 (1.2)	
Cancer stage					
Metastatic disease	4.7 (3.1)	Reference	0.025	Reference	0.72
Neoadjuvant/adjuvant	6.4 (3.4)	1.7 (0.8)		−0.2 (0.6)	
Secondary PRO manuscript					
No	3.5 (2.1)	Reference	<0.0001	Reference	<0.0001
Yes	9.0 (1.6)	5.5 (0.4)		5.6 (0.5)	

[a]0–11 scale. The estimates shown indicate the incremental benefit observed compared with the reference level. Any positive value indicates benefit compared with reference, while any negative value indicates detriment compared with reference.
PRORQS, patient-reported outcome reporting quality score; RCT, randomized, controlled trial.

inmain manuscripts reporting RCTs was usually small. An improvement of the adherence to the 2013 PROs-specific CONSORT extension might be observed on manuscript published after the publication of these guidelines. A striking finding is our study was that the PROs-specific publication resulted in a significant and clinically relevant increase in the quality of PROs reporting, being nearly 2.5-fold better than those where PROs was reported as part of the primary publication and having mean scores of 9.0 out of possible 11 on our PRORQS scale. We could speculate that this was a function of both an adequate emphasis on PROs by these study investigators and the ability to devote focus entirely on PROs in such secondary publications. Although being not mentioned in the CONSORT extension guidelines, one could argue that publications of all pivotal oncology RCTs should ideally provide a short report of PROs in the primary manuscript, with a more exhaustive subsequent publication in a dedicated PROs-focused manuscript.

Our analysis included all cancer types, at various stages, from curative setting to palliative chemotherapy. Even if these covariates were not associated with the quality of PRO reporting in multivariate analysis, the nature of PRO is expected to vary across the situations. The high number of inclusion criteria chosen in the present study allows a comprehensive analysis of PRO reporting in oncology, but might be a source of heterogeneity in analyses. Although being potentially subject to publication biases related to the limited number of assessed journals, our analysis is certainly a picture of the current reporting quality in oncology, as most of significant RCTs manuscripts are published in a few leading journals. Our analysis was limited to published studies, and therefore is potentially subject to publication

bias. Indeed, it is known that some RCTs, especially those with non-significant results, are never published [17]. In addition, some RCTs may have been poorly designed, or manuscripts may have been so poorly written that they were rejected for publication. Such manuscripts would likely have low PRORQS. Furthermore, although we report on the adequacy of reporting, as defined by the number of CONSORT-mandated items reported, we are unable to comment on the accuracy of reporting because we were unable to compare the publications to the actual trial protocols.

In conclusion, our findings show that methods and results related to PROs were often poorly reported. The main exception was where PROs were reported in separate PROs-dedicated manuscripts. As adequate reporting of PROs is essential for the successful extrapolation of HRQL data from clinical trials to clinical practice, use of the 2013 PROs CONSORT extensions should be encouraged and their effects on PROs reporting reassessed.

funding

JP is the recipient of a grant from the Nuovo-Soldati Research Foundation.

disclosure

The authors have declared no conflicts of interest.

references

1. US Food and Drug Administration. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. 2009; http://www.fda.gov/cder/guidance/index.htm (28 July 2014, date last accessed).
2. Niraula S, Seruga B, Ocana A et al. The price we pay for progress: a meta-analysis of harms of newly approved anticancer drugs. J Clin Oncol 2012; 30: 3012–3019.
3. Lipscomb J, Reeve BB, Clauser SB et al. Patient-reported outcomes assessment in cancer trials: taking stock, moving forward. J Clin Oncol 2007; 25: 5133–5140.
4. Gnanasakthy A, Mordin M, Clark M et al. A review of patient-reported outcome labels in the United States: 2006 to 2010. Value Health 2012; 15: 437–442.
5. Basch E. The missing voice of patients in drug-safety reporting. N Engl J Med 2010; 362: 865–869.
6. Johnston BC, Patrick DL, Busse JW et al. Patient-reported outcomes in meta-analyses—part 1: assessing risk of bias and combining outcomes. Health Qual Life Outcomes 2013; 11: 109.
7. Lohr KN, Zebrack BJ. Using patient-reported outcomes in clinical practice: challenges and opportunities. Qual Life Res 2009; 18(1): 99–107.
8. Alemayehu D, Cappelleri JC. Conceptual and analytical considerations toward the use of patient-reported outcomes in personalized medicine. Am Heal Drug Benefits 2012; 5(5): 310–317.
9. Efficace F, Osoba D, Gotay C et al. Has the quality of health-related quality of life reporting in cancer clinical trials improved over time? Towards bridging the gap with clinical decision making. Ann Oncol 2007; 18: 775–781.
10. Calvert M, Brundage M, Jacobsen PB et al. The CONSORT patient-reported outcome (PRO) extension: implications for clinical trials and practice. Health Qual Life Outcomes 2013; 11: 184.
11. Brundage M, Bass B, Davidson J et al. Patterns of reporting health-related quality of life outcomes in randomized clinical trials: implications for clinicians and quality of life researchers. Qual Life Res 2011; 20: 653–664.
12. Calvert M, Blazeby J, Altman DG et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013; 309: 814–822.
13. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c332.
14. Péron J, Maillet D, Gan HK et al. Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: a systematic review. J Clin Oncol 2013; 31: 3957–3963.
15. Péron J, Pond GR, Gan HK et al. Quality of reporting of modern randomized controlled trials in medical oncology: a systematic review. J Natl Cancer Inst 2012; 104: 982–989.
16. US Food and Drug Administration. Guidance for industry clinical trial endpoints for the approval of cancer drugs and biologics guidance for industry clinical trial endpoints for the approval of cancer drugs and biologics. 2007; http://www.fda.gov/cder/guidance/index.htm (28 July 2014, date last accessed).
17. Easterbrook PJ, Berlin JA, Gopalan R et al. Publication bias in clinical research. Lancet 1991; 337: 867–872.

Un des messages principaux de cet article est qu'une majorité des articles rapportant des essais de phase III en oncologie médicale ne font pas référence à un CRP. Lorsqu'un CRP est utilisé, il est le plus souvent rapporté dans le manuscrit principal (71%). Une publication secondaire dédiée aux critères de jugement rapportés par les patients n'est retrouvée que dans 29% des cas.

Dans cette revue, il existait une très importante variabilité dans la qualité du rapport des éléments méthodologiques nécessaires pour évaluer la validité des résultats et des conclusions relatifs aux CRPs. De façon intéressante, la qualité du rapport était nettement plus élevée lorsqu'une publication secondaire dédiée aux CRPs était utilisée. Il est possible que ce résultat soit une conséquence d'un intérêt plus important des investigateurs pour les CRPs, associé à une moindre limitation en espace de rédaction dans les manuscrits.

Ces résultats nous amènent à recommander de décrire une première fois, de façon succincte, les méthodes et les résultats relatifs aux CRPs dans le manuscrit principal rapportant un essai de phase III. Puis, nous recommandons de rapporter de façon plus exhaustive l'ensemble des méthodes et des résultats relatifs CRPs dans une publication secondaire dédiée. Cette seconde publication semblant nécessaire pour qu'une évaluation de la validité interne des résultats soit réalisable. De plus cette seconde publication pourrait également être le lieu d'une analyse globale de l'effet des traitements, intégrant les critères de jugement classiques (survie, évolution de la taille tumorale, événement indésirable), et les critères de jugement rapportés par les patients (échelles de symptôme, qualité de vie).

Chapitre III

III. Développement méthodologique des comparaisons par paire généralisées

III.1. La procédure standard

III.1.a La procédure standard – principe général

La méthode des comparaisons par paire généralisées a été proposée en 2010 par Marc Buyse dans le journal *Statistics in Medicine* [1]. Il s'agit d'une extension de la statistique de test U définie par Mann et Whitney [10], qui permet de comparer deux groupes de patients en fonction d'un critère de jugement continu. Par la suite nous considérerons un groupe de n patients exposés au traitement en cours d'investigation (groupe T) et un groupe de m patients servant de contrôles (groupe C). Les patients peuvent être évalués sur un critère de jugement, possiblement mesuré de façon répétée, ou sur plusieurs critères de jugement hiérarchisés.

Les comparaisons par paire nécessitent de considérer l'ensemble de toutes les $m.n$ paires de patients, l'un étant issu du groupe T et l'autre du groupe C. Une paire de patient est classée 'favorable' au traitement si le patient issu du groupe T a un meilleur résultat thérapeutique que le patient issu du groupe C. Une paire est classée 'défavorable' au traitement si le patient issu du groupe T a un moins bon résultat thérapeutique que le patient issu du groupe C. Lorsque les deux patients ont des résultats thérapeutiques comparables, la paire est classée 'neutre'. Lorsqu'il n'est pas possible de déterminer lequel des deux patients au sein d'une paire a le meilleur résultat thérapeutique (du fait de données manquantes ou de données censurées), la paire est classée 'non informative'. Lorsque deux patients sont comparés au sein d'une paire, la définition d'un 'meilleur résultat' doit être définie de façon explicite. Cette définition d'un 'meilleur résultat' doit correspondre à une définition cliniquement pertinente.

III.1.b La procédure standard – classement des paires selon le type de critère de jugement

Par la suite, nous noterons x_i la valeur du critère de jugement X pour le $i^{ème}$ patient issu du groupe T. Pour des raisons qui deviendront claires plus loin dans l'exposé de cette méthode, nous noterons y_j la valeur du même critère de jugement noté cette fois Y pour le $j^{ème}$ patient issu du groupe C.

Le premier type de critère de jugement utilisable dans une procédure de comparaison par paire est le critère de jugement binaire. Un succès est indiqué par $X = 1$ (ou $Y = 1$), et un échec est indiqué par $X = 0$ (ou $Y = 0$). Le classement de chacune des paires de patients sur un critère de jugement binaire est non ambigu (tableau III.1). Il n'y a donc pas de paire non informative, sauf en présence de données manquantes.

Tableau III-1. Classement des paires sur un critère de jugement de type binaire

Paire	Classement
$x_i = 1$ et $y_j = 0$	Favorable
$x_i = 1$ et $y_j = 1$	Neutre
$x_i = 0$ et $y_j = 0$	Neutre
$x_i = 0$ et $y_j = 1$	Défavorable
x_i ou y_j manque	Non informative

Considérons maintenant un critère de jugement de type continu qui est mesuré par une variable X (ou Y). Nous considérerons ici un critère de jugement continu pour lequel une valeur élevée de X (ou Y) est préférable à une valeur basse de X (ou Y). Dans certains cas, il est légitime de considérer que la différence entre les valeurs x_i de X et y_j de Y doit dépasser une valeur seuil (notée τ) pour être considérée comme significative. La valeur de τ peut dépendre de la précision de la mesure de X (ou Y), ou être le reflet d'une différence minimale

cliniquement significative. Le classement de chacune des paires de patients sur un critère de jugement continu est présenté dans le tableau III.2.

Tableau III-2. Classement des paires sur un critère de jugement de type continu

Paire	Classement
$x_i - y_j > \tau$	Favorable
$\lvert x_i - y_j \rvert < \tau$	Neutre
$x_i - y_j < -\tau$	Défavorable
x_i ou y_j manque	Non informative

Considérons un critère de jugement de type temps jusqu'à événement. Dans ce cas la variable *X* (ou *Y*) peut être censurée à droite. Nous noterons x_i^0 la valeur du temps jusqu'à événement du $i^{ème}$ patient issu du groupe *T*, et y_j^0 la valeur du temps jusqu'à événement du $j^{ème}$ patient issu du groupe *C*. Les événements ne sont pas observés pour tous les individus du fait des censures à droite. Les temps jusqu'à observation sont notés $x_i = \min(x_i^0, u_i)$, et $y_j = \min(y_j^0, v_j)$, où u_i et v_j sont les temps jusqu'à censure des patients *i* et *j* respectivement.

Les indicatrices d'événement sont définies par:

$$\gamma_i = \begin{cases} 1 \; si \; x_i = x_i^0 \\ 0 \; si \; x_i < x_i^0 \end{cases} \Big\} \text{dans le groupe T}$$

$$\varepsilon_j = \begin{cases} 1 \; si \; y_j = y_j^0 \\ 0 \; si \; y_j < y_j^0 \end{cases} \Big\} \text{dans le groupe C}$$

Comme lors de l'analyse d'un critère de jugement continu, il est parfois légitime de considérer que la différence entre les valeurs x_i^0 de *X* et y_j^0 de *Y* doit dépasser une valeur seuil (notée τ) pour être considérée comme significative. Le classement de chacune des paires de patients sur un critère de jugement de type temps jusqu'à événement selon la procédure standard est présenté dans le tableau III.3.

Tableau III-3. Classement des paires sur un critère de jugement de type temps jusqu'à événement selon la procédure standard

$(\gamma_i, \varepsilon_j)$	$x_i - y_j > \tau$	$x_i - y_j < -\tau$	$\lvert x_i - y_j \rvert \leq \tau$
$(1,1)$	Favorable	Défavorable	Neutre
$(0,1)$	Favorable	Non informative	Non informative
$(1,0)$	Non informative	Défavorable	Non informative
$(0,0)$	Non informative	Non informative	Non informative

III.1.c La procédure standard – priorisation de critères de jugement multiples

Multiples critères de jugement

Les comparaisons par paire peuvent être étendues à plusieurs critères de jugement, à condition qu'ils soient hiérarchisés en priorités successives. Une stratégie de classement des paires en fonction de chacun des critères inclus dans l'analyse de l'effet thérapeutique global (binaire, continu, temps jusqu'à événement) doit être définie – voire la section III.1.b –. Une façon naturelle d'intégrer dans l'analyse plusieurs critères de jugement est d'analyser dans un premier temps l'ensemble des paires sur le critère de jugement de première priorité. Les paires neutres ou non informatives sur le critère de jugement de première priorité sont alors analysées sur le critère de jugement de priorité inférieure (tableau III.4). Ce principe peut être répété pour L critères de jugement hiérarchisés ($l = 1,\ldots,L$).

Tableau III-4. Comparaisons par paire généralisées pour deux critères de jugement priorisés

Critère de jugement de priorité supérieure	Critère de jugement de priorité inférieure	Classement final de la paire
favorable	ignorée	favorable
défavorable	ignorée	défavorable
neutre/non informative	favorable	favorable
neutre/non informative	défavorable	défavorable
neutre/non informative	neutre/non informative	neutre/non informative

Multiples seuils de significativité clinique

Un même critère de jugement de type continu ou de type temps jusqu'à événement peut être inclus de façon répétée à plusieurs priorités d'une procédure de comparaison par paire, à condition de faire varier les seuils de significativité clinique. La valeur des seuils utilisés pour un même critère de jugement doit être maximale pour la priorité supérieure, et diminuer pour les priorités inférieures. Un seuil de significativité clinique exigeant peut ainsi être défini à un niveau de priorité élevé, et des seuils plus modestes définis à des niveaux de priorités inférieures. L'analyse d'un critère de jugement avec deux seuils de significativité clinique $\tau_1 > \tau_2$ se fait de façon similaire à l'analyse de deux critères de jugement présentée dans la section précédente (tableau III.5). Ce principe peut être répété pour L seuils de significativité clinique (l = 1,…,L).

Tableau III-5. Comparaisons par paire généralisées pour un critère de jugement avec deux seuils de significativité clinique $\tau_1 > \tau_2$

Seuil τ_1 priorité supérieure	Seuil τ_2 priorité inférieure	Classement finale de la paire
favorable	ignorée	favorable
défavorable	ignorée	défavorable
neutre/non informative	favorable	favorable
neutre/non informative	défavorable	défavorable
neutre/non informative	neutre/non informative	neutre/non informative

Une analyse par comparaison par paire généralisée peut intégrer plusieurs critères de jugement, avec pour chacun de ces critères de jugement plusieurs seuils de significativité clinique. Une contrainte naturelle étant que pour un même critère de jugement, la priorité de niveau supérieur doit être associée à un seuil plus élevé.

III.1.d La procédure standard – estimation et test de l'effet du traitement

Un score $p_{ij}(l)$ est défini pour le $l^{ème}$ critère de jugement priorisé (l = 1,…,L) et pour chaque paire de patients incluant un patient i (i = 1,…,n) issu du groupe T et un patient j (j = 1,…,m) issu du groupe C.

$$p_{ij}(l) = \begin{cases} 1 & \text{si la paire est favorable} \\ -1 & \text{si la paire est défavorable} \\ 0 & \text{si la paire est neutre, non informative ou déjà classée (ignorée)} \end{cases}$$

La 'propension au succès' $\delta(l)$ pour le $l^{ème}$ critère de jugement est la différence entre le nombre de paires classées favorables au traitement et le nombre de paires classées défavorables au traitement divisée par le nombre total de paires. La 'propension au succès' est une traduction française du terme utilisé en anglais 'chance of a better outcome' (terme équivalent au terme 'proportion in favor of treatment' précédemment utilisé par Marc Buyse [1]).

$$\delta(l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(l)}{n.m}$$

En présence d'une stratification, le calcul de $\delta(l)$ est réalisé au sein de chacune des K strates ($k = 1, \dots, K$) :

$$\delta(l) = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(l)}{\sum_{k=1}^{K} n_k \,.\, m_k}$$

La 'propension cumulée au succès' pour le $l^{ème}$ critère de jugement est donnée par :

$$\Delta(l) = \sum_{\lambda=1}^{l} \delta(\lambda)$$

La 'propension cumulée au succès' pour le critère de jugement de la dernière priorité est appelée 'propension globale au succès' et est notée Δ. Un test de permutation peut être utilisé pour tester l'hypothèse $H_0: \Delta = 0$. Le test de permutation repose sur la simulation d'un grand nombre (noté S) de jeux de données identiques au jeu de données étudié [11]. Les données de chaque patient sont conservées, à l'exception du groupe de traitement (groupe T ou groupe C) qui est alloué de façon aléatoire. La probabilité d'allocation dans chacun des groupes de traitement est équivalente à la proportion de patients inclus dans ces groupes. La stratégie d'allocation des groupes doit être identique à la stratégie de randomisation de l'essai étudié (randomisation simple, randomisation stratifiée, etc…). Dans les S jeux de données simulés, l'appartenance au groupe thérapeutique est aléatoire. La valeur de la propension cumulée au

succès calculée pour chacun de ces jeux de données ne diffère donc de 0 que par le fait du hasard. L'intervalle de confiance à $(1-\alpha)$ pourcent de la 'propension cumulée au succès' observée (notée Δ_{obs}) est calculée à partir de la distribution empirique de Δ sous l'hypothèse nulle de la façon suivante. Notons $\Delta_{\alpha/2}$ la valeur de Δ_s $(s = 1, \dots, S)$ qui est supérieure à $\frac{\alpha}{2}$ pourcent de l'ensemble des valeurs de Δ_s. Notons ensuite $\Delta_{1-\alpha/2}$ la valeur de Δ_s qui est inférieure à $\frac{\alpha}{2}$ pourcent de l'ensemble des valeurs de Δ_s. L'intervalle de confiance à $(1-\alpha)$ pourcent de Δ_{obs} est $\left[\Delta_{obs} + \Delta_{\alpha/2}; \Delta_{obs} + \Delta_{1-\alpha/2}\right]$.

Le niveau de significativité statistique (P-value) associé avec Δ_{obs} peut également être calculé à partir de la distribution empirique de Δ sous l'hypothèse nulle. Notons s_1 le nombre de valeurs de Δ_s $(s = 1, \dots, S)$ obtenues par permutation pour lesquelles $\Delta_s \geq \Delta_{obs}$. Notons ensuite s_2 le nombre de valeurs de Δ_s pour lesquelles $|\Delta_s| \geq |\Delta_{obs}|$. La P-value associée avec Δ_{obs} est égale à s_1/S pour un test unilatéral, et s_2/S pour un test bilatéral. Le test de la propension au succès sur un critère de jugement binaire a été montré comme équivalent à l'approximation de Monte Carlo du test de Fisher exact [1]. Le test sur un critère de jugement continu a été montré comme équivalent à un test de Mann-Whitney-Wilcoxon lorsque $\tau = 0$ [1], [10]. Enfin, le test de la propension au succès sur un critère de jugement de type temps jusqu'à événement a été montré comme équivalent à un test Wilcoxon généralisé par Gehan lorsque $\tau = 0$ [1], [12].

III.2. Extension de la procédure pour les données de type temps jusqu'à événement

III.2.a. Limites de la procédure standard pour les données de type temps jusqu'à événement

La procédure standard d'analyse des variables de type temps jusqu'à événement dans les comparaisons par paire généralisées a été définie dans la section III.1.b. Une limite de cette méthode est que les paires de patients qui ne sont pas directement classables (favorable, défavorable ou neutre) du fait des censures sont considérées comme non informatives, et ceci quelles que soient les valeurs des temps jusqu'aux censures. Pour illustrer cette limite, prenons l'exemple d'une paire de patients $\{i, j\}$ issus du groupe T et du groupe C respectivement. Si $\gamma_i = \varepsilon_j = 0$, c'est-à-dire si les deux patients sont censurés aux temps x_i et y_j, alors la participation au score de la paire $\{i, j\}$ selon la procédure standard est $p_{ij} =$

0, quelles que soient les valeurs de x_i et y_j. Pourtant si x_i est très grand et y_j très petit, la valeur de p_{ij} devrait intuitivement être positive car $\hat{\Delta}= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}}{n.m}$ et Δ est la différence entre la probabilité pour une paire d'être favorable au groupe T et la probabilité pour une paire d'être défavorable au groupe T. Cette dernière équation sera par la suite écrite de façon simplifiée $\Delta= \mathbb{P}[X > Y] - \mathbb{P}[Y > X]$. Cette écriture simplifiée a pour but de simplifier la compréhension du paramètre Δ. Cette limite de la procédure standard entraine deux risques principaux. Le premier risque est de diminuer la puissance du test de permutation cherchant à rejeter l'hypothèse nulle $H_0: \Delta= 0$. Le second risque est de fournir une estimation biaisée de Δ, notée $\hat{\Delta}$. En effet, la valeur de $\hat{\Delta}$ tend vers 0 à mesure que le taux de censure augmente et que le taux de paires non informatives du fait des censures augmente.

Nous proposons de calculer la participation au score de chaque paire (p_{ij}) non directement classable via une estimation de $\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right] - \mathbb{P}\left[y_j^0 > x_i^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$. Ce calcul est donc basé sur l'estimation de la probabilité pour chaque paire de patient d'être classée favorable, défavorable ou neutre si les événements étaient observés. Pour calculer p_{ij}, nous utiliserons l'estimation de la fonction de survie selon la méthode de Kaplan et Meier, et les couples (x_i, γ_i) et (y_j, ε_j). Trois méthodes de calcul ont été proposées définissant trois extensions de la méthode des comparaisons par paire généralisées. La première suit la philosophie du test de Wilcoxon généralisé par Peto et Peto, lorsque $\tau = 0$ [13]. La seconde est équivalente au test de Wilcoxon généralisé par Efron, lorsque $\tau = 0$ [14]. La troisième est une modification de l'extension dite de Efron et est nommée procédure dite de Péron.

III.2.b. Extension dite de Peto et Peto

Cette extension de la procédure standard pour analyser les données de type temps jusqu'à événement a été la première à être proposée. L'objectif commun avec les deux autres extensions est de calculer p_{ij} via une estimation de $\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right] - \mathbb{P}\left[y_j^0 > x_i^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$.

En suivant la philosophie de la modification du test d'Efron proposée par Latta et le test de Wilcoxon généralisé par Peto et Peto, cette extension utilise les observations des patients issus des groupes T et C pour estimer une fonction de survie conjointe. En effet, sous l'hypothèse nulle, la distribution d'une observation issue du groupe T est identique à la distribution d'une

observation issue du groupe C : $S(t) = \mathbb{P}[x_i^0 \geq t] = \mathbb{P}\left[y_j^0 \geq t\right]$. L'estimation de $\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$ et de $\mathbb{P}\left[y_j^0 > x_i^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$ peut alors être réalisée à partir de l'estimation produit-limite $\hat{S}(t)$ de la fonction de survie basé sur l'ensemble des observations. La description de cette extension dite de Peto et Peto a fait l'objet d'un article qui a été soumis pour publication au journal *Statistics in Medicine*. Cet article est en cours de revue et est présentée ci-dessous tel qu'il a été soumis au journal.

An extension to generalized pairwise comparisons for prioritized outcomes with censoring

Julien Péron [1,2], Marc Buyse [3,4], Brice Ozenne [1,2], Laurent Roche [1,2], Pascal Roy [1,2].

1. Service de biostatistiques, Centre Hospitalier Lyon-Sud, Institut du Cancer des Hospices Civils de Lyon, F-69310, Pierre-Bénite, France

2. CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Université Lyon 1, Villeurbanne, France

3. International Drug Development Institute (IDDI), Raleigh, NC, USA

4. Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Hasselt, Belgium

Running Head

Generalized pairwise comparisons with censoring

Abstract

Generalized pairwise comparisons have been proposed to permit a comprehensive assessment of several prioritized outcomes between two groups of observations (Buyse 2010). We propose an extension of generalized pairwise comparisons for time to event outcomes that takes into account the time to censored observations. We show how pairwise scores can be calculated from the Kaplan-Meier estimates of the survival function in the presence of right-censored data. These scores are used to estimate the chance of a better outcome with treatment than with control, which is defined as $\Delta = \mathbb{P}[X > Y] - \mathbb{P}[Y > X]$ where the outcome is captured by the variable X in the treatment group and by the variable Y in the control group. A randomization test can be used to test the null hypothesis $\Delta = 0$, and to calculate a confidence interval for Δ. The extended procedure for generalized pairwise comparisons is shown to be more efficient than the standard procedure. When several outcomes are prioritized in a single assessment of the overall treatment effect (i.e. benefit-risk assessment), we show that the estimation of the chance of a better outcome varies only slightly with the censoring pattern. Since the way the censoring occurs is independent of the parameters of interest (benefits and risks of an investigational treatment), the censoring rate on the survival outcome should not have a large impact on the estimation of the chance of a better outcome. Finally we report the results of a benefit-risk assessment using the extended procedure on one illustrative dataset in the setting of advanced pancreatic cancer.

Keywords: pairwise comparisons ; multivariate analysis ; randomized trial ; survival outcome

Introduction

In randomized clinical trials aiming at assessing the efficacy of experimental therapies, the choice of the primary endpoint is critical. However, identifying a single primary outcome that allows a comprehensive assessment of the treatment efficacy may be challenging. Efficacy assessment using multiple outcomes is appealing [1]. In such cases, the main outcomes are commonly combined into one single composite endpoint [2]. However composite endpoints have some limitations [3]. One important limitation arises when the components of the composite endpoint are not in agreement in assessing the treatment efficacy, specifically when some components reflect treatment benefit while others capture harm [4]. In this case, analysis of the composite endpoint may be misleading and a benefit-risk balance analysis is required. Generalized pairwise comparisons and the win ratio have been proposed to permit a comprehensive assessment of several prioritized outcomes between two groups of observations [5, 6, 7]. We focus here on generalized pairwise comparisons [4]. The key difference between generalized pairwise comparisons and composite endpoints is that the former prioritize the various outcomes based on their clinical relevance, while the latter is based on the time course of the various outcomes, and considers the first one to occur. Another attractive feature of generalized pairwise comparisons is that the various outcomes can be of any type: discrete, continuous, or time-to-event variables. For example, an immediately observed treatment response can be analyzed together with a time to failure.

Essentially, generalized pairwise comparisons consist of forming all pairs of observations taking one patient from the treatment group and one patient from the control group [4]. When a time-to-event variable is considered, pairs of observations were previously considered uninformative in the presence of censored data. This was not optimal because the time at which an observation was censored was not taken into account. It induced a loss of power of the test, and it altered the relation between the estimated effect size (which we call hereafter

the “chance of a better outcome”) and the true effect of the treatment in the absence of censoring [8]. Buyse showed that generalized pairwise comparisons are equivalent to Gehan’s modification of the Wilcoxon test when the comparison of two groups of patients uses a single time-to-event outcome [5]. Efron proposed to take into account the time to censored observations using the Kaplan and Meier estimates of the survival function based on the observations [9]. Efron’s test has been shown to be more powerful than Gehan’s test in most situations [9]. Peto and Peto developed an approach close to the one developed by Efron [10], and they showed that, in the presence of censoring, Gehan’s scores and Wilcoxon’s scores are not asymptotically equivalent, although Gehan’s scores was presented as a generalization of the Wilcoxon scores [11]. Peto and Peto’s generalization of the Wilcoxon score addressed this issue, by taking account of the time to censored observations using the survival estimates under the null hypothesis (combined distribution of the survival times). In simulations studies, Peto and Peto’s test was shown to perform better in the case of heavy censoring, when sample sizes are unequal and when the censoring mechanisms differ greatly between groups [12].

In this paper, we propose an extension of the generalized pairwise comparisons for prioritized outcomes that take further into account the time of censored observations.

In section 2, we describe briefly the procedure of generalized pairwise comparisons, as it was initially proposed [5]. In this procedure, several outcomes, including one or more time-to-event outcomes, can be included in the overall analysis of treatment effect. Prespecified thresholds can also be used, when the difference between two variables (continuous or time-to-event) needs to exceed a clinically relevant threshold to be considered meaningful [5, 13]. In section 3, we detail the extension of the procedure for censored observations. In section 4, the extended procedure is illustrated through simulations of randomized trials. The power of the test using the extended procedure is compared with the standard procedure, and with other traditional statistical tests used in survival analyses. The sensitivity to censoring of the effect-

size estimation is also investigated. Section 5 illustrates the results of the extended procedure on one actual data of a randomized trial. Section 6 briefly presents the software used to compute generalized pairwise comparisons, and section 7 discusses the benefits and the limitations of the extended procedure of generalize pairwise comparisons in contrast with the standard procedure and with other multivariate analyses.

1. Generalized pairwise comparisons of prioritized outcomes

2.1. General Overview

Generalized pairwise comparisons apply to the situation of two groups of individuals to be compared in terms of one or more outcomes. We assume that one group of *n* individuals is exposed to a treatment (labeled '*T*"), and the other group of *m* individuals serves as a control (labeled '*C*'). Considering one outcome, we denote x_i the value of the outcome *X* for the i^{th} individual in the group *T*, and y_j the value of the same outcome denoted Y for the j^{th} individual in the group *C*. Pairwise comparisons require consideration of all possible pairs of individuals, one taken from group *T* and the other taken from group *C*. Stratified pairwise comparisons can also be performed. The outcomes of the two individuals forming a pair are compared. The pair is said to be 'favorable' if the outcome of the individual in group *T* is better than the outcome of the individual in group *C*, 'unfavorable' if the outcome of the individual in group *T* is worse than the outcome of the individual in group *C*, 'neutral' if there is no difference between the outcomes of the patients, and 'uninformative' if the two outcomes cannot be ordered because of missing data or censoring. It is possible to extend the procedure to *L* outcomes by prioritizing the variables that capture them. The highest priority is assigned to the variable considered the most clinically relevant. A natural way of handling uninformative pairs because of missing data or censored observations is to consider the outcomes in descending order of priority: whenever a pair is uninformative or neutral for an outcome of higher priority, the outcomes of lower priority are examined (Table 1).

A pairwise scoring indicator $p_{ij}(l)$ for the l^{th} outcome measure (l = 1,…,L) is defined for the pair formed by the i^{th} individual (i = 1,…,n) in group T, and the j^{th} individual (j = 1,…,m) in group C :

$$p_{ij}(l) = \begin{cases} 1 & \text{if the pair is favorable} \\ -1 & \text{if the pair is unfavorable} \\ 0 & \text{if the pair is neutral} \end{cases}$$

The 'chance of a better outcome' (called 'proportion in favor of treatment' by Buyse [5]) for the l^{th} outcome is the net difference between the number of favorable pairs and the number of unfavorable pairs divided by the total number of pairs.

$$\hat{\delta}(l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}(l)}{n.m}$$

The cumulative chance of a better outcome for the l^{th} outcome is $\Delta(l) = \mathcal{E}\left(\sum_{\lambda=1}^{l}\hat{\delta}(\lambda)\right)$ estimated by $\hat{\Delta}(l) = \sum_{\lambda=1}^{l}\hat{\delta}(\lambda)$. The cumulative chance of a better outcome for the outcome of lowest priority is called the overall chance of a better outcome and denoted Δ. A randomization test can be used to test the null hypothesis H_0: Δ= 0, and a randomization test-based confidence interval for $\hat{\Delta}$ can be calculated using the empirical distribution of the $\hat{\Delta}_r$obtained by permutation under H_0.

2.2. Time-to-event variables

We denote x_i^0 the value of the time-to-event outcome X of the i^{th} individual in group T, and and y_j^0 the value of the same outcome denoted Y for the j^{th} individual in the group C. In the case of right censoring, the event is not observed for all individuals, and the observable variables are $x_i = \min(x_i^0, u_i)$, and $y_j = \min\left(y_j^0, v_j\right)$, where u_i and v_j denote the censoring times for individuals i and j, respectively.

Let further define the event statuses:

$$\delta_i = \begin{cases} 1 \; if \; x_i = \; x_i^0 \\ 0 \; if \; x_i < \; x_i^0 \end{cases} \Big\} \text{in group T}$$

$$\varepsilon_j = \begin{cases} 1 \; if \; y_j = \; y_j^0 \\ 0 \; if \; \; y_j < \; y_j^0 \end{cases} \Big\} \text{in group C}$$

Generalized pairwise comparisons can use a threshold τ to reflect meaningful differences in outcomes captured by a time-to-event variable. The standard procedure of pairwise scoring for time-to-event variable is to classify as 'uninformative' all the pairs not known to be favorable or unfavorable because of censoring (table 2).

When only one time-to-event outcome is considered and $\tau = 0,$ this approach is equivalent to Gehan's extension of the Wilcoxon test for censored data [5, 14].

3. Extension of the procedure for censored observations

3.1. Pairwise comparisons with threshold reflecting clinical relevance

When dealing with time-to-event variables, the procedure described previously has two main limitations. First, the estimation of the chance of a better outcome varies according to the pattern of censoring imposed on the observations [11]. Second, it ignores part of the available information, resulting in a loss of power in many situations [9]. For example, pairs with two censored observations are considered uninformative, irrespective of the relative magnitudes of u_i and v_j. An optimal procedure would compute the pairwise score p_{ij} as an estimation of $\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right] - \mathbb{P}\left[y_j^0 > x_i^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$. When a pair can be decidedly classified as favorable or unfavorable, p_{ij} takes the value 1 or -1, respectively. Efron proposed an extension of the Wilcoxon test in the special case where $\tau = 0$ based on $\hat{S}_T(t)$ and $\hat{S}_C(t)$, the Kaplan and Meier estimates of the survival function $S_T(t) = \mathbb{P}\,[x_i^0 \geq t]$ and $S_C(t) = \mathbb{P}\left[y_j^0 \geq t\right]$. Indeed the conditional probability $\mathbb{P}\,[x_i^0 \geq t \,|\, x_i < t, \delta_i = 0\,]$ can be

estimated by $\hat{S}_T(t)/\hat{S}_T(x_i)$. With Efron's extension, the pairwise score can be calculated as presented in table 3.

Efron's test was shown to be more powerful than Gehan's test in some cases [15], but its use has been limited in practice because of computational difficulties. Other approaches, such as the log-rank test and Peto and Peto's extension of the Wilcoxon test, were usually preferred, practically, for the comparison of two survival curves [11]. Following Latta's modification of Efron's test and Peto and Peto's test, we assume that under the null hypothesis the individuals in group T and in group C have the same survival function $S(t) = \mathbb{P}[x_i^0 \geq t] = \mathbb{P}[y_j^0 \geq t]$. $\mathbb{P}[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j]$, and $\mathbb{P}[y_j^0 > x_i^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j]$ can then be estimated using the Kaplan-Meier estimate $\hat{S}(t)$ of the survival function based on all observations.

Under these conditions, $\mathbb{P}[x_i^0 > y_j^0 | x_i > \, s, y_j > s] = 0.5$, when $\delta_i = 0$, and $\varepsilon_j = 0,$ and then one can show that if $x_i - \, y_j < \, \tau,$

$$\mathbb{P}[x_i^0 > y_j^0 + \tau | x_i^0 > \, x_i, \, y_j^0 > y_j] = 0.5 \cdot \frac{\mathbb{P}[x_i^0 > y_j + \tau]}{\mathbb{P}[x_i^0 > x_i]}$$

$$= 0.5 \cdot \frac{\hat{S}(y_j + \tau)}{\hat{S}(x_i)}$$

The pairwise score $p_{ij} = \, \mathbb{P}[x_i^0 > y_j^0 + \tau] - \mathbb{P}[y_j^0 > x_i^0 + \tau]$ for each combination of $\{x_i, \, y_j, \tau, \, \delta_i, \varepsilon_j\}$ can then be estimated as shown as shown in table 4.

3.2. Pairwise comparisons with several prioritized outcomes

Generalized pairwise comparisons permit the simultaneous analysis of *L* outcomes, as long as a hierarchical order can be defined for these outcomes. Let us consider the simple case of two prioritized outcomes when the first priority outcome is captured by a time-to-event variable

and the two captured outcomes are different. A pair of individual analyzed on the first priority outcome can be classified as 'definitely informative', 'neutral', or 'unclassified'. The term 'uninformative' should not be used anymore because the time to censoring is now included in the analysis. Pairs unclassified on the outcome with higher priority are to be analyzed on the outcome of lower priority. We denote x_{il}^0 and y_{jl}^0 the time to event and δ_{il} and ε_{jl} the censoring indicator for the of the l^{th} outcomes (l = 1,…,L) of the i^{th} individual in group T and of the j^{th} individual in group C, respectively. The score assigned to the pair on the second priority outcome should be weighted by the probability that the pair would be classified neutral if the events were observed for the two individuals: $\omega_{ij}(l=2) = \mathbb{P}\left[(|x_{i1}^0 - y_{j1}^0| < \tau_1)|\, x_{i1},\, y_{j1},\, \delta_{i1},\, \varepsilon_{j1}\right]$. The process can easily be extended to L prioritized outcomes (table 5). In the simple case of two prioritized outcomes, if we suppose that the two outcomes are independent, then $\mathbb{P}\left[\left(|x_{i1}^0 - y_{j1}^0| < \tau_1|\, x_{i1},\, y_{j1},\, \delta_{i1},\, \varepsilon_{j1}\right) \cap \left(|x_{i2}^0 - y_{j2}^0| < \tau_2|\, x_{i2},\, y_{j2},\, \delta_{i2},\, \varepsilon_{j2}\right)\right] = \mathbb{P}\left[|x_{i1}^0 - y_{j1}^0| < \tau_1|\, x_{i1},\, y_{j1},\, \delta_{i1},\, \varepsilon_{j1}\right] \times \mathbb{P}\left[|x_{i2}^0 - y_{j2}^0| < \tau_2|\, x_{i2},\, y_{j2},\, \delta_{i2},\, \varepsilon_{j2}\right]$.

