

Analyser et communiquer les critères de survie dans les essais randomisés

Understanding and communicating survival outcomes based on randomised trials

Julien Péron^{1,2,3,4,5}
Brice Ozenne^{6,7}

¹ Hospices civils de Lyon
Service d'oncologie médicale
165, chemin du Grand-Revoyet
69310 Pierre-Bénite
France
<julien.peron@chu-lyon.fr>

² Hospices civils de Lyon
Service de biostatistique et bio-informatique
69003 Lyon
France

³ Université de Lyon
69000 Lyon
France

⁴ Université Lyon 1
69100 Villeurbanne
France

⁵ CNRS, UMR 5558
Laboratoire de biométrie et biologie évolutive
Équipe biostatistique-santé
69100 Villeurbanne
France

⁶ Neurobiology Research Unit
Rigshospitalet
Copenhague
Danemark
<brice.ozenne@orange.fr>

⁷ University of Copenhagen
Department of Public Health
Section of Biostatistics
Copenhague
Danemark

Remerciements et autres mentions :

Financement : les travaux de recherche de B.O. sont financés par les fondations Lundbeck (R231-2016-3236) et Marie-Curie (NEUROMODEL 746850).

Liens d'intérêts : les auteurs déclarent ne pas avoir de lien d'intérêt.

RÉSUMÉ

Les critères de survie sont les critères de jugement principaux les plus utilisés dans les essais randomisés en oncologie. Leur utilisation pour le design, l'analyse et la communication des résultats est relativement uniformisée et repose sur le calcul des probabilités de survie selon la méthode de Kaplan-Meier, l'évaluation de la différence de survie entre les groupes de traitement en estimant le *hazard ratio*, puis la réalisation d'un test du log-rank afin de tester statistiquement les différences de survie observées.

L'objectif de cet article est d'aider les cliniciens à comprendre les limites principales de chacune des méthodes classiques, et de présenter les principales méthodes alternatives. Effectivement, l'effet d'un traitement sur un critère de survie est certainement mieux décrit en utilisant plusieurs mesures de cet effet.

Les critères de jugement alternatifs qui seront décrits incluent la survie spécifique, l'incidence cumulée et la survie nette. Différentes mesures de l'effet d'un traitement seront abordées incluant la différence entre les moyennes de survie restreintes ou non restreintes, le *win ratio*, et le bénéfice net. Enfin, plusieurs alternatives au test du log-rank seront décrites.

Nous encourageons les statisticiens, les méthodologistes et les investigateurs à inclure plusieurs mesures de l'effet des traitements dans les rapports d'essais randomisés, en prenant en compte le fait que différentes situations cliniques peuvent nécessiter le recours à différentes mesures.

● **Mots clés** : analyse de survie ; analyse statistique ; essais cliniques.

ABSTRACT

Survival endpoints are the most frequent primary end points used in Phase III oncology trials. The evaluation of these end points is important to provide information for clinical practice. The current use of survival endpoints in the design, analysis, and communication of results of trials has been fairly standardised and survival probabilities are most often estimated using the Kaplan-Meier method. The overall treatment effect is then assessed using the estimated Hazard ratio and differences in observed survival are then tested statistically using a log-rank test.

The aim of this article is to help clinicians understand the main strengths and limitations of traditional and novel measures used to denote the effect of treatment in randomised trials. However, although different measures can be used to describe the effect of treatment on these end points, we believe that any treatment benefit in a given trial is best reported using various absolute and relative measures.

Alternative endpoints capturing survival data that will be described in this article include specific survival, cumulative incidence, and net survival. Several novel measures and statistical tests designed to assess the effect of

treatments on survival will also be described, including the difference between restricted mean survival, the Win ratio, and net survival benefit.

We encourage statisticians and clinical scientists to include various measures on the outcomes of treatment in reports of Phase III trials, taking into account the fact that different clinical situations may call for different measures of treatment effect.

● **Key words:** survival analysis; statistical analysis; clinical trial.