When the variables capturing the two prioritized outcomes are identical with two different thresholds $\tau_1 > \tau_2$, given the natural relation between the two variable, the pairwise score for the second priority outcome is $p_{\mathrm{ij}}(\mathrm{l}=2) = \mathbb{P}\left[(x_{i2}^0 - y_{j2}^0 \in [\tau_2, \tau_1])|\, x_{i2},\, y_{j2},\, \delta_{i2},\, \varepsilon_{j2}\right] - \mathbb{P}\left[(y_{j2}^0 - x_{i2}^0 \in [\tau_2, \tau_1])|\, x_{i2},\, y_{j2},\, \delta_{i2},\, \varepsilon_{j2}\right]$. The weight assigned to the score should be $\omega_{ij}(l=2) = 1$, because the $\mathbb{P}\left[|x_{i1}^0 - y_{j1}^0| > \tau_1|\, x_{i1},\, y_{j1},\, \delta_{i1},\, \varepsilon_{j1}\right]$ is already taken into account in the definition of $\mathrm{p_{ij}}(\mathrm{l}=2)$.

For lower priority outcomes of different nature, the calculation of $\omega_{\mathrm{ij}}(\mathrm{l}>2)$ is based only on the smallest threshold τ_2 because $\mathbb{P}\left[\left(|x_{i1}^0 - y_{j1}^0| < \tau_1|\, x_{i1},\, y_{j1},\, \delta_{i1},\, \varepsilon_{j1}\right) \cap \left(|x_{i2}^0 - y_{j2}^0| < \tau_2|\, x_{i2},\, y_{j2},\, \delta_{i2},\, \varepsilon_{j2}\right)\right] = \mathbb{P}\left[|x_i^0 - y_j^0| < \tau_2|\, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$. The process can be extended for l^{th} prioritized outcomes, when the same variable is included in more than one priority with

decreasing thresholds using a priority indicator $I_{l\lambda}$ defined for the l^{th} outcome and for the λ^{th} higher priority outcomes ($\lambda = 1, \dots, l-1$) (table 6):

$$I_{l\lambda} = \begin{cases} 0 \text{ if the outcome variable included at priorities } \lambda \text{ and } l \text{ are different} \\ 1 \text{ if the same outcome variable is included at priorities } \lambda \text{ and } l \end{cases}$$

4. Simulation study

We conducted an extensive simulation study to examine the performance of the extended generalized pairwise comparison procedure. For all scenarios, 1000 datasets were simulated with 200 patients divided in two groups of equal size. Outcome variables were simulated for each patient. Time-to-event variables were simulated using an exponential distribution, with a scale parameter $\lambda_C = 1$ in group C. The scale parameter in group T varied among scenarios. In all scenarios, we assumed that the censoring times were distributed uniformly, with the same parameter value in both treatment groups. We aim to compare the power of the test to reject the null hypothesis compared to the standard procedure and to other classical tests for survival. The comparison was performed in the case of proportional and non-proportional hazards.

4.1. Scenario 1: One time-to-event outcome and proportional hazards

In the first scenario, treatment groups were compared on an unique time-to-event outcome. The hazard for survival in group T was assumed to be proportional to the hazard for survival in group C (scale parameter in group T : $\lambda_T = \text{HR} \times \lambda_C$, where HR was the hazard ratio for survival). For each simulated dataset, the time-to-event outcome was compared between the two groups by the standard generalized pairwise comparison procedure, by the extended generalized pairwise comparison procedure, by the Peto and Peto's test, and by the log-rank test. The proportion of datasets for which the p-value was lower than 0.05 was an estimation

of the α-risk when HR = 1 and an estimation of the power of the test when HR ≠ 1. For all values of HR and of censoring rates, the log-rank test was more powerful when hazards were proportional, which is a well-known property of this test. The superiority of the log-rank test was most pronounced when the censoring rate was low (Figure 1A). The extended generalized pairwise comparison procedure was more powerful than the standard procedure, and had the same performance as Peto and Peto's test. When there was no difference between treatment group (HR = 1), the α-risk was very close to 0.05 (ranging from 0.045 to 0.061) for the four tests in all scenarios, and none performed better. Ideally, the estimation of the chance of a better outcome should not be modified by the censoring rate. Its theoretical value is 0.333 when HR=0.5, and is 0.176 when HR=0.7 [8]. The decrease in the estimated chance of a better outcome was less pronounced when the extended procedure was used compared to the standard procedure (Figure 1B).

4.2. Scenario 2: One time-to-event outcome and non-proportional hazards with attenuation of treatment effect over time

In the second scenario, treatment groups were again compared on only one time-to-event outcome. However, the hazard ratio was not constant over time and increased progressively in four steps from HR at time=0 to 1 at the end of follow-up ($HR_1 = HR$ if $t < \frac{T_{0.5}}{2}$ where $T_{0.5}$ is the time at which half of the patients in group T would have presented the event if HR was proportional ; $HR_2 = 0.75 \times HR + 0.25$ if $\frac{T_{0.5}}{2} < t < T_{0.5}$; $HR_3 = 0.5 \times HR + 0.5$ if $T_{0.5} < t < 3.\frac{T_{0.5}}{2}$; $HR_4 = 0.25 \times HR + 0.75$ if $3.\frac{T_{0.5}}{2} < t < 2.T_{0.5}$; and $HR_5 = 1$ if $t > 2.T_{0.5}$; figure A in appendix). Here the log-rank test was less powerful than any of the Wilcoxon family test. The loss in power observed with the log-rank test was more pronounced when the censoring rate was low. The extended generalized pairwise comparison procedure had the same performance as Peto and Peto's test and was slightly inferior to the standard procedure (Figure 2A). When there was no difference between treatment group (HR = 1), the α-risk

was very close to 0.05 for the four tests (ranging from 0.036 to 0.058), and again none performed better. The decrease in chance of a better outcome in the case of high censoring was less pronounced with the extended procedure (Figure 2B). In the inverse scenario, with non-proportional hazards and delayed treatment effect, the log-rank test was the most efficient, and the extended generalized pairwise comparison procedure performed better than the standard procedure in the presence of censored data (Figure B in appendix).

4.3. Scenario 3: One time-to-event outcome and one binary outcome

The time-to-event outcome distribution was simulated as described in scenario 1, and an additional binary outcome was simulated via a binomial distribution. The treatment effect size was assumed to be identical for the two outcomes (based on the same absolute value of the chance of a better or worse outcome when analyzed separately and in the absence of censoring), but in opposite directions (the time-to-event outcome favored the group T with a hazard ratio of 0.7, which translates to a theoretical chance of a better outcome equal to $\Delta_{time-to-event} = 0.176$; and the binary outcome favored the group C with $\Delta_{binary} = -0.176$). In generalized pairwise comparisons, the first priority outcome was the time-to-event outcome, and the second priority outcome was the binary outcome. The survival threshold for clinical relevance was set at 1 time unit for illustrative purpose. Simulations were repeated for various censoring rates. The weight of the second priority binary outcome increased largely for high censoring rate when the standard procedure was used, modifying then the estimation of the chance of a better outcome. This counter-intuitive and undesirable relation was far less important when the extended procedure was used (figure 3).

5. Analysis of an illustrative dataset

The NCIC CTG PA.3 trial was an international study in which patients with advanced pancreatic cancer were randomized to receive gemcitabine in combination with either erlotinib or placebo as first-line treatment [15]. The primary outcome was Overall Survival

(OS). Toxicity was a secondary outcome. In this trial, 569 patients were stratified by center, performance status (Eastern Cooperative Oncology Group [ECOG] 0 or 1 vs. 2), and extent of disease (locally advanced vs. metastatic). OS was significantly better in the combination treatment, but the benefits were of modest magnitude (HR for overall survival (OS) = 0.82, 95% CI, 0.69 to 0.99 ; P = 0.038). The censoring rate for OS was 15%. The frequency of all grades and grade ≥ 3 treatment-related adverse events (AEs) was higher for the erlotinib and gemcitabine group (90% and 31%, respectively) compared with the placebo and gemcitabine group (76% and 20%, respectively). A benefit-risk balance analysis using the standard procedure of generalized pairwise comparisons showed an unfavorable overall effect of erlotinib [16]. The first priority outcome used in the main analysis of the benefit-risk balance was OS with a threshold $\tau_1 = 2$ months [16]. The second priority outcome was treatment-related AEs, with patients experiencing the lower grade related AE considered to have had a more favorable outcome. The overall chance of a better outcome was calculated at -3.6% (95% CI, -14.2% to 7.1%; P=0.51) against the erlotinib arm with the standard procedure, and at 1.2% (95% CI, -11.5% to 14.1%; P=0.86) slightly in favor of the erlotinib arm with the extended procedure (Figure 4). The evaluation of the benefit-risk balance depending of the OS threshold, performed as a sensitivity analysis, differed between the two procedures, the extended procedure making better use of the survival data and hence leaning more in favor of erlotinib.

6. Software

An R package (BuyseTest) was developed by one author (BO) and extensively controlled by another author (JP). It includes both the standard and the extended procedure for generalized pairwise comparisons. It is available upon request, and will soon be proposed to the CRAN for larger diffusion.

7. Discussion

Generalized pairwise comparisons offer an intuitive and efficient alternative approach to standard non-parametric tests in randomized trials [5]. The pairwise comparison of time-to-event variables should be based on survival estimates. Indeed the extended procedure was shown to be more efficient to detect time-to-event differences than the standard procedure of generalized pairwise comparisons. When the extended procedure was performed on one single time-to-event variable with a null threshold, the test was shown to have the same power as the Peto and Peto's generalization of the Wilcoxon test. As expected, it was less powerful than the log-rank test when hazards were proportional or when the treatment effect was delayed, and more powerful than the log-rank test in the case of early survival differences.

Generalized pairwise comparisons may prove useful to consider simultaneously several variables, for example in benefit-risk assessment or when the benefits of an investigational treatment are captured by more than one variable. In the late case, composite endpoints are often defined to capture all relevant events in a single variable [5]. In advanced cancer, for example, progression-free survival is defined as the time to disease progression or death, whichever occurs first. A clear limitation of progression-free survival is that it ignores the time to death after disease progression, and it does not incorporate other potentially relevant outcomes such as toxic effects of the treatment. In contrast, using generalized pairwise comparisons, overall survival can be defined as the first priority outcome, and progression-free survival as a second priority outcome. The weight of each endpoint in the overall analysis can be easily reported and interpreted. Additionally, an outcome related to treatment toxicity can further be considered to account for treatment side-effects. Other multivariate analysis procedures, such as Overall Treatment Utility (OTU) or Quality Adjusted Life Year (QALY), have been proposed [17–19] to perform the simultaneous analysis of benefit and risk of innovative treatment. However the respective weights of the different treatment effects may

be difficult to justify and to report for both methods, which may have limited their use in clinical trial routine.

In this paper we have shown how generalized pairwise comparisons can be used to analyze two outcomes, including time-to-event outcomes, in order to perform a benefit-risk assessment of a new treatment. We used simulations to study the performance of the chance of a better outcome with treatment than with control estimated with a time-to-event first priority outcome (simulated to capture treatment benefit) and a binary second priority outcome (simulated to capture treatment toxicity). The chance of a better outcome, which captures the benefit-risk balance of the new treatment, was shown to vary greatly with the chosen value for the survival threshold. This was expected, because the choice of this threshold directly reflects the weight given to the survival outcome in the overall analysis. For example, when the survival threshold is set a zero, any survival difference is considered as clinically relevant, and the weight of the second priority outcome is then null in the extended procedure for generalized pairwise comparisons. In the standard procedure, by contrast, the weight of the second priority outcome is not null because of pairs that are uninformative for the first priority outcome. When the survival threshold gets larger, the weight of the survival outcome in the overall analysis is expected to go to zero. Since the way the censoring occurs is independent of the parameters of interest (related to the benefit or the risks of an investigational treatment), the censoring rate on the survival outcome should not have a large impact on the benefit-risk balance of the new treatment when the extended procedure is used.

One limitation of the extended procedure is that the estimations of the survival probability were based on the observations made in both treatment groups. Under the null hypothesis, the distribution of survival times is similar in the two groups. The combination of all the observations together to estimate the survival function is then reasonable when the goal is to reject the null hypothesis. However, under the alternative hypothesis, this procedure

introduces a systematic underestimation in the estimation of the probability index $\mathbb{P}[X > Y]$ and of the chance of a better outcome with treatment than with control $\Delta = \mathbb{P}[X > Y] - \mathbb{P}[Y > X]$ in the presence of censoring. This issue might be fixed by using separate estimations of the survival functions by treatment group [20]. However this approach would be far more computationally intensive and would require substantial run times even with modern computers [20], hence it was not deemed suited for a practical use of the procedure. The bias was low for low censuring rates and the validity of the test of the null hypothesis was not affected by this issue. It should therefore be a minor issue in the interpretation of generalized pairwise comparison results. However, caution is required about the estimated effect size in case of heavy censoring.

The extended procedure addresses some drawbacks of the standard procedure when generalized pairwise comparisons are based on time-to-event variables. Generalized pairwise comparisons offer unprecedented possibilities of assessing treatment effects based on several prioritized outcomes, including time-to-event variables.

References

1. Jensen SM, Pipper CB, Ritz C. Evaluation of multi-outcome longitudinal studies. Stat Med. 2015;(in press). doi:10.1002/sim.6461.

2. Sankoh AJ, Li H, D'Agostino RB. Use of composite endpoints in clinical trials. Stat Med. 2014;33(27):4709-4714. doi:10.1002/sim.6205.

3. Wittkop L, Smith C, Fox Z, et al. Methodological issues in the use of composite endpoints in clinical trials: examples from the HIV field. Clin Trials. 2010;7(1):19-35. doi:10.1177/1740774509356117.

4. Kleist P. Composite Endpoints for Clinical Trials. Int J Pharm Med. 2007;21(3):187-198. doi:10.2165/00124363-200721030-00001.

5. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med. 2010;29(30):3245-3257. doi:10.1002/sim.3923.

6. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012;33(2):176-82. doi:10.1093/eurheartj/ehr352.

7. Luo X, Tian H, Mohanty S, Tsai WY. An alternative approach to confidence interval estimation for the win ratio statistic. Biometrics. 2014. doi:10.1111/biom.12225.

8. Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators. Clin Trials. 2008;5(6):641-642. doi:5/6/641 [pii] 10.1177/1740774508098328.

9. Efron B. The two sample problem with censored data. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967;(4): 831–853.

10. Latta RB. Generalized Wilcoxon statistics for the two sample problem with censored data. Biometrika. 1977;63(3):633-635.

11. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. J R Stat Soc A. 1972;135(2):185-198.

12. Latta RB. A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data. Journal of the American Statistical Association. 1981;76:713-719.

13. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA. 2014;312(13):1342-1343. doi:10.1001/jama.2014.13128.

14. Gehan EA. A generalized two-sample Wilcoxon test for doubly censored data. Biometrika. 1965;52(3):650-653.

15. Moore MJ, Goldstein D, Hamm J, et al. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol. 2007;25(15):1960-1966. doi:JCO.2006.07.9525 [pii] 10.1200/JCO.2006.07.9525.

16. Péron J, Roy P, Ding K, Parulekar W, Roche L, Buyse M. Assessing the benefit–risk of new treatments using generalised pairwise comparisons: the case of erlotinib in pancreatic cancer. Br J Cancer. 2015;(in press). doi: 10.1038/bjc.2015.55.

17. Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health. 2009;12 Suppl 1:S5-9. doi:10.1111/j.1524-4733.2009.00515.x.

18. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull. 2010;96(1):5-21. doi:10.1093/bmb/ldq033.

19. Seymour MT, Thompson LC, Wasan HS, et al. Chemotherapy options in elderly and frail patients with metastatic colorectal cancer (MRC FOCUS2): an open-label, randomised factorial trial. Lancet. 2011;377(9779):1749-1759. doi:10.1016/S0140-6736(11)60399-1.

20. Koziol JA, Jia Z. The concordance index C and the Mann-Whitney parameter Pr(X>Y) with randomly censored data. Biom J. 2009;51(3):467-474. doi:10.1002/bimj.200800228.

Outcome with higher priority	Outcome with lower priority	Pair is
favorable	ignored	favorable
unfavorable	ignored	unfavorable
uninformative/neutral	favorable	favorable
uninformative/neutral	unfavorable	unfavorable
uninformative/neutral	uninformative/neutral	uninformative/neutral

Table 1. Generalized pairwise comparisons for two prioritized outcomes

$(\delta_i, \varepsilon_j)$	$x_i - y_j \geq \tau$	$x_i - y_j \leq -\tau$	$\lvert x_i - y_j \rvert < \tau$
$(1,1)$	1	-1	0
$(0,1)$	1	0	0
$(1,0)$	0	-1	0
$(0,0)$	0	0	0

Table 2. Value of p_{ij} for a time-to-event outcome when some pairs are considered uninformative because of censoring

$(\boldsymbol{\delta_i}, \boldsymbol{\varepsilon_j})$	$\boldsymbol{x_i > y_j}$	$\boldsymbol{x_i < y_j}$
$(\mathbf{1}, \mathbf{1})$	1	-1
$(\mathbf{0}, \mathbf{1})$	1	$2 \cdot \frac{\hat{S}_T(y_j)}{\hat{S}_T(x_i)} - 1$
$(\mathbf{1}, \mathbf{0})$	$1 - 2 \cdot \frac{\hat{S}_C(x_i)}{\hat{S}_C(y_j)}$	-1
$(\mathbf{0}, \mathbf{0})$	$1 - 2 \cdot \frac{\hat{S}_C(x_i)}{\hat{S}_C(y_j)} - 2 \int_{x_i}^{\infty} \frac{\hat{S}_T(t)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$	$-2 \int_{y_j}^{\infty} \frac{\hat{S}_T(t)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) - 1$

Table 3. Value of $\boldsymbol{p_{ij}}$ for a time-to-event outcome using Efron's extension of the Wilcoxon test

$(\boldsymbol{\delta_i, \varepsilon_j})$	$\boldsymbol{x_i - y_j > \tau}$	$\boldsymbol{x_i - y_j < -\tau}$	$\boldsymbol{\lvert x_i - y_j \rvert < \tau}$
$(\mathbf{1}, \mathbf{1})$	1	-1	0
$(\mathbf{0}, \mathbf{1})$	1	$\frac{\hat{S}(y_j + \tau) + \hat{S}(y_j - \tau)}{\hat{S}(x_i)} - 1$	$\frac{\hat{S}(y_j + \tau)}{\hat{S}(x_i)}$
$(\mathbf{1}, \mathbf{0})$	$1 - \frac{\hat{S}(x_i + \tau) + \hat{S}(x_i - \tau)}{\hat{S}(y_j)}$	-1	$-\frac{\hat{S}(x_i + \tau)}{\hat{S}(y_j)}$
$(\mathbf{0}, \mathbf{0})$	$1 - 0.5 \cdot \left(\frac{\hat{S}(x_i + \tau) + \hat{S}(x_i - \tau)}{\hat{S}(y_j)}\right)$	$0.5 \cdot \left(\frac{\hat{S}(y_j + \tau) + \hat{S}(y_j - \tau)}{\hat{S}(x_i)}\right) - 1$	$0.5 \cdot \left(\frac{\hat{S}(y_j + \tau)}{\hat{S}(x_i)} - \frac{\hat{S}(x_i + \tau)}{\hat{S}(y_j)}\right)$

Table 4. Value of p_{ij} for a time-to-event outcome integrating a threshold τ to reflect a clinically relevant difference between two outcomes

Priority	Variable type	$\omega_{ij}(l)$	$p_{ij}(l)$
1	Time to event	1	$\mathbb{P}\left[(x_{i1}^0 > y_{j1}^0 + \tau_1) \middle\| x_{i1},\, y_{j1},\, \delta_{i1},\, \varepsilon_{j1}\right] - \mathbb{P}\left[(y_{j1}^0 > x_{i1}^0 + \tau_1) \middle\| x_{i1},\, y_{j1},\, \delta_{i1},\, \varepsilon_{j1}\right]$
l	Any type	$\mathbb{P}\left[\bigcap_{\lambda=1}^{l-1} \lvert x_{i\lambda}^0 - y_{j\lambda}^0\rvert < \tau_\lambda \middle\| x_{i\lambda},\, y_{j\lambda},\, \delta_{i\lambda},\, \varepsilon_{j\lambda}\right]$	$\mathbb{P}\left[(x_{il}^0 > y_{jl}^0 + \tau_l) \middle\| x_{il},\, y_{jl},\, \delta_{il},\, \varepsilon_{jl}\right] - \mathbb{P}\left[(y_{jl}^0 > x_{il}^0 + \tau_l) \middle\| x_{il},\, y_{jl},\, \delta_{il},\, \varepsilon_{jl}\right]$

Table 5. Value of the score $p_{ij}(l)$ and of the weight $\omega_{ij}(l)$ assigned to each pair of individuals in the case of l prioritized different outcomes

Priority	Variable	$\omega_{ij}(l)$	$p_{ij}(l)$
l	p^{th}occurrence of a time to event outcome	$\mathbb{P}\left[\bigcap_{\lambda=1,\, I_{l\lambda}\neq 1}^{(l-1)} \lvert x_{i\lambda}^0 - y_{j\lambda}^0\rvert < \tau_\lambda \rvert\, x_{i\lambda},\, y_{j\lambda},\, \delta_{i\lambda},\, \varepsilon_{j\lambda}\right]$	$\mathbb{P}\left[(x_{il}^0 - y_{jl}^0 \in [\tau_l, min(\tau_\lambda \mid I_{l\lambda}=1)]) \rvert\, x_{il},\, y_{jl},\, \delta_{il},\, \varepsilon_{jl}\right] -$ $\mathbb{P}\left[(y_{jl}^0 - x_{il}^0 \in [\tau_l, min(\tau_\lambda \mid I_{l\lambda}=1)]) \rvert\, x_{il},\, y_{jl},\, \delta_{il},\, \varepsilon_{jl}\right]$

Table 6. Value of the score $p_{ij}(l)$ and of the weight $\omega_{ij}(l)$ assigned to each pair of individuals in the case of L prioritized outcomes, in the case where the l^{th} outcome is a time-to-event outcome which was possibly included in the analysis at a higher priority with higher threshold value

Figure and Figure Legend:

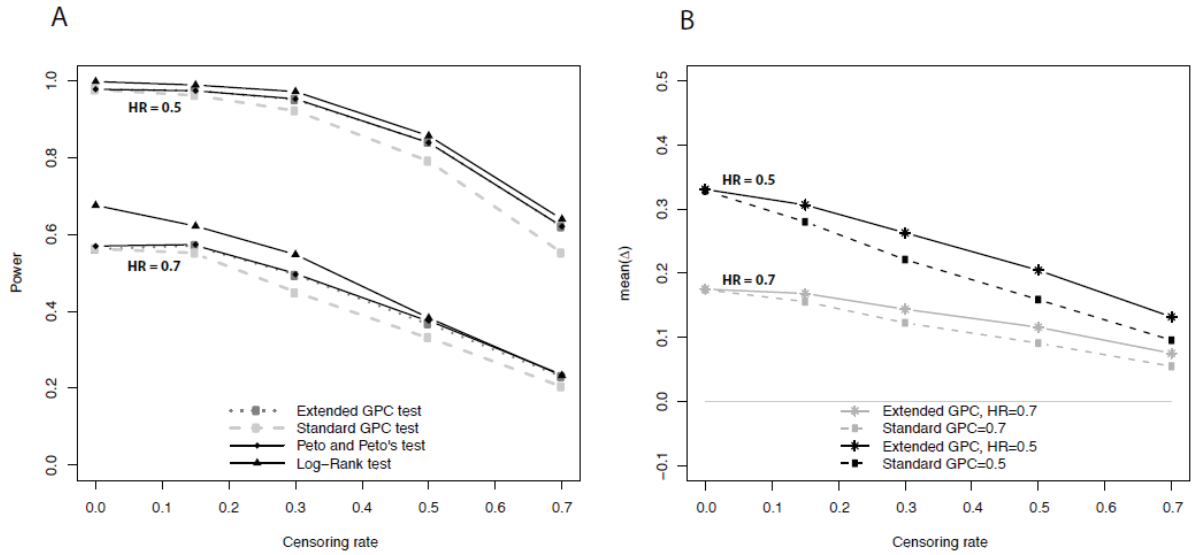


Figure 1. Title: Power (A) and chance of a better outcome (B) of several tests in the case of proportional hazards

Footnotes: Δ = chance of a better outcome; HR = Hazard ratio ; GPC = Generalize pairwise comparisons

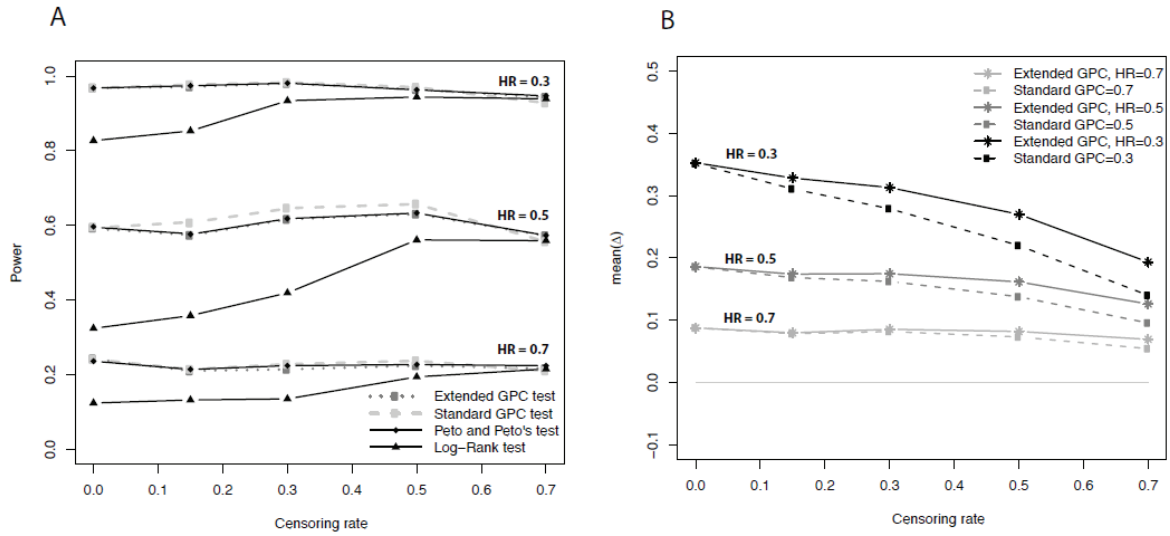


Figure 2. Title: Power (A) and chance of a better outcome (B) of several tests in the case of non-proportional hazards with attenuation of treatment effect over time

Footnotes: Δ = chance of a better outcome; HR = Hazard ratio ; GPC = Generalize pairwise comparisons

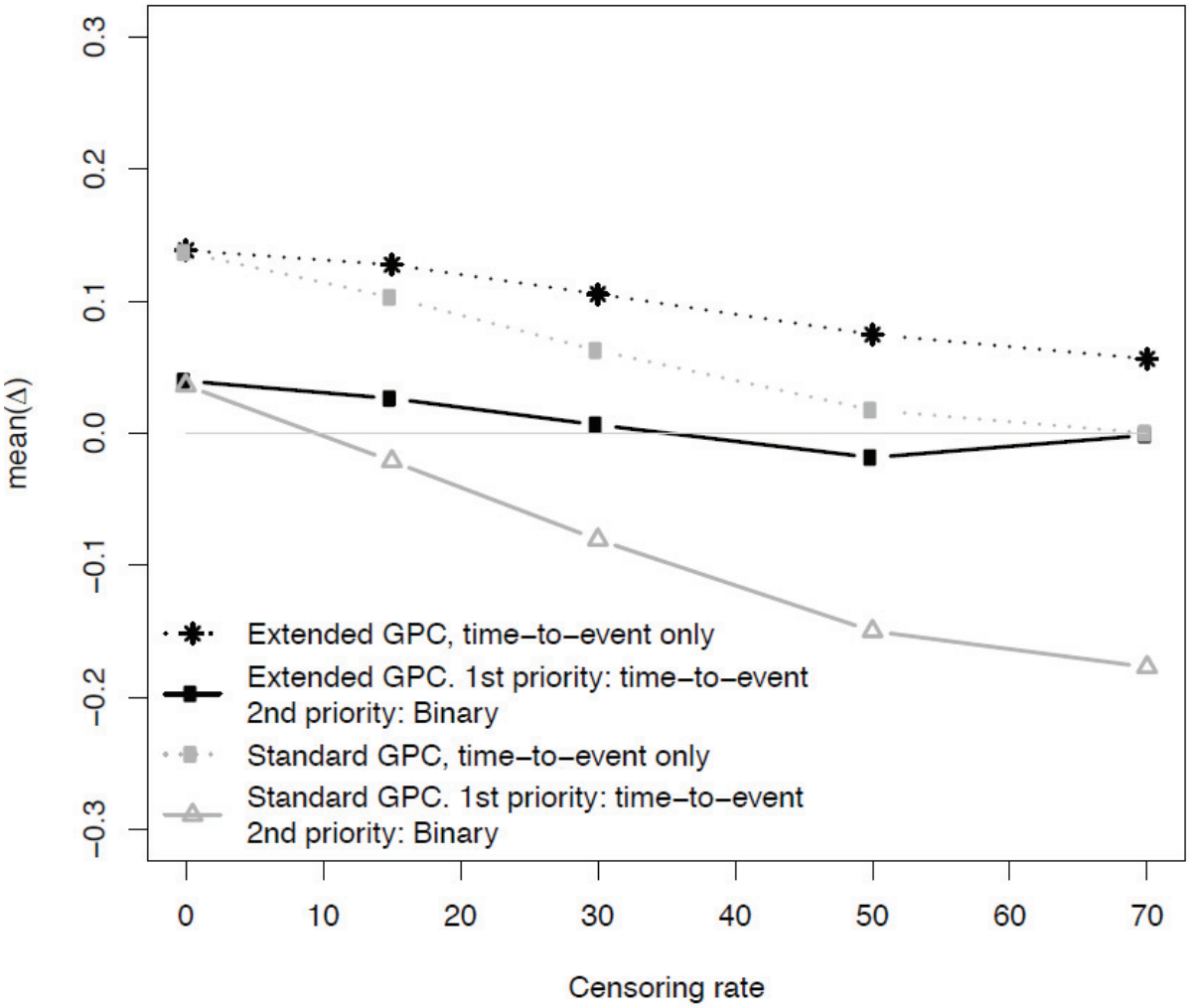


Figure 3. Title : Chance of a better outcome according to the censoring rate when a time-to-event outcome is in favor of the treatment group and a binary outcome is in favor of the control group, and when the survival threshold for clinical relevance is equal to 1 time unit.

Footnotes: Δ = chance of a better outcome; GPC = Generalize pairwise comparisons

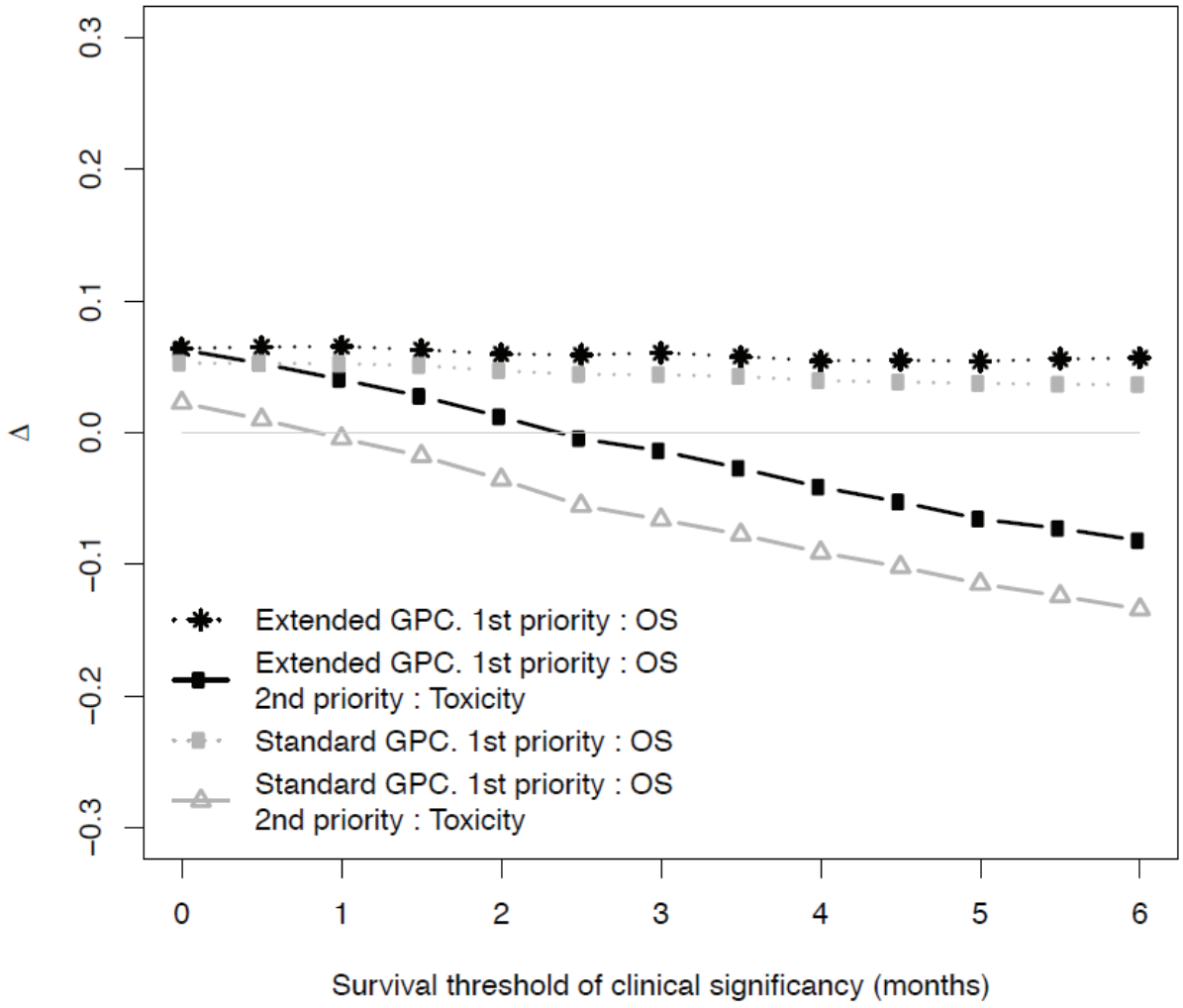
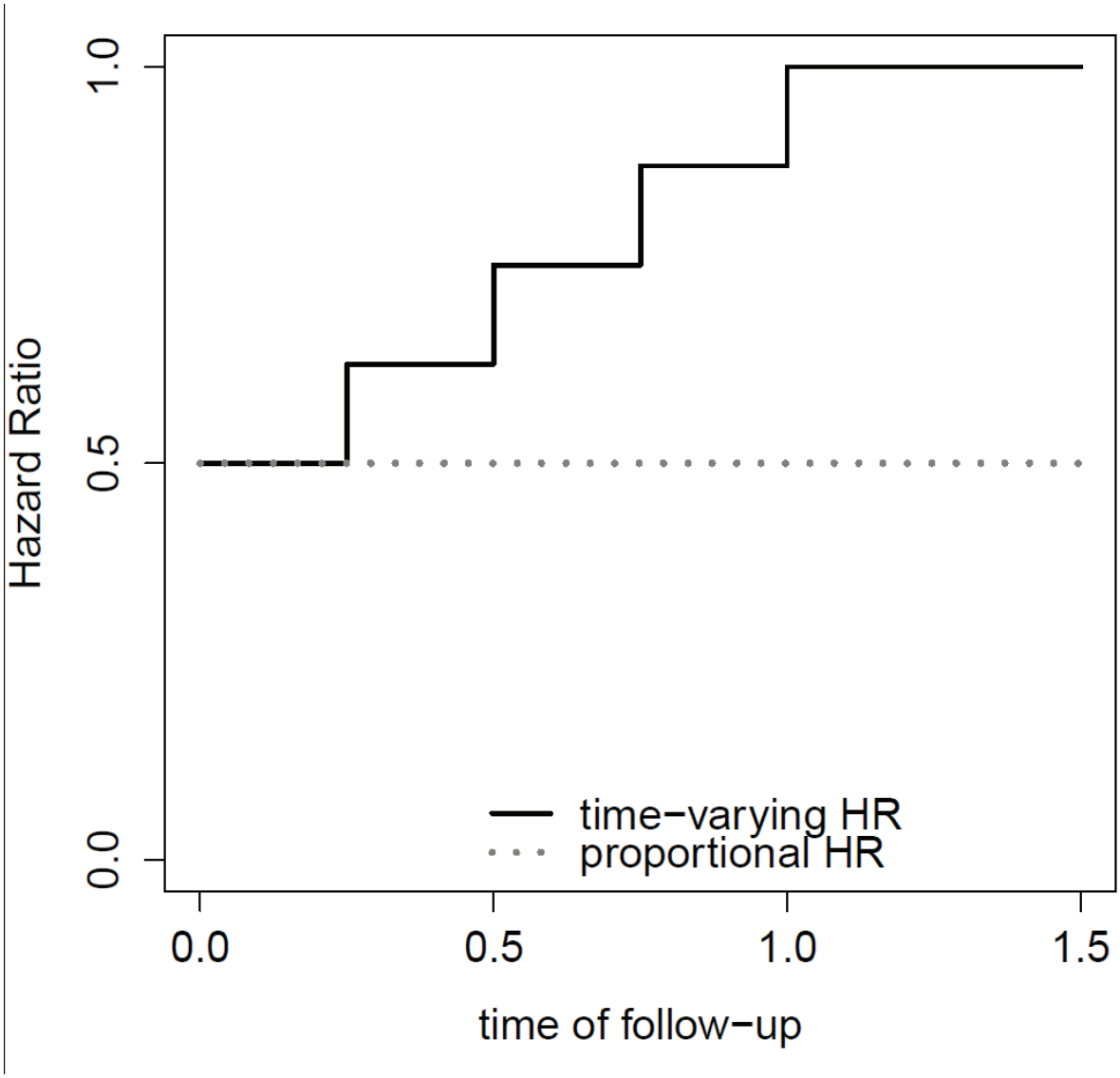


Figure 4. Title : Chance of a better outcome for erlotinib according to the survival threshold for clinical relevance in the PA.3 trial.

Footnotes: Δ = chance of a better outcome; GPC = Generalize pairwise comparisons ; OS = Overall survival.

Appendix :



HR : Hazard ratio in the simulated scenario. $T_{0.5}$: the time at which half of the patients in group T would have presented the event if HR was proportional. Then $T_{0.5} = \frac{\ln(2)}{\lambda_T}$.

Figure A. Hazard ratio versus time used in the scenario 2 of the simulation study: non-proportional hazards with attenuation of treatment effect over time

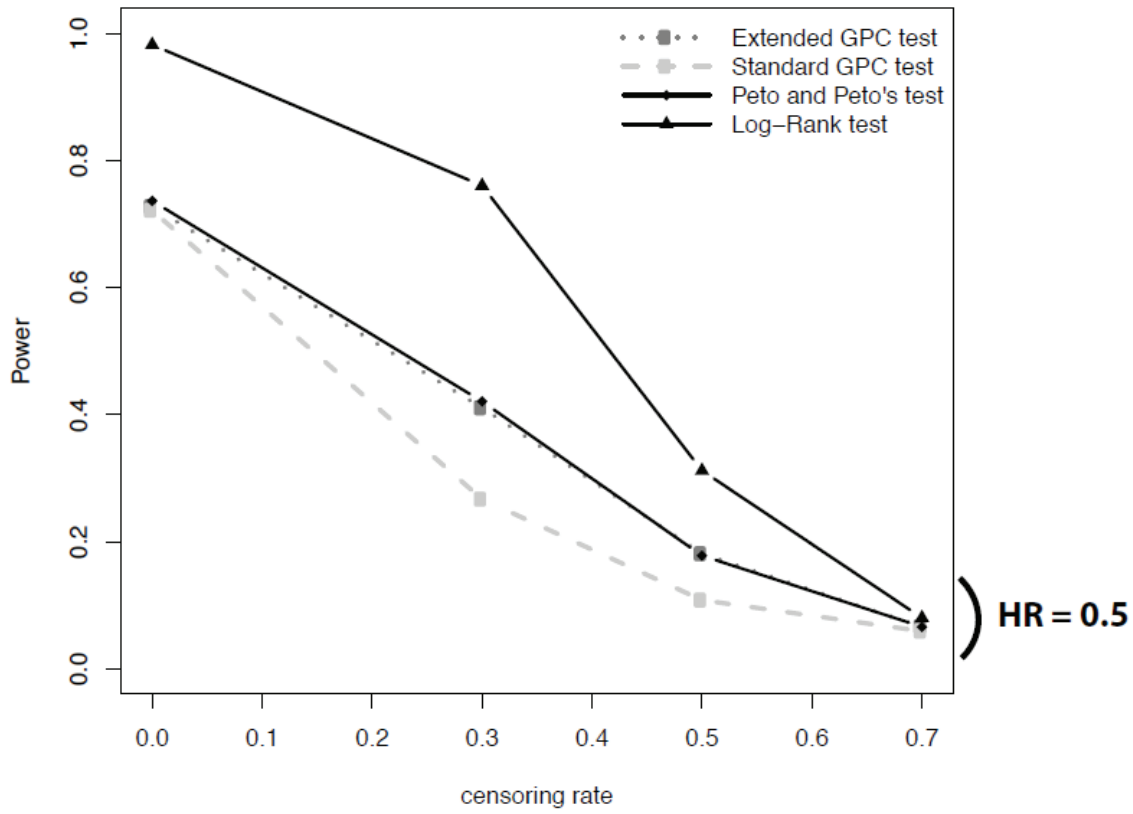


Figure B. Power of several tests in the case of uniform censoring and non-proportional hazards with delayed of treatment effect

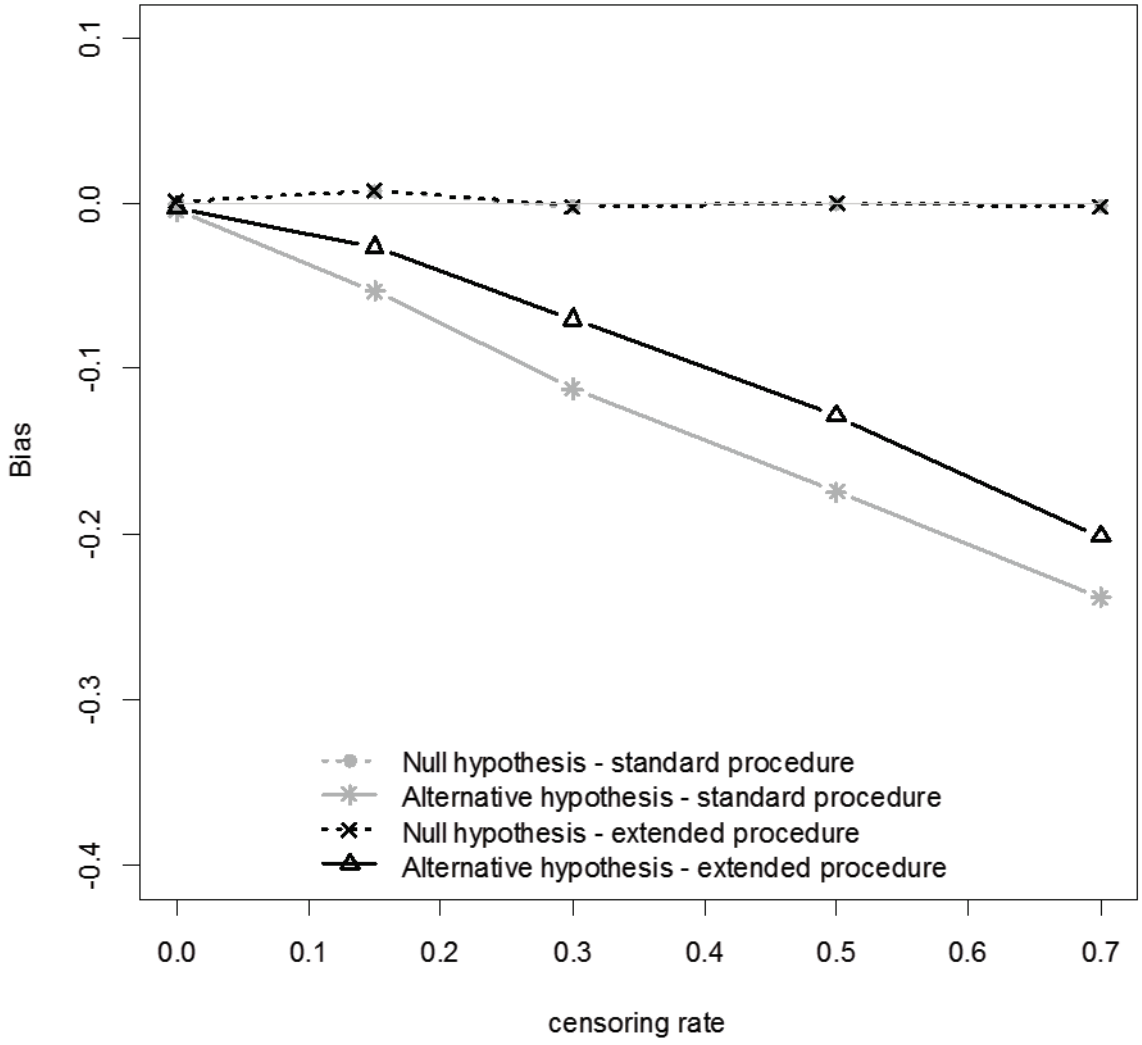


Figure C. Bias of the chance of a better outcome estimate, depending on the censoring rate and of the hazards proportionality.

Footnotes : Null hypothesis = hazard ratio at 0 ; Alternative hypothesis = hazard ratio at 0.5 in a proportional hazards scenario

Dans ce manuscrit, la prise en compte des temps jusqu'à censure par l'extension dite de Peto et Peto a permis d'augmenter la puissance du test de permutation par rapport à la procédure standard des comparaisons par paire. Cette augmentation de puissance du test était attendue car la prise en compte des données censurées est ici très proche de la généralisation du test de Wilcoxon par Peto et Peto. Le test de Wilcoxon généralisé par Peto et Peto était plus puissant que le test de Wilcoxon généralisé par Gehan dans la majorité des situations [13], [15]. La proximité entre l'extension dite de Peto et Peto proposée dans cet article et le test de Wilcoxon généralisé par Peto et Peto est apparente dans la figure 1 du manuscrit, puisque les puissances de ces deux approches sont similaires. Le risque de première espèce calculée était très proche de 5 % dans l'ensemble des scénarios simulés, et quelle que soit la procédure de comparaison par paire utilisée.

Une limite majeure de l'extension dite de Peto et Peto est le biais dans l'estimation de la propension au succès lors de l'analyse d'une variable de type temps jusqu'à événement et en présence de censure. Dans une étude de simulation, le biais d'un estimateur peut être estimé comme la différence entre la valeur théorique et la moyenne de cet estimateur, qui correspond dans notre cas, à la moyenne des valeurs estimées. Le biais de la propension au succès dans une étude avec M jeux de données simulés est donc:

$$Biais = \frac{1}{M}\sum_{i=1}^{M}\hat{\Delta}_i - \Delta^*$$

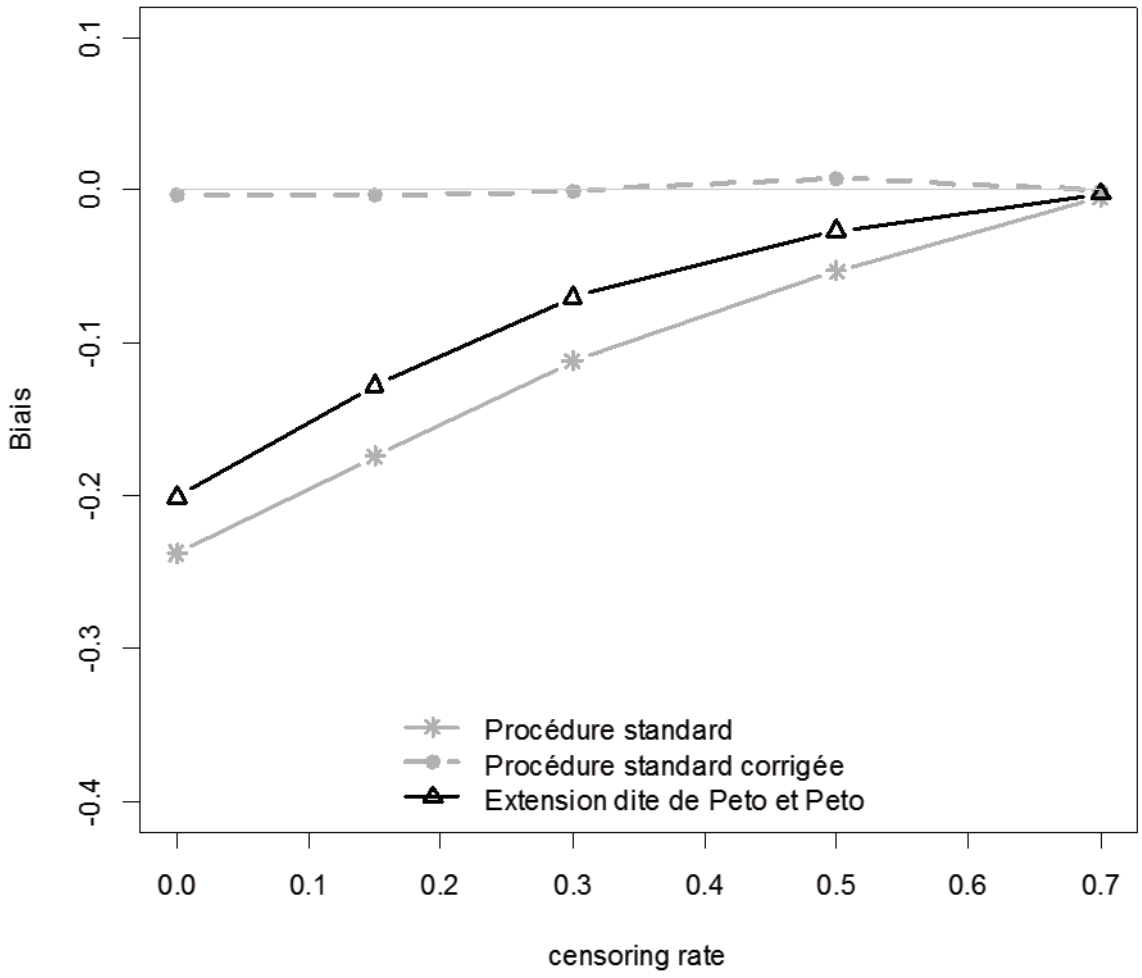
La figure 1 du manuscrit et la figure C de l'annexe illustrent la présence d'un biais dans l'estimation de Δ^*, lorsque $\Delta^* \neq 0$ et en présence de censures. En présence de paires non directement informatives du fait des censures, $\hat{\Delta}$ tend vers 0 lorsque le taux de paires non informatives tend vers 100%. Le biais est plus important lorsque Δ^* est estimée par la procédure standard. Néanmoins dans le cas de la procédure standard, le biais est calculable dans le cas particulier où un seul critère de jugement est inclus dans l'analyse des comparaisons par paire et lorsque $\tau = 0$. Notons f_{sd} la proportion de paires informatives selon la procédure standard. Il est alors possible de calculer une propension au succès corrigée pour chaque essai simulé :

$$\hat{\Delta}corr_i = \frac{\hat{\Delta}_i}{f_{sd_i}}$$

La figure C du manuscrit a donc été reproduite en rajoutant la correction de l'estimation de la propension au succès pour la procédure standard (figure III.1). $\hat{\Delta}corr$ permet d'estimer Δ^* de

façon non biaisée. Néanmoins la correction n'est pas applicable lorsque plusieurs critères de jugement sont analysés simultanément à des priorités successives. Cette correction n'est donc pas utilisable dans le cadre d'une utilisation pratique des comparaisons par paire généralisées. La procédure étendue dite de Peto et Peto ne dispose pas de ce type de correction du fait du calcul de la participation au score p_{ij} des paires non directement informatives.

Figure III-1. Biais de la propension au succès observé lorsqu'un seul critère de jugement de type temps jusqu'à événement est analysé



Lorsque l'hypothèse nulle est vraie, l'analyse de la figure 1 du manuscrit montre que $\hat{\Delta}$ est non biaisée, que l'estimation soit réalisée avec la procédure standard ou avec la procédure étendue dite de Peto et Peto. En conclusion la procédure étendue dite de Peto et Peto est

utilisable et performante pour tester l'hypothèse nulle $\Delta = 0$, mais ne permet pas d'estimer correctement Δ lorsque $\Delta \neq 0$ et en présence de censures.

III.2.c. Extension dite de Efron

Le développement de l'extension dite de Efron a été motivé par l'observation du biais dans l'estimation de la propension au succès avec l'extension dite de Peto et Peto. L'hypothèse était que ce biais était secondaire à l'utilisation d'une estimation conjointe de la fonction de survie à partir des observations de l'ensemble des patients, quel que soit le groupe de traitement. Le calcul de p_{ij} sous cette hypothèse aboutit intuitivement à une sous-estimation systématique de la différence d'effet thérapeutique entre les deux groupes. Dans l'extension dite de Efron, la participation au score de chaque paire de patient repose sur x_i, y_j, δ_i, ε_j, ainsi que sur $\hat{S}_T(t)$ et $\hat{S}_C(t)$ - les estimations des fonctions de survie par la méthode de Kaplan et Meier pour les patients issus du groupe T ($S_T(t) = \mathbb{P}\ [x_i^0 \geq t]$) et du groupe C ($S_C(t) = \mathbb{P}\left[y_j^0 \geq t\right]$) respectivement -. L'objectif est le même que celui qui motivait le développement de l'extension dite de Peto et Peto. Pour chaque paire de patients $\{i, j\}$ non directement classable du fait de censures, il s'agit de calculer p_{ij}, estimation de $\mathbb{P}\left[x_i^0 > y_j^0 + \tau |\, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right] - \mathbb{P}\left[y_j^0 > x_i^0 + \tau |\, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$. Dans cette extension, l'estimation des fonctions de survie traite la plus large observation de chaque groupe comme un événement. Les δ_i et ε_j correspondant aux x_i et y_j les plus élevées prennent donc la valeur 1. Cette méthode, proposée par Efron, permet de s'affranchir de la portion non estimable des fonctions de survie lorsque la dernière observation est une censure. On peut alors estimer $\mathbb{P}\left[x_i^0 > y_j^0 + \tau |\, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ (tableau III.6) et $\mathbb{P}\left[y_j^0 > x_i^0 + \tau |\, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ (tableau III.7), ainsi que la participation au score de chaque paire p_{ij} (tableau III.8).

Tableau III-6. Estimation de $\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$ dans l'extension dite de Efron

$(\delta_i, \varepsilon_j)$	$x_i - y_j > \tau$	$x_i - y_j < -\tau$	$\lvert x_i - y_j \rvert < \tau$
$(1,1)$	1	0	0
$(0,1)$	1	$\frac{\hat{S}_T(y_j+\tau)}{\hat{S}_T(x_i)}$	$\frac{\hat{S}_T(y_j+\tau)}{\hat{S}_T(x_i)}$
$(1,0)$	$1 - \frac{\hat{S}_C(x_i-\tau)}{\hat{S}_C(y_j)}$	0	0
$(0,0)$	$1 - \frac{\hat{S}_C(x_i-\tau)}{\hat{S}_C(y_j)} - \int\limits_{\substack{t>x_i-\tau \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t+\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$	$- \int\limits_{\substack{t>y_j \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t+\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$	$- \int\limits_{\substack{t>y_j \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t+\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$

Tableau III-7. Estimation de $\mathbb{P}\left[y_j^0 > x_i^0 + \tau | x_i, y_j, \delta_i, \varepsilon_j\right]$ dans l'extension dite de Efron

$(\delta_i, \varepsilon_j)$	$x_i - y_j > \tau$	$x_i - y_j < -\tau$	$\lvert x_i - y_j \rvert < \tau$
$(1,1)$	0	1	0
$(0,1)$	0	$1 - \frac{\hat{S}_T(y_j - \tau)}{\hat{S}_T(x_i)}$	0
$(1,0)$	$\frac{\hat{S}_C(x_i + \tau)}{\hat{S}_C(y_j)}$	1	$\frac{\hat{S}_C(x_i + \tau)}{\hat{S}_C(y_j)}$
$(0,0)$	$\frac{\hat{S}_C(x_i + \tau)}{\hat{S}_C(y_j)} + \int_{\substack{t > x_i + \tau \\ t \in \{y_j\} \\ \varepsilon_j = 1}}^{\infty} \frac{\hat{S}_T(t - \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$	$1 + \int_{\substack{t > y_j \\ t \in \{y_j\} \\ \varepsilon_j = 1}}^{\infty} \frac{\hat{S}_T(t - \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$	$\int_{\substack{t > x_i + \tau \\ t \in \{y_j\} \\ \varepsilon_j = 1}}^{\infty} \frac{\hat{S}_T(t - \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) + \frac{\hat{S}_C(x_i + \tau)}{\hat{S}_C(y_j)}$

Tableau III-8. Calcul de p_{ij} dans l'extension dite de Efron

$(\delta_i, \varepsilon_j)$	$x_i - y_j > \tau$	$x_i - y_j < -\tau$	$\lvert x_i - y_j \rvert < \tau$
$(1,1)$	1	-1	0
$(0,1)$	1	$\frac{\hat{S}_T(y_j+\tau)+\hat{S}_T(y_j-\tau)}{\hat{S}_T(x_i)} - 1$	$\frac{\hat{S}_T(y_j+\tau)}{\hat{S}_T(x_i)}$
$(1,0)$	$1 - \frac{\hat{S}_C(x_i+\tau)+\hat{S}_C(x_i-\tau)}{\hat{S}_C(y_j)}$	-1	$-\frac{\hat{S}_C(x_i+\tau)}{\hat{S}_C(y_j)}$
$(0,0)$	$1 - \frac{\hat{S}_C(x_i-\tau)+\hat{S}_C(x_i+\tau)}{\hat{S}_C(y_j)} - \int_{\substack{t>x_i-\tau \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t+\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) - \int_{\substack{t>x_i+\tau \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t-\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$	$- \int_{\substack{t>y_j \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t+\tau)+\hat{S}_T(t-\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) - 1$	$- \int_{\substack{t>y_j \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t+\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) - \int_{\substack{t>x_i+\tau \\ t\in\{y_j\} \\ \varepsilon_j=1}}^{\infty} \frac{\hat{S}_T(t-\tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) - \frac{\hat{S}_C(x_i+\tau)}{\hat{S}_C(y_j)}$

L'estimateur de Kaplan et Meier est une fonction en escalier, continue à droite et discontinue à gauche des événements seulement. On peut donc calculer :

$$\int_{t_a}^{\infty} \hat{S}_T(t)\, d\hat{S}_C(t) = \sum_{\substack{y_j> \\ t_a}}^{\infty} \hat{S}_T(t).\left(\hat{S}_C(t+1) - \hat{S}_C(t)\right)$$

Un défaut théorique de l'extension dite de Efron est la transformation des données de δ_i et ε_j correspondant aux x_i et y_j les plus élevés lorsque ceux-ci ont la valeur 0. Cette transformation est intuitivement inacceptable lorsque les probabilités de survie estimées sont non négligeables aux temps précédents les x_i et y_j les plus élevés. Ce défaut théorique nous a amené à proposer l'extension dite de Péron. Les évaluations des extensions dites de Efron et de Péron ont été réalisées de façons simultanées, et seront présentées conjointement dans le sous-chapitre III.2.d.

III.2.d. Extension dite de Péron

Dans cette extension, dite de Péron, l'estimation des probabilités de survie n'est pas réalisée après les x_i et y_j les plus élevées lorsque δ_i et/ou ε_j sont égales à 0. Pour chaque paire non directement classable, la probabilité qu'elle soit classée favorable, défavorable ou neutre à une priorité l est calculée. Néanmoins la somme de ces probabilités n'est pas obligatoirement égale à 1, car il existe également une probabilité que la paire soit classée non informative, en conséquence directe des zones de temps où la fonction de survie n'est pas estimable par la méthode de Kaplan et Meier. De ce fait l'égalité $\widehat{\mathbb{P}}\left[x_i^0 > y_j^0 | \, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right] = 1 - \widehat{\mathbb{P}}\left[y_j^0 > x_i^0 | \, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ n'est plus vraie lorsque $\mathbb{P}\left[x_i^0 > y_j^0 | \, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ et $\mathbb{P}\left[y_j^0 > x_i^0 | \, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ sont estimés par la méthode de Péron. L'estimation de $\mathbb{P}\left[x_i^0 > y_j^0 | \, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ se réalise de façon identique à la façon décrite dans l'extension dite de Efron (tableau III.6). L'estimation de $\mathbb{P}\left[y_j^0 > x_i^0 | \, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ doit être réalisée de façon symétrique telle que présentée dans le tableau III.9. La méthode de calcul de $\widehat{\mathbb{P}}\left[x_i^0 > y_j^0 | \, x_i,\, y_j,\, \delta_i,\, \varepsilon_j\right]$ est présentée en annexe A de cette thèse.

Tableau III-9. Estimation de $\mathbb{P}\left[y_j^0 > x_i^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$ dans l'extension dite de Péron

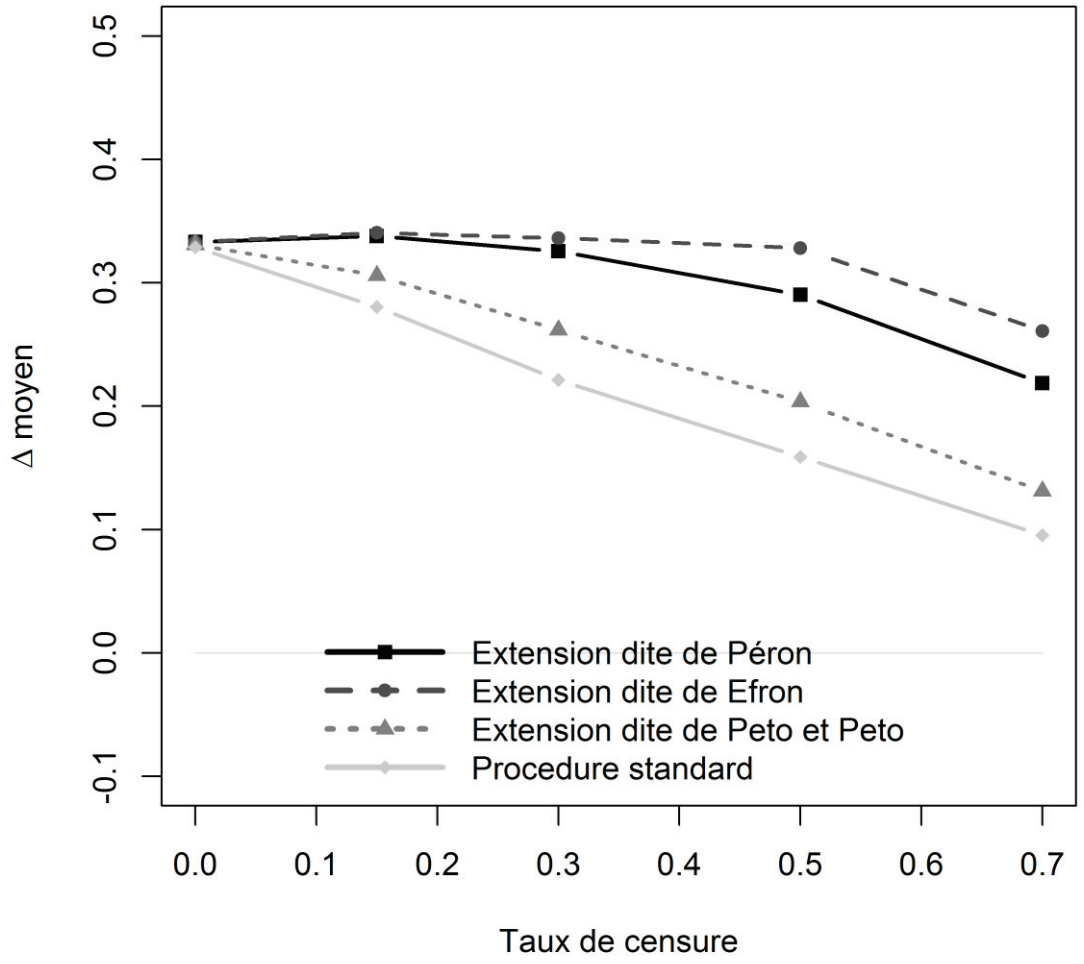
$(\delta_i, \varepsilon_j)$	$x_i - y_j > \tau$	$x_i - y_j < -\tau$	$\lvert x_i - y_j \rvert < \tau$
$(\mathbf{1}, \mathbf{1})$	0	1	0
$(\mathbf{0}, \mathbf{1})$	0	$1 - \frac{\hat{S}_T(y_j - \tau)}{\hat{S}_T(x_i)}$	0
$(\mathbf{1}, \mathbf{0})$	$\frac{\hat{S}_C(x_i + \tau)}{\hat{S}_C(y_j)}$	1	$\frac{\hat{S}_C(x_i + \tau)}{\hat{S}_C(y_j)}$
$(\mathbf{0}, \mathbf{0})$	$-\int_{\substack{t > x_i \\ t \in \{x_i\} \\ \delta_i = 1}}^{\infty} \frac{\hat{S}_C(t + \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_T(t)$	$1 - \frac{\hat{S}_T(y_j - \tau)}{\hat{S}_T(x_i)} - \int_{\substack{t > y_j - \tau \\ t \in \{x_i\} \\ \varepsilon_j = 1}}^{\infty} \frac{\hat{S}_C(t + \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_T(t)$	$-\int_{\substack{t > x_i \\ t \in \{x_i\} \\ \delta_i = 1}}^{\infty} \frac{\hat{S}_C(t + \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_T(t)$

La procédure d'évaluation de l'extension dite de Peto et Peto décrite dans l'article inclus dans le sous-chapitre III.2.b a été répétée afin d'évaluer les extensions dites de Efron et de Péron. Une étude de simulation a été réalisée. Pour chacun des scénarios présentés, 1000 jeux de données ont été simulés, incluant 200 patients répartis en deux groupes de traitement de même taille. Une variable de type temps jusqu'à événement était simulée pour chaque patient en utilisant la distribution exponentielle. Le paramètre d'échelle était fixé à 1 pour les patients du groupe C, et variait pour les patients du groupe T selon les scénarios. Dans tous les scénarios, les temps de censures étaient distribués de façon uniforme, avec la même distribution dans les deux groupes de traitement. Les valeurs des paramètres de la distribution uniforme variaient afin de faire varier le taux de censure effectif. Les objectifs de cette étude de simulation étaient de rechercher et de quantifier un biais dans l'estimation de la propension au succès, d'évaluer la puissance du test de l'hypothèse nulle, et de quantifier le risque de première espèce α.

Scenario 1: Une variable de type temps jusqu'à événement – taux instantanés d'événement proportionnels

Dans le premier scenario, les groupes de traitement sont comparés sur une seule variable de type temps jusqu'à événement. Le taux instantané de décès dans le groupe T était simulé de façon proportionnel au taux instantané de décès dans le groupe C (paramètre d'échelle dans le groupe T : $\lambda_T = \mathrm{HR} \times \lambda_C$, où HR est le rapport des taux instantanés de décès). Pour chaque jeu de données simulé, la variable de type temps jusqu'à événement était comparée entre les deux groupes en utilisant la procédure standard des comparaisons par paire, l'extension dite de Peto et Peto, l'extension dite de Efron, et l'extension dite de Péron. La propension au succès était estimée à partir des quatre procédures, en prenant comme seuil de significativité clinique $\tau = 0$. La moyenne des estimations de la propension au succès est représentée sur la figure III.2.

Figure III-2. Estimation de la propension au succès selon les quatre procédures de comparaison par paire pour une variable de type temps jusqu'à événement. Les taux instantanés de décès des groupes T et C sont proportionnels, et le rapport des taux instantanés de décès est de 0.5.



L'estimation de la propension au succès est moins dépendante du taux de censure lorsqu'elle est réalisée selon l'extension dite de Efron. L'estimation de la propension au succès par la méthode dite de Péron est également moins dépendante du taux de censure par rapport à l'extension de Peto et Peto et à la procédure standard.

Pour les quatre procédures, l'estimation de Δ^* tend vers zéro lorsque le taux de censure est élevé, même si cette tendance est moins forte pour les extensions dites de Efron et de Péron. Dans le sous-chapitre III.2.b nous avons vu qu'il est possible de calculer une propension au succès corrigée à partir de la propension au succès estimée par la procédure standard et de f_{sd} la proportion de paires informatives. Ceci est valable dans le cas particulier

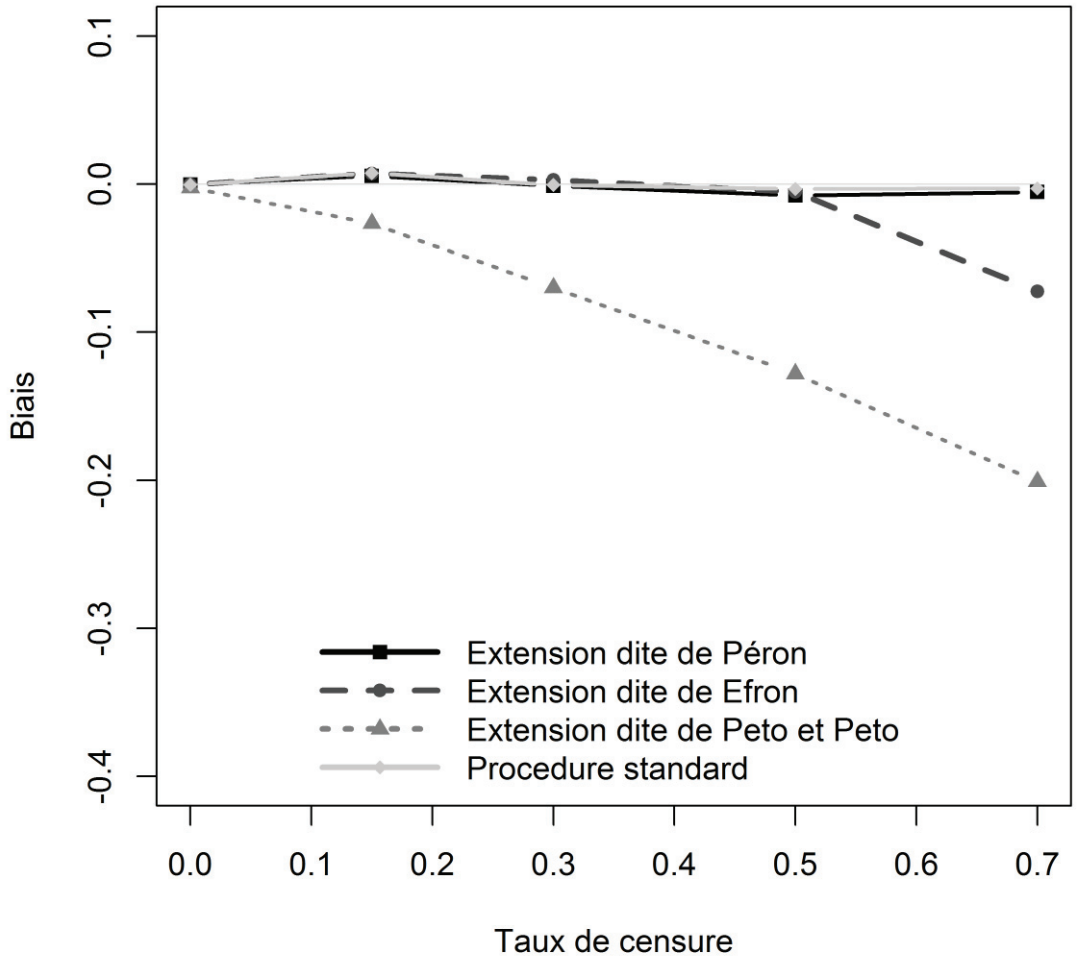
où un seul critère de jugement est inclus dans l'analyse des comparaisons par paire et lorsque $\tau = 0$. De la même façon, il est possible de corriger la propension au succès estimée par l'extension dite de Péron à partir de la proportion f_{Per} de paires informatives.

$$\hat{\Delta}corr_i = \frac{\hat{\Delta}_i}{f_{Per}}$$

Il n'y a pas de méthode de correction équivalente pour l'extension dite de Peto et Peto et l'extension dite de Efron. On peut noter que $f_{Peto} = f_{Efron} = 1$.

Le biais de l'estimation de la propension corrigée au succès est nul lorsque l'estimation est réalisée par la procédure standard ou la procédure dite de Péron, même lorsque le taux de censure est élevé. Le biais de l'estimation de la propension au succès par l'extension dite de Efron apparait lorsque les taux de censure sont élevés. En utilisant l'extension dite de Peto et Peto, l'estimation de la propension au succès est biaisée même pour des taux de censure faibles (figure III.3). Néanmoins les corrections ne sont pas applicables lorsque plusieurs critères de jugement priorisés sont inclus dans une procédure de comparaison par paire. Ces corrections ne sont donc pas utilisables dans le cadre d'une utilisation pratique des comparaisons par paire généralisées.

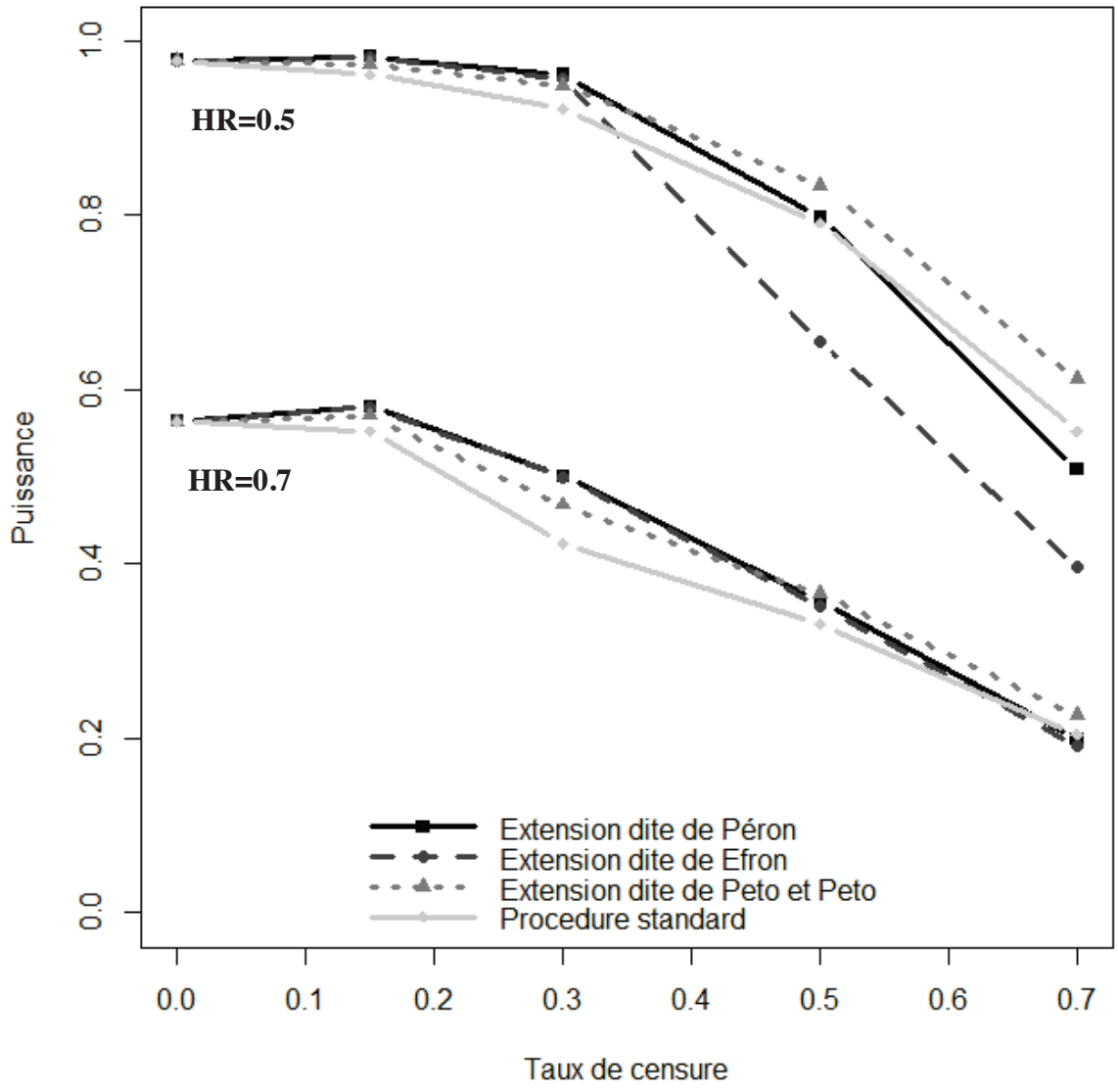
Figure III-3. Biais de la propension au succès corrigée estimée par la procédure standard et l'extension dite de Péron, ainsi que de la propension au succès estimée par les extensions dites de Efron et de Peto et Peto. Les groupes T et C sont comparés sur une variable de type temps jusqu'à événement. Les taux instantanés de décès sont proportionnels, et le rapport des taux instantanés de décès est de 0.5.