Introduction

Les critères de jugement évaluant le temps entre la randomisation et la survenue d'un événement (décès, progression, récurrence, etc.) sont les critères de jugement principaux les plus utilisés dans les essais randomisés en oncologie [1-3]. Dans cet article, nous les appellerons, de façon générique, les critères de survie. La méthode utilisée pour rapporter les critères de jugement a un impact démontré sur l'appréciation d'un essai clinique par les cliniciens [4, 5]. Il est donc important d'avoir une compréhension claire des différentes méthodes utilisées et utilisables pour analyser et rapporter les critères de survie.

La procédure la plus communément utilisée pour analyser et rapporter un critère de survie est composée de trois étapes. La première étape consiste à définir précisément et calculer la distribution du critère de survie dans la population des patients de l'étude. Par exemple, pour la survie globale, elle est simplement définie comme le temps entre la randomisation et le décès, quelle qu'en soit la cause. Le calcul de la probabilité de survie à chaque temps de suivi t est typiquement réalisé en utilisant la méthode de Kaplan-Meier et est rapporté de façon graphique (courbes de survie par groupe de traitement en fonction du temps de suivi). Puis, lors de la seconde étape, l'objectif est d'évaluer la différence de survie entre les groupes de traitement en utilisant soit une mesure résumée, telle que la différence entre les proportions de survie à un temps de suivi donné ou la différence entre les médianes de survie, soit une mesure basée sur l'ensemble des temps d'événement telle que le *hazard ratio* (HR). Enfin, la troisième étape consiste à tester si la différence de survie entre les groupes est statistiquement différente de 0, typiquement en utilisant un test du log-rank. Des analyses ajustées ou stratifiées peuvent également être réalisées, mais ne seront pas abordées dans cet article.

Chacune des méthodes citées a été largement évaluée et est certainement adaptée à de nombreuses situations. L'objectif de cet article est de présenter les limites de chacune des méthodes classiques et les principales méthodes alternatives. Les oncologues pourront ainsi comprendre les avantages et inconvénients principaux des différentes méthodes, afin de pouvoir les utiliser dans

leurs recherches et les interpréter de façon adéquate lorsqu'elles sont utilisées.

Définir le critère de jugement et l'estimer (tableau 1)

Survie sans événement

Il s'agit du temps jusqu'à l'observation d'un type d'événement. Il faut donc définir un ou plusieurs événements d'intérêt : cela peut être le décès (on parle alors de survie globale), le décès ou la progression/récurrence de la maladie (critère composite) ou tout autre critère de type binaire définissant un événement (toxicité, etc.). La survie à un temps t correspond alors à la proportion de patients pour lesquels aucun événement n'est observé jusqu'au temps t .

Le plus souvent, on ne dispose pas d'un suivi complet pour l'ensemble des patients. En effet, et c'est là la particularité des analyses de survie, le suivi des patients qui n'ont pas présenté l'événement d'intérêt peut s'arrêter du fait d'un arrêt de l'étude (date de point), de patients perdus de vue, ou d'un événement faisant que le suivi ne peut plus être réalisé après sa survenue (par exemple, évaluation de la survie sans progression sous un traitement non réalisable en cas de changement de traitement pour une raison autre que la progression). On dit alors que les données sont « censurées » à la dernière date où le patient a pu être observé sans avoir présenté l'événement d'intérêt.

La méthode de Kaplan-Meier permet de calculer la probabilité de survie en fonction du temps de suivi, en prenant en compte les données censurées sous l'hypothèse que les censures sont non informatives. Les censures sont non informatives si un patient censuré au temps t a le même risque d'événement après le temps t que les patients qui n'ont pas été censurés. Le calcul de la probabilité de survie par la méthode de Kaplan-Meier n'est réalisé qu'au temps d'événement, ce qui explique la forme en marches d'escalier des courbes de survie. La « taille » de la marche d'escalier correspond au nombre d'événements à l'instant t divisé par le nombre de patients encore suivis à l'instant juste avant t .

Une prudence particulière est nécessaire lors de l'interprétation de la partie droite des courbes de survie. Alors qu'il s'agit souvent de la partie de la courbe de survie qui

Tableau 1. Avantages et inconvénients des différents critères de jugement.
Table 1. Advantages and disadvantages of various judgement criteria.