Pour chaque réglage des paramètres de simulation, la proportion de jeux de données pour lesquels la p-valeur était inférieure à 0.05 permettait d'estimer le risque α lorsque HR = 1 et d'estimer la puissance du test lorsque HR ≠ 1. Le risque de première espèce α était estimé autour de 5% pour toutes les valeurs du taux de censure et quelle que soit la procédure de comparaison par paire utilisée. En présence d'un taux de censure modéré, entre 15 et 30%, les procédures de comparaison par paires étendues dites de Efron et de Péron étaient plus puissantes que la procédure standard, et légèrement plus puissante que la procédure étendue dite de Peto et Peto. Pour des valeurs élevées des taux de censure, la procédure dite de Efron

était la moins puissante. L'extension dite de Peto et Peto était alors légèrement plus puissante que la procédure standard et que l'extension dite de Péron (Figure III.4)

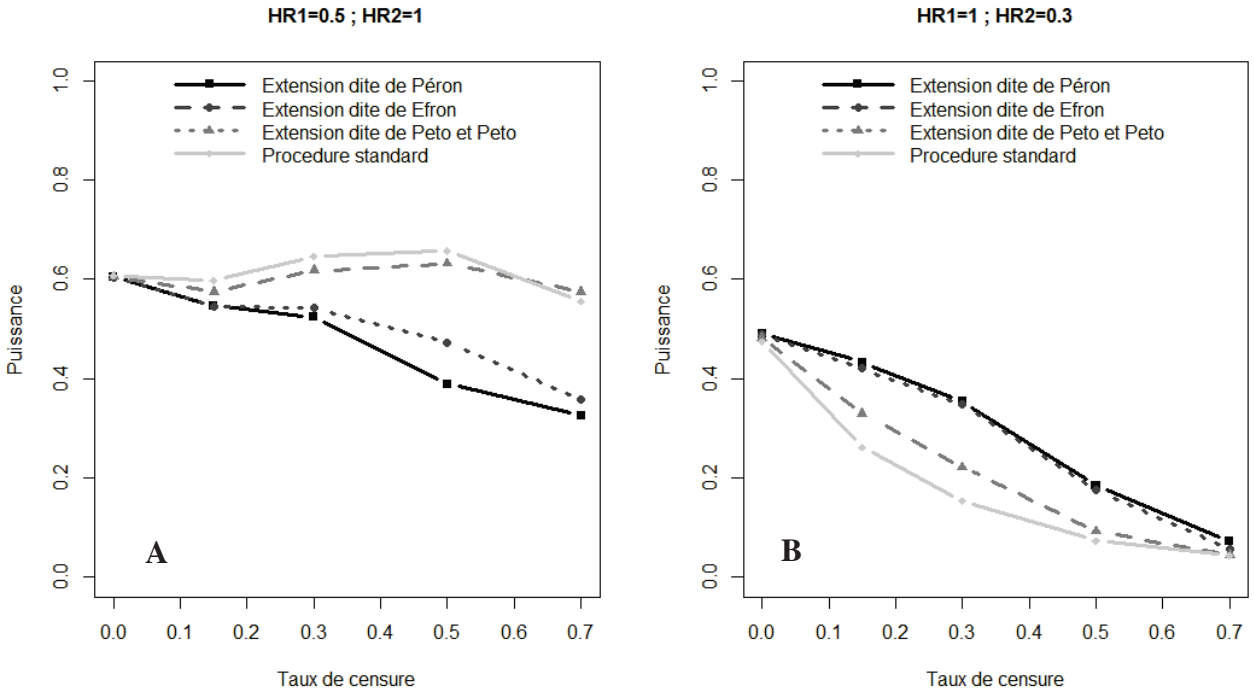
Figure III-4. Puissance des quatre procédures de comparaison par paire pour une variable de type temps jusqu'à événement. Les taux instantanés de décès sont proportionnels, et le rapport des taux instantanés de décès (HR) est de 0.5 ou de 0.7.



Scenario 2: Une variable de type temps jusqu'à événement – taux instantanés d'événement non proportionnels

Dans ce second scénario, les groupes de traitement étaient à nouveau comparés à partir d'une seule variable de type temps jusqu'à événement. Les taux instantanés de décès n'étaient pas proportionnels, et le rapport des taux instantanés de décès (HR) évoluait progressivement en quatre étapes, de HR1 pour $t = 0$ à HR2 à la fin du suivi. ($HR = HR_1$ pour $t < \frac{T_{0.5}}{2}$ où $T_{0.5}$ est le temps auquel la moitié des patients du groupe T ont présentés l'événement ; $HR = 0.75 \times HR_1 + 0.25 \times HR_2$ pour $\frac{T_{0.5}}{2} < t < T_{0.5}$; $HR = 0.5 \times HR_1 + 0.5 \times HR_2$ si $T_{0.5} < t < 3.\frac{T_{0.5}}{2}$; $HR = 0.25 \times HR_1 + 0.75 \times HR_2$ si $3.\frac{T_{0.5}}{2} < t < 2.T_{0.5}$; et $HR = HR_2$ si $t > 2.T_{0.5}$). $HR_1 < HR_2$ lorsque l'objectif est de simuler une d'atténuation dans le temps de l'effet thérapeutique. A l'inverse, lorsque l'objectif est de simuler un effet thérapeutique différé dans le temps, $HR_1 > HR_2$. Lorsque l'effet thérapeutique était atténué dans le temps (figure III.5.A), la procédure standard et l'extension dite de Peto et Peto étaient les plus puissantes en présence de censures. Le manque de puissance des extensions dites de Efron et de Péron dans ce scénario était certainement lié à une meilleure prise en compte des probabilités de survie à des temps avancés de suivi, temps auxquels le rapport des taux instantanés de décès était égal à 1. A l'inverse, lorsque l'effet thérapeutique était différé dans le temps (figure III.5.B), les extensions dites de Efron et de Péron étaient les plus puissantes en présence de censures.

Figure III-5. Puissance des quatre procédures de comparaison par paire pour une variable de type temps jusqu'à événement. Les taux instantanés de décès sont non proportionnels. L'effet du traitement est précoce (A), ou différé dans le temps (B).



En conclusion de ce chapitre, les extensions dites de Péron et de Efron ont des performances relativement proches pour analyser les variables de type temps jusqu'à événement. Ces deux extensions sont plus puissantes que la procédure standard en cas d'effet thérapeutique différé dans le temps, et légèrement plus puissantes en cas d'effet proportionnel. En cas de taux de censure très élevés la procédure dite de Efron est mise en défaut. Dans ce cas, l'extension dite de Péron fournit après correction une estimation non biaisée de la propension au succès, et permet de rejeter l'hypothèse nulle avec une meilleure puissance. L'extension dite de Peto et Peto est performante pour rejeter l'hypothèse nulle. Son profil de puissance est légèrement meilleur que celui de la procédure standard. Néanmoins l'estimation de la propension au succès par l'extension dite de Peto et Peto présente un biais non corrigeable sous l'hypothèse alternative. Un des intérêts majeurs de la procédure de comparaisons par paire étant d'estimer un paramètre cliniquement pertinent, nous recommandons d'utiliser l'extension dite de Péron pour analyser les variables de type temps jusqu'à événement.

III.2.e. Relation avec d'autres mesures de l'effet d'un traitement

Lorsqu'un seul critère de jugement est analysé, Buyse a montré que la propension au succès était équivalente à des paramètres classiques de mesure de l'effet d'un traitement [1].

Dans le cas d'une variable binaire, la propension au succès correspond à la différence absolue du risque, $pT - pC$, où pT et pC sont les probabilités de succès dans les groupes traitement et contrôle respectivement. Le test de permutation pour la propension au succès est une approximation de Monte Carlo du test exact de Fisher [1].

Dans le cas d'une variable continue, il existe une relation directe entre la propension au succès Δ et la statistique W_{MW} de Mann-Whitney.

$$W_{MW} = \frac{1}{2} \cdot n \cdot m \cdot (1 - \Delta)$$

Δ correspond à la statistique U utilisé dans le test de Wilcoxon :

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

Le test de permutation pour la propension au succès est alors une approximation de Monte Carlo du test exact de Wilcoxon [1].

Dans le cas d'un critère de jugement de type temps jusqu'à événement, les liens entre les procédures proposées et les généralisations du test de Wilcoxon décrites par Gehan, Peto et Peto, et par Efron ont été discutés plus haut dans ce chapitre.

Harrell et al [16] ont introduit dans la communauté biomédicale le *C*-index dans un objectif d'évaluer les performances prédictives sur une variable de type temps jusqu'à événement d'une variable continue. Ce paramètre est une mesure de la probabilité de concordance entre deux observations bivariées (x_i, z_i) et (x_j, z_j).

$$\mathbb{P}(x_i > x_j | z_i > z_j)$$

Son utilisation est large pour évaluer les performances prédictives d'un modèle. Considérons pour rester cohérent avec les notations précédemment utilisées que x_i est la valeur observée d'un critère de jugement de type temps jusqu'à événement pour le $i^{ème}$ sujet (i=1, …, m) du groupe T, et y_j est la valeur observée du même un critère de jugement de type temps jusqu'à événement pour le $j^{ème}$ sujet (j=1, …, n) du groupe C. Considérons maintenant

que z_i est une indicatrice du groupe de traitement qui prend la valeur 0 si le patient est issu du groupe contrôle et 1 si le patient est issu du groupe traitement. Le lien entre la probabilité de concordance et la comparaison de deux distributions de survie par la propension au succès est alors :

$$\Delta = 2 \cdot \mathbb{P}(x_i > y_j | z_i = 1, z_j = 0) - 1$$

Le calcul de la statistique C de Harrell se calcule en considérant l'ensemble de toutes les $m.n$ paires de patients, l'un étant issu du groupe T et l'autre du groupe C. Les paires pour lesquelles le temps jusqu'à événement le plus court correspond à une censure sont exclues du calcul. Pour les paires restantes, un score de 1 est attribué si $x_i > y_j$, et un score de 0 si $x_i < y_j$. C est alors la somme des scores divisée par le nombre de paires évaluables. En l'absence de censure, C correspond à la statistique U de Mann-Whitney. En présence de censures à droite, même non-informatives, Koziol et al ont montré que la valeur de C dépendait de la distribution des censures [17]. La statistique D de Somer [18] est liée au C de Harrell par :

$$D = 2 \cdot C - 1.$$

Gönen et Heller ont proposé d'utiliser le CPE (*concordance probability estimate*) pour estimer la probabilité de concordance en présence de censures sous l'hypothèse d'un rapport des taux instantanés de décès proportionnel. Pour estimer la probabilité de concordance, Gönen et Heller n'utilisent pas directement les observations, mais les estimations des coefficients de régression dans un modèle de Cox à taux proportionnels. Ils ont montré par simulation que leur estimation de la probabilité de concordance n'était pas dépendante de la distribution des temps de censure, à l'inverse du C de Harrell [19]. Néanmoins leur estimation de la probabilité de concordance nécessite que les paramètres estimés dans le modèle de Cox soient corrects, et donc que l'hypothèse des taux proportionnels soit vérifiée.

La propension au succès est une mesure générale de l'effet thérapeutique qui est indépendante de la nature de la variable considérée. Δ a une relation directe avec l' « index de probabilité » proposé par plusieurs auteurs [20]. L'index de probabilité est noté $\mathbb{P}(X > Y)$ et est défini comme la probabilité qu'un patient pris au hasard dans le groupe traitement ait un meilleur résultat thérapeutique qu'un patient pris au hasard dans le groupe contrôle.

$$\Delta = 2 \cdot \mathbb{P}(X > Y) - 1$$

Un avantage de la propension au succès sur l'index de probabilité réside dans son interprétation. Par exemple un index de probabilité $\mathbb{P}(X > Y) = 0.5$ doit être interprété comme une absence d'effet thérapeutique. Cette situation correspond à $\Delta = 0$, qui semble une façon plus directe d'indiquer une absence d'effet.

Le Win ratio a été proposé en 2012 par Pocock et al dans le même objectif que les comparaisons par paire généralisées [21]. Deux approches sont proposées. Dans la première approche du Win ratio, chaque patient du groupe traitement est apparié à un patient du groupe contrôle, en fonction d'un score de risque. Dans la seconde approche, l'ensemble des paires incluant un patient du groupe T et un patient du groupe C est considéré. Deux critères de jugement sont considérés et hiérarchisés. Chaque paire est classée favorable au traitement, défavorable au traitement ou neutre selon une stratégie de classement proche de celle qui est réalisée dans la procédure standard des comparaisons par paire généralisées. Le Win ratio est le rapport entre le nombre de paires classées favorables et le nombre de paires classées défavorables. Les auteurs proposent une solution pour calculer l'intervalle de confiance à 95% et la P-value associée au Win ratio. La principale différence entre la seconde approche du Win ratio et la propension au succès réside certainement dans la construction du paramètre. La propension au succès est la différence entre la proportion de paires favorables et défavorables, alors que le Win ratio est le rapport de ces deux proportions. La propension au succès a une interprétation qui peut sembler plus aisée, du fait de son rapport avec des statistiques déjà largement utilisées. La métrique utilisée s'étend entre -1 et 1, et l'absence d'effet thérapeutique correspond à une propension au succès à 0. Le Win ratio est une métrique qui s'étend de 0 à l'infini, l'absence d'effet thérapeutique correspond à un Win ratio à 1. Un Win ratio à 2 signifie qu'il y a deux fois plus de succès dans le groupe traitement que dans le groupe témoin. Lors d'un essai incluant 10 patients dans chaque bras, si 2 succès sont observés dans le bras traitement et si un succès est observé dans le bras contrôle, le Win ratio sera estimé à 2.25. Lors d'un autre essai incluant le même nombre de patients, si 8 succès sont observés dans le bras traitement et 7 succès sont observés dans le bras contrôle, le Win ratio sera alors estimé à 1.71. Dans les deux cas il y a un succès supplémentaire dans le groupe traitement par rapport au groupe contrôle. La propension au succès est estimée à 10% dans les deux scénarios, et semble décrire de façon plus interprétable l'effet du traitement. Les autres différences entre les deux méthodes sont la variété des critères pouvant être inclus dans

la méthode des comparaisons par paire, le développement des extensions permettant de prendre en compte les données censurées, et l'utilisation d'une méthode de permutation pour calculer l'intervalle de confiance de la propension au succès et la P-valeur qui lui est associée. Néanmoins il ne s'agit pas de différences intrinsèques et ces méthodes pourraient être intégrées à la construction du Win ratio.

III.3. Evaluation de l'ampleur d'effet thérapeutique

L'utilisation des comparaisons par paire permet d'estimer la propension au succès. Une propriété intéressante de ce paramètre est qu'il est un reflet du bénéfice clinique des patients, lorsque le ou les critères de jugement analysés sont cliniquement pertinents. Δ est la différence entre la probabilité pour une paire d'être favorable au groupe T et la probabilité pour une paire d'être défavorable au groupe T. Cette dernière équation étant écrite de façon simplifiée $\Delta = \mathbb{P}[X > Y] - \mathbb{P}[Y > X]$. La propension au succès est donc directement liée à la probabilité pour un patient pris au hasard dans le groupe T d'avoir un meilleur résultat thérapeutique qu'un patient pris au hasard dans le groupe C, moins la probabilité inverse. Cette définition permet d'expliciter la signification clinique de la propension au succès.

Lors de l'analyse d'une variable unique de type temps jusqu'à événement, la propension au succès est également appelée 'propension à une meilleure survie'. Dans l'article suivant, l'objectif était de montrer comment la propension à une meilleure survie pouvait être utilisée afin de rapporter explicitement l'ampleur d'un effet thérapeutique sur une variable de type temps jusqu'à événement. Cette méthode est particulièrement intéressante lorsque le rapport des taux instantanés de décès est non proportionnel. La représentation graphique de la propension à une meilleure survie en fonction du seuil de bénéfice minimal cliniquement significatif permet d'identifier un effet thérapeutique de grande ampleur. La méthode permet également de quantifier la proportion de patients qui ont un bénéfice à long terme du traitement (ou qui sont guéris par le traitement). La propension à une meilleure survie pour un seuil élevé de bénéfice minimal cliniquement significatif est une estimation de cette proportion moins la probabilité opposée. Cet article est en cours de revue dans le *Journal of the American Medical Association Oncology*. Les notations utilisées sont parfois différentes des notations utilisées ailleurs dans cette thèse afin de rendre l'article plus accessible à un public de lecteurs non habitués aux formules mathématiques.

The chance of a better survival in randomized clinical trials of new therapies

Running head: chance of a better survival

Julien Péron [1,2], Pascal Roy [1,2] , Brice Ozenne [1,2], Laurent Roche [1,2], Marc Buyse [3,4]

[1]Hospices Civils de Lyon, Service de Biostatistique, 165 Chemin du Grand Revoyet, Bt 4D. 69495 Pierre-Benite CEDEX, France.

[2] Université Lyon 1, CNRS UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Villeurbanne, France.

[3] International Drug Development Institute (IDDI), Raleigh, NC, USA.

[4] Interuniversity Institute for Biostatistics and statistical Biosinformatics (I-Biostat), Hasselt University, Hasselt, Belgium.

Corresponding author :

Dr Julien Péron

Service de Biostatistique, Centre Hospitalier Lyon-Sud

165 Chemin du Grand Revoyet

F-69310, Pierre-Bénite, France.

Tel: (+33) 6 16 25 89 91

Fax: (+33) 4 78 86 57 74

E-mail: julien.peron@chu-lyon.fr

Funding: Dr Julien Péron is the recipient of a grant from the Nuovo-Soldati Research Foundation

Abstract

Background – Time to events or "survival endpoints" are common in randomized trials in oncology, and are commonly analyzed under the assumption of proportional hazards (PH). However, the PH assumption is not always met, and this may lead to erroneous or misleading conclusions. We show here that a different measure of treatment effect, called "the chance of a better survival" may be useful to report the magnitude of the difference between groups, especially when the PH assumption is violated.

Methods – The chance of a better survival by at least m months, where m months is considered clinically relevant, is defined as the difference between the probability that a random patient in the experimental arm has a survival longer by at least m months than a random patient in the control group, and the probability of the opposite situation. The chance of a better survival is equal to zero if treatment does not differ from control. It ranges from -100% (if all patients in the control group fare better than patients in the treatment group) to +100% (in the opposite situation). The chance of a better survival can be estimated for different values of m using generalized pairwise comparisons. We simulated datasets for realistic trials under various scenarios of proportional or non-proportional survival hazards, and plotted the Kaplan-Meier survival curves as well as the chance of a better survival as a function of m.

Results – When proportional hazards hold, the chance of a better survival goes to zero as m increases. In contrast, when treatment effects are delayed, the chance of a longer survival benefit increases. In the best case scenario where some patients are cured by treatment, the chance of a long survival benefit tends to the cure rate.

Conclusion - The chance of a better survival is an intuitive measure of treatment benefit that has direct relevance to patients and health care providers. It can prove especially

useful when the assumption of proportional hazards is violated in the analysis of survival endpoints.

Keywords:

Statistics as topic; treatment outcome; survival analysis; randomized controlled trial

Introduction

Survival endpoints, such as the time from treatment initiation to cancer recurrence, progression, or death, are widely used in oncology trials. The treatment effect in survival analysis is usually quantified and reported using the hazard ratio (HR), a relative measure of difference between two survival curves. However, one of the assumptions for computing a meaningful HR is that hazard functions are proportional over time. When this proportional hazards (PH) assumption of is not met, the computed HR does not reliably reflect the treatment benefit, as the true HR is changing over time. [1, 2] In addition to focusing on the HR, researchers often attempt to determine the effect of treatment on some absolute scale. One frequently adopted solution is to compare the percentage of patients free of event or the mean survival at specific time-points. [3] However, this solution provides a limited measure of benefit since it ignores all the events that occur after this time point and makes the estimates sensitive to the selection of this time-point. [4, 5] Another solution is to examine the differences in the hazard functions or in survival rate differences between given time-points. [6, 7] The latter solution is less restrictive than the former, but it ignores the events that occur before and after the selected time-points and remains sensitive to the time-points selection. Hence, other statistical approaches are warranted to compare and describe the survival experience of patients in clinical trials, and some have been proposed by various investigators. [8–11]

The probabilistic index has been proposed as an alternative measure of treatment benefit. It is defined as the probability that a random patient in the experimental group has a better outcome than a random patient in the control group. [11] The relationship between the probabilistic index and the HR was investigated by Moser & McCann [12] and Buyse [13].One drawback of the probabilistic index is that it is equal to 0.5 when treatment does not differ from control. An intuitive extension of the probabilistic index is the chance of a better survival used here, which is equal to zero when treatment does not differ from control. In this

paper, we focus on the advantages of using the chance of a better survival through simulated datasets for randomized clinical trials under different scenarios for the treatment effect.

Methods

The chance of a better survival

The chance of a better survival is defined as the probability that a random patient in the experimental group survives by at least m months longer than a random patient receiving the control intervention minus the probability of the opposite situation. Note that m can be equal to zero, in which case any survival difference is considered clinically relevant. The chance of a better survival can be computed and its significance tested for any value of m using the method of generalized pairwise comparisons of prioritized outcomes, which is briefly summarized in Appendix A [13].

Simulation of randomized trial datasets

We simulated five scenarios of survival differences. In the first scenario, the hazards were proportional between the two treatment groups, with a hazard ratio of 0.75. In the others scenarios, the hazards were non-proportional. In scenario 2, the hazard ratio was increasing over time from 0.4 to 1 (early survival differences). This scenario might apply to the survival outcomes observed in most trials evaluating cytotoxic chemotherapy or many molecular targeted therapies for metastatic solid cancers [14]. In scenario 3, the hazard ratio was decreasing over time (late survival differences). This scenario might apply to the survival outcomes observed with modern immunotherapies in solid cancer [15]. In scenario 4, 10% of the patients were cured by the treatment and the other patients had no effect of the treatment. This scenario might apply to the survival outcomes observed with allografts in pediatric trials [16]. Finally, in scenario 5, half of the patient had a benefit from the treatment and the other half

had a detrimental effect from the treatment. This scenario might apply to the survival outcomes observed when molecular targeted therapy is compared with cytotoxic chemotherapy among all-comers when only 50% of the patients respond to the targeted therapy [17]. In each simulated dataset, the overall hazard ratio for survival was 0.75 when estimated through a Cox model. For each scenario, one dataset was generated including two treatment groups, each with 600 patients. The simulation parameters are reported in Appendix B. For each dataset, the chance of a better survival was calculated and plotted for values of m ranging from 0 to 40 months.

Generalized pairwise comparisons were performed with the package BuyseTest in the R software, available upon request.

Results

Findings from the simulated datasets

In the first illustrative dataset, the survival curves separated harmoniously. The median survival was 9.3 months and 10.6 months in the group C and in the group T respectively. The chance of a better survival was 13% (95%CI 6.5 to 19.4, P<0.001) when any survival difference was considered clinically relevant (m=0 month); this means that a random patient in the treatment group would have a 13% chance of a longer survival than a random patient in the control group. However the chance of a better survival decreased for long-term survival differences. When only survival differences larger than 20 months were considered relevant (m = 20), the chance of a better survival was very close to zero (0.5%, 95%CI -0.1 to 1.1, P=0.094). In the second illustrative dataset, corresponding to the scenario with early survival differences, the chance of a better survival was 23% (95%CI 16.8 to 28.9, P<0.001) when any survival benefit was considered clinically relevant (m=0 month), but it decreased quickly and was close to zero (0.6%, 95%CI 0.1 to 1.0, P=0.003) when only survival differences larger

than 20 months were considered relevant (m=20 months). In the illustrative datasets 3 and 4, corresponding to delayed survival differences and a 10% cure rate respectively, the chance of a better survival remained high for large values of m (e.g. m > 20). In scenario 3, the chance of a better survival decreased very slowly, while in the scenario 4 the chance of a better survival remained stable at 10% (95%CI -7.5 to 12.5, P<0.001 for m = 40 months), even when very large survival differences were considered. In the fifth illustrative dataset, the survival curves crossed near the 11th month of follow-up. When all the survival differences were considered relevant, the chance of a better survival was negative (-6.9%, 95%CI -14.0 to -0.5, P=0.047). However when only large survival differences were considered, the chance of a better survival became strongly positive (8.9%, 95%CI 6.7 to 11.1, P<0.001).

Discussion

In the current study, the computation of the chance of a better survival allowed an overall assessment of the treatment benefit, even when the assumption of proportional hazards was violated. To our knowledge, the concept of 'chance of a better survival' for a specified minimal relevant survival difference has not yet been used in clinical research, even though it allows a meaningful and simple estimation of the treatment effect. This proportion may be interpreted as the net difference between the probability that a random patient in the experimental arm has a survival longer by at least m months than a random patient in the control group, and the probability of the opposite situation, m being a specified minimal clinically relevant difference in survival. A rapid and informative assessment of treatment effect could be derived from the graph (Figure 1) that represents the chance of a better survival as a function of the minimal clinically relevant difference m. Indeed, the analysis of the extreme left part of the graph, which corresponds to a null value of m, informs about the respective probabilities that the new treatment prolongs or reduces the survival time for a

patient. This measure of the treatment effect derives from the probabilistic index $\mathbb{P}[X > Y]$ proposed by others [11, 12, 18] and has been shown to be a robust and meaningful non-parametric measure of the effect size for longitudinal data [11]. Then, for higher values of m, the chance of a better survival informs about the probabilities of inducing long-term survival increases or decreases.

The method should not be considered simply as a new test for survival analyses, but as a new method to assess graphically the distribution of survival benefits. However it is possible to test statistically the chance of a better survival with a randomization test. [13, 19, 20] When a single test is used at a specific critical value of m, no adjustment for test multiplicity is needed. When multiple tests are carried out, an adjustment for test multiplicity should be made as previously described. [21]

The method was informative to identify and quantify the proportion of patients with long-term benefit of the treatment (or patients cured by the treatment). The chance of a better survival for high minimal relevant difference m is a direct estimation of this probability minus the opposite probability.

Some researchers have pointed the limits of one-time survival differences or of median survival to estimate the absolute benefit of a treatment in randomized trials. This is particularly apparent when hazards are not proportional. Tan and Murphy proposed to use the 'average duration of life gained' to summarize difference in treatment effect. This statistic is equal to the area between the survival curves. Of note, this statistic is also equal to the area under the curve of the chance of a better survival by at least m months over all values of m. [9, 22–24]

The method was also informative when the Kaplan-Meier curves crossed. The fifth dataset was simulated using the parameters reported by the Iressa Pan-Asia Study (IPASS) investigators. [17] In the IPASS trial, previously untreated patients with advanced lung

adenocarcinoma were randomized to receive either gefitinib (a tyrosine-kinase inhibitor of the epidermal growth factor receptor, EGFR) or a chemotherapy combination of carboplatin and paclitaxel. Activating EGFR mutations are now known to be predictive of benefit from gefitinib. [25] Although this predictive role was unknown when the trial was initiated, the frequency of such mutations was high (approximately half) because the study patients were Asians and a large proportion were non/light smokers, two features knowingly associated with activating mutations. [26] In the corresponding simulated dataset, Kaplan-Meier curves crossed and the assumption of proportional hazards was then obviously violated. It followed that the estimation of the HR (0.75) was biased and potentially misleading. The analysis of the chance of a better survival identified a clear benefit in favor of the treatment (gefitinib) when only long-term survival differences were considered. However, the control group did slightly better when any survival difference was considered relevant ($m = 0$ month). The analysis of the chance of a better survival as a function of m, applied to the simulated IPASS data, may be summarized in two sentences: 1) with gefitinib, the probability to shorten progression-free survival (PFS) was slightly higher than the probability of prolonging it, but the disadvantages were often of short magnitude; and 2) the probability of inducing long-term PFS benefit with gefitinib was much more important than that of inducing long-term PFS detriment. This method may help physicians inform an individual patient about the effects to be expected from a new treatment.

We have shown how the chance of a better survival for specified minimal clinically relevant differences in survival may be helpful when the hazards are not proportional. Such non-proportional hazards may result from two main mechanisms: (1) interactions between the treatment effect and patient or disease features, and (2) variation of the treatment effect over time (e.g., in trials comparing transplantation with non-transplantation strategies). In such cases, Kaplan-Meier curves (hereafter curves) often display unusual shapes (e.g., curves do

not separate uniformly), and standard comparison techniques may lead to erroneous conclusions, [27, 28] while an analysis of the chance of a better survival provide a heuristic interpretation for the treatment difference in time-to-event outcomes. In the context of a single time-to-event endpoint, generalized pairwise comparisons prioritized on several successive thresholds should be considered to estimate the chance of a better survival, even when the proportional hazards assumption is violated.

Conflict of Interest Statement

The Authors declare no conflict of interest.

Acknowledgement

None

Funding

Dr Julien Péron is the recipient of a grant from the Nuovo-Soldati research foundation. The funding source had no role in the study design, in the collection, analysis, or interpretation of data.

References

1. Hattori S, Henmi M: Estimation of treatment effects based on possibly misspecified Cox regression. Lifetime Data Anal 18:408–33, 2012

2. Schoenfeld DA: Sample-size formula for the proportional-hazards regression model. Biometrics 39:499–503, 1983

3. Royston P, Parmar MKB: Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 13:152, 2013

4. Klein JP, Logan B, Harhoff M, et al: Analyzing survival curves at a fixed point in time. Stat Med 26:4505–4519, 2007

5. Royston P, Parmar MKB: An approach to trial design and analysis in the era of non-sal hazards of the treatment effect. Trials 15:314, 2014

6. Zhao L, Tian L, Uno H, et al: Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. Clin Trials 9:570–7, 2012

7. Logan BR, Klein JP, Zhang MJ: Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. Biometrics 64:733–740, 2008

8. Royston P, Parmar MKB, Altman DG: Visualizing length of survival in time-to-event studies: a complement to Kaplan-Meier plots. J Natl Cancer Inst 100:92–7, 2008

9. Seruga B, Pond GR, Hertz PC, et al: Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. Ann Oncol 23:2977–82, 2012

10. Coory M, Lamb KE, Sorich M: Risk-difference curves can be used to communicate time-dependent effects of adjuvant therapies for early stage cancer. J Clin Epidemiol 67:966–72, 2014

11. Acion L, Peterson JJ, Temple S, et al: Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. Stat Med 25:591–602, 2006

12. Moser BK, McCann MH: Reformulating the hazard ratio to enhance communication with clinical investigators. Clin Trials 5:248–252, 2008

13. Buyse M: Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med 29:3245–3257, 2010

14. Johnston S, Pippen J, Pivot X, et al: Lapatinib combined with letrozole versus letrozole and placebo as first-line therapy for postmenopausal hormone receptor-positive metastatic breast cancer. J Clin Oncol 27:5538–46, 2009

15. Maio M, Grob J-J, Aamdal S, et al: Five-Year Survival Rates for Treatment-Naive Patients With Advanced Melanoma Who Received Ipilimumab Plus Dacarbazine in a Phase III Trial. J Clin Oncol , 2015

16. Oliansky DM, Rizzo JD, Aplan PD, et al: The role of cytotoxic therapy with hematopoietic stem cell transplantation in the therapy of acute myeloid leukemia in children: an evidence-based review. Biol Blood Marrow Transplant 13:1–25, 2007

17. Mok TS, Wu YL, Thongprasert S, et al: Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med 361:947–957, 2009

18. Grissom RJ: Probability of the superior outcome of one treatment over another.

19. Edgington P. ESO: Randomization Tests. Chapman Hall/CRC New York , 2007

20. Good P: Resampling Methods: A Practical Guide to Data Analysis (3rd edn). Birkhauser New York , 2006

21. Buyse M: Reformulating the hazard ratio to enhance communication with clinical investigators. Clin Trials 5:641–642, 2008

22. Datta S: Estimating the mean life time using right censored data. Stat Methodol 2:65–69, 2005

23. Ajani JA: The area between the curves gets no respect: is it because of the median madness? J Clin Oncol 25:5531, 2007

24. Tan LB, Murphy R: Shifts in mortality curves: saving or extending lives? . Lancet 354:1378–81, 1999

25. Lynch TJ, Bell DW, Sordella R, et al: Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350:2129–39, 2004

26. Shigematsu H, Lin L, Takahashi T, et al: Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. J Natl Cancer Inst 97:339–46, 2005

27. Uno H, Claggett B, Tian L, et al: Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol 32:2380–5, 2014

28. Bellera C, MacGrogan G, Debled M, et al: Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC Med Res Methodol 16:10–20, 2010

Figure 1.

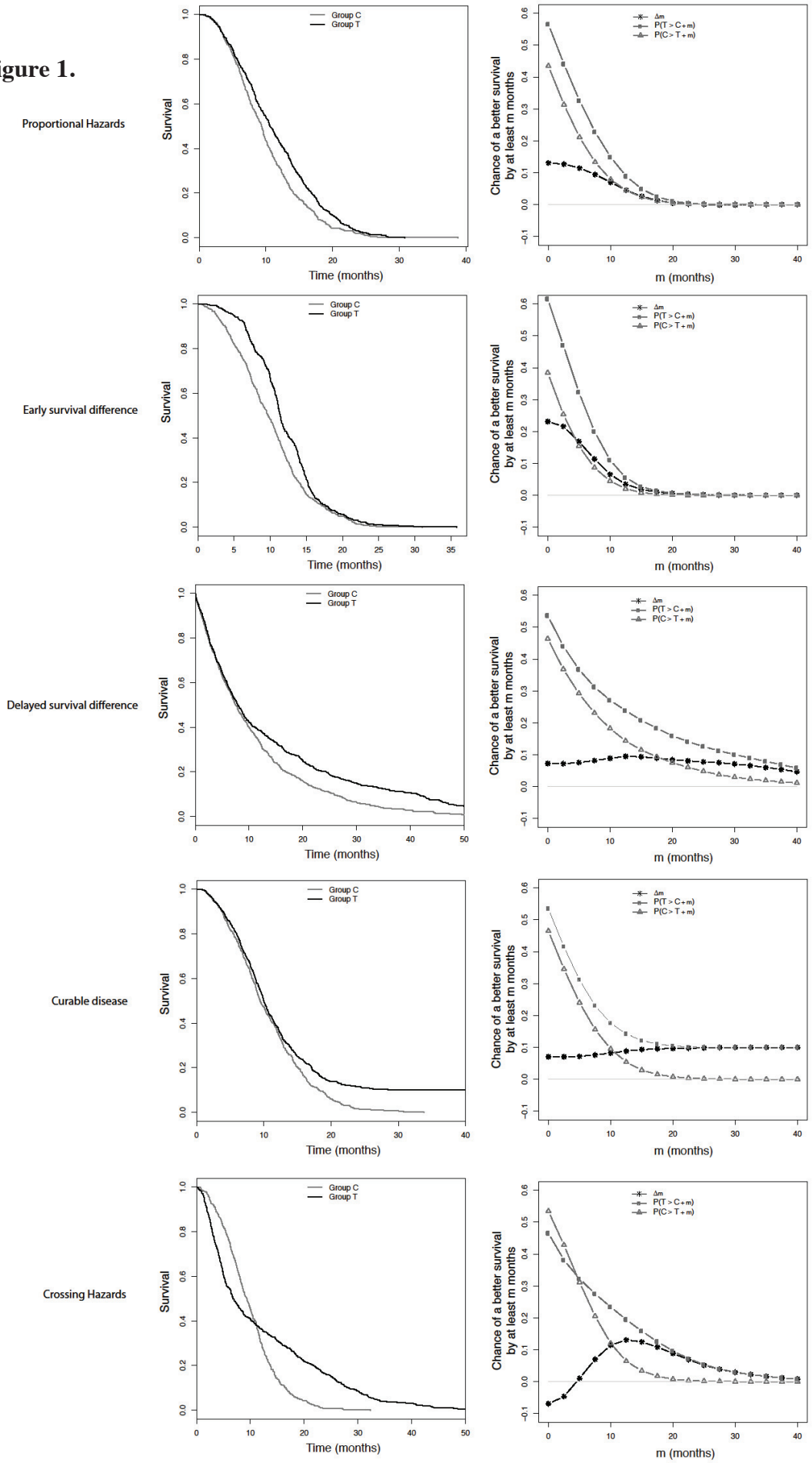

Proportional Hazards
Early survival difference
Delayed survival difference
Curable disease
Crossing Hazards
Survival
Time (months)
Group C
Group T
Chance of a better survival by at least m months
m (months)
Δm
P(T > C+m)
P(C > T+m)

Figure 1 title: Graphical representation of survival benefits in a scenario of proportional hazards and four scenarios of non-proportional hazards. (A) Kaplan-Meier estimates of survival functions over time, (B) chance of a better survival by at least m months.

Footnotes: Δ_{m} = chance of a better outcome by at least m months ; $\mathbb{P}[T > C + m]$ = the probability that a random patient in the experimental group survives by at least m months longer than a random patient receiving the control intervention ; $\mathbb{P}[C > T + m]$ = the probability that a random patient in the control group survives by at least m months longer than a random patient in the experimental group.