Critères	Avantages	Inconvénients
Survie sans événement	<ul style="list-style-type: none"> ● Interprétation intuitive ● Presque toujours utilisée 	<ul style="list-style-type: none"> ● Si plusieurs types d'événement (critères combinés), donne plus de poids au plus fréquent/plus précoce, et non au plus important ● Dilution de l'effet thérapeutique si celui-ci n'est attendu que sur un seul des types d'événement (par exemple, décès lié au cancer) ● Biaisée en cas de censure informative/risque compétitif
Survie spécifique	<ul style="list-style-type: none"> ● Exprime un bénéfice clinique direct pour le patient ● Ne prend en compte que la mortalité liée à la maladie 	<ul style="list-style-type: none"> ● Biaisée si les risques compétitifs (décès non liés à la maladie) ne sont pas pris en compte ● Subjectivité dans la détermination des causes de décès
Incidence cumulée	<ul style="list-style-type: none"> ● Intérêt direct pour les patients issus de la même population ● Adaptée en cas de risque compétitif 	<ul style="list-style-type: none"> ● Dépend de l'incidence des risques compétitifs ● Non utilisable pour comparer des populations différentes ● Subjectivité dans la détermination des causes de décès
Survie nette	<ul style="list-style-type: none"> ● Ne prend en compte que la mortalité liée à la maladie ● Non basée sur l'attribution des causes de décès par les cliniciens/investigateurs ● Permet une comparaison entre essais/pays 	<ul style="list-style-type: none"> ● Nécessite de connaître la mortalité toute cause attendue ● Peu utilisable dans les essais cliniques du fait des critères d'inclusion restrictifs ● Ne correspond pas à la réalité des patients (élimination artificielle des autres causes de décès)

intéresse le plus les cliniciens, les estimations des probabilités de survie sont très incertaines dans cette partie car le nombre de patients restant à risque à ces temps de suivi avancé est souvent limité. La confiance dans l'estimation de la probabilité de survie doit se faire en utilisant soit les intervalles de confiance s'ils sont représentés, soit le nombre de patients encore suivis au temps d'intérêt.

La méthode de Kaplan-Meier a l'avantage de ne faire aucune hypothèse sur la distribution des temps d'événements : on parle d'approche « non paramétrique ». Toutefois, si on veut mesurer l'impact d'un certain nombre de covariables sur la survie, d'autres méthodes sont nécessaires. La première alternative est de faire l'hypothèse que les temps d'événements suivent une certaine distribution et d'estimer le ou les paramètres de cette dernière : on parle alors d'approche « paramétrique ». La seconde est de formuler certaines hypothèses qui vont permettre de modéliser l'impact des covariables sur les courbes de survie sans avoir à entièrement spécifier une distribution pour cette dernière. Ces modèles, qui présentent un compromis entre les deux approches précédentes, sont qualifiés de « semi-paramétriques ». Le modèle de Cox à risques proportionnels est le plus utilisé d'entre eux.

Survie spécifique

Une alternative à la survie globale parfois utilisée est la survie spécifique. Ce critère est notamment utilisé

lorsque la population est âgée ou présente des comorbidités importantes et que, par conséquent, les patients sont à haut risque de décéder d'une autre cause que le cancer. La survie spécifique se différencie de la survie globale car elle prend en compte comme événement les seuls décès liés à une cause spécifique, typiquement le cancer étudié. Une difficulté réside dans le choix de l'attribution des causes de décès, notamment pour les décès toxiques [1].

L'estimation de la survie spécifique est complexe lorsque les patients morts d'une autre cause n'ont pas le même risque de mourir du cancer que ceux qui ne sont pas morts d'une autre cause. Cela est vrai pour tous les critères de type survie ne prenant pas en compte les décès pour autre cause comme événement. Or, l'hypothèse d'indépendance entre les décès liés au cancer et les décès pour autre cause est souvent peu réaliste.

Imaginons une situation particulière où le traitement anticancéreux est en fait un traitement qui évite les décès cardiovasculaires, et où il existe une corrélation forte entre le risque de mourir du cancer et le risque de mourir d'accident cardiovasculaire. Le fait de censurer les patients décédant d'un accident cardiovasculaire dans le groupe témoin, alors qu'ils seraient souvent rapidement morts du cancer s'ils n'avaient pas eu l'accident cardiovasculaire, va donner la fausse impression que la survie spécifique du groupe témoin est meilleure que la survie spécifique du groupe traité.

Risques compétitifs et incidence cumulée

Ainsi, dans certaines circonstances, la survenue d'un événement dit compétitif peut modifier (risque compétitif non exclusif), voire empêcher (risque compétitif exclusif), la survenue de l'événement d'intérêt [6]. C'est le cas, dans l'exemple abordé ci-dessus, lorsqu'un patient décède d'une autre cause que le cancer, alors que l'événement d'intérêt est le décès lié au cancer. C'est aussi le cas lorsqu'un patient décède sans avoir présenté de rechute alors que l'événement d'intérêt est la rechute tumorale. Dans ces deux cas, l'événement d'intérêt n'a plus aucune chance de survenir après la survenue de l'événement compétitif, qui est donc exclusif. Lorsque l'événement d'intérêt est la rechute locorégionale, la rechute à distance est un événement compétitif non exclusif, dans le sens où sa survenue va conduire à de nouveaux traitements anticancéreux qui vont modifier la probabilité de rechute locorégionale.

L'incidence cumulée d'un événement est la probabilité que cet événement ait lieu avant un temps t . Avec la méthode de Kaplan-Meier, l'incidence cumulée d'un événement correspond à 1 moins la probabilité de survie. Il s'agit donc d'une courbe ascendante allant de 0 à 100 %, là où la courbe de survie est une courbe descendante allant de 100 à 0 %.

En présence d'événements compétitifs non exclusifs (exemple de la rechute locorégionale et de la rechute à distance), il est possible d'analyser la rechute locorégionale en ignorant les rechutes à distance (méthode *Ignore*), ou en censurant les observations à la date de survenue de la rechute à distance (méthode *Censure*). Ces deux méthodes sont intuitives, mais conduisent à surestimer l'incidence cumulée des survies spécifiques si les événements ne sont pas indépendants. Or l'hypothèse d'indépendance est rarement vraie. Dans notre exemple, cela revient à dire que les traitements mis en œuvre pour traiter une rechute à distance ne modifient pas le risque de rechute locale.

« La méthode d'Aalen-Johanson permet de calculer l'incidence cumulée »

La méthode *Inclure* doit donc être utilisée dans le cas de risques compétitifs si l'hypothèse d'indépendance n'est pas acceptable. Dans cette méthode, tous les événements (d'intérêt ou compétitifs) sont pris en compte dans l'analyse. Le calcul de l'incidence cumulée spécifique de l'événement d'intérêt est ainsi une courbe qui n'atteint pas 100 % si l'événement compétitif est exclusif (l'événement d'intérêt n'a aucune chance de survenir après l'événement compétitif). En effet, un patient décédé sans avoir présenté de rechute ne présentera jamais de rechute, même après un temps de suivi infini. En revanche, la somme des incidences cumulées des différents types d'événements correspond au complément de la survie sans événement (SSE). La méthode d'Aalen-Johanson permet de calculer l'incidence cumulée en

présence de risque compétitif de façon analogue à la méthode de Kaplan-Meier pour le calcul de la probabilité de survie. En effet, c'est une méthode non paramétrique, qui fait également « l'hypothèse d'indépendance » entre les censures et les événements.

En l'absence de risque compétitif, la courbe d'incidence cumulée peut également être utilisée, et permettra de rapporter de façon plus claire les probabilités d'apparition d'un événement dans le cas où le taux d'événement est faible (peu d'événements).