Appendix A – Generalized Pairwise Comparisons

A full description of generalized pairwise comparisons has been previously published and an extension has been proposed for survival endpoints. We restrict our discussion to the analysis of data from randomized trials comparing an experimental group to a control. Pairwise comparisons are carried out on all possible pairs of patients, one from the experimental group (group T) and the other from the control group (group C). Let x_i and y_j be the times to the event of a patient i from group T (i = 1,…,$n1$) and a patient j from group C (j = 1,…,$n2$). A pair is classified as 'favorable' if the outcome of the patient i is higher than that of the patient j by at least m months, i.e., $x_i - y_j > m$; 'unfavorable' in the opposite situation, i.e., $x_i - y_j < -m$; and 'neutral' in all other cases, i.e. $-m < x_i - y_j < m$. The probabilities for a random pair to be favorable and unfavorable are then denoted by $\mathbb{P}[x_i > y_j + m]$ and $\mathbb{P}[y_j > x_i + m]$, respectively. For each pair, a pairwise score $p_{ij} = \mathbb{P}[x_i > y_j + m] - \mathbb{P}[y_j > x_i + m]$ is calculated. When there is no censored observation, pairs can be decidedly classified as favorable, neutral or unfavorable, and p_{ij} takes the value 1, 0, or -1, respectively. With censoring, the pairwise score can be estimated from the Kaplan-Meier survival estimates.

$$\Delta = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}}{n1.n2}$$

is the chance of a better survival. When $m = 0$, the chance of a better survival is a generalization of the Wilcoxon test statistic. A confidence interval for the chance of a better survival, and a test of statistical significance can be computed using a randomization test.

Appendix B - Details on the simulation parameters

Scenario 1 : proportional hazards

Survival times in the control group followed a Weibull distribution with a shape parameter equal to 2 ($k = 2$) and a scale parameter equal to 11.5 ($\lambda_C = 11.5$). Survival times in the experimental group followed a Weibull distribution with a shape parameter equal to 2 ($k = 2$) and a scale parameter equal to 13.4 ($\lambda_T = 13.4$).

Scenario 2 : early survival differences

Survival times in the control group followed a Weibull distribution with a shape parameter equal to 2 ($k = 2$) and a scale parameter equal to 11.5 ($\lambda_C = 11.5$). The hazard ratio was set at 0.4 between 0 and 4 months. It increased to 0.55 between 4 and 8 months, to 0.7 between 8 and 12 months and to 0.85 between 12 and 16 months and to 1 (no effect) thereafter.

Scenario 3 : delayed survival differences

Survival times in the control group followed an exponential distribution with a rate parameter equal to 0.0866 ($\lambda_C = 0.0866$). The hazard ratio was set at 1 (no effect) between 0 and 4 months. It decreased to 0.875 between 4 and 8 months, to 0.75 between 8 and 12 months, to 0.625 between 12 and 16 months and to 0.5 thereafter.

Scenario 4 : curable disease

Survival times in the control and in the experimental groups followed a Weibull distribution with a shape parameter equal to 2 ($k = 2$) and a scale parameter equal to 11.5 ($\lambda_C = 11.5$). Ten percent of the patients selected at random in the experimental group were assumed cured (very long survival times).

Scenario 5 : crossing hazards

Survival times in the control group followed a Weibull distribution with a shape parameter equal to 2 ($k = 2$) and a scale parameter at 11.5 ($\lambda_C = 11.5$). The survival times in the experimental groups were generated from two distributions. The survival times of half of the patients were generated using a proportional hazard ratio equal to 2.5 (detrimental effect of the experimental treatment), and the survival times of the other half of the patients were generated using a proportional hazard ratio at 0.5 (benefit of the experimental treatment).

Cet article permet d'illustrer comment la propension au succès reflète de façon cliniquement pertinente un effet thérapeutique. Seul un critère de jugement de type temps jusqu'à événement était inclus dans la procédure des comparaisons par paire généralisées. L'estimation de la propension au succès, appelée ici propension à une meilleure survie, a été réalisée en utilisant différents seuils de bénéfice minimal cliniquement significatif. Une analyse réalisée avec un seuil de bénéfice minimal cliniquement significatif nul correspondait à une analyse de l'ensemble des différences de survie induites, quelles que soient leurs amplitudes. Le choix d'un seuil élevé correspondait à une analyse restreinte aux différences de survie induites de grandes amplitudes. Un avantage de cette méthode par rapport à une analyse fine des courbes de survie estimées par la méthode de Kaplan et Meier est la possibilité de quantifier et de tester statistiquement un effet thérapeutique à long terme. En effet, l'analyse des courbes de survie permet également de suspecter un bénéfice à long terme lorsque l'on observe une séparation des courbes de survie tardivement pendant le suivi. Néanmoins la pertinence statistique d'un tel écart peut être mise en défaut lorsque le nombre de patients encore en cours de suivi est faible (du fait des censures ou des événements).

La propension à une meilleure survie peut être utilisée même lorsque les courbes de survie ont une évolution non proportionnelle. Dans cette situation, les paramètres classiquement utilisés pour résumer la différence entre les fonctions de survie (le rapport des taux instantanés de décès, la différence des médianes de survie, ou la différence de probabilité de survie à un temps donné) peuvent être mis en défaut. L'utilisation de la propension à une meilleure survie est alors une solution qui permet de refléter directement le bénéfice attendu d'un traitement.

Chapitre **IV**

IV. Analyse de la balance bénéfice-risque des traitements en utilisant les comparaisons par paire généralisées

IV.1. Les méthodes d'évaluation de la balance bénéfice-risque des traitements

Les autorités d'enregistrement des produits de santé, européennes comme américaines, recommandent de réaliser une analyse de la balance bénéfice-risque des nouveaux traitements [3], [5]. Néanmoins aucune procédure d'évaluation qualitative ou quantitative de cette balance bénéfice-risque n'est actuellement recommandée. Une enquête réalisée par l'European Medicines Agency (EMA) auprès de professionnels d'autorités nationales compétentes d'évaluation des médicaments (France, Pays-Bas, Espagne, Suède et Royaume-Unis) a permis de mettre en évidence une hétérogénéité dans la définition d'un bénéfice et d'un risque thérapeutique [16]. Cette enquête a mis en évidence une perception très hétérogène de l'effort et du temps nécessaire pour évaluer le bénéfice d'un traitement par rapport au risque d'un traitement. L'évaluation de la balance bénéfice-risque était perçue comme difficile, faisant partie du domaine du jugement d'experts.

Il semble acceptable que la décision finale d'utiliser ou non un traitement en fonction de la balance bénéfice-risque estimée appartienne à un jugement humain, et non à la conclusion d'une approche purement quantitative. Cette évaluation qualitative des bénéfices et des risques d'un traitement peut être réalisée par les patients eux-mêmes, par les médecins, ou par les autorités compétentes d'évaluation des médicaments. En effet une évaluation qualitative reste indispensable afin d'adapter la décision aux caractéristiques de l'individu ou du groupe d'individus pour lequel la décision est prise. Néanmoins dans la suite de ce chapitre, seules les méthodes d'analyse quantitative de la balance bénéfice-risque seront développées [17]. Ces méthodes d'analyse quantitative ont pour objectif d'assister et d'informer le jugement humain. Le sujet sera restreint aux méthodes principales permettant d'évaluer la balance bénéfice-risque d'un traitement comparé à un contrôle dans un essai randomisé.

Les méthodes classiques d'analyse quantitative de la balance bénéfice-risque des traitements évalués au sein d'essais randomisés ont pour point commun de ramener sur une métrique commune les variables évaluant les bénéfices et les variables évaluant les risques. Certaines métriques proposées sont relativement simples. Par exemple Seymour et al ont proposé d'utiliser une métrique appelée 'Overall Treatment Utility' (OTU) qui permet de combiner des critères de jugement objectifs et subjectifs, relatifs aux bénéfices et aux événements indésirables du traitement [18]. Dans leur étude, le bénéfice clinique et la tolérance/acceptabilité du traitement étaient mesurés 12 semaines après le début du traitement. Le bénéfice clinique était 'positif' en l'absence de progression tumorale clinique ou radiologique et 'négatif' sinon. La tolérance/acceptabilité était 'positive' en l'absence d'événement indésirable grave déclaré par l'investigateur, ou par le patient dans un questionnaire auto-administré, et 'négative' sinon. L'OTU était 'favorable' en présence d'un bénéfice clinique 'positif' et d'une tolérance/acceptabilité 'positive'. L'OTU était 'intermédiaire' en présence d'un composant 'négatif', et était 'mauvais' en présence des deux composants 'négatifs'. L'OTU permettait donc de définir une variable à trois catégories reflétant la balance bénéfice-risque des traitements. L'OTU a l'avantage d'être simple à interpréter, mais ne prend en compte qu'une petite partie des bénéfices thérapeutiques attendus. Elle ne permet pas d'étudier le poids respectif des différents critères de jugement, puisque ces composants (bénéfice clinique et la tolérance/acceptabilité) sont déjà des critères composites. De plus l'utilisation d'une variable à trois classes limite certainement les performances de test cherchant à comparer deux groupes de patient en termes d'OTU.

Le nombre de sujets à traiter (NNT) et le nombre de sujets pour observer un événement indésirable (NNH pour '*number needed to harm*') sont des mesures simples pour évaluer la balance bénéfice-risque d'un traitement dans le cadre d'un essai comparatif [19]. Le NNT est le nombre de sujets à traiter pour observer un succès thérapeutique. Il dépend des conditions de l'essai et notamment du bras contrôle. Le NNH est le nombre de sujets à traiter pour observer un événement indésirable grave. Le rapport $\frac{\mathrm{NNT}}{\mathrm{NNH}}$ est une mesure simple du nombre de succès thérapeutiques obtenus par événement indésirable grave. Cette mesure est critiquable car elle donne le même poids à un succès thérapeutique et à un événement indésirable grave quelles que soient les situations cliniques et les définitions d'un succès thérapeutique et d'un événement indésirable grave. De plus elle utilise des critères composites résumant le succès thérapeutique et la toxicité qui ne permettent probablement pas d'apprécier la réalité des effets thérapeutiques des traitements.

Les analyses multi-critères définissent une classe générale de modèles permettant de guider les décisions à partir de plusieurs critères de jugement. L'objectif général des analyses multi-critères est d'ordonner les décisions possibles de la plus à la moins préférable [20]. Les analyses multi-critères comportent 8 étapes (Tableau IV.1). Il s'agit d'un cadre d'analyse très général, dans lequel la balance entre le bénéfice et le risque des traitements est le plus souvent réalisée au niveau populationnel et non pas individuel [21].

Tableau IV-1. Description des 8 étapes d'une analyse multi-critères, et adaptation de la démarche théorique à l'analyse de la balance bénéfice-risque d'un traitement évalué dans un essai randomisé

	Définition théorique de l'étape	Application à l'analyse de la balance bénéfice-risque dans essai comparatif
Etape 1	Définir le contexte de décision	Définir un bénéfice thérapeutique et un risque thérapeutique dans la situation thérapeutique étudiée
Etape 2	Définir les options disponibles	Définir les traitements dans les groupes thérapeutiques comparés
Etape 3	Identifier les objectifs, les critères de jugement	Définir les critères de jugement, avec utilisation d'éventuels critères de jugement composites
Etape 4	'Scoring' : Evaluer l'effet de chaque option sur chaque critère de jugement	Analyser séparément l'ensemble des critères de jugement cliniquement pertinent
Etape 5	'Weighting' : Attribuer un poids à chaque critère de jugement afin de pondérer son impact en fonction de son importance	Attribuer un poids à chaque critère de jugement
Etape 6	Combiner les scores et les poids pour obtenir une estimation de l'effet global	-
Etape 7	Examiner les résultats	-
Etape 8	Conduire des analyses de sensibilité	-

Plusieurs méthodes appartenant au cadre des analyses multi-critères ont été proposées pour analyser la balance bénéfice-risque des traitements à un niveau individuel. Gelber et al ont proposé d'utiliser comme critère de jugement le temps de survie sans symptôme de la

maladie et sans toxicité [22]. Ce critère de jugement a été ensuite généralisé sous le nom de Q-TWiST (*quality-adjusted Time Without Symptoms of disease progression or Toxicity of treatment*) [23]. Le temps de survie de chaque patient inclus dans un essai est partitionné en plusieurs catégories représentant un état de santé (temps passé sans symptôme ni toxicité ; temps passé avec des symptômes de la maladie ; et temps passé avec toxicité). Un poids correspondant à l'utilité de chaque état de santé est déterminé, et est utilisé pour pondérer le temps moyen passé dans chacun des états. Cette méthode permet de prendre en considération à la fois le bénéfice en survie, en survie sans symptôme, et la toxicité des traitements. De plus une représentation graphique élégante a été proposée afin de décrire les temps passés dans chaque état. Cette méthode présente néanmoins certaines limites. Les critères de jugement utilisés pour mesurer l'efficacité sont la survie et la survie sans symptôme. Ces critères de jugement sont adaptés à la plupart des situations en cancérologie, mais peuvent être moins pertinents dans d'autres situations. Le choix des utilités attribuées à chaque état de santé est difficile. La quantification de la valeur relative d'une journée passée avec des symptômes ou avec une toxicité par rapport à une journée en bonne santé n'est pas réalisée en pratique courante par les cliniciens, par les patients, ni les autorités d'évaluation des médicaments. Cette différence conceptuelle entre le Q-TWiST et le fonctionnement du jugement humain explique probablement sa faible utilisation en pratique courante [24]. De plus, comme l'ensemble des méthodes d'analyse quantitative de la balance bénéfice-risque, le Q-TWiST est confronté à la faible standardisation des variables mesurant la sévérité et la durée de la toxicité des traitements (qui peuvent être déclarées par les médecins, par les patients, ou être des données biologiques) [25].

D'autres auteurs ont proposé d'estimer séparément pour chaque patient un paramètre résumant l'efficacité du traitement e_i, et un paramètre résumant la toxicité du traitement r_i [26]. Cette méthode est nommée *Global Risk-Benefit method*. Une mesure du bénéfice ajusté au risque est réalisée par $e_i^* = e_i - f\, r_i$; la constante f étant le reflet du poids des toxicités par rapport à l'efficacité. Cette méthode est peu adaptée lorsque l'efficacité du traitement est mesurée par un critère de type temps jusqu'à événement. Elle ne permet pas de prendre en compte de façon adaptée de multiples événements indésirables concurrents. De plus, la constante f a une importance majeure sur le résultat de l'analyse, et sa détermination est au moins aussi complexe que les utilités utilisées dans une analyse de Q-TWiST.

Des méthodes d'analyse multi-critère ont également été proposées pour balancer les bénéfices et les risques d'un traitement au niveau strictement populationnel. Un modèle de

balance bénéfice-risque a été proposé par Felli et al sous le nom de Benefit-Risk Assessment Model (BRAM) [21]. Ce modèle a été développé en prenant la perspective d'une industrie pharmaceutique. L'ensemble des bénéfices et des risques évalués ont été listés, et organisés de façon hiérarchique. Les bénéfices ont été divisés en trois sous-sections : efficacité, effets sur la vie des patients, et acceptabilité. Les risques ont également été divisés en trois sous-sections : toxicité, acceptabilité, et mésusage. Une utilité est ensuite attribuée à chaque sous-section, puis à chaque critère composant les sous-sections. L'ensemble des critères ont été recodés sur une métrique continue entre 0 et 1 afin de compenser les différences entre les unités de mesures. Cette méthode a permis de comparer au niveau populationnel plusieurs traitements en termes de balance bénéfice-risque. L'avantage de cette méthode est sa grande souplesse. Elle permet d'intégrer de multiples critères de jugement, et de comparer de multiples traitements entre eux. Néanmoins elle présente plusieurs limites. L'analyse étant faite au niveau populationnel, elle ne prend pas en compte les éventuelles corrélations entre différents critères de jugement. Le choix des utilités présente les mêmes difficultés que la méthode Q-TWiST. La transformation des critères de jugement sur une échelle entre 0 et 1 n'est pas intuitive, par exemple lors de l'utilisation de critères de type temps jusqu'à événement.

En conclusion de ce sous-chapitre, plusieurs méthodes d'évaluation quantitative de la balance bénéfice-risque des traitements ont été proposées. Certaines, comme la méthode BRAM, permettent de classer les traitements en ordre de préférence, et sont ainsi adaptées à informer la décision des industries pharmaceutiques ou des autorités d'évaluation des médicaments. La méthode Q-TWiST permet de comparer les traitements entre eux en identifiant dans l'analyse le temps passé avec toxicité, le temps passé avec symptôme et le temps passé en bonne santé. Néanmoins le paramètre servant à comparer les groupes thérapeutiques est d'interprétation difficile et ne permet donc pas d'informer pleinement les cliniciens et les patients sur la balance bénéfice-risque attendue.

Nous proposons d'utiliser la propension au succès pour réaliser l'évaluation de la balance bénéfice-risque d'un traitement évalué dans un essai contrôlé randomisé. La méthode des comparaisons par paire s'intègre dans le cadre général des analyses multicritères. La sélection des critères de jugement pertinents est identique à celle des méthodes décrites plus haut. Par contre, la façon de pondérer l'importance relative de chaque critère est très différente. La méthode requiert de définir une stratégie de classement permettant de définir un succès thérapeutique, un échec thérapeutique et une équivalence thérapeutique lorsque l'on

compare les critères de jugement de deux patients au sein d'une paire. Afin de réaliser ce classement, il est possible de définir autant de priorités que nécessaire. Chaque critère de jugement priorisé peut être associé à un seuil de bénéfice minimal cliniquement significatif. Cette façon de pondérer l'importance relative de chacun des critères au sein des paires de patients semble dans une certaine mesure plus naturelle que la pondération des critères de jugement en fonction de leurs utilités au niveau populationnel ou même individuel. L'utilisation des comparaisons par paire pour analyser la balance bénéfice-risque des traitements va être illustrée dans le sous-chapitre IV.2. Le premier traitement évalué est l'erlotinib en association avec la gemcitabine pour traiter les patients atteints de cancers du pancréas avancés ou métastatiques. Cette étude a fait l'objet d'une publication dans le *British Journal of Cancer* en 2015 [27]. Le second traitement évalué est l'association de 5-fluorouracile, d'oxaliplatine, de leucovorine et d'irinotecan (FOLFIRINOX) évalué en comparaison à la gemcitabine pour traiter les patients atteints de cancers du pancréas métastatiques. Cette étude a fait l'objet d'un article qui est en cours de soumission au *Journal of the National Cancer Institute*.

IV.2. Evaluation de la balance bénéfice-risque dans le cancer du pancréas métastatique

IV.3.a. Evaluation de la balance bénéfice-risque de l'erlotinib

FULL PAPER

BJC

British Journal of Cancer (2015), 1–6 | doi: 10.1038/bjc.2015.55

Keywords: statistics as topic; treatment outcome; pancreatic cancer; erlotinib; randomised controlled trial

Assessing the benefit–risk of new treatments using generalised pairwise comparisons: the case of erlotinib in pancreatic cancer

J Péron*[,1,2], P Roy[1,2], K Ding[3], W R Parulekar[3], L Roche[1,2] and M Buyse[4]

[1]*Service de biostatistiques, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, Pierre-Bénite F-69310, France;* [2]*CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Université Lyon 1, Villeurbanne, France;* [3]*NCIC Clinical Trials Group, Queen's University, Kingston, Ontario, Canada and* [4]*International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium*

Background: Efficacy and safety are the two considerations when characterising the effects of a new therapy. We sought to apply an innovative method of assessing the benefit–risk balance using data from a completed randomised controlled trial that compared erlotinib *vs* placebo added to gemcitabine in patients with advanced pancreatic cancer (NCIC CTG PA.3).

Methods: We applied generalised pairwise comparisons with several prioritised outcome measures (e.g., one or more benefit outcomes and one or more risk outcomes). Here, the first priority outcome was overall survival (OS) time. Differences in OS that exceeded 2 months were considered clinically meaningful. The second priority outcome was toxicity. The overall treatment effect was quantified using the proportion in favour of erlotinib, which can be interpreted as the net proportion of patients who have a better overall outcome with erlotinib as compared with placebo. Sensitivity analyses were performed.

Results: In this trial 569 patients were randomly assigned in a 1 : 1 ratio to receive gemcitabine plus either erlotinib or a matched placebo. Overall, the method indicated no statistically significant overall treatment effect in favour of erlotinib; if anything, the point estimate of the net proportion leaned in favour of the placebo group (overall proportion in favour of erlotinib = −3.6%, 95% CI, −14.2– 7.1%; $P = 0.51$). The net proportion was never in favour of the erlotinib group throughout all sensitivity analyses.

Conclusions: Generalised pairwise comparisons make it possible to assess the benefit–risk balance of new treatments using a single statistical test for any number of prioritised outcomes. The benefit–risk assessment was not in favour of adding erlotinib to gemcitabine for the treatment of patients with advanced pancreatic cancer.

When characterising a treatment effect, efficacy and safety are the primary considerations. In the reporting of clinical trials, efficacy and safety outcomes are usually reported independently, no formal overall evaluation of the treatment effect is performed (Péron *et al*, 2012, 2013). Both US Food and Drug Administration and the European Medicines Agency have stressed the importance of a more structured and transparent approach to benefit–risk assessment (BRA) in the evaluation of new therapies (Committee for Medicinal Products for Human Use (CHMP), 2008; Food and Drug Administration, 2011).

Patients with advanced pancreatic cancer have a poor prognosis and the standard first-line regimen is cytotoxic chemotherapy (gemcitabine in monotherapy or in combination with nab-paclitaxel or a combination of 5-fluorouracil, oxaliplatin and irinotecan for patients with good performance status) (Burris *et al*, 1997; Conroy *et al*, 2011). The NCIC Clinical Trials Group Study PA.3 (NCIC CTG PA.3) phase III trial investigated the addition of erlotinib to gemcitabine in patients with advanced pancreatic cancer (Moore *et al*, 2007). Both survival and progression-free survival were significantly better for the combination treatment but the overall benefits were of modest magnitude (HR for overall survival (OS) = 0.82, 95% CI, 0.69–0.99; $P = 0.038$). The excess toxicity, the unfavourable cost-effectiveness observed with the combination with erlotinib, (Miksad *et al*, 2007; Tam *et al*, 2013)

*Correspondence: Dr J Péron; E-mail: julien.peron@chu-lyon.fr

Received 7 November 2014; revised 31 December 2014; accepted 12 January 2015

and the absence of a biomarker predictive of erlotinib efficacy, (da Cunha Santos *et al*, 2010; Boeck *et al*, 2013) led to a poor uptake of this regimen in the oncology community (Verslype *et al*, 2007; Saif, 2008; Choi *et al*, 2012).

No systematic assessment of the benefit–risk balance of erlotinib combination has been performed in the setting of advanced pancreatic cancer. We report here such an assessment based on the method of generalised pairwise comparisons (Buyse, 2010). This method extends the non-parametric Mann–Whitney–Wilcoxon test for a single outcome in the absence of censored data. It allows one to calculate and test the overall benefit of a new treatment based on any number of prioritised outcomes, some reflecting benefit from the intervention (e.g., survival or time to progression) and the others reflecting harm (e.g., treatment-related toxicities and side effects).

MATERIALS AND METHODS

Overview. The NCIC CTG PA.3 trial was an international study that randomised patients with advanced pancreatic cancer to receive gemcitabine in combination with either erlotinib or placebo as first-line treatment. The primary outcome was OS. Progression-free survival (PFS) and toxicity were secondary outcomes.

In this trial, 569 patients were stratified by center, performance status (Eastern Cooperative Oncology Group 0 or 1 *vs* 2) and extent of disease (locally advanced *vs* metastatic), and randomly assigned in a 1 : 1 ratio to receive gemcitabine plus either erlotinib or a matched placebo. Progression was evaluated using Response Evaluation Criteria in Solid Tumors (V1.0) every 8 weeks. Toxicity was assessed at every visit using the National Cancer Institute Common Toxicity Criteria version 2.0.

Generalised pairwise comparisons. We applied generalised pairwise comparisons extended to several outcome measures (a benefit outcome, and a risk outcome). A full description of generalised pairwise comparisons has been previously published (Buyse, 2010). In brief, pairwise comparisons require consideration of all possible pairs of patients, one taken from the erlotinib arm and the other taken from the placebo arm. Pairwise comparisons are easily stratified for the stratification factors used in the randomisation process. The outcomes of these two patients are compared according to the first priority outcome. The pair is said to be 'favourable' if the outcome of the patient in the erlotinib arm is better than the outcome of the patient in the placebo arm, 'unfavourable' if the outcome of the patient in the erlotinib arm is worse than the outcome of the patient in the placebo arm and 'uninformative' if it cannot be determined which of the two patients has a better outcome (e.g., because of censoring, because the two observations are equal or because the difference of outcomes does not reach a pre-specified threshold value). Such a pairwise comparison is carried out for all pairs of patients, and the difference between the proportion of favourable pairs and the proportion of unfavourable pairs is calculated for the first priority outcome. This difference is called the proportion in favour of treatment for the first priority outcome (Buyse, 2008; Moser and McCann, 2008).

For pairwise comparisons that are uninformative for the first priority outcome, the second priority outcome is used in turn to classify the pair as favourable, unfavourable or uninformative (Table 1). After consideration of the second priority outcome, the 'overall proportion in favour of treatment' is calculated to provide an overall assessment of both the benefit and the risks of the treatment, suitably prioritised.

Standard analysis of efficacy and toxicity. A log-rank test adjusted for stratification factors at baseline was used to compare treatment groups in terms of survival. Worst grade adverse events (AE) that were at least possibly related to the study treatment ('treatment-related AEs') were reported by treatment group. All analyses were performed on all randomly assigned patients as per the intent-to-treat principle.

Table 1. Generalised pairwise comparisons for two prioritised outcomes

First priority outcome	Second priority outcome	Pair is
Favourable	Ignored	Favourable
Unfavourable	Ignored	Unfavourable
Uninformative	Favourable	Favourable
Uninformative	Unfavourable	Unfavourable
Uninformative	Uninformative	Uninformative

Main analysis of the benefit–risk balance. The first priority outcome used in the main analysis was OS. Only pairs of patients with differences in OS exceeding 2 months were considered informative, because smaller differences in OS were not considered clinically meaningful. The second priority outcome was treatment-related AEs, with patients experiencing the lower grade-related AE considered to have had a more favourable outcome. Treatment arms were compared using the overall proportion in favour of the erlotinib group (Δ[erlotinib]). A randomisation test stratified by performance status and extent of disease at diagnosis was performed to test the null hypothesis (H_0: Δ[erlotinib] = 0). The contribution of each outcome to Δ[erlotinib] was calculated.

Sensitivity analyses. The impact of the choice of outcomes, thresholds and priority on the results was assessed in sensitivity analyses. First, the main analysis was repeated with various thresholds for the minimal OS difference considered as clinically meaningful, ranging from 0 (any difference in OS considered clinically meaningful) to 6 months. Second, the toxicity outcome was defined as a binary variable where only grade $\geqslant 3$ AEs were considered. Third, a subgroup analysis was performed among patients treated with 100 mg per day of erlotinib, the actual recommended dose. Finally, a wide range of scenarios integrating OS, PFS and AE grades with several successive thresholds were built to provide a comprehensive assessment of the treatment effects. For each scenario, the overall proportion in favour of the erlotinib group was calculated.

RESULTS

Efficacy outcome. The main analysis of efficacy and safety was conducted after 486 deaths (239 on erlotinib and gemcitabine and 247 on placebo and gemcitabine) and has already been reported (Moore *et al*, 2007). Overall survival was significantly longer in the erlotinib and gemcitabine arm with an estimated HR of 0.82 (95% CI, 0.69–0.99; $P = 0.011$; log-rank test stratified for performance status, extent of disease). Median survival times were 6.24 months *vs* 5.91 months for the erlotinib and gemcitabine *vs* placebo and gemcitabine groups, respectively.

Four hundred and ninety-nine patients had developed progressive disease or had died at the end of the trial. Progression-free survival was significantly longer in the erlotinib and gemcitabine arm with an estimated HR of 0.77 (95% CI, 0.64–0.92; $P = 0.004$; median, 3.75 months *vs* 3.55 months).

Toxicity outcomes. Two hundred eighty-two patients on the erlotinib and gemcitabine arm and 280 on the placebo and gemcitabine arm received at least one dose of study medication and were available for the assessment of toxicity.

The frequency of all grade and grade ⩾3 treatment-related AEs was higher for the erlotinib and gemcitabine group (90% and 31%, respectively) compared with the placebo and gemcitabine group (76% and 20%, respectively) (Table 2). The increase in grade ⩾3 AEs was especially notable for rash (6% *vs* 0%).

Benefit–risk assessment. The proportion in favour of the erlotinib group was +4.7%, 95% CI, −5.6–14·6% (thus favouring erlotinib) for the first priority outcome (OS) but −8.3%, 95% CI, −14.2–7.1% (thus favouring placebo) for the second priority outcome (toxicity) among patients uninformative on the OS outcome. Overall, the net proportion favoured non-significantly the placebo group (overall Δ[erlotinib] = −3.6, 95% CI, −14.2–7.1; $P = 0.51$), suggesting an unfavourable benefit–risk balance of erlotinib added to gemcitabine (Table 3).

Table 2. Worst grade toxicity by treatment group

Worst grade related AE	Erlotinib group ($n = 282$)	Placebo group ($n = 280$)
Grade 1	48 (17.0%)	69 (24.6%)
Grade 2	118 (41.8%)	89 (31.8%)
Grade 3	72 (25.5%)	47 (16.8%)
Grade 4	11 (3.9%)	6 (2.1%)
Grade 5	4 (1.4%)	3 (1.1%)

Abbreviation: AE = adverse events.

Table 3. Main analysis of the benefit–risk balance of erlotinib and gemcitabine combination

	Proportion of pairs (%)		Difference
Priority	Erlotinib > placebo	Placebo > erlotinib	Δ[erlotinib]
OS (threshold = 2 months)	37.0	32.3	4.7
Worst related AE grade	7.5	15.7	−8.3
Overall	44.5	48.1	−3.6 ($P = 0.51$)

Abbreviations: > = better than; AE = adverse events; Δ[erlotinib] = proportion in favour of the erlotinib group; OS = overall survival.

Sensitivity analyses. The analysis was repeated with various values for the OS threshold, varying between 0 and 6 months. When the OS threshold was set at 0 month, meaning that any difference in OS was considered meaningful, the overall analysis was not statistically significant (overall proportion in favour of erlotinib = 2.3, 95% CI, −8.1–12.7; $P = 0.67$). This setting gave a large weight to the first priority OS outcome, because any survival improvement was considered clinically significant, regardless of AEs. As the OS threshold increased, the overall assessment leaned more and more in favour of the placebo group. It reached statistical significance in favour of erlotinib for values of the OS threshold >5 months (Figure 1).

The analysis was repeated using a threshold of two AE grades for the second priority toxicity outcome (hence, in this analysis, a difference of one grade or less was not considered clinically meaningful). Again, the analysis tended to favour the placebo group but remained non-significant statistically (Table 4).

When only Grade ⩾3 AEs were considered in the second priority toxicity outcome, the overall proportion in favour of erlotinib was again low for OS threshold under 2 months (+1.5, 95% CI, −8.5–11.4; $P = 0.77$) and became negative for OS thresholds larger than 2.5 months (Figure 2). The analyses never reached statistical significance for the tested OS thresholds (up to 6 months).

When skin rashes were excluded from the list of AEs analyzed in the second priority outcome, the overall analysis was not in favour of erlotinib (overall proportion in favour of erlotinib = −0.3, 95% CI, −9.1–8.4; $P = 0.94$) (Table 5). A subgroup analysis was performed according to the occurrence of a grade ⩾2 rash in the erlotinib group. The benefit–risk of erlotinib in the subgroup of patients experiencing grade ⩾2 rashes was statistically significantly favourable (Δ[erlotinib] = 13.7; $P = 0.032$), and it was statistically significantly unfavourable in the subgroup of patients with grade 0 or 1 rashes (Δ[erlotinib] = −13.8; $P = 0.016$) (Table 6).

In the subgroup of the 521 patients treated with 100 mg per day of erlotinib, the main analysis of benefit–risk once again was not in favour of the erlotinib (overall proportion in favour of erlotinib = −2.7, 95% CI, −13.6–8.1; $P = 0.62$).

Comprehensive sensitivity analyses of the benefit–risk were carried out using various thresholds for OS, PFS and worst AE grade. Some scenarios with clinically meaningful choices of end point prioritisation and of thresholds are presented in Table 7.

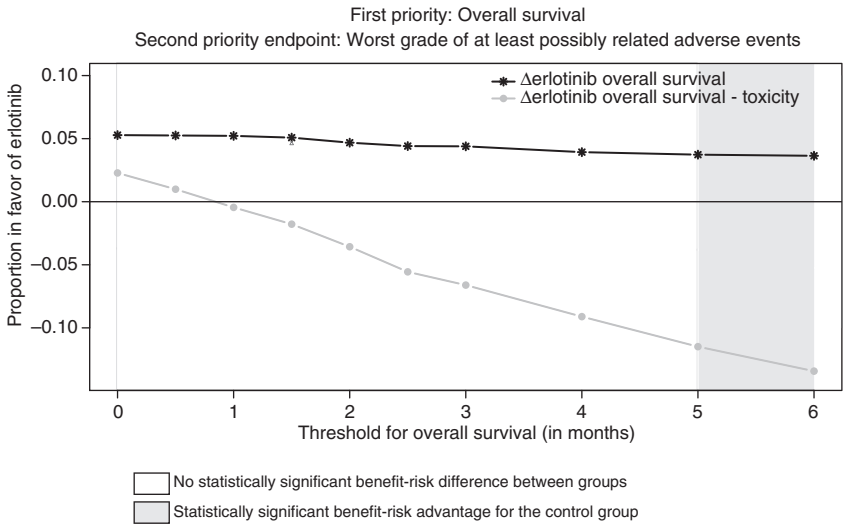


Figure 1. Benefit–risk of erlotinib according to the minimum survival benefit considered clinically meaningful. Proportion in favour of erlotinib according to the minimum survival benefit considered clinically meaningful. First priority outcome: overall survival. Second priority outcome: worst grade of at least possibly related adverse events. Solid black line with asterisks: proportion in favour of erlotinib according to the first priority outcome (OS) only. Solid light-grey line with points: overall proportion in favour of erlotinib.

For none of the scenarios considered was the overall benefit risk assessment in favour of erlotinib.

DISCUSSION

We have used generalised pairwise comparisons, prioritised on several outcomes, to perform an assessment of the benefit–risk balance of adding erlotinib to gemcitabine for the treatment of patients with advanced pancreatic cancer. These analyses showed that the OS benefit in favour of erlotinib diminished when using increased thresholds for the OS benefit and/or adding AEs in an assessment of the net benefit of the combination. The benefit risk assessment did not favour adding erlotinib in the main analysis, and this result was confirmed in all sensitivity analyses.

The method of generalised pairwise comparisons gives higher priority to the outcome considered clinically more important – in this case, overall survival was considered more important than any grade of toxicity. The method can incorporate both a priority and a threshold for each of the outcomes considered (in this instance, OS and treatment-related toxicities), and as such it reflects the thinking process of clinicians and decision makers, who try to assess the net effect of a new treatment on several outcomes considered to be of clinical importance. As such, the method may be particularly informative in health technology assessment.

Several methods have been proposed to help the scientific assessment of the benefit–risk balance of interventions. These methods are most frequently designed to weigh relevant efficacy and safety data into a single construct (Committee for Medicinal Products for Human Use (CHMP), 2008). QALY is a measurement of health status that assigns a weight in each period of time according to the quality of life during this period (Weinstein *et al*, 2009). It might be used to adjust a gain in survival to an increased level of toxicity by assigning a smallest weight to the time of survival with significant toxicity. However, it requires clearly defined health states, as well as weights for each state, which might be difficult to establish when planning a trial. This limitation makes QALY difficult to use as a primary end point to evaluate therapeutic interventions, and a more suitable tool for medico-economic evaluation (Whitehead and Ali, 2010). Other methods such as Overall Treatment Utility (OTU) can be used to combine subjective and objective measures of the treatment effect into a single composite end point. However the respective weights of the different treatment effects included in OTU may be difficult to justify and to report (Seymour *et al*, 2011).

The method of generalised pairwise comparisons only requires the priority of each outcome to be defined. Sensitivity analyses are useful to confirm the conclusion of the main analysis. Indeed, the conclusion may rest entirely on arbitrary (though arguably relevant) choices made regarding outcome priorities and thresholds values (if any). Most clinicians and patients would agree that

Table 4. Sensitivity analysis of the benefit–risk balance of the erlotinib and gemcitabine combination – only differences in treatment-related AEs of at least two grades are considered clinically meaningful

	Proportion of pairs (%)		Difference
Priority	Erlotinib > placebo	Placebo > erlotinib	Δ[erlotinib]
OS (threshold = 2 months)	37.0	32.3	4.7
Worst related AE grade (threshold = 2 grades)	3.0	8.4	−5.3
Overall	40.1	40.7	−0.6 (P = 0.90)

Abbreviations: > = better than; AE = adverse events; Δ[erlotinib] = proportion in favour of the erlotinib group; OS = overall survival.

Table 5. Sensitivity analysis of the benefit–risk balance of the erlotinib and gemcitabine combination – skin rashes are excluded from the list of adverse events

	Proportion of pairs (%)		Difference
Priority	Erlotinib > placebo	Placebo > erlotinib	Δ[erlotinib]
OS (threshold = 2 months)	37.0	32.3	4.7
Worst related AE grade[a]	9.1	14.1	−5.0
Overall	46.1	46.4	−0.3 (P = 0.94)

Abbreviations: > = better than; AE = adverse events; Δ[erlotinib] = proportion in favour of the erlotinib group; OS = overall survival.
[a]Skin rashes are excluded from the list of adverse events.

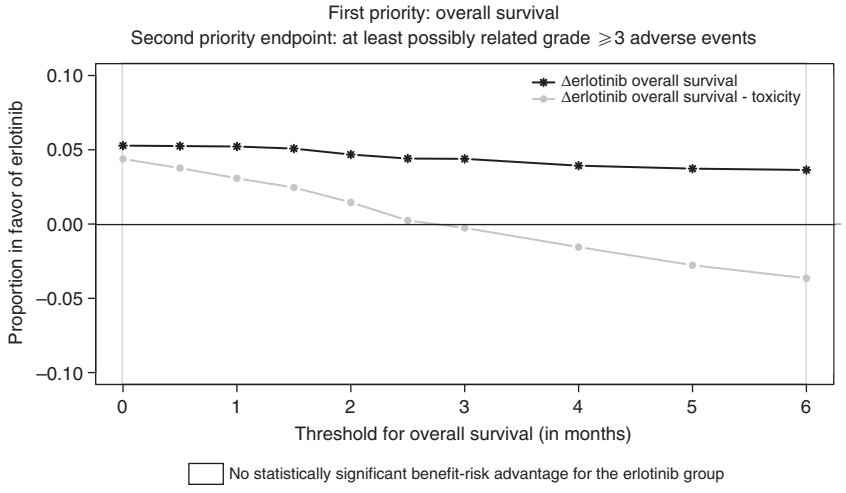


Figure 2. Benefit–risk of erlotinib according to the minimum survival benefit considered clinically meaningful. Only grade 3–4 adverse events are included in this analysis. Proportion in favour of erlotinib according to the minimum survival benefit considered clinically meaningful. First priority outcome: overall survival. Second priority outcome: presence of a grade 3–4 at least possibly related adverse events. Solid black line with asterisks: proportion in favour of erlotinib according to the first priority (OS) outcome only. Solid light-grey line with points: overall proportion in favour of erlotinib.

small gains in survival cannot be considered as a positive outcome if such gains are obtained at the expense of severe toxicities. However, determining the minimal survival benefit threshold for which most patients would accept to experience a treatment-related AE is very complex. It may depend on the type of AE and its grade, and it may vary considerably from patient to patient. Survival benefits may be offset by severe and/or long-term AEs. Investigators can now use generalised pairwise comparisons to test the benefit–risk balance of investigational therapies, depending on the level of tolerable toxicity that is deemed acceptable for a given magnitude of survival benefit. Various scenarios for the threshold of survival benefit and the grades of AEs are reported in the Table 7. Throughout all the scenarios, the benefit–risk balance leaned against erlotinib, which does provide some confirmation of the results of the main analysis. Moreover the clinical impact of AEs may vary a lot depending of the type of AEs, even among AEs of the same grade. When skin rashes were excluded from the list of relevant adverse events, the benefit risk assessment of erlotinib was close to zero.

Table 6. Analysis of the benefit–risk balance of the erlotinib and gemcitabine combination, according to the occurrence of a grade ⩾2 rash in the erlotinib group

	Δ[erlotinib]	
Priority	**Grade ⩾2 rash in the erlotinib group**	**Grade 0–1 rash in the erlotinib group**
OS (threshold = 2 months)	31.2	−11.0
Worst related AE grade	−17.5	−2.8
Overall	13.7 (P = 0.032)	−13.8 (P = 0.016)

Abbreviations: AE = adverse events; Δ[erlotinib] = proportion in favour of the erlotinib group; OS = overall survival.

Relevant toxicity criteria could potentially vary from trial to trial. For example, a risk assessment could focus on predefined AEs of special interest, on all severe AEs, on severe treatment-related AEs, or on AEs leading to drug discontinuation. (Ioannidis *et al*, 2004) For the PA.3 trial, the frequency of lethal AEs or of AEs leading to treatment discontinuation was low, as well as the frequency of grade 3–4 AEs.

Generalised pairwise comparisons are useful to perform a quantitative assessment of the benefit–risk balance of a new treatment as compared with a standard therapy. Such an assessment is especially useful when overall efficacy differences are small, and no subset of patients has been identified as being more likely to benefit from treatment. In such cases, generalised pairwise comparisons provide a clinically intuitive way of comparing patients with respect to all important efficacy and toxicity outcomes, with full flexibility as to the priority of each outcome, and a threshold of clinical significance. In particular, when some patients benefit from treatment at the price of a given toxicity (e.g., severe treatment-related rash after administration of a tyrosine kinase inhibitor), the prioritisation of their outcomes

Table 7. Further sensitivity analyses of the benefit–risk balance of the erlotinib and gemcitabine combination, using different priorities and threshold values for the outcomes of interest

		Proportion of pairs (%)		Difference
Priority	**Threshold**	**Erlotinib > placebo**	**Placebo > erlotinib**	**Δ[erlotinib]**
1. OS	6 months	16.8	13.1	3.6
2. PFS	6 months	3.0	1.8	1.2
3. Worst related AE grade[a]	3 grades	2.3	5.8	−3.5
4. OS	3 months	11.2	10.8	0.4
5. PFS	3 months	3.4	2.7	0.7
6. Worst related AE grade[a]	2 grades	2.3	6.0	−3.7
7. OS	0 months	9.5	9.1	0.4
8. PFS	0 months	0.5	0.6	−0.1
9. Worst related AE grade[a]	1 grade	0.2	0.5	−0.3
Overall		49.2	50.4	−1.2 (P = 0.82)
1. OS	4 months	25.7	21.8	3.9
2. PFS	4 months	4.5	2.6	1.9
3. Worst related AE grade[a]	2 grades	5.0	11.9	−6.9
4. OS	2 months	6.1	5.8	0.3
5. PFS	2 months	2.0	1.9	0.1
6. Worst related AE grade[a]	1 grade	3.1	4.7	−1.6
7. OS	0 months	2.1	2.1	0.0
8. PFS	0 months	0.2	0.2	0.0
Overall		48.7	50.9	−2.2 (P = 0.67)
1. Worst related AE grade[a]	3 grades	3.2	8.8	−5.6
2. OS	4 months	22.8	18.9	3.9
3. PFS	4 months	4.0	2.5	1.6
4. Worst related AE grade[a]	2 grades	3.3	7.8	−4.5
5. OS	2 months	6.1	5.8	0.3
6. PFS	2 months	2.0	1.9	0.1
7. Worst related AE grade[a]	1 grade	3.1	4.7	−1.6
8. OS	0 months	2.1	2.1	0.0
9. PFS	0 months	0.2	0.2	0.0
Overall		46.9	52.7	−5.8 (P = 0.27)

Abbreviations: > = better than; AE = adverse events; Δ[erlotinib] = proportion in favour of the erlotinib group; OS = overall survival; PFS = progression-free survival.
[a]Only differences beyond the threshold value in treatment-related AEs are included in this toxicity assessment.

naturally ensures that the benefit trumps the toxicity in the overall assessment of the benefit–risk balance.

ACKNOWLEDGEMENTS

Dr Julien Péron is the recipient of a grant from the Nuovo-Soldati Research Foundation. The study was not funded.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Julien Péron and Marc Buyse: design of the study, data collection and analysis, and writing and approval of the manuscript. Guarantor: Pascal Roy and Laurent Roche: design of the study, data analysis, and writing and approval of the manuscript. Keyue Ding and Wendy R Parulekar: data collection and analysis and writing and approval of the manuscript.

REFERENCES

Boeck S, Jung A, Laubender RP, Neumann J, Egg R, Goritschan C, Ormanns S, Haas M, Modest DP, Kirchner T, Heinemann V (2013) KRAS mutation status is not predictive for objective response to anti-EGFR treatment with erlotinib in patients with advanced pancreatic cancer. *J Gastroenterol* **48**: 544–548.

Burris HA, Moore MJ, Andersen J, Green MR, Rothenberg ML, Modiano MR, Cripps MC, Portenoy RK, Storniolo AM, Tarassoff P, Nelson R, Dorr FA, Stephens CD, Von Hoff DD (1997) Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. *J Clin Oncol* **15**: 2403–2413.

Buyse M (2008) Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials* **5**: 641–642.

Buyse M (2010) Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med* **29**: 3245–3257.

Choi M, Razzaque S, Kim R (2012) Systemic therapy of advanced pancreatic cancer: has the landscape changed? *Clin Adv Hematol Oncol* **10**: 442–451.

Committee for Medicinal Products for Human Use (CHMP) (2008) Report of the CHMP working group on benefit-risk assessment models and methods. http//www.ema.europa.eu. Last accessed March 2014.

Conroy T, Desseigne F, Ychou M (2011) FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med* **364**: 1817–1825.

Da Cunha Santos G, Dhani N, Tu D, Chin K, Ludkovski O, Kamel-Reid S, Squire J, Parulekar W, Moore MJ, Tsao MS (2010) Molecular predictors of outcome in a phase 3 study of gemcitabine and erlotinib therapy in patients with advanced pancreatic cancer: National Cancer Institute of Canada Clinical Trials Group Study PA.3. *Cancer* **116**: 5599–5607.

Food and Drug Administration (2011) PDUFA Reauthorization performance goals and procedures fiscal years 2013 through 2017. [Internet]. http//www.fda.gov/downloads/ForIndustry/User- Fees/Prescription DrugUserFee/UCM270412.pdf. Last accessed March 2014.

Ioannidis JPA, Evans SJW, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D. CONSORT Group (2004) Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* **141**: 781–788.

Miksad RA, Schnipper L, Goldstein M (2007) Does a statistically significant survival benefit of erlotinib plus gemcitabine for advanced pancreatic cancer translate into clinical significance and value? *J Clin Oncol* **25**: 4506–4507, author reply 4508.

Moore MJ, Goldstein D, Hamm J, Figer A, Hecht JR, Gallinger S, Au HJ, Murawa P, Walde D, Wolff RA, Campos D, Lim R, Ding K, Clark G, Voskoglou-Nomikos T, Ptasynski M, Parulekar W. National Cancer Institute of Canada Clinical Trials Group (2007) Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *J Clin Oncol* **25**: 1960–1966.

Moser BK, McCann MH (2008) Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials* **5**: 248–252.

Péron J, Maillet D, Gan HK, Chen EX, You B (2013) Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: a systematic review. *J Clin Oncol* **31**: 3957–3563.

Péron J, Pond GR, Gan HK, Chen EX, Almufti R, Maillet D, You B (2012) Quality of reporting of modern randomized controlled trials in medical oncology: a systematic review. *J Natl Cancer Inst* **104**: 982–989.

Saif MW (2008) Is there a standard of care for the management of advanced pancreatic cancer? Highlights from the Gastrointestinal Cancers Symposium, Orlando, FL, USA. January 25-27, 2008. *JOP* **9**: 91–98.

Seymour MT, Thompson LC, Wasan HS, Middleton G, Brewster AE, Shepherd SF, O'Mahony MS, Maughan TS, Parmar M, Langley RE. FOCUS2 InvestigatorsNational Cancer Research Institute Colorectal Cancer Clinical Studies Group (2011) Chemotherapy options in elderly and frail patients with metastatic colorectal cancer (MRC FOCUS2): an open-label, randomised factorial trial. *Lancet* **377**: 1749–1759.

Tam VC, Ko YJ, Mittmann N, Cheung MC, Kumar K, Hassan S, Chan KK (2013) Cost-effectiveness of systemic therapies for metastatic pancreatic cancer. *Curr Oncol* **20**: e90–e106.

Verslype C, Van Cutsem E, Dicato M, Cascinu S, Cunningham D, Diaz-Rubio E, Glimelius B, Haller D, Haustermans K, Heinemann V, Hoff P, Johnston PG, Kerr D, Labianca R, Louvet C, Minsky B, Moore M, Nordlinger B, Pedrazzoli S, Roth A, Rothenberg M, Rougier P, Schmoll HJ, Tabernero J, Tempero M, van de Velde C, Van Laethem JL, Zalcberg J (2007) The management of pancreatic cancer. Current expert opinion and recommendations derived from the 8th World Congress on Gastrointestinal Cancer, Barcelona, 2006. *Ann Oncol* **18**: 1–10.

Weinstein MC, Torrance G, McGuire A (2009) QALYs: the basics. *Value Health* **12**: 5–9.

Whitehead SJ, Ali S (2010) Health outcomes in economic evaluation: the QALY and utilities. *Br Med Bull* **96**: 5–21.

La conclusion principale de cet article est que la balance bénéfice-risque de l'erlotinib associé à un traitement par gemcitabine dans le cancer du pancréas avancé ou métastatique n'est pas favorable. Au moment de la réalisation de cette étude, les extensions de la méthode des comparaisons par paire dites de Peto, de Efron et de Péron n'avaient pas encore été développées. Comme cela a été présenté dans le chapitre III de cette thèse, l'estimation de la propension au succès estimée à partir de la procédure standard et en présence de données censurées est biaisée dans le sens d'une sous-estimation systématique de l'effet thérapeutique, positif ou négatif. Nous avons donc répété l'analyse principale rapportée dans le *British Journal of Cancer* en utilisant l'extension dite de Péron. Les résultats de cette analyse supplémentaire sont rapportés dans le tableau IV.2.

Tableau IV-2. Analyse principale de la balance bénéfice-risque de l'erlotinib associé à la gemcitabine en utilisant l'extension dite de Péron.

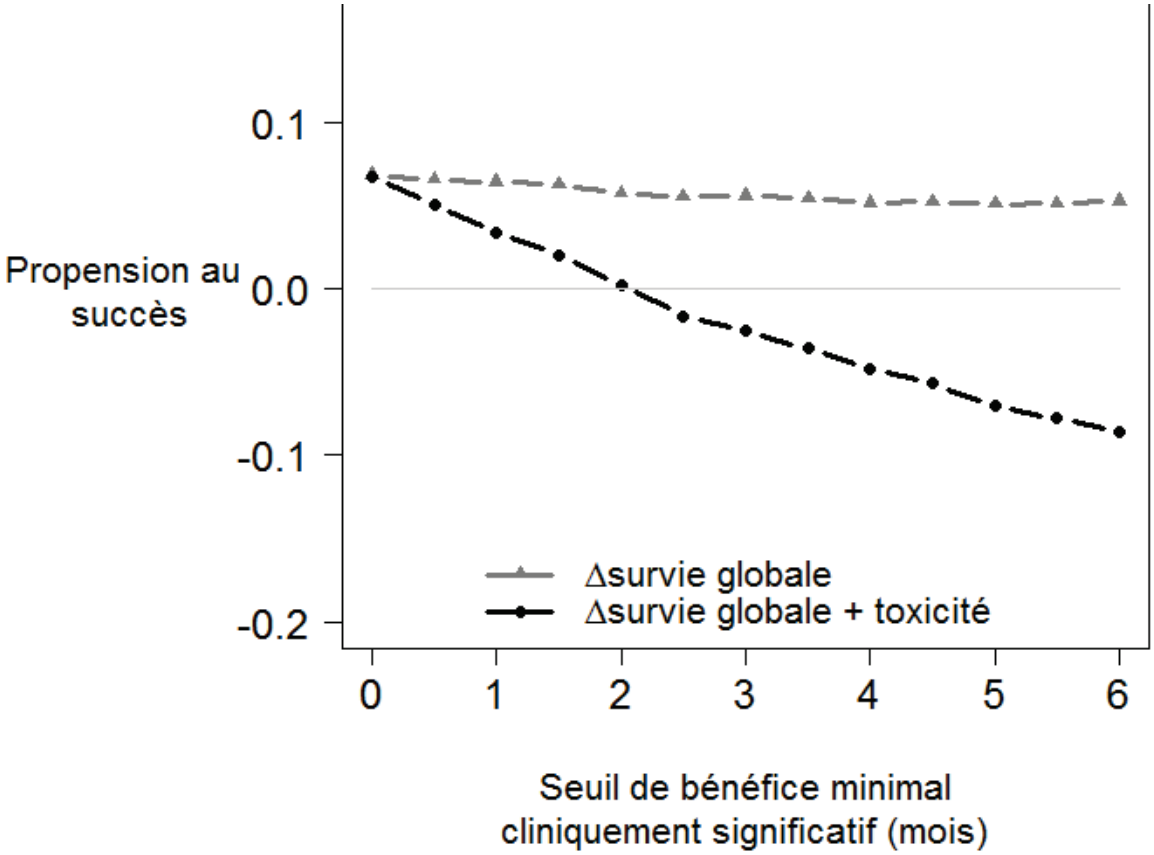
Priority	**Erlotinib > Placebo**	**Placebo > Erlotinib**	**Δerlotinib**
1 : Survie (Seuil = 2 mois)	40.3%	34.5%	5.8%
2 : Grade max d'EI	6.8%	12.4%	-5.6%
Global	**47.1%**	**46.9%**	**0.2% (P=.96)**

Légende du tableau IV.2 : EI=Evénement Indésirable ; Δerlotinib=propension au succès d'un patient traité par erlotinib.

La propension au succès estimée par l'extension dite de Péron fait une meilleure utilisation des données de survie globale. En effet les paires qui ne sont pas directement classables sur le critère de survie globale ne sont pas directement analysées sur le critère de jugement de seconde priorité. Pour chaque paire non directement classable, la probabilité que la paire soit favorable, défavorable, ou neutre sur le critère de survie est estimée. La propension globale au succès est alors estimée à 0.2% (Intervalle de confiance (IC) à 95% = -10.3%-10.9%, P=0.96). Cette estimation est supérieure à celle qui avait été réalisée en utilisant la procédure standard, mais la conclusion globale de l'étude n'est pas modifiée. En effet aucun effet significativement favorable ou défavorable de l'erlotinib en terme de balance bénéfice-risque n'a été identifié, quelle que soit la procédure utilisée. Les analyses de sensibilité rapportées dans le *British Journal of Cancer* ont également été répétées en utilisant l'extension dite de Péron (Figure IV.1, tableau IV.3). L'analyse rapportée dans la figure IV.1

est l'analogue de celle qui est rapportée en figure 1 dans l'article publié. Le seuil de bénéfice minimal cliniquement significatif en survie globale variait entre 0 et 6 mois. La propension globale au succès, intégrant la survie globale en critère de première priorité et la toxicité en critère de seconde priorité diminuait à mesure que le seuil augmentait. Néanmoins sur tout l'éventail de seuil analysé, la propension au succès estimée n'était jamais statistiquement significativement différente de 0.

Figure IV-1. Analyse bénéfice-risque de l'erlotinib en fonction du seuil de bénéfice minimal cliniquement significatif.



Légende de la figure IV.1 : La ligne grise correspond à la propension au succès en analysant uniquement la survie globale ($\Delta_{\text{survie globale}}$). La ligne noire correspond à la propension globale au succès ($\Delta_{\text{survie globale + toxicité}}$).

Tableau IV-3. Analyse de sensibilité de la balance bénéfice-risque de l'erlotinib associé à la gemcitabine en utilisant l'extension dite de Péron.

Priorité	Seuil	Erlotinib> Placebo	Placebo> Erlotinib)	∆erlotinib
1 : SG	6 mois	22.7%	16.3%	6.4%
2 : Grade max d'EI	≥ 3 grades	2.2%	5.6%	-3.4%
3 : SG	3 mois	12.0%	10.3%	1.7%
4 : Grade max d'EI	≥ 2 grades	2.9%	6.0%	-3.1%
5 : SG	0 mois	11.4%	10.6%	0.8%
6 : Grade max d'EI	≥ 1 grade	0%	0.1%	0%
Global		**51.2%**	**48.9%**	**2.3% (P=.62)**

Légende du tableau IV.3 : SG=Survie globale. EI=Evénement indésirable.

Le tableau IV.3 rapporte une analyse de sensibilité permettant d'évaluer de façon plus précise la balance bénéfice-risque de l'erlotinib. Dans cette analyse, une différence en survie d'au moins 6 mois était considérée cliniquement significative, quel que soit le profil de toxicité observé chez les patients. Lorsque deux patients comparés au sein d'une paire avaient une différence de survie située entre 3 et 6 mois, une différence de sévérité en termes d'événements indésirables d'au moins 3 grades était considérée inacceptable. Le raisonnement était ensuite poursuivi avec plusieurs seuils de différence minimale cliniquement significative sur la survie et sur la sévérité des événements indésirables observés. Comme ce qui avait été observé en utilisant la procédure standard, les analyses de sensibilité conduites avec l'extension dite de Péron ne retrouvaient jamais d'effet favorable de l'erlotinib en termes de balance bénéfice-risque.

IV.3.b. Evaluation de la balance bénéfice-risque du FOLFIRINOX

Depuis l'essai NCIC-PA3, qui évaluait l'association de l'erlotinib à la gemcitabine dans le cancer du pancréas avancé ou métastatique, deux traitements ont été évalués et approuvés sur les données d'essais randomisés positifs. L'association de chimiothérapie FOLFIRINOX (5-fluorouracile, oxaliplatine, irinotecan, leucovorine) a montré dans l'essai

PRODIGE 4 sa supériorité en termes de survie sans progression (rapport des taux instantanés de décès ou de progression, 0.47; IC 95%, 0.37 à 0.59; P<0.001) et de survie globale (rapport des taux instantanés de décès, 0.57; IC 95%, 0.45 to 0.73; P<0.001) [28]. Cependant, du fait de la toxicité du traitement par FOLFIRINOX et de la sélection de patients en bon état général dans l'essai PRODIGE 4, certains auteurs remettent en question sa balance bénéfice-risque [29]. Plus récemment, un essai a montré que la combinaison de nab-paclitaxel et de gemcitabine était supérieure à la gemcitabine seule en termes de survie sans progression et de survie globale [30]. Le FOLFIRINOX et le nab-paclitaxel représentent donc deux alternatives en traitement de première ligne du cancer du pancréas métastatique chez les patients en bon état général. Du fait de l'absence d'essai comparant directement ces deux options thérapeutiques, il est intéressant d'évaluer leurs balances bénéfice-risque respectives par rapport à la gemcitabine. Nous rapportons ici le résultat d'une évaluation de la balance bénéfice-risque du FOLFIRINOX dans l'essai PRODIGE 4 en utilisant l'extension dite de Péron des comparaisons par paire généralisées. Ce manuscrit est en cours de soumission au *European Journal of Cancer*.

Title : An assessment of the benefit-risk balance of FOLFIRINOX in metastatic pancreatic adenocarcinoma.

Running title: benefit-risk of FOLFIRINOX in pancreatic cancer

Julien Péron[1,2], Pascal Roy [1], Brice Ozenne [1], Laurent Roche [1], Marc Buyse [3]

1. Service de biostatistiques, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, F-69310, Pierre-Bénite, France, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Villeurbanne, France

2. Medical oncology department, Centre Hospitalier Lyon-Sud, Institut de Cancérologie des Hospices Civils de Lyon –IC-HCL), CITOHL, F-69310, Pierre-Bénite, France

3. International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium

Corresponding author:

Julien Péron, MD

Service de Biostatistique, Centre Hospitalier Lyon-Sud

165 Chemin du Grand Revoyet

F-69310, Pierre-Bénite, France.

Tel: (+33) 6 16 25 89 91

Fax: (+33) 4 78 86 57 74

E-mail: julien.peron@chu-lyon.fr

Conflict of interest statement: We have no conflict of interest to disclose

Abstract

Objective

Efficacy and safety are the two considerations when characterizing the effects of a new therapy. We sought to assess the benefit-risk balance using data from a completed randomized controlled trial that compared FOLFIRINOX versus gemcitabine in patients with metastatic pancreatic adenocarcinoma (The Prodige 4 - ACCORD 11/0402 trial).

Methods

We used generalized pairwise comparisons. This innovative statistical method permits the simultaneous analysis of several prioritized outcome measures (e.g. one or more benefit outcomes and one or more risk outcomes). Here, the first priority outcome was survival time (OS). Differences in OS that exceeded two months were considered clinically relevant. The second priority outcome was toxicity. The overall treatment effect was quantified using the chance of a better outcome with FOLFIRINOX, which can be interpreted as the net probability for a random patient treated in the FOLFIRINOX group to have a better overall outcome than a random patient in the gemcitabine group. The chance of a better outcome with FOLFIRINOX ranges from +1 (if all patients fare better with FOLFIRINOX than with gemcitabine) to -1 in the opposite situation. Sensitivity analyses were performed.

Results

In this trial 342 patients received either FOLFIRINOX or gemcitabine. The chance of a better outcome favored strongly and significantly the FOLFIRINOX group (24.7%, 95% CI, 11.1% to 38.0%; P<.001), suggesting a favorable benefit-risk balance of FOLFIRINOX versus gemcitabine. The positive benefit-risk balance of FOLFIRINOX was observed throughout all sensitivity analyses.

Conclusions

Generalized pairwise comparisons are useful to perform a quantitative assessment of the benefit-risk balance of new treatments. It provides a clinically intuitive way of comparing patients with respect to all important efficacy and toxicity outcomes, with full flexibility as to the priority of each outcome, and to the thresholds of clinical relevance. Overall the benefit-risk balance of FOLFIRINOX was strongly positive.

Keywords :

Statistics as topic ; treatment outcome ; pancreatic cancer ; chemotherapy ; randomized controlled trial

Introduction

Efficacy and safety are the primary considerations when characterizing a treatment effect. Both US Food and Drug Administration and the European Medicines Agency have stressed the importance of a more structured and transparent approach to benefit–risk assessment (BRA) in the evaluation of new therapies [1, 2]. In oncology clinical trials, efficacy and safety outcomes are usually analyzed and reported independently [3, 4].

Patients with metastatic pancreatic cancer have a poor prognosis and the historical first line regimen is gemcitabine [5]. Several new systemic therapies have been investigated in combination with gemcitabine in randomized trials. Among them the NCIC Clinical Trials Group Study PA.3 phase III trial investigated the addition of erlotinib to gemcitabine. Both overall survival (OS) and progression-free survival (PFS) were significantly better for the combination treatment [6]. However a benefit-risk assessment was performed using generalized pairwise comparison and was not in favor of adding erlotinib to gemcitabine for the treatment of patients with advanced pancreatic cancer [7].

In the last few years, two chemotherapy combination regimens have shown in randomized trials to improve largely patients' outcomes. FOLFIRINOX (5-fluorouracil, oxaliplatin, irinotecan, leucovorin) has shown superiority over gemcitabine in both PFS (hazard ratio for disease progression, 0.47; 95% CI, 0.37 to 0.59; P<0.001) and OS (hazard ratio for death, 0.57; 95% confidence interval [CI], 0.45 to 0.73; P<0.001) [8]. However, there is controversy as to whether the survival benefits of the FOLFIRINOX combination outweigh the associated toxicities [9]. More recently, a trial comparing a combination of nab-paclitaxel and gemcitabine versus gemcitabine alone demonstrated a statistically significant survival benefit for this new doublet, introducing another option for the management of advanced pancreatic cancer [10]. With the introduction of these therapeutic options, and the lack of randomized trials that directly compare all available treatments, it was of interest to compute an assessment of the benefit-risk balance of FOLFIRINOX. We report here such an assessment based on the method of generalized pairwise comparisons [11]. This method extends the Mann-Whitney-Wilcoxon test for a single outcome in the absence of censored data. It allows one to calculate and test the overall effect of a new treatment based on any number of prioritized outcomes, some reflecting benefit from the intervention (e.g., survival or time to progression), and the others reflecting harms (e.g., treatment-related toxicities and side effects).

Methods

Overview

The Prodige 4 - ACCORD 11/0402 trial randomized patients with metastatic pancreatic cancer to a combination chemotherapy regimen consisting of oxaliplatin, irinotecan, fluorouracil, and leucovorin (FOLFIRINOX) as compared with gemcitabine as first-line therapy. The primary outcome was OS. PFS and toxicity were secondary outcomes.

In this trial, 342 patients were stratified according to center, performance status (0 vs. 1), and primary tumor localization (the head vs. the body or tail of the pancreas), and randomly assigned in a 1:1 ratio to receive FOLFIRINOX or gemcitabine. Progression was evaluated using Response Evaluation Criteria in Solid Tumors (V1.0) every 2 months. Toxicity was assessed at every visit using the National Cancer Institute Common Toxicity Criteria version 3.0.

Generalized pairwise comparisons

We applied generalized pairwise comparisons extended to several outcome measures (a benefit outcome, and a risk outcome). A full description of the method has been previously published [1]. Briefly, pairwise comparisons require consideration of all possible pairs of patients, one taken from the FOLFIRINOX group, and the other taken from the gemcitabine group. Pairwise comparisons are stratified for the stratification factors used in the randomization process.

The outcomes of these two patients are compared according to the first priority outcome. The pair is said to be ‘favorable’ if the outcome of the patient in the FOLFIRINOX group is better than the outcome of the patient in the gemcitabine group and ‘unfavorable’ if the outcome of the patient in the FOLFIRINOX group is worse than the outcome of the patient in the gemcitabine group. The pair is said to be ‘uninformative’ if it cannot be determined which of the two patients has a better outcome (e.g., because of censoring, or because of missing data), and to be ‘neutral’ when the two observations are equal, or when the difference of outcomes does not reach a pre-specified threshold of clinical significance. When a pair is uninformative on a survival outcome as a result of censoring, the probabilities for this pair to be favorable, unfavorable, or neutral are calculated from the time to the censored observations, and from the Kaplan-Meier estimates of the survival functions. Such a pairwise comparison is carried out for all pairs of patients, and the difference between the probability

for a pair to be favorable and the probability to be unfavorable pairs is calculated for the first priority outcome. This difference is called the chance of a better outcome for the first priority outcome [12, 13].

For pairwise comparisons that are neutral or uninformative for the first priority outcome, the second priority outcome is used in turn to classify the pair as favorable, unfavorable, neutral, or uninformative (Table 1). After consideration of the second priority outcome, the "chance of a better overall outcome" is calculated to provide an overall assessment of both the benefit and the risks of the treatment, suitably prioritized.

First priority outcome	Second priority outcome	Pair is
favorable	ignored	favorable
unfavorable	ignored	unfavorable
neutral/uninformative	favorable	favorable
neutral/uninformative	unfavorable	unfavorable
neutral/uninformative	neutral/uninformative	neutral/uninformative

Table 1: Generalized pairwise comparisons for two prioritized outcomes

Standard analysis of efficacy and toxicity

A log-rank test stratified for stratification factors at baseline was used to compare treatment groups in terms of survival. The incidence of worst grade adverse events that were at least possibly related to the study treatment ("treatment-related AEs") were compared using a stratified chi-square test. Biological adverse events were excluded from the overall analysis of the treatment toxicity. All analyses were performed on all randomly assigned patients as per the intent-to-treat principle.

Main analysis of the benefit-risk balance

The first priority outcome used in the main analysis was OS. Only differences in OS exceeding two months were considered clinically significant. The second priority outcome was treatment-related AEs, with patients experiencing the lower AE grade considered to have had a more favorable outcome. Treatment groups were compared using the chance of a better outcome with FOLFIRINOX over gemcitabine (Δ[FOLFIRINOX]). A randomization test stratified by performance status and extent of disease at diagnosis was performed to test the null hypothesis (H_0 : Δ[FOLFIRINOX] = 0). The contribution of each outcome to Δ[FOLFIRINOX] was calculated.

Sensitivity analyses

The impact of the choice of outcomes, thresholds, and priority on the results was assessed in sensitivity analyses. First, the main analysis was repeated with various thresholds for the minimal OS difference considered as clinically significant, ranging from 0 (any difference in OS considered clinically meaningful) to 6 months. Second, the toxicity outcome was defined as a binary variable where only grade ≥ 3 AEs were considered. Finally, a wide range of scenarios integrating OS, PFS, and AE grades with several successive thresholds were built in order to provide a comprehensive assessment of the treatment effects. For each scenario, the chance of a better overall outcome in the FOLFIRINOX group was estimated and tested for statistical significance.

Statistical analyses

All analyses reported in this manuscript were performed using the extended procedure of generalized comparisons, using the method “Peron” of the package BuyseTest in the R software (available from the first author).

Results

Efficacy outcome

The main analysis of efficacy and safety was conducted after 273 deaths (126 in the FOLFIRINOX group and 147 in the gemcitabine group) and has already been reported [31]. Overall survival was significantly longer in the FOLFIRINOX group with an estimated HR of 0.57 (95% CI, 0.45 to 0.72; P<0·001; log-rank test stratified for performance status and primary tumor localization). Median survival times were 11.1 months versus 6.7 months for the FOLFIRINOX versus gemcitabine groups, respectively.

Three hundred and seventeen patients had developed progressive disease or had died at the end of the trial. Progression-free survival was significantly longer in the FOLFIRINOX and gemcitabine group with an estimated HR of 0·47 (95% CI, 0·37 to 0·59; P<0·001 ; median, 6.3 months versus 3.2 months).

Toxicity outcomes

The frequency grade ≥ 3 treatment-related clinical AEs was higher for the FOLFIRINOX group (69%) compared with the gemcitabine group (60%), but the difference was not statistically significant (P=0.083) (Table 2). The increase in grade ≥ 3 AEs was especially notable for the neurologic adverse events (11.1% versus 2.3%, P=0.0028), gastrointestinal adverse events (33.9% versus 24.6%, P=0.042), infectious adverse events (10.5% versus 5.3%, P=0.096), and general adverse events (28.7% versus 22.8%, P=0.19).

Worst grade related AE	FOLFIRINOX group (n=171)	Gemcitabine group (n=171)
Grade 0	6 (3.5%)	2 (1.2%)
Grade 1	7 (4.1%)	5 (2.9%)
Grade 2	40 (23.4%)	62 (36.3%)
Grade 3	81 (47.4%)	67 (39.2%)
Grade 4	36 (21.1%)	34 (19.9%)
Grade 5	1 (0.6%)	1 (0.6%)

AE = Adverse Events

Table 2: Worst grade toxicity by treatment group

Benefit-risk assessment

The chance of a better overall survival in the FOLFIRINOX group (first priority outcome with a threshold of clinical significance at 2 months) was 27.4%, 95% CI, 14.2% to 40.6% (thus favoring FOLFIRINOX), and the chance of a better toxicity (second priority outcome) was -2.7% (thus favoring gemcitabine) among patients neutral on the OS outcome. The chance of a better overall outcome favored significantly the FOLFIRINOX group (Overall Δ[FOLFIRINOX] = 24.7%, 95% CI, 11.1% to 38.0%; P<.001), suggesting a favorable benefit-risk balance of FOLFIRINOX versus gemcitabine (Table 3).

Priority	Pairwise probabilities (%)		Δ[FOLFIRINOX]
	FOLFIRINOX > Gemcitabine	Gemcitabine > FOLFIRINOX	
1 : OS (threshold = 2 months)	54.4%	26.9%	27.4%
2 : Worst related AE grade	4.8%	7.5%	-2.7%
Overall	59.2%	34.4%	**24.7% (P<.001)**

OS = Overall Survival ; AE = Adverse Events ; Δ[FOLFIRINOX] = Chance of a better outcome in the FOLFIRINOX group.

Table 3 : Main analysis of the benefit-risk balance of FOFLIRINOX versus gemcitabine

Sensitivity analyses

The analysis was repeated with various values for OS threshold, varying between 0 and 6 months. When the OS threshold was set at 0 month, meaning that any difference in OS was considered clinically significant, the overall analysis was statistically significant (chance of a better overall outcome with FOLFIRINOX = 27.0%, 95% CI, 12.7% to 40.1% ; P<.001). Even when only differences in OS larger than 6 months were considered clinically significant (threshold for OS = 6 months), the benefit-risk balance favored significantly the FOLFIRINOX group (chance of a better overall outcome with FOLFIRINOX = 16.8%, 95% CI, 3.4% to 29.6% ; P=.013) (Figure 1).

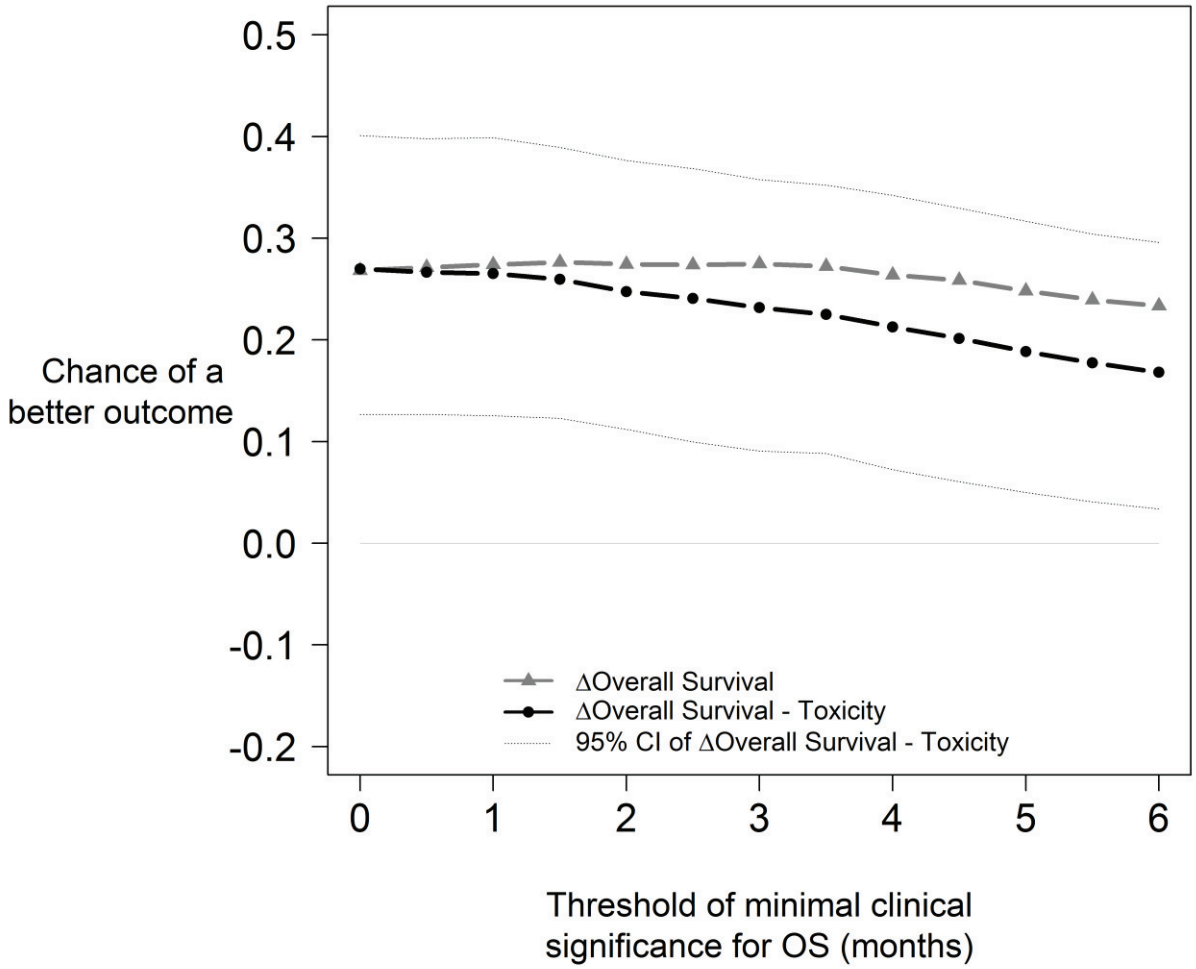


Figure 1 :

Title: Benefit-risk of FOLFIRINOX according to the minimum survival benefit considered clinically significant.

Legend: Chance of a better outcome with FOLFIRINOX according to the minimum survival benefit considered clinically meaningful. First priority outcome: Overall survival. Second priority outcome: Worst grade of at least possibly related adverse events. Solid black

line with asterisks : Chance of a better survival with. Solid light-grey line with points : Chance of a better overall outcome with FOLFIRINOX.