La survie nette

La survie nette est la survie qui serait observée si la seule cause de décès possible était le cancer étudié [7]. Elle ne correspond pas à la réalité mais, en s'affranchissant des différences de mortalité due à d'autres causes que le cancer, elle permet des comparaisons entre pays et entre périodes de temps. Cette mesure est largement utilisée en épidémiologie. La cause du décès, rapportée par les investigateurs ou cliniciens, n'est pas utilisée pour calculer la survie nette (cela correspond à la survie spécifique citée précédemment). La survie nette est calculée en utilisant la mortalité pour toute cause observée dans une étude et la mortalité pour toute cause attendue dans une population comparable de patients mais ne présentant pas de cancer (âge, sexe, pays, etc.). La survie nette est ainsi difficile à calculer dans les essais cliniques du fait du caractère le plus souvent restrictif des critères d'inclusion des essais (patients ayant un bon état général, peu de comorbidité). Il est en effet difficile de disposer d'estimations de la mortalité toute cause attendue pour une population comparable à la population incluse dans un essai.

Mesurer l'effet du traitement (tableau 2)

Après avoir évalué le critère de jugement dans chacun des bras de traitement, il reste à quantifier l'effet du traitement.

Parmi les mesures de l'effet du traitement, il faut distinguer les mesures absolues et relatives. Les mesures absolues calculent une différence, comme une différence de médiane de survie. Les mesures relatives calculent un rapport, comme le HR. Il est communément admis que les mesures absolues sont plus informatives du point de vue des patients [8]. Néanmoins, elles doivent également être interprétées avec précaution quand il s'agit d'informer un patient individuel, car la plupart de ces mesures rapportent l'effet du traitement sur un patient « moyen » inclus dans l'essai, et ne correspondent donc pas au bénéfice individuel attendu pour un patient donné.

Le hazard ratio

En épidémiologie, le risque relatif est la mesure la plus utilisée pour évaluer un changement de risque d'un événement. Il s'agit du rapport du risque de survenue

Tableau 2. Avantages et inconvénients des différentes mesures de l'effet thérapeutique.
Table 2. Advantages and disadvantages of various treatment effect measures.

Mesures	Avantages	Inconvénients
<i>Hazard ratio</i>	<ul style="list-style-type: none"> • Presque toujours rapporté • Prend en compte l'information des courbes de survie du début à la fin du suivi 	<ul style="list-style-type: none"> • Notion non triviale qui peut être difficile à comprendre et à interpréter pour le patient • Interprétation difficile si les <i>hazards</i> ne sont pas proportionnels • Change en fonction du suivi si les <i>hazards</i> ne sont pas proportionnels
La différence entre les proportions de survie à un temps t	<ul style="list-style-type: none"> • Facile à lire sur les courbes de survie 	<ul style="list-style-type: none"> • Dépend du choix de t • Perte d'information • Reflète mal l'effet du traitement si les <i>hazards</i> ne sont pas proportionnels
La différence entre les médianes de survie	<ul style="list-style-type: none"> • Facile à lire sur les courbes de survie • Facile à mémoriser 	<ul style="list-style-type: none"> • Interprétation difficile pour un patient individuel • Médianes pas toujours atteintes • Dépend de la fréquence des évaluations si le critère n'est pas la survie globale • Perte d'information • Intervalle de confiance large • Reflète mal l'effet du traitement si les <i>hazards</i> ne sont pas proportionnels
La différence entre les moyennes de survie restreintes	<ul style="list-style-type: none"> • Prend en compte l'information des courbes de survie jusqu'au temps t^* • Interprétable même si les <i>hazards</i> ne sont pas proportionnels • Interprétation intuitive : différence entre les aires sous les courbes de survie 	<ul style="list-style-type: none"> • Rarement rapportée • Interprétation difficile si les courbes de survie sont éloignées de 0 au temps de suivi maximum
La différence entre les moyennes de survie non restreintes	<ul style="list-style-type: none"> • Concept intuitif pour un patient individuel • Facile à mémoriser • Prend en compte l'information des courbes de survie du début à la fin du suivi • Interprétable même si les <i>hazards</i> ne sont pas proportionnels 	<ul style="list-style-type: none"> • Presque jamais rapportée • Nécessite une modélisation de la courbe de survie si l'estimation de celle-ci n'atteint pas 0 en fin de suivi • Résultat peu basé sur les données réellement observées si les données ne sont pas matures (survie loin de 0 en fin de suivi) • Non interprétable en cas de guérison ou de risques compétitifs
Modèle de Cox dépendant du temps	<ul style="list-style-type: none"> • Permet de décrire l'évolution temporelle du <i>hazard ratio</i> • Interprétation possible lorsque les <i>hazards</i> ne sont pas proportionnels 	<ul style="list-style-type: none"> • Peu utilisé dans les essais cliniques • Nécessite de pré-spécifier le modèle pour pouvoir l'utiliser comme analyse principale
Modèle de guérison	<ul style="list-style-type: none"> • Permet de décrire l'effet du traitement lorsqu'il existe une fraction de patients guéris • Interprétation possible lorsque les <i>hazards</i> ne sont pas proportionnels 	<ul style="list-style-type: none"> • Plusieurs modèles disponibles • Peu utilisé dans les essais cliniques • Nécessite de pré-spécifier le modèle pour pouvoir l'utiliser comme analyse principale • Nécessite un suivi prolongé pour observer la proportion de patients guéris

Tableau 2. (Suite).
Table 2. (Continued).

Mesures	Avantages	Inconvénients
Bénéfice net	<ul style="list-style-type: none"> • Concept intuitif pour un patient individuel • Expression du bénéfice en termes de gain (ou perte) absolu en durée de survie • Prend en compte l'information des courbes de survie du début à la fin du suivi • Interprétable même si les <i>hazards</i> ne sont pas proportionnels • Peut être utilisé pour analyser simultanément plusieurs critères de jugement, en priorisant le plus important cliniquement 	<ul style="list-style-type: none"> • Récemment proposé, donc peu d'expérience en pratique • Ne permet pas d'ajuster sur des covariables continues
<i>Win ratio</i>	<ul style="list-style-type: none"> • Prend en compte l'information des courbes de survie du début à la fin du suivi • Interprétable même si les <i>hazards</i> ne sont pas proportionnels • Peut être utilisé pour analyser simultanément plusieurs critères de jugement, en priorisant le plus important cliniquement 	<ul style="list-style-type: none"> • Peu interprétable pour un patient individuel • Récemment proposé, donc peu d'expérience en pratique • Ne permet pas d'ajuster sur des covariables continues

d'un événement dans le groupe exposé sur le risque de survenue de l'événement dans le groupe non exposé. Le risque relatif ne prend pas en compte le temps précis de l'événement. Dans les essais cliniques, et particulièrement pour les événements qui vont survenir chez la majorité, voire l'ensemble des patients, le temps de survenue est important à considérer. Le risque relatif a donc été remplacé par d'autres mesures comme le HR. Le HR est, en pratique, une mesure systématiquement rapportée de l'effet des traitements dans les essais comparatifs. Ce paramètre sera donc souvent utilisé afin d'illustrer les avantages et limites des méthodes alternatives.

Le HR peut être traduit en français par le rapport des taux instantanés d'événement. Le taux instantané d'événement à un temps t dans un groupe de traitement est le « risque instantané » de survenue de l'événement à l'instant t sachant qu'il n'est pas survenu avant le temps t . Il est ainsi important de considérer que le HR correspond au « risque relatif instantané » d'événement dans un groupe de traitement par rapport au groupe de référence. Il devrait donc être interprété comme une réduction ou augmentation du risque d'événement sur une certaine période de temps.

En général, le HR peut donc varier avec le temps.

Néanmoins, sous l'hypothèse que les taux instantanés d'événement sont proportionnels, le HR est constant au cours du temps et peut donc être exprimé par une valeur

numérique unique. Sous cette hypothèse, le HR peut être estimé par un modèle de Cox à taux proportionnels [9, 10].

L'hypothèse de proportionnalité n'est pas tenable dans deux situations distinctes lorsque :

- la population de l'essai est un mélange de patients avec un effet thérapeutique différent ;
- l'effet du traitement varie au cours du temps. Un effet thérapeutique variant au cours du temps est typiquement observé lorsqu'une stratégie chirurgicale curative est comparée à un traitement médical palliatif, ou lorsqu'une immunothérapie anticancéreuse, ayant un effet prolongé chez une partie des patients, est comparée à une chimiothérapie cytotoxique ayant un effet de courte durée chez la totalité des patients [11].