When related AEs were considered as a binary outcome (occurrence of at least one grade ≥ 3 related adverse event versus no grade ≥ 3 related adverse event), the chance of a better overall outcome favored significantly the FOLFIRINOX group (25.3%, 95% CI, 11.8% to 38.8% ; P<.001 , table 4, and figure A in the appendix).

Priority	Pairwise probabilities (%)		
	FOLFIRINOX > Gemcitabine	Gemcitabine > FOLFIRINOX	Δ[FOLFIRINOX]
1 : OS (threshold = 2 months)	54.4%	26.9%	27.4%
2 : Worst related AE grade ≥ 3	3.1%	5.2%	-2.1%
Overall	57.4%	32.1%	**25.3% (P<.001)**

OS = Overall Survival ; AE = Adverse Events ; Δ[FOLFIRINOX] = Chance of a better outcome in the FOLFIRINOX group.

Table 4 : Sensitivity analysis of the benefit-risk balance of FOFLIRINOX versus gemcitabine – the occurrence of a grade ≥ 3 related adverse event was analyzed as a binary outcome

When biological adverse events were included in the overall analysis of the benefit-risk balance, the chance of a better overall outcome with FOLFIRINOX varied only slightly (24.2%, 95% CI, -10.7% to 37.6% ; P<0.001).

Comprehensive sensitivity analyses of the benefit-risk were carried out using various thresholds for OS, and worst adverse event grade. Some scenarios with clinically meaningful choices of endpoint prioritization and of thresholds are presented in Table 5. All the scenarios considered favored the FOLFIRINOX group in term of benefit-risk balance.

1

Priority	Threshold	Pairwise probabilities		
		FOLFIRINOX > Gemcitabine	FOLFIRINOX > Gemcitabine	Δ[FOLFIRINOX]
1 : OS	9 months	26.1%	7.7%	18.4%
3 : Worst related AE grade*	3 grades	2.3%	1.2%	1.1%
4 : OS	6 months	11.2%	6.1%	5.1%
6 : Worst related AE grade*	2 grades	2.7%	4.7%	-2.0%
7 : OS	3 months	10.3%	6.3%	4.0%
9 : Worst related AE grade*	1 grade	4.2%	8.1%	-3.9%
Overall		56.8%	34.2%	**22.6% (P<.001)**
1 : Worst related AE grade*	3 grades	3·2%	1.8%	1.4%
2 : OS	6 months	36.1%	13.3%	22.8%
3 : PFS	6 months	4.6%	1.1%	3.5%
4 : Worst related AE grade*	2 grades	2.5%	4.0%	-1.5%
5 : OS	3 months	8.3%	5.5%	2.8%
6 : PFS	3 months	3.2%	1.2%	2.0%
7 : Worst related AE grade*	1 grade	3.1%	5.4%	-2.3%
8 : OS	0 months	3.2%	3.3%	-0.1%
9 : PFS	0 months	0.0%	0.0%	0.0%
Overall		64.3%	35.7%	**28.6% (P<.001)**

2 AE = Adverse Events ; OS = Overall Survival ; PFS = Progression Free Survival ;
3 Δ[FOLFIRINOX] = Chance of a better outcome in the FOLFIRINOX group.

Table 5: Further sensitivity analyses of the benefit-risk balance FOFLIRINOX versus gemcitabine, using different priorities and threshold values for the outcomes of interest

Discussion

We have used generalized pairwise comparisons using several outcomes to perform an assessment of the benefit-risk balance of FOLFIRINOX versus gemcitabine for first-line treatment of patients with metastatic pancreatic cancer. The main analysis of the benefit-risk balance, as well as all the sensitivity analyses, was strongly in favor of the FOLFIRINOX.

A similar analysis of the benefit-risk balance was previously conducted on the NCIC PA.3 phase III trial. This trial investigated the addition of erlotinib to gemcitabine in patients with advanced pancreatic cancer [6]. Both survival and progression-free survival were significantly better in the erlotinib group, but the overall benefits were of modest magnitude. Moreover the addition of erlotinib was associated with an increased frequency of all grade and grade $\geq$ 3 treatment-related AEs. Overall, the benefit-risk balance, assessed with generalized comparison was not in favor of the erlotinib. It should be noted that the procedure of generalized pairwise comparison used to assess the benefit-risk balance of erlotinib in the published report [7] was the standard procedure. The difference between the standard procedure and the extended procedure used in the current study is the method to handle censored observations for survival outcomes. The benefice-risk balance of erlotinib in the PA.3 trial was reassessed using the extended procedure used in this study. The priorities and threshold values for the main analysis were the same as those used in this study. The chance of a better overall outcome with erlotinib was 0.2% (95% CI, -10.1% to 10.7% ; P=.96). With the extended procedure, the benefit-risk balance of erlotinib was not in favor of the erlotinib group, the conclusion was then the same as the one based on the standard procedure.

More recently, the treatment options for metastatic pancreatic adenocarcinomas have increased with the approval of nab-paclitaxel (albumin-bound paclitaxel) as first line therapy in combination with gemcitabine. In the randomised phase III trial (MPACT) which compared gemcitabine plus nab-paclitaxel versus gemcitabine alone (n = 861), the combination treatment was demonstrated to improve OS (median OS: 8.5 months versus 6.7 months, HR 0.72, $p < 0.0001$), and PFS (median PFS: 5.5 versus 3.7 months, HR 0.69, $p < 0.0001$) [10]. In

the absence of head-to-head clinical trials comparing FOLFIRINOX and the combination of nab-paclitaxel and gemcitabine, the best treatment approach to use for untreated metastatic pancreatic adenocarcinoma is not known. We believe that the two regimens should be compared in terms of benefit-risk balance, because of their different toxicity profiles. Generalized pairwise comparison could be used to perform such comparison in a prospective randomized trial. An indirect comparison of the benefit-risk balance of the two regimens could also be performed through a retrospective analysis of their respective trials, because the comparative groups were gemcitabine alone in the three randomized trials evaluating either FOLFIRINOX or the nab-paclitaxel plus gemcitabine combination [8, 10, 14]. All these trials included patients in good performance status and were conducted in first-line.

One advantage of generalized pairwise comparison in the assessment of the benefit-risk balance is that it gives higher priority to the outcome considered clinically more important. The method can analyze simultaneously any number of outcomes. Each prioritized outcome is associated with a threshold of clinical significance, and as such it reflects the thinking process of clinicians and decision makers, who try to assess the net effect of a new treatment on several outcomes considered to be of clinical importance. Moreover, a single outcome can be included repeatedly at several priorities with different thresholds values.

Other methods have been proposed to help the scientific assessment of the benefit-risk balance of interventions [2]. QALY is a measurement of survival that assigns a weight in each period of time according to the quality of life of this period [15]. It might be used to adjust a gain in survival to an increased level of toxicity by assigning a smallest weight to the time of survival with significant toxicity. However it requires clearly defined health states, as well as weights for each state, which might be difficult to establish when planning a trial. The use of QALY as a primary endpoint in clinical trials has been limited for this reason. QALY is often considered more suited for medico-economic evaluation [16]. Overall Treatment Utility (OTU) can be used to combine subjective and objective measures of the treatment effect into a single composite endpoint. However the respective importance of the different treatment effects included in OTU may be difficult to understand and to report [17].When assessing the benefit-risk balance with generalized pairwise comparisons, sensitivity analyses are useful to assess the robustness of the main analysis conclusion. Indeed, the conclusion may rest entirely on arbitrary (though arguably relevant) choices made regarding outcome priorities and thresholds values. Most clinicians and patients would agree that small gains in survival cannot be considered as a positive outcome if such gains are obtained at the expense of severe toxicities.

However, the minimal survival benefit threshold for which most patients would accept to experience a treatment-related adverse event is often unknown. Investigators can now use generalized pairwise comparisons to test the benefit-risk balance of investigational therapies, depending on the level of tolerable toxicity that is deemed acceptable for a given magnitude of survival benefit. That is the purpose of the sensitivity analyses reported in figure 1 and table 5. Each scenario reported in these analyses could be chosen as the most relevant scenario by investigators or patients, depending on their expectation on a treatment efficacy and their tolerance to adverse events. Throughout all the scenarios, the benefit-risk balance favored the FOLFIRINOX group. In other trials investigating other regimens, the sensitivity analyses might lean to opposite conclusions. In such case, generalized pairwise comparisons could be used to help clinicians and patients in the choice of the best treatment depending on the patient own expectations of the treatment effect.

Generalized pairwise comparisons are useful to perform a quantitative assessment of the benefit-risk balance of a new treatment as compared with a standard therapy. It provides a clinically intuitive way of comparing patients with respect to all important efficacy and toxicity outcomes, with full flexibility as to the priority of each outcome, and a threshold of clinical significance. The benefit-risk balance of FOLFIRINOX versus gemcitabine in the Prodige 4 - ACCORD 11/0402 trial was strongly positive.

References

1. Drug USF and: Guidance for Industry and Food and Drug Administration Staff - Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approvals and De Novo Classifications [Internet]Available from: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm267829.htm

2. Committee for Medicinal Products for Human Use (CHMP) E: Report of the CHMP working group on benefit-risk assessment models and methods. http//www.ema Eur Guidel CHMP (last accessed 18 july 2012) , 2008

3. Péron J, Maillet D, Gan HK, et al: Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: a systematic review. J Clin Oncol 31:3957–63, 2013

4. Péron J, Pond GR, Gan HK, et al: Quality of reporting of modern randomized controlled trials in medical oncology: a systematic review. J Natl Cancer Inst 104:982–9, 2012

5. Burris HA, Moore MJ, Andersen J, et al: Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. J Clin Oncol 15:2403–13, 1997

6. Moore MJ, Goldstein D, Hamm J, et al: Erlotinib Plus Gemcitabine Compared With Gemcitabine Alone in Patients With Advanced Pancreatic Cancer : A Phase III Trial of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol 25:1960–1966, 2007

7. Péron J, Roy P, Ding K, et al: Assessing the benefit–risk of new treatments using generalised pairwise comparisons: the case of erlotinib in pancreatic cancer. Br J Cancer (in press), 2015

8. Conroy T, Desseigne F, Ychou M: FOLFIRINOX versus Gemcitabine for Metastatic Pancreatic Cancer. N Engl J Med 364:1817–1825, 2011

9. Faris JE, Blaszkowsky LS, McDermott S, et al: FOLFIRINOX in locally advanced pancreatic cancer: the Massachussetts General Hospital Cancer Center experience. Oncologist 18:543–548, 2013

10. Von Hoff DD, Ervin T, Arena FP, et al: Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. N Engl J Med 369:1691–703, 2013

11. Buyse M: Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med 29:3245–3257, 2010

12. Moser BK, McCann MH: Reformulating the hazard ratio to enhance communication with clinical investigators. Clin Trials 5:248–252, 2008

13. Buyse M: Reformulating the hazard ratio to enhance communication with clinical investigators. Clin Trials 5:641–642, 2008

14. Singhal MK, Kapoor A, Bagri PK, et al: 617PD- A phase III trial comparing FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. Ann Oncol 25:iv210–iv253, 2014

15. Weinstein MC, Torrance G, McGuire A: QALYs: the basics. Value Health 12 Suppl 1:S5–9, 2009

16. Whitehead SJ, Ali S: Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull 96:5–21, 2010

17. Seymour MT, Thompson LC, Wasan HS, et al: Chemotherapy options in elderly and frail patients with metastatic colorectal cancer (MRC FOCUS2): an open-label, randomised factorial trial. Lancet 377:1749–59, 2011

Acknowledgement

Dr Julien Péron is the recipient of a grant from the Nuovo-Soldati Research Foundation.

Appendix

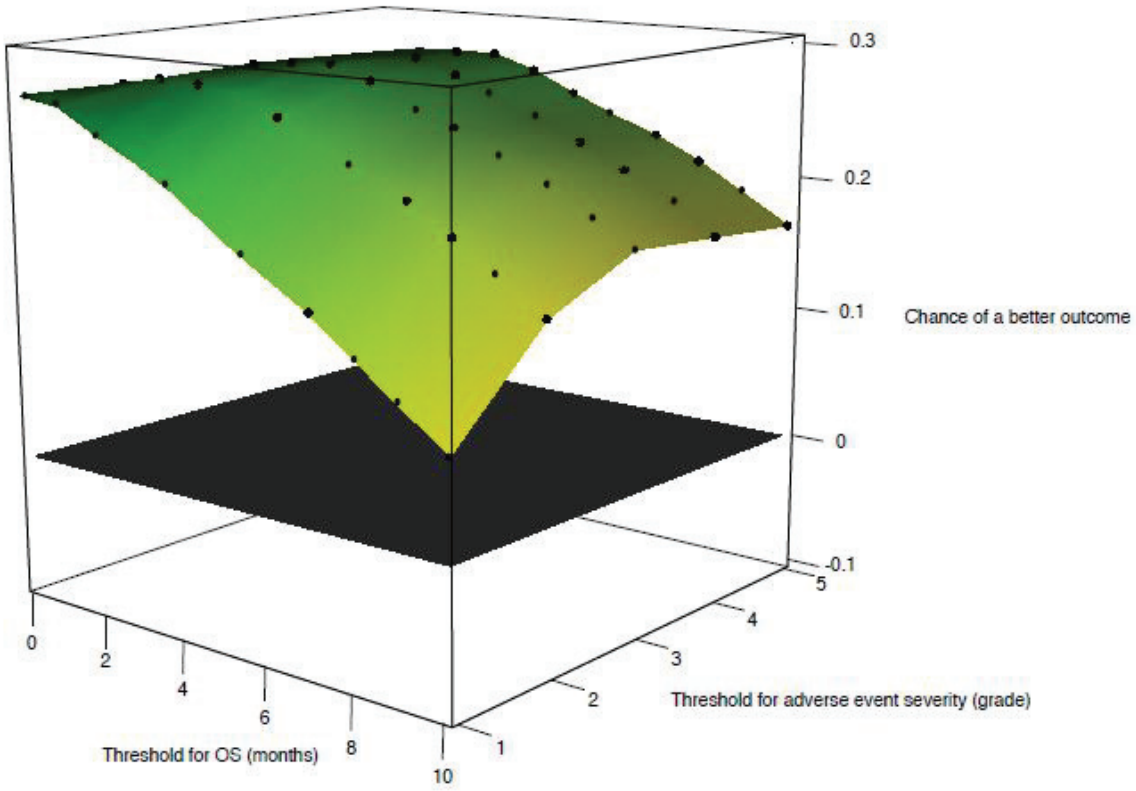


Figure A :

Title: Benefit-risk of FOLFIRINOX according to the survival threshold of clinical significance and to the adverse event severity threshold (in number of adverse event grades).

Legend: Chance of a better overall outcome with FOLFIRINOX according to the minimum survival benefit considered clinically significant and to the minimal difference in adverse event grade considered clinically significant. First priority outcome: Overall survival. Second priority outcome: Worst grade of at least possibly related adverse events. In green, the chance of a better overall outcome is strongly in favor of the FOLFIRINOX. In yellow the chance of a better overall outcome is mildly in favor of the FOLFIRINOX

Authors' contribution :

Julien Péron : Design of the study. Data collection and analysis. Writing and approval of the manuscript. Guarantor.

Pascal Roy : Design of the study. Data analysis. Writing and approval of the manuscript.

Brice Ozenne : Design of the study. Data analysis. Writing and approval of the manuscript.

Laurent Roche : Design of the study. Data analysis. Writing and approval of the manuscript.

Marc Buyse : Design of the study. Data collection and analysis. Writing and approval of the manuscript.

Conflict of interest statement:

The authors have no conflict of interest to disclose

Role of funding source:

The study was not funded. Julien Péron was the recipient of a grant from the Nuovo-Soldati Research Foundation.

Ethics committee approval: not applicable

La conclusion principale de cet article est que la balance bénéfice-risque du FOLFIRINOX est très favorable. En effet la propension globale au succès est estimée entre 20 et 30% dans l'analyse principale et dans l'ensemble des analyses de sensibilité pertinentes réalisées. Cette conclusion encourage à traiter par FOLFIRINOX les patients en première ligne thérapeutique d'un adénocarcinome du pancréas métastatique. Néanmoins la population de l'étude était restreinte aux patients présentant des métastases, en bon état général, et sans comorbidité ni élévation de la bilirubine. Il n'est donc pas possible de généraliser l'évaluation de la balance bénéfice-risque à l'ensemble des patients atteints de cancers du pancréas avancés. Cette méthode d'évaluation de la balance bénéfice-risque pourrait être réalisée selon une procédure similaire dans l'essai évaluant le nab-paclitaxel associé à la gemcitabine [30]. Les deux combinaisons thérapeutiques (FOLFIRINOX et nab-paclitaxel associé à la gemcitabine) devraient idéalement être comparées directement en termes de balance bénéfice-risque au travers d'un essai randomisé. Néanmoins la réalisation d'un tel essai étant peu probable, une comparaison indirecte de la balance bénéfice-risque de ces deux traitements pourrait être réalisée de façon rétrospective. En effet, les essais randomisés de phase III ayant évalué ces combinaisons thérapeutiques avaient utilisé le même bras contrôle. De plus ces deux essais étaient réalisés en première ligne thérapeutique chez des patients avec un état général conservé. Nous espérons donc pouvoir accéder aux données issues de l'essai MPACT ayant évalué le nab-paclitaxel associé à la gemcitabine afin de réaliser cette analyse comparative indirecte.

Chapitre V

V. Utilisation des comparaisons par paire généralisées en alternative aux critères de jugement composites

Les critères de jugement centrés sur le patient reflètent ce qu'il perçoit de son bien-être ou de sa survie, et permettent ainsi une évaluation directe des bénéfices cliniques d'une intervention thérapeutique [6]. De manière générale, en oncologie, ces critères comprennent la qualité de vie relative à la santé et surtout la survie globale, historiquement considérée comme le critère de référence et le plus convaincant en termes d'efficacité. Néanmoins, du fait de l'augmentation du nombre de traitements efficaces et de l'amélioration des soins de support, il devient de plus en plus difficile de démontrer un allongement statistiquement significatif de la survie globale. En effet, l'effet du traitement peut être dilué par les traitements reçus ultérieurement, notamment lorsque les patients du bras contrôle reçoivent le traitement expérimental après une première progression. De plus, une analyse ne prenant en compte que le temps de survie globale apparait comme restrictive. Même une analyse ne couvrant que les bénéfices attendus d'un traitement intègre à la fois des notions de quantité de vie et de qualité de vie. Une amélioration de la qualité de vie des patients est le plus souvent obtenue lorsque le traitement expérimental a un meilleur profil de toxicité que le traitement standard, une meilleure acceptabilité, ou lorsqu'il permet de diminuer ou retarder les symptômes de la maladie.

Des critères composites, comme la survie sans progression ou la survie sans détérioration significative de la qualité de vie, sont utilisés en cancérologie afin de prendre en compte l'effet du traitement sur la survie et son effet sur la progression de la maladie tumorale. Ces critères permettent d'augmenter le nombre d'événements inclus dans le critère de jugement, et ainsi d'augmenter la puissance du test de la différence en survie sans événement, lorsqu'un effet thérapeutique positif est attendu sur l'ensemble des événements. Les critères composites sont mis en défaut lorsque l'effet d'un traitement est hétérogène sur les différents événements inclus dans le critère composite, et lorsque l'importance clinique de ces événements est inégale [7]. Par exemple une différence observée en survie sans progression ne garantit pas une différence en termes de survie globale. Lorsque les décès sont des événements minoritaires, il est même possible d'observer une amélioration de la survie sans progression en faveur d'un traitement malgré une augmentation de la fréquence des décès. Dans l'article

présenté ci-dessous, les comparaisons par paire généralisées sont proposées pour analyser de façon conjointe plusieurs critères de jugement sur lesquels un bénéfice thérapeutique est attendu. L'article prend l'exemple d'une analyse combinée de la survie globale et de l'évolution de la qualité de vie relative à la santé, car il s'agit de deux critères de jugement centrés sur les patients. Néanmoins, il est possible d'utiliser la même méthode pour analyser de façon conjointe la survie globale et la survie sans progression, ou encore la survie globale et la toxicité si celle-ci est attendue moins élevée dans le groupe traitement. Cet article est en cours d'élaboration et n'a pas encore été soumis à une revue scientifique.

Using generalized pairwise comparisons can reduce the sample size in clinical trials when more than one endpoint are improved by the treatment

J. Péron [1,2] ; M. Buyse [3] ; B. Ozenne [1,2] ; L. Roche [1,2] ; E. Pujade-Lauraine [4] ; P. Roy [1,2]

1. Service de biostatistiques, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, F-69310, Pierre-Bénite, France;

2. CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Villeurbanne, France.

3. International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium

4. University Paris Descartes, Assistance Publique-Hôpitaux de Paris, Hôpital Hôtel Dieu, Paris, France.

Introduction:

Major concerns have been raised regarding the choice of endpoints in oncology clinical trials. Because Overall Survival (OS) is an objective, unambiguously defined endpoint with clear clinical meaning, OS is nowadays the main primary endpoint in oncology phase III clinical trials. However, the required length and sample size of studies that use OS as the primary endpoint may be important in order to reach an acceptable power [1]. Many randomized trial designs use composite endpoints, such as progression-free survival [2] or symptom-free survival. Composite endpoints include higher event rates, leading to better power and statistical efficiency. However results obtained with composite endpoints may be misleading if the number of events related to the component of the greatest importance is small [3, 4]. The observed effect of the composite does not necessarily reflect the effects of the most important component. For example, a difference in the time to a composite event (death and another event) in favor of a treatment group does not guarantee a difference in the time to death in favor of the same group; even worse, the time to death may well be globally more favorable in the control group. One way to avoid such concerns is to demonstrate prior that the combined endpoint is a reliable surrogate of OS. However, only few composite endpoints have been validated as surrogates [5].

Another way is to restrict the analysis to endpoints that directly measure the clinical benefit of patients. An endpoint based on the available data on OS and a symptom-centered endpoint (health-related quality of life (QoL), functional, or symptom score) could be useful to provide efficient and reliable information on treatment efficacy, when both OS and the symptom-centered endpoint are thought to be potentially improved by the treatment. The combined analysis of OS and a symptom-centered endpoint is challenging. Some authors proposed the time spent in a health state to be weighted by the utility score given to that health state, and derive an integrated measurement of quantity and quality of life [6]. However the quality adjusted survival is of difficult interpretation, and is then often restricted to cost-efficacy analyses [7]. Symptom-related score and OS could hardly be the two components of a standard composite endpoint (e. g. time from randomization to death or clinically significant deterioration in QoL) , because investigational treatments are likely to have different effect size on the two components of such composite endpoint. Moreover most patients would not give equal value to a fatal issue compared with the occurrence of a symptom, even a severe one.

We propose here a procedure based on generalized pairwise comparisons [8, 9] allowing the calculation and the test of the overall treatment benefit based on two or more prioritized outcomes: for example one variable capturing survival data and one variable related to patients' symptoms. In this manuscript, we describe the principles of the method. Then we report a simulation study, comparing the power of the compound test with traditional tests on survival and symptom score. An illustrative application of the method was performed on the Calypso trial, a GINECO (Groupe d'Investigateurs Nationaux pour l'Etude des Cancers Ovariens) study [10]. Lastly we report sample size calculation based on efficacy hypothesis derived from the calypso trial results.

Methods:

Overview:

We applied generalized pairwise comparisons extended to two outcome measures (one survival outcome and one symptom-derived score). A complete description of generalized pairwise comparisons has been previously published [9]. Briefly, pairwise comparisons require consideration of all pairs of individuals, one taken from the arm of investigational treatment (group T) and the other taken from the control arm (group C). The outcomes of these two individuals are first compared on the first priority survival outcome. The probability that the difference in survival exceed a prespecified threshold of clinical significance (τ_{surv}) is calculated for each pair. This probability is 0, or 1 in the absence of censoring, and could take any value between 0 and 1 when one of the observations is censored. The probability that the difference in survival is neutral (ranges between $-\tau_{\mathrm{surv}}$ and $+\tau_{\mathrm{surv}}$) is also calculated for each pair, and is used to weight the analysis of the second priority symptom-centered endpoint. Whenever there is a non-null probability that a pairwise comparison is neutral on survival, the second priority outcome (symptom-centered endpoint) is analyzed (Table 1). Again, the difference in symptom-centered endpoint should exceed a prespecified threshold of clinical significance (τ_{symptom}) to be considered informative. An overall score is calculated for each pair, ranging between -1 and 1. A score of -1 meaning that the outcomes of the patient from group T are decidedly inferior compared with the outcomes of the patient from group C. A score of 0 meaning that the outcomes of the two patients are equivalent. A score of 1 meaning that the outcomes of the patient from group T are decidedly better than the outcomes of the patient in the group C.

Let's call "chance of a better overall outcome" the net difference between the mean probability for a random pair to be favorable to the group T minus the mean probability for a random pair to be favorable to the group C. The chance of a better overall outcome reflects the overall effect of the treatment. It has an intuitive interpretation, since it reflects the probability for a random patient from group T to perform better than a random patient from group C minus the opposite probability, depending on the definition of a better outcome by investigators (Table 2).

Survival	Symptom-related score	Pair is
favorable	ignored	favorable
unfavorable	ignored	unfavorable
neutral	favorable	favorable
neutral	unfavorable	unfavorable
neutral	neutral	neutral

Table 1: Generalized pairwise comparisons for two priorities in the absence of censoring

Priority	Outcomes	Threshold	Chance of a better outcome
1	Survival	τ_{surv}	Chance of a better survival
2	Symptom-centered endpoint	τ_{symptom}	Chance of a better overall outcome

Legend: τ_{surv} = Survival threshold of clinical significancy ; τ_{symptom} = Symptom-related score threshold of clinical significancy

Table 2: Definition of priorities

Simulation study

Performances of the above method were compared with the isolated analyses of survival or symptom-centered endpoint through simulation of randomized trials datasets, in which survival and/or symptom-related scores were improved by the investigational treatment. Two hundreds patients per trial, equally assigned between two treatment groups, were supposed to be uniformly included over a period of 10 months. Times to death in the control group were generated using an exponential distribution corresponding to a median survival of 10 months. One symptom-centered score was generated by patient, using a normal distribution with a standard deviation of 25 [11, 12]. The mean symptom-centered score was 0

in the group C. The investigational treatment was supposed to have independent effects on survival and symptom-centered score. Nine scenarios were simulated, with various proportional hazard ratio (HR) values for survival (HR=0.6, 0.7, and 1.0) and three mean symptom-centered score in group T (+7, +10, and +15). The analyses were repeated with various censoring rates for survival. For each scenario, 1000 independent trial datasets were generated.

For each trial, the Log-Rank test for survival was performed. The magnitude of response in symptom-centered score considered to be clinically significant was 10 [7,13]. The proportion of patients per arm who reported "improved", "stable" and "worsened" symptom-centered scores was calculated, and a Fisher test was performed to test for statistically significant differences between the three categories of response. Then the compound test, using generalized pairwise comparisons, was performed. Various values were chosen for τ_{surv} ranging from 0 to 4 months. τ_{symptom} was set at 10. For each scenario and for each censoring rate, the power of a test was the proportion of trials resulting in rejection of the null hypothesis (lack of treatment effect). At the 95% confidence level, it was the proportion of tests which yielded a p-value of .05 or less.

Illustration on the Calypso Trial

From April 2005 to September 2007, a total of 976 patients with histologically proven ovarian cancer with recurrence more than 6 months after first- or second-line platinum and taxane-based therapies were randomly assigned to carboplatin plus pegylated liposomal doxorubicin (CD) or carboplatin plus paclitaxel (CP) for at least 6 cycles [10]. Primary end point was progression-free survival (PFS); quality of life (QoL) was a secondary endpoint [14]. Survival distributions between arms were compared with the Log-Rank test. Global health status score evolution between 0 and 3 months was categorized in three classes as described above, and a Fisher test was performed to test for statistically significant differences between the three categories of response. The median follow-up was 52 months and ended 83 months after the inclusion of the first patient. In the generalized pairwise comparison procedure, the first priority outcome was PFS and the second priority outcome was the global health status evolution between 0 and 3 months. The compound test was performed with values of τ_{PFS} ranging from 0 to 4 months and τ_{QoL} = 10. Using inclusion dates, date of

QOL questionnaires completion, and dates of death or progression, a sensitivity analysis was performed to explore the results that would have been observed after shorter follow-up times (increased censoring rates).

Sample size calculation based on Calypso trial results

Using the median PFS observed in the calypso trial (9.4 months and 11.5 months in group CP and CD respectively), the mean evolution in global Health Status/QoL score between baseline and 3 months (2.6 points in group CD and -2.2 in group CP), and the standard deviation for the global Health Status/QoL score evolution (25 points), we calculated the sample size necessary to reproduce the same trial with power of 80% and a two-sided alpha risk of 5%. Patients were to be included uniformly during 30 months, and the time of the analysis was to be performed 6 months after the last inclusion (expected censoring rate on PFS = 30%). Survival time followed an exponential distribution, and the hazards were proportional. The global Health Status/QoL score evolution missing rate was set at 20%. The allocation ratio was to be 1:1. The sample size calculation was based on several statistical tests: generalized pairwise comparison on PFS alone with a threshold of minimal significant difference set at 2 months; generalized pairwise comparison on QOL alone with a threshold of minimal significant difference set at 10 points; generalized pairwise comparison on PFS as a first priority outcome and on QOL as a second priority outcome; Fisher test on QOL; and Log-Rank test on PFS.

Software

All statistical analyses were performed using the R software [15]. The generalized pairwise comparisons were performed using the BuyseTest package. Sample size calculations were performed using the BuysePower function of the BuyseTest package.

Results:

Simulation study

In a first time, simulated trials were analyzed after 10 months of follow-up. Censoring rate for survival ranged between 42% and 46% depending on the hazard ratio. Powers of the several tests (Log-Rank tests on survival, Fisher test on symptom-related response score and compound tests) were calculated for all the 9 scenarios. When τ_{surv} = 0 the compound test gave a null weight to the second priority symptom-derived outcome. The compound test

was more powerful than the log-rank test based on survival alone. The gain in power increased with the benefit in symptom-related outcome. Then, analyzes were focused on the scenario with mean hazard ratio for survival of 0.7 and mean symptom-derived score response of +10 in group T (Figure 1). The tests performances were evaluated as a function of the follow-up time. The power of the tests based on survival increased progressively while the censoring rate decreased. The power of the compound tests was consistent for any censoring rate. The power gain over the Log-Rank test was stronger when the censoring rate was high (figure 1).

	Tests on survival	Test on symptom score	Compound tests		
Scenario	Log-rank	Fisher	$\tau_{surv} = 0$	$\tau_{surv} = 2$	$\tau_{surv} = 4$
HR = 0.6 ; symptom = +7	74.3	29.8	77.7	84.4	88.2
HR = 0.6 ; symptom = +10	74.3	58.1	81.0	89.9	93.3
HR = 0.6 ; symptom = +15	74.3	92.3	86.6	95.8	97.8
HR = 0.7 ; symptom = +7	46.4	29.8	51.5	62.1	66.3
HR = 0.7 ; symptom = +10	46.4	58.1	57.0	71.3	77.5
HR = 0.7 ; symptom = +15	46.4	92.3	64.6	83.8	91.4
HR = 1 ; symptom = +7	5.4*	29.8	5.2	8.5	11.7
HR = 1 ; symptom = +10	5.4*	58.1	6.1	11.8	19.2
HR = 1 ; symptom = +15	5.4*	92.3	7.5	19.7	34.4

Legend: HR = Hazard ratio for OS ; symptom = mean improvement in symptom-related score ; *the proportion of tests rejecting the null hypothesis is the risk α ; τ_{surv} = Survival threshold of clinical significancy

Table 3. Power of several tests in various scenarios of survival and quality of life benefit magnitude.

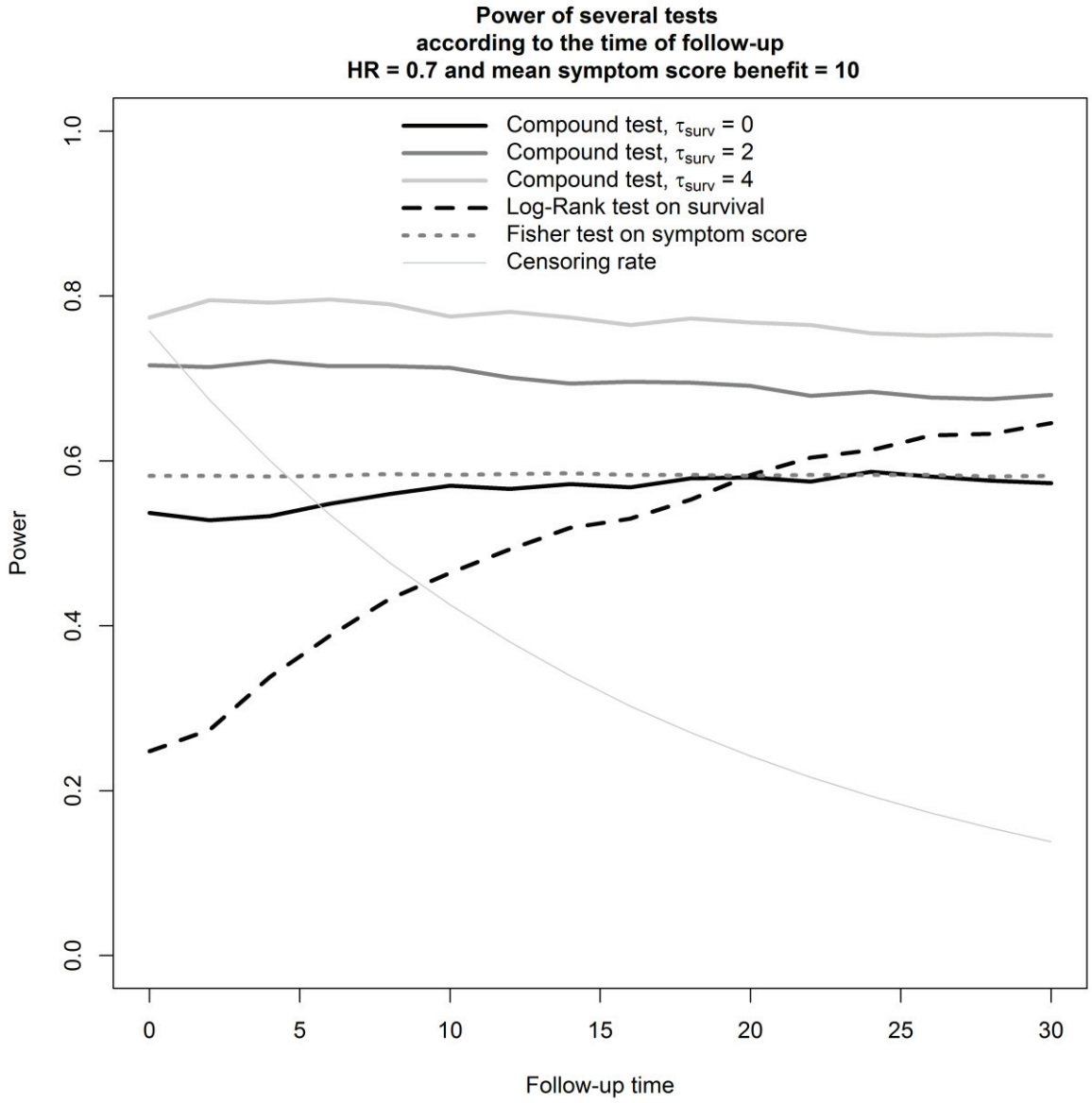


Figure 1: Powers of tests based on survival alone (Log-Rank test, compound test with $\tau = 0$), quality of life alone (Fisher test) , or on both survival and quality of life (compound test with $\tau > 0$).

Compound analysis of progression-free survival and quality of life in the Calypso Trial

At the mature analysis of data from the Calypso trial (median follow-up = 52 months), treatment with CD was associated with a significantly better PFS (Log-Rank test p-value = 0.016) and a nearly significant better global Health Status/QoL (Fisher test p-value = 0.070). When analyses were realized earlier during the follow-up, the number of included patients was lower, and the rate of missing or censored values for PFS and QoL was higher. It might have decreased the power of the tests on PFS and QoL (Table 5). For example, 12 months after the start of inclusions, the number of included patients was 209 (21%). Using pairwise

comparisons, the chance of a better outcome estimated on both PFS and QoL was consistently higher than the chance of a better outcome estimated on PFS only. The estimation of the chance of a better outcome varied only slightly even when the analysis date was as early as 12 months after the start of inclusions. The P-value associated with the compound analysis was also consistently lower than the P-value of tests based on the PFS alone or based on the QoL alone. (Table 5).

Test	Date of analysis							
	12 months (n=209)				18 months (n=451)			
	CD > CP*	CD < CP*	Chance of a better outcome	P	CD > CP*	CD < CP*	Chance of a better outcome	P
GPC								
- PFS alone (τ_{PFS}=2 months)	48.6	31.9	16.7	0.030	46.9	33.4	13.5	0.010
- QoL alone (τ_{QoL}=10 points)	6.9	3.2	3.7	0.031	11.1	6.6	4.5	0.011
- PFS and QoL (τ_{PFS}=2 months, τ_{QoL}=10 points)	50.1	32.4	17.7	0.021	49.3	34.9	14.4	0.0068
Log-Rank test on PFS	—	—	—	0.23	—	—	—	0.099
Fisher test on QoL	—	—	—	0.17	—	—	—	0.015
	24 months (n=778)				Final analysis (n=976)			
	CD > CP*	CD < CP*	Chance of a better outcome	P	CD > CP*	CD < CP*	Chance of a better outcome	P
GPC								
- PFS alone (τ_{PFS}=2 months)	46.4	34.2	12.2	0.0031	46.5	33.7	12.8	<0.001
- QoL alone (τ_{QoL}=10 points)	13.8	9.5	4.3	0.0095	22.6	18.7	3.9	0.094
- PFS and QoL (τ_{PFS}=2 months, τ_{QoL}=10 points)	49.3	36.2	13.1	0.0018	51.3	37.5	13.8	<0.001
Log-Rank test on PFS	—	—	—	0.037	—	—	—	0.016
Fisher test on QoL	—	—	—	0.006	—	—	—	0.070

Legend: GPC = Generalized pairwise comparison ; PFS = Progression Free Survival ; CD = carboplatin plus pegylated liposomal doxorubicin ; CP = carboplatin plus paclitaxel ; QoL = Quality of Life ; Cumulative Δ = cumulative chance of a better outcome; τ_{PFS} = Progression-free survival threshold of clinical significance ; τ_{QoL} = Quality of life score threshold of clinical significance. * CD > CP is the probability that a patient in the group carboplatin plus pegylated liposomal doxorubicin has a better outcome than a patient in the group carboplatin plus paclitaxel. CD < CP is the opposite probability.

Tables 4: Testing the treatment effect based on progression-free survival and global Health Status/QoL score at several times of follow-up in the CALYPSO trial.

Sample size calculation for a new trial identical to Calypso trial

The number of patients needed to reproduce the Calypso trial with a power of 80% and a two-sided alpha risk of 5% was 1110 if the primary test was the log-rank test on PFS. When PFS was to be compared between groups using a generalized pairwise comparison procedure, the estimated sample size was 1160 patients. This number dropped to 840 patients when both PFS and QoL score were included in the generalized pairwise comparison procedure.

Test	Sample size
GPC based on PFS alone (τ_{PFS}=2 months)	N=1160
GPC based on QoL alone (τ_{QoL}=10 points)	N=1130
GPC based on PFS and QoL (τ_{PFS}=2 months ; τ_{QoL}=10 points)	N=840
Log-Rank test on PFS	N=1110
Fisher test on QoL	N=1660

Legend: GPC = Generalized pairwise comparison ; PFS = Progression Free Survival; QoL = Quality of Life

Table 5 : Sample size calculation using the parameters of the Calypso trial, based on progression-free survival alone, global Health Status/QoL score alone, or both outcomes in one compound test

Discussion:

Pairwise comparisons, generalized to several successive outcomes, allow an overall assessment of treatment benefits in randomized clinical trials. The several outcome variables should capture different natures of treatment benefit. An appropriate assessment of several endpoints should take into account their respective importance, the magnitude of the treatment effect on each endpoint, and the correlation structure of the several treatment effects.

Composite endpoints are often used to reduce sample size and to capture the overall impact of therapeutic interventions. The use of composite endpoints may mislead if the number of events in the component of greater relevance is small, and if the effect differs in magnitude or in direction across components [16]. For example, a treatment with negative effect on survival but with positive effect on symptoms is likely to be considered positive by the use of

composite endpoint (e.g. time from randomization to death or clinically significant deterioration in symptom-related score), while most investigators and patients would assign a higher weight to the survival loss.

One advantage of the generalized pairwise comparisons procedure over composite endpoints is the possibility to give higher priorities to the most important endpoint (such as overall survival) over less relevant endpoint (such as length of hospitalization, or temporary adverse events). Only if there is no clinically significant difference on the outcome of first priority, the analysis of treatment effect is to be performed on the second priority outcome. The relative weight of each endpoint in the overall assessment of the treatment effect depends on the thresholds of clinical significance. Generalized pairwise comparisons belong to the general class of multi-criteria decision analyses [17]. The main difference with other methods previously proposed [18, 19] is that the weighting process of the several endpoints is closer to the intuitive process made by patients, clinicians or decision makers. Also the estimate of the chance of a better overall outcome has a strong clinical meaning. It reflects the probability for a random patient from the group T to have a better overall outcome than a random patient from the group C, minus the opposite probability.

Recommendations on randomized clinical trial reporting distinguish the reporting of benefit and the reporting of harms of investigational treatments [20]. Treatment effect might therefore be reported on two sections: a comprehensive analysis of all the relevant positive effects; and a clear and concise analysis of the safety concerns. However, unfavorable endpoint, such as adverse events rate, could also be included in an overall analysis of all the treatment effects when it is deemed necessary.

It is possible to estimate the chance of a better outcome for each outcome included in a generalized pairwise comparison procedure. Hence, the participation of each outcome to the overall treatment effect is easy to report. For each prioritized outcome, it is possible to test the null hypothesis ($\mathcal{H}_0$: chance of a better overall outcome$= 0$) with a randomization test [9].

Generalized pairwise comparison intends to estimate the overall probability that a random patient chosen in arm T has a better overall outcome that a random patient chosen in arm C, according to all clinically meaningful endpoints. For that reason, only endpoints directly

related to patient clinical status should be included. For many symptom-related scales, a minimal perceptible difference can be specified. For example, it has been reported that the smallest change perceptible by the patients for the QLQ-C30 global QOL scale was 10, and that change of more than 20 were considered to be large [7]. A pair should therefore be classified as ”favorable” or “defavorable” only if the difference between the two QoL score evolutions from baseline exceeds the smallest change known to be clinically significant. Throughout the manuscript, the symptom-derived score is a continuous variable, while the case can easily be extended to binary outcomes or to multiple time-to-event outcomes.

When the treatment effect is assessed through one single survival outcome with a threshold of clinical significance set at zero, the test of the null hypothesis is equivalent to a Efron’s generalized Wilcoxon test [21]. This test is known to be less powerful than the log-rank test under the assumption of proportional hazards [22], but more powerful than the log-rank in the case of early survival differences [23]. The test of the chance of a better overall outcome is sensitive to the ordering of the outcomes and to the setting of the significance thresholds. These choices should then be specified a priori, and analyses are strongly recommended.

Prioritized pairwise comparisons could be considered as an available method capable to catch the overall treatment effect based on several prioritized outcomes. It might reduce the sample size of randomized trials, when the treatment is supposed to have a positive effect on several outcomes. Results of this approach are easy to interpret when adequately reported. It reflects the overall probability to induce a clinically meaningful benefit.

References

1. Saad ED, Buyse M: Overall survival: patient outcome, therapeutic objective, clinical trial end point, or public health measure? J Clin Oncol 30:1750–1754, 2012

2. Kleist P: Composite Endpoints for Clinical Trials. Int J Pharm Med 21:187–198, 2007

3. Ferreira-González I, Busse JW, Heels-Ansdell D, et al: Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 334:786, 2007

4. Freemantle N, Calvert M, Wood J, et al: Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 289:2554–9, 2003

5. Burzykowski G. Buyse, M. TM: The Evaluation of Surrogate Endpoints. 2005

6. Gelber RD, Gelman RS, Goldhirsch A: A quality-of-life-oriented endpoint for comparing therapies. Biometrics 45:781–95, 1989

7. Osoba D, Bezjak A, Brundage M, et al: Analysis and interpretation of health-related quality-of-life data from clinical trials: basic approach of The National Cancer Institute of Canada Clinical Trials Group. Eur J Cancer 41:280–7, 2005

8. Péron J, Roy P, Ding K, et al: Assessing the benefit-risk of new treatments using generalised pairwise comparisons: the case of erlotinib in pancreatic cancer. Br J Cancer , 2015

9. Buyse M: Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med 29:3245–3257, 2010

10. Pujade-Lauraine E, Wagner U, Aavall-Lundqvist E, et al: Pegylated liposomal Doxorubicin and Carboplatin compared with Paclitaxel and Carboplatin for patients with platinum-sensitive ovarian cancer in late relapse. J Clin Oncol 28:3323–3329, 2010

11. De Marinis F, Pereira JR, Fossella F, et al: Lung Cancer Symptom Scale outcomes in relation to standard efficacy measures: an analysis of the phase III study of pemetrexed versus docetaxel in advanced non-small cell lung cancer. J Thorac Oncol 3:30–6, 2008

12. Farivar SS, Liu H, Hays RD: Half standard deviation estimate of the minimally important difference in HRQOL scores? Expert Rev Pharmacoecon Outcomes Res 4:515–23, 2004

13. Sloan JA, Frost MH, Berzon R, et al: The clinical significance of quality of life assessments in oncology: a summary for clinicians. Support Care Cancer 14:988–98, 2006

14. Brundage M, Gropp M, Mefti F, et al: Health-related quality of life in recurrent platinum-sensitive ovarian cancer--results from the CALYPSO trial. Ann Oncol 23:2020–7, 2012

15. R CoreTeam: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., 2014

16. Ferreira-González I, Permanyer-Miralda G, Busse JW, et al: Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol 60:651–7; discussion 658–62, 2007

17. Dodgson J, Spackman M, Pearman A, et al: Multi-criteria analysis: a manual. Department of Transport, Local Government and the Regions, London, 2009

18. Felli JC, Noel RA, Cavazzoni PA: A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives. Med Decis Making 29:104–15

19. Chuang-Stein C: A new proposal for benefit-less-risk analysis in clinical trials. Control Clin Trials 15:30–43, 1994

20. Ioannidis JPA, Evans SJW, Gøtzsche PC, et al: Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 141:781–8, 2004

21. Efron B: The two sample problem with censored data. Fifth Berkeley Symp Math Stat Prob Division o, 1967

22. Letón E, Zuluaga P: Relationships among tests for censored data. Biom J 47:377–87, 2005

23. Lee ET, Desu MM, Gehan EA: A Monte Carlo Study of the Power of Some Two-Sample Tests. Biometrika 62:425–432, 1975

La conclusion principale de cet article est que la propension au succès peut être utilisée pour évaluer conjointement plusieurs critères de jugement sur lesquels un effet thérapeutique positif est attendu. Lorsque les effets thérapeutiques sur les différents critères de jugement ne sont pas complétement corrélés, la puissance du test de l'hypothèse nulle basée sur la propension globale au succès est alors plus élevée que la puissance des tests basés sur chacun des critères de jugement isolément. Cette conclusion est assez logique, mais a pu être démontrée ici par des études de simulation d'essais randomisés, et en utilisant les données d'un essai randomisé dans le carcinome ovarien, l'essai CALYPSO [32]. Afin d'évaluer l'impact qu'aurait eu le choix d'une évaluation conjointe de la survie sans progression et de la qualité de vie lors de la mise en place de l'essai CALYPSO, un calcul du nombre de sujets nécessaires pour réaliser le même essai une deuxième fois a été réalisé. Ce calcul du nombre de sujets nécessaires a utilisé les paramètres observés dans l'essai CALYPSO comme hypothèses. Le nombre de sujets à inclure permettant de rejeter l'hypothèse nulle avec une puissance de 80% et un risque α bilatéral de 5% a été calculé pour plusieurs tests. Huit cent quarante patients devraient être inclus si le test utilisé est la propension globale au succès. Le nombre de sujet à inclure calculé est nettement plus élevé pour tous les autres tests étudiés basés sur la survie sans progression ou sur la qualité de vie isolément. Les comparaisons par paire généralisées peuvent donc être utilisées afin de limiter le nombre de sujets à inclure dans les essais randomisés lorsqu'un effet thérapeutique positif est attendu sur plusieurs critères de jugement pertinents.

Chapitre VI

VI. Conclusions et perspectives

VI.1. Conclusions

Les essais contrôlés randomisés sont une source d'information majeure sur l'effet des traitements. Il s'agit d'une des sources d'information principale des agences d'évaluation et d'enregistrement des médicaments, et parfois de la seule source d'information pour un médecin ou un patient évaluant le bénéfice et le risque attendu d'un nouveau médicament. En oncologie médicale, l'effet des traitements est le plus souvent évalué sur de nombreux critères de jugement. Les plus souvent évalués sont les critères en lien avec le volume tumoral et le temps de contrôle tumoral, la survie globale, les symptômes, la qualité de vie relative à la santé, et les événements indésirables liés aux traitements. Les chapitres I et II de cette thèse rapportent une revue systématique des méthodes actuellement utilisées pour analyser et rapporter les événements indésirables et les critères de jugement rapportés par les patients dans les essais de phase III récemment publiés en oncologie médicale. Une hétérogénéité importante des méthodes de recueil et d'analyse des données a été mise en évidence. De plus de nombreux manuscrits ne contenaient pas certains éléments essentiels à la compréhension et à l'évaluation par le lecteur des méthodes utilisées. La grande majorité des essais inclus dans la revue utilisait comme critère de jugement principal soit un critère objectif comme la survie globale, soit un critère évalué par les cliniciens comme les critères centrés sur la tumeur. Les critères de jugement rapportés par les patients et les événements indésirables étaient rapportés de façon parallèle. Des analyses quantitatives combinant ces différents types de critère de jugement n'ont été que rarement rapportées. La balance entre les différents effets thérapeutiques était par contre souvent discutée de façon qualitative dans les manuscrits. Une utilisation plus large des méthodes d'analyse intégratives de ces différents critères de jugement pourrait permettre d'améliorer l'évaluation de l'effet global des traitements.

La méthode des comparaisons par paire généralisées permet d'analyser simultanément plusieurs critères de jugement à condition qu'ils soient hiérarchisés en priorités successives. Elle permet d'estimer la propension globale au succès, reflet de la probabilité pour un patient traité dans le bras expérimental d'avoir un meilleur résultat thérapeutique global qu'un patient traité dans le groupe contrôle. La première priorité étant attribuée au critère de jugement le

plus important. Un seuil de bénéfice minimal cliniquement significatif est attribué à chaque priorité lorsque les variables évaluant les critères de jugement sont de type continu ou de type temps jusqu'à événement. Ce procédé permet ainsi de moduler le poids accordé à chaque variable de façon cliniquement pertinente. Dans le troisième chapitre de cette thèse, la méthode des comparaisons par paire a été étendue afin de réaliser une estimation non biaisée de la propension au succès lors de l'analyse des critères de type temps jusqu'à événement.

Dans les quatrième et cinquième chapitres, deux utilisations de la propension globale au succès sont illustrées. Lorsqu'à la fois des critères d'efficacité thérapeutique et des critères de toxicité sont intégrés dans l'analyse, la propension au succès permet d'évaluer la balance bénéfice-risque des traitements. A l'inverse, lorsqu'un effet thérapeutique positif est attendu sur plusieurs critères de jugement non complètement corrélés, la propension au succès permet d'évaluer globalement les bénéfices thérapeutiques. Dans ce cas, l'analyse conjointe des effets thérapeutiques permet d'augmenter la puissance du test de l'hypothèse nulle par rapport aux analyses individuelles de chacun des critères de jugement.

La méthode a été implémentée dans le cadre de cette thèse dans le logiciel R [33]. Le package BuyseTest inclut la procédure standard décrite par Marc Buyse [1], ainsi que les différentes extensions permettant d'analyser des variables de type temps jusqu'à événement décrites dans le chapitre III. L'analyse peut inclure des variables de stratification. Elle fournit en résultat une estimation de la propension au succès pour chaque critère hiérarchisé en priorités successives, et la propension globale au succès. L'intervalle de confiance de chacun de ces paramètres est estimé par une méthode de permutation. L'hypothèse nulle est testée par permutation. Le guide utilisateur du package BuyseTest est inclus en annexe de cette thèse.

VI.2. Perspectives

La méthode des comparaisons par paire généralisées permet donc d'analyser de façon simultanée et cliniquement pertinente plusieurs critères de jugement. D'autres méthodes ont été proposées pour réaliser une analyse conjointe de plusieurs critères de jugement. La méthode Q-TWiST (*quality-adjusted Time Without Symptoms of disease progression or Toxicity of treatment*) permet de réaliser une évaluation conjointe de la survie globale, de la survie sans maladie symptomatique et de la toxicité des traitements. Cette méthode a été décrite dans le chapitre IV de cette thèse. Les deux méthodes ont comme objectif commun de réaliser une évaluation de la balance bénéfice-risque des traitements. La méthode de pondération des différents critères de jugement étant très différente, une comparaison directe

de ces deux méthodes sur des données réelles d'essais randomisés ainsi que sur des données simulées pourra être réalisée. Cela permettra d'identifier leurs similitudes, et leurs avantages respectifs. La méthode des comparaisons par paire généralisées est plus souple que la méthode Q-TWiST dans le sens où elle permet d'analyser autant de critères de jugement que nécessaire, et ceci quelles que soient leurs natures (binaires, continus ou de type temps jusqu'à événement).

Un autre champ d'utilisation potentiel des comparaisons par paire est l'évaluation médico-économique comparative de deux traitements. En effet l'estimation du coût des traitements peut être incluse dans une procédure de comparaison par paire. La propension globale au succès serait alors le reflet de la probabilité qu'un patient pris au hasard dans le groupe de traitement expérimental ait une meilleure balance coût-efficacité qu'un patient pris au hasard dans le groupe contrôle.

Une limite actuelle de la méthode est le temps de calcul nécessaire du fait de l'utilisation d'un test de permutation afin de calculer l'intervalle de confiance et le niveau de significativité associé à la propension au succès. L'utilisation de la loi des grands nombres pour construire l'intervalle de confiance et le test de significativité statistique de la propension au succès a été proposée [40]. Cette méthode permet de réduire considérablement la quantité de calcul en comparaison avec la méthode de permutation. Il sera nécessaire de confirmer les performances de cette méthode dans le contexte des extensions proposées permettant de prendre en compte les données censurées.

L'estimation de la propension au succès étant sensible à la définition faite par les investigateurs du succès thérapeutique, la procédure d'analyse devrait être définie a priori et incluse dans le plan d'analyses statistiques des essais randomisés. Une procédure principale d'analyse devra être définie pour chaque essai, ainsi que la nature des analyses de sensibilité programmées.

Bibliographie

[1] M. Buyse, “Generalized pairwise comparisons of prioritized outcomes in the two-sample problem” *Stat Med*, vol. 29, no. 30, pp. 3245–3257, 2010.

[2] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations.” *J. Amer. Stat. Assn.*, vol. 53, pp. 457–481, 1958.

[3] U. S. Food and Drugs Administration, “PDUFA Reauthorization performance goals and procedures fiscal years 2013 through 2017” *(last accessed 18 july 2015)*, 2011.

[4] U. S. Food and Drugs Administration, “Guidance for Industry Use in Medical Product Development to Support Labeling Claims Guidance for Industry” no. December, 2009.

[5] European Committee for Medicinal Products for Human Use (CHMP), “Report of the CHMP working group on benefit-risk assessment models and methods.” *(last accessed 25 june 2015)*, 2008.

[6] F. Fiteni, V. Westeel, X. Pivot, C. Borg, D. Vernerey, and F. Bonnetain, “Endpoints in cancer clinical trials.” *J. Visc. Surg.*, vol. 151, no. 1, pp. 17–22, Feb. 2014.

[7] P. Kleist, “Composite Endpoints for Clinical Trials” *Int. J. Pharm. Med.*, vol. 21, no. 3, pp. 187–198, 2007.

[8] J. Péron, D. Maillet, H. K. Gan, E. X. Chen, and B. You, “Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: a systematic review.” *J. Clin. Oncol.*, vol. 31, no. 31, pp. 3957–63, Nov. 2013.

[9] O. Bylicki, H. K. Gan, F. Joly, D. Maillet, B. You, and J. Péron, “Poor patient-reported outcomes reporting according to CONSORT guidelines in randomized clinical trials evaluating systemic cancer therapy.” *Ann. Oncol.*, vol. 26, no. 1, pp. 231–7, Jan. 2015.

[10] H. B. Mann and D. R. Whitney, “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other” *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[11] E. S. O. Edgington P., “Randomization Tests” *Chapman Hall/CRC New York*, 2007.

[12] E. A. Gehan, “A generalized two-sample Wilcoxon test for doubly censored data” *Biometrika*, vol. 52, no. 3, pp. 650–653, 1965.

[13] R. Peto and J. Peto, “Asymptotically efficient rank invariant test procedure” *J. R. Stat. Soc. A*, vol. 135, no. 2, pp. 185–198, 1972.

[14] B. Efron, “The two sample problem with censored data.” *Fifth Berkeley Symp. Math. Stat. Prob.*, vol. Division o, 1967.

[15] R. B. Latta, “A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data” Mar. 1981.

[16] F. E. Harrell, “Evaluating the Yield of Medical Tests” *JAMA J. Am. Med. Assoc.*, vol. 247, no. 18, pp. 2543-2546, May 1982.

[17] J. A. Koziol and Z. Jia, “The concordance index C and the Mann-Whitney parameter Pr(X>Y) with randomly censored data.” *Biom. J.*, vol. 51, no. 3, pp. 467–74, Jun. 2009.

[18] R. H. Somers, “A new asymmetric measure of association for ordinal variables.” American Sociological Review Vol. 27, No. 6, pp. 799-811, 1962

[19] M. Gönen and G. Heller, “Concordance probability and discriminatory power in proportional hazards regression” *Biometrika*, vol. 92, no. 4, pp. 965–970, Dec. 2005.

[20] L. Acion, J. J. Peterson, S. Temple, and S. Arndt, “Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects” *Stat Med*, vol. 25, no. 4, pp. 591–602, 2006.

[21] S. J. Pocock, C. A. Ariti, T. J. Collier, and D. Wang, “The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities.” *Eur. Heart J.*, vol. 33, no. 2, pp. 176–82, Jan. 2012.

[22] EMA Benefit-Risk Methodology Project Team, “Benefit-risk methodology project Work package 1 report: description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network” 2011.

[23] J. J. Guo, S. Pandey, J. Doyle, B. Bian, Y. Lis, and D. W. Raisch, “A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group” *Value Health*, vol. 13, no. 5, pp. 657–66, Aug. 2010.

[24] M. T. Seymour, L. C. Thompson, H. S. Wasan, G. Middleton, A. E. Brewster, S. F. Shepherd, M. S. O’Mahony, T. S. Maughan, M. Parmar, and R. E. Langley, “Chemotherapy options in elderly and frail patients with metastatic colorectal cancer (MRC FOCUS2): an open-label, randomised factorial trial.” *Lancet*, vol. 377, no. 9779, pp. 1749–59, May 2011.

[25] A. Laupacis, D. L. Sackett, and R. S. Roberts, “An assessment of clinically useful measures of the consequences of treatment.” *N. Engl. J. Med.*, vol. 318, no. 26, pp. 1728–33, Jun. 1988.

[26] J. Dodgson, M. Spackman, A. Pearman, and L. Phillips, “Multi-criteria analysis: a manual.” Department for Communities and Local Government: London, 22-Jul-2009.

[27] J. C. Felli, R. A. Noel, and P. A. Cavazzoni, “A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives.” *Med. Decis. Making*, vol. 29, no. 1, pp. 104–15, Jan. 2009.

[28] R. D. Gelber, R. S. Gelman, and A. Goldhirsch, “A quality-of-life-oriented endpoint for comparing therapies.” *Biometrics*, vol. 45, no. 3, pp. 781–95, Sep. 1989.

[29] R. D. Gelber, A. Goldhirsch, and B. F. Cole, “Evaluation of effectiveness: Q-TWiST. The International Breast Cancer Study Group.” *Cancer Treat. Rev.*, vol. 19 Suppl A, pp. 73–84, Jan. 1993.

[30] G. Duru, J. P. Auray, A. Béresniak, M. Lamure, A. Paine, and N. Nicoloyannis, “Limitations of the methods used for calculating quality-adjusted life-year values.” *Pharmacoeconomics*, vol. 20, no. 7, pp. 463–73, Jan. 2002.

[31] W. R. Tate and G. H. Skrepnek, “Quality-adjusted time without symptoms or toxicity (Q-TWiST): patient-reported outcome or mathematical model? A systematic review in cancer.,” *Psychooncology.*, vol. 24, no. 3, pp. 253–61, Mar. 2015.

[32] C. Chuang-Stein, “A new proposal for benefit-less-risk analysis in clinical trials.” *Control. Clin. Trials*, vol. 15, no. 1, pp. 30–43, Feb. 1994.

[33] J. Péron, P. Roy, K. Ding, W. R. Parulekar, L. Roche, and M. Buyse, “Assessing the benefit-risk of new treatments using generalised pairwise comparisons: the case of erlotinib in pancreatic cancer.” *Br. J. Cancer*, Feb. 2015.

[34] T. Conroy, F. Desseigne, M. Ychou, O. Bouché, R. Guimbaud, Y. Bécouarn, A. Adenis, J.-L. Raoul, S. Gourgou-Bourgade, C. de la Fouchardière, J. Bennouna, J.-B. Bachet, F. Khemissa-Akouz, D. Péré-Vergé, C. Delbaldo, E. Assenat, B. Chauffert, P. Michel, C. Montoto-Grillot, and M. Ducreux, “FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer.” *N. Engl. J. Med.*, vol. 364, no. 19, pp. 1817–25, May 2011.

[35] J. E. Faris, L. S. Blaszkowsky, S. McDermott, A. R. Guimaraes, J. Szymonifka, M. A. Huynh, C. R. Ferrone, J. A. Wargo, J. N. Allen, L. E. Dias, E. L. Kwak, K. D. Lillemoe, S. P. Thayer, J. E. Murphy, A. X. Zhu, D. V. Sahani, J. Y. Wo, and J. W. Clark, “FOLFIRINOX in locally advanced pancreatic cancer: the Massachussetts General Hospital Cancer Center experience” *Oncologist*, vol. 18, no. 5, pp. 543–548, 2013.

[36] D. D. Von Hoff, T. Ervin, F. P. Arena, E. G. Chiorean, J. Infante, M. Moore, T. Seay, S. A. Tjulandin, W. W. Ma, M. N. Saleh, M. Harris, M. Reni, S. Dowden, D. Laheru, N. Bahary, R. K. Ramanathan, J. Tabernero, M. Hidalgo, D. Goldstein, E. Van Cutsem, X. Wei, J. Iglesias, and M. F. Renschler, “Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine.” *N. Engl. J. Med.*, vol. 369, no. 18, pp. 1691–703, Oct. 2013.

[37] T. Conroy, F. Desseigne, and M. Ychou, “FOLFIRINOX versus Gemcitabine for Metastatic Pancreatic Cancer” *N. Engl. J. Med.*, vol. 364, pp. 1817–1825, 2011.

[38] E. Pujade-Lauraine, U. Wagner, E. Aavall-Lundqvist, V. Gebski, M. Heywood, P. A. Vasey, B. Volgger, I. Vergote, S. Pignata, A. Ferrero, J. Sehouli, A. Lortholary, G. Kristensen, C. Jackisch, F. Joly, C. Brown, N. Le Fur, and A. du Bois, “Pegylated liposomal Doxorubicin and Carboplatin compared with Paclitaxel and Carboplatin for patients with platinum-sensitive ovarian cancer in late relapse” *J Clin Oncol*, vol. 28, no. 20, pp. 3323–3329, 2010.

[39] R CoreTeam, “R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.” 2014.

[40] I. Bebu and J. M. Lachin, “Large sample inference for a win ratio analysis of a composite outcome based on prioritized components” *Biostatistics*, epub ahead of print, 2015.

Annexe A : **méthode de calcul de $\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$ dans les extensions dites de Efron et de Péron**

Situation : $\begin{cases} x_i > y_j + \tau \\ \delta_i = 0 \\ \varepsilon_j = 0 \\ x_i^0 > x_i \\ y_j^0 > y_j \end{cases}$

$\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i^0 > x_i, y_j^0 > y_j\right]$

$= \mathbb{P}\left[x_i > y_j^0 + \tau | \, x_i^0 > x_i, y_j^0 > y_j\right] + \mathbb{P}\left[(x_i^0 > y_j^0 + \tau) \cap (x_i < y_j^0 + \tau) | \, x_i^0 > x_i, y_j^0 > y_j\right]$

$= 1 - \mathbb{P}\left[y_j^0 > x_i - \tau | \, x_i^0 > x_i, y_j^0 > y_j\right] + \frac{\mathbb{P}\left[(x_i^0 > y_j^0 + \tau) \cap (x_i < y_j^0 + \tau) \cap (x_i^0 > x_i) \cap (y_j^0 > y_j)\right]}{\mathbb{P}(x_i^0 > x_i) \cdot \mathbb{P}(y_j^0 > y_j)}$

Comme $\mathbb{P}\left[(x_i^0 > y_j^0 + \tau) \cap (x_i^0 > x_i) \cap (y_j^0 + \tau > x_i)\right] = \mathbb{P}\left[(x_i^0 > y_j^0 + \tau) \cap (y_j^0 + \tau > x_i)\right]$

Et comme $x_i - \tau > y_j$, alors $\mathbb{P}\left[(y_j^0 > x_i - \tau) \cap (y_j^0 > y_j)\right] = \mathbb{P}\left[y_j^0 > x_i - \tau\right]$

Alors :

$\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i^0 > x_i, y_j^0 > y_j\right] = 1 - \mathbb{P}\left[y_j^0 > x_i - \tau | \, x_i^0 > x_i, y_j^0 > y_j\right] + \frac{\mathbb{P}\left[(x_i^0 > y_j^0 + \tau) \cap (x_i < y_j^0 + \tau)\right]}{\mathbb{P}(x_i^0 > x_i) \cdot \mathbb{P}(y_j^0 > y_j)}$.

Notons $S_T(t) = \mathbb{P}\left[x_i^0 \geq t\right]$ et $S_C(t) = \mathbb{P}\left[y_j^0 \geq t\right]$. Lorsque $S_T(t)$ et $S_C(t)$ sont estimé par les estimateurs de Kaplan et Meier $\hat{S}_T(t)$ et $\hat{S}_C(t)$, alors :

$$\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i^0 > x_i, y_j^0 > y_j\right] =$$

$$1 - \frac{\hat{S}_C(x_i - \tau)}{\hat{S}_C(y_j)} - \int_{\substack{t > x_i + \tau \\ t \in \{y_j\} \\ \varepsilon_j = 1}}^{\infty} \frac{\hat{S}_T(t + \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t)$$

$$= 1 - \frac{\hat{S}_C(x_i - \tau)}{\hat{S}_C(y_j)} - \sum_{\substack{t > x_i + \tau \\ t \in \{y_j\} \\ \varepsilon_j = 1}}^{\infty} \frac{\hat{S}_T(t + \tau)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} \cdot \left(\hat{S}_C(t + 1) - \hat{S}_C(t)\right)$$

Cette méthode est répétée pour calculer $\mathbb{P}\left[x_i^0 > y_j^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$ et $\mathbb{P}\left[y_j^0 > x_i^0 + \tau | \, x_i, \, y_j, \, \delta_i, \, \varepsilon_j\right]$ dans les autres situations que celle présentée dans cette annexe.

Annexe B : **Paquet BuyseTest sous R, documentation**

Generalized Pairwise Comparisons

Description

Performs Generalized Pairwise Comparisons for binary, continuous and time to event data.

Developpers : Brice Ozenne and Julien Péron

Usage

```
BuyseTest(data,treatment,endpoint,threshold=NULL,strata=NULL,censoring=NULL
,type,method = "Peron", n.bootstrap=0,prob.alloc=NULL,alternative =
"two.sided",seed=10,cpus=1,trace=3)
```

Arguments

`data`	A `data.frame` containing the variables.
`treatment`	the name of the treatment variable identifying the control and the experimental group. *character*.
`endpoint`	the name of the endpoint variable(s). *character vector*.
`threshold`	the thresholds, one for each endpoint variable. *numeric vector*. Default is `NULL` indicating no threshold.
`strata`	the name of the strata variable(s). *numeric vector*. Default is `NULL` indicating only one strata.
`censoring`	the name of the censoring variable(s), one for each endpoint. *character vector*. Default is `NULL`.
`type`	the type of each endpoint. *character vector*. Can be `"binary"`, `"continuous"` or `"timeToEvent"`.
`method`	paires with censored data can be either classified as uniformative (`"Gehan"`) or compared regarding the predicted survival using a common survival curve for treated and control patients (`"Peto"`) or a separate survival curve for treated and control patients (`"Efron"`or `"Peron"`).
`n.bootstrap`	the number of bootstrap samples used for computing the confidence interval and the p.values. *integer*. Default is `0` meaning no bootstrap (and thus only ponctual estimation).
`prob.alloc`	the resampling probability for assignement to the experimental group in the bootstrap samples. *double*. Default is `NULL` indicating to use the proportion of patients in the experimental group.
`alternative`	a *character* specifying the alternative hypothesis. Must be one of `"two.sided"`, `"greater"` or `"less"`. Default is `"two.sided"`.
`seed`	the seed to consider for the bootstrap. *integer*. Default is `10`.

`cpus` the number of CPU to use. *integer*. Default is `1`.

`trace` Should the execution of the function be traced ? *integer*. Default is `3`.

Details

The variable corresponding to `treatment` in data must have only two levels (e.g. `0` and `1`).

Arguments `endpoint`, `threshold`, `censoring` and `type` must have the same length. `threshold` must be `NA` for binary endpoints and positive for continuous or time to event endpoints. `censoring` must be `NA` for binary or continuous endpoints and indicate a variable in data for time to event endpoints.

Short forms for argument `type` are `"bin"` (binary endpoint), `"cont"` (continuous endpoint), `"TTE"` (time to event endpoint).

The number of bootstrap replications (argument `n.bootstrap`) must be specified to enable the computation of the confidence intervals and the p.value. A large number of bootstrap samples (e.g. `n.bootstrap=10000`) are needed to obtain accurate CI and p.value. See (Buyse et al., 2010) for more details.

`3` corresponds to complete tracing, `2` make message from silent parallelization messages , `1` to only trace the bootstrap and `0` to remain silent.

Argument `cpus` can be set to `"all"` to use all available cpus. The parallelization relies on the *snowfall* package (function *sfClusterApplyLB*) et the detection of the number of cpu on the `detectCores` function from the *parallel* package

Neutral pairs correspond to pairs for which the difference between the endpoint of the control observation and the endpoint of the treatment observation is (in absolute value) below the threshold. Uninformative pairs correspond to pairs for which the censoring prevend from classifying them into favorable, unfavorable or neutral.

Neutral or uninformative pairs for an endpoint with priority `m` are, when available, analysed on the endpoint with priority `m-1`.

Value

An `R` object of class `BuyseRes`.

References

Marc Buyse (2010) Generalized pairwise comparisons of prioritized endpoints in the two-sample problem *Statistics in Medicine* **vol. 29** 3245-3257

See Also

`summary,BuyseRes-method` for a summary of the results of generalized pairwise comparison.
`BuyseRes-class` for a presentation of the `BuyseRes` object.
`constStrata` to create a strata variable from several clinical variables.

Examples

```
#### real example ; survival endpoint####

data(veteran,package="survival")
BuyseTest_veteran <-
  BuyseTest(data=veteran,endpoint="time",treatment="trt",
  type="timeToEvent",censoring="status",threshold=0,n.bootstrap=10000,
  method="Peron")
summary_BuyseTest_veteran <- summary(BuyseTest_veteran)

#### simulated example ; one survival endpoint and one continuous endpoint,
with parallel computing####

n.Treatment <- 500
n.Control <- 500
lambda.Treatment <- 0.75
lambda.Control <- 0.75
lambda.Censoring <- 0.5
mu.Treatment <- 4
mu.Control <- 0

set.seed(10)
data_test <- data.frame(treatment=c(rep(1,n.Treatment),rep(0,n.Control)))
data_test$strata <- rbinom(n.Treatment+n.Control,size=4,prob=0.5)

data_test$EventTime <- c(rexp(n.Treatment,rate=lambda.Treatment),
rexp(n.Control,rate=lambda.Control))
data_test$CensoringTime <- c(rexp(n.Treatment,rate=lambda.Censoring),
rexp(n.Control,rate=lambda.Censoring))

data_test$Survendpoint <-
apply(data_test[,c("EventTime","CensoringTime")],1,min)
data_test$event <-
apply(data_test[,c("EventTime","CensoringTime")],1,which.min)==1
data_test$event <- as.numeric(data_test$event)

data_test$continuous <- c(rnorm(n.Treatment,mean=mu.Treatment),
                          rnorm(n.Control,mean=mu.Control))

BuyseTest_testMixte <- BuyseTest(data=data_test,
                  endpoint=c("Survendpoint","continuous"),
                  treatment="treatment",
                  censoring=c("event",NA),
                  strata="strata",
                   type=c("timeToEvent","continuous"),
                   threshold=c(1,0.5),
                   method="Peron",
                   n.bootstrap=1000,cpus="all")

summary_BuyseTest_testMixte <- summary(BuyseTest_testMixte)
```

Résumé
Dans les essais randomisés conduits en oncologie médicale, l'effet des traitements est le plus souvent évalué sur plusieurs critères de jugement, dont un ou plusieurs critères de type temps jusqu'à événement. Une analyse globale de l'effet d'un traitement intègre les résultats observés sur l'ensemble des critères de jugement pertinent.
Un des objectifs de notre travail était de réaliser une revue systématique de la littérature évaluant les méthodes de recueil, d'analyse et de rapport des événements indésirables et des critères de jugement rapportés par les patients dans les essais de phase III en oncologie médicale. Cette revue a mis en évidence une grande hétérogénéité des méthodes utilisées. De plus les rapports des essais omettaient souvent certaines informations indispensables pour évaluer la validité des résultats rapportés en toxicité ou sur les critères de jugement rapportés par les patients.
Un autre objectif de cette thèse était de développer une extension de la méthode des comparaisons par paire généralisées permettant d'évaluer de façon non biaisée la propension au succès en présence de censure lorsqu'un des critères de jugement est de type temps jusqu'à événement.
Cette thèse avait également pour objectif de montrer comment les comparaisons par paire pouvaient être utilisées afin d'évaluer la balance bénéfice-risque de traitements innovants dans les essais randomisés. De la même façon, la propension globale au succès permet d'évaluer le bénéfice thérapeutique global lorsqu'un effet positif est attendu sur plusieurs critères de jugement.

Mots-clé : analyse de survie ; essais contrôlés randomisés ; cancer ; comparaison par paire ; analyse multi-critère ; analyse statistique ; balance bénéfice-risque

Title : A multicriteria analysis of the chance of a better outcome in randomized trials using generalized pairwise comparisons with survival data.

Summary
In medical oncology randomized trials, treatment effect is usually assessed on several endpoints, including one or more time-to-event endpoints. An overall analysis of the treatment effect may include the outcomes observed on all the relevant endpoints.
A systematic review of medical oncology phase III trials was conducted. We extracted the methods used to record, analyze and report adverse events and patient-reported outcomes. Our findings show that some methodological aspects of adverse events or patient-reported outcomes collection and analysis were poorly reported. Even when reported, the methods used were highly heterogeneous.
Another objective was to develop an extension of the generalized pairwise comparison procedure for time-to-event variables. The extended procedure provides an unbiased estimation of the chance of a better outcome even in presence of highly censored observations.
Then, we show how the chance of an overall better outcome can be used to assess the benefit-risk balance of treatment in randomized trials. When a benefit is expected on more than one endpoint, the chance of an overall better outcome assesses the overall therapeutic benefit. The test of the null hypothesis is more powerful than the test based on one single endpoint.

Keywords : Survival analysis ; randomized trials ; cancer ; pairwise comparisons ; multicriteria analysis ; statistical analysis ; benefit-risk balance

Intitulé et adresse du laboratoire
UMR CNRS 5558-Laboratoire de Biométrie et Biologie Evolutive - Département Biostatistiques et Modélisation pour la Santé et l'Environnement - Equipe Biostatistiques-Santé
165 chemin du Grand Revoyet - Bâtiment 4D - 69495 Pierre-Bénite

THÈSES MAZENOD

☏ 06 01 99 75 70

contact@imprimerie-mazenod.com

www.thesesmazenod.fr