La différence entre les proportions de survie à un temps t

La comparaison des probabilités de survie à un instant t peut être directement lue sur les courbes de survie estimées par la méthode de Kaplan-Meier. L'avantage de cette méthode est qu'elle est simple à appréhender. Néanmoins, le choix du temps d'analyse a un impact crucial sur l'analyse, qui ignore les temps d'événements précis, et ignore tous les événements qui surviennent après le temps t . Il y a donc une perte importante

d'information. Même dans le scénario de proportionnalité des taux instantanés d'événement, la différence entre les proportions de survie à un temps t variera en fonction de t . La valeur sera nulle pour $t = 0$, augmentera progressivement jusqu'à une valeur maximale et diminuera ensuite progressivement jusqu'à 0 pour les longs temps de suivi. Il n'y a donc pas de relation directe entre le HR et une différence de proportions de survie même sous l'hypothèse de proportionnalité. Le même problème touche le nombre de sujets à traiter (NST), qui est l'inverse de la différence absolue des probabilités de décès (le NST au temps t est 1 divisé par cette différence au temps t). Le NST, quoique peu compréhensible par les patients, est une mesure importante pour la médecine fondée sur les preuves, mais présente les mêmes limites que la différence de proportion de décès à un temps t .

La différence entre les médianes de survie

La médiane de survie est la durée de suivi pour laquelle 50 % des sujets ont présenté l'événement d'intérêt. La différence entre les médianes de survie des deux groupes de patients est une mesure qui est très largement utilisée dans la communauté oncologique. Cette mesure est facile à calculer, probablement la plus facile à mémoriser, et est appréciée par les cliniciens. Néanmoins, elle présente les mêmes limites que celles citées pour la différence entre les proportions de survie à un temps t et, évidemment, n'est calculable que lorsque la médiane de survie est atteinte avant le dernier temps de suivi disponible. À noter que le rapport entre les médianes de survie est égal au HR uniquement lorsque la distribution des temps de survie est exponentielle et que l'hypothèse de proportionnalité est vérifiée. Ces deux hypothèses sont difficiles à justifier en réalité. Une autre limite à l'utilisation des médianes de survie survient lorsque le critère de jugement n'est pas la survie globale, et dépend d'examens périodiques (par exemple, par imagerie). Dans ce cas, l'estimation de la médiane de survie dépendra de la fréquence de ces examens. Lorsque le critère de jugement principal est la survie sans progression, et que les temps de progression sont globalement courts (traitement tardif d'un cancer agressif au stade métastatique), il est classique que les médianes de survie soient identiques entre deux groupes de traitement alors que le traitement est efficace (HR nettement inférieur à 1). La médiane de survie correspond alors à la date de la première ou seconde évaluation prévue dans le protocole. Enfin, l'estimation de la médiane de survie est très sensible à la proportion de patients perdus de vue avec un temps de suivi court, tout en étant peu efficace (intervalle de confiance large).

La différence entre les moyennes de survie restreintes

La distribution des temps de survie étant en général non normale, la moyenne des temps de survie a longtemps

été négligée comme mesure de la tendance centrale dans les analyses de survie. Néanmoins, la comparaison des moyennes de survie est cliniquement interprétable et assez intuitive. Si la courbe de survie atteint 0, la survie moyenne peut être calculée en utilisant l'aire sous la courbe de survie. Toutefois, ce n'est que rarement le cas en pratique. Il est alors possible d'estimer la moyenne de survie restreinte, en restreignant (ou tronquant) le suivi jusqu'à un temps t^* , et en analysant les données que jusqu'à un temps donné t^* . Le temps t^* peut être choisi de façon arbitraire, mais il s'agit habituellement du temps d'événement maximum du groupe de traitement où il est le plus bas. La moyenne de survie restreinte correspond à l'aire sous la courbe de survie jusqu'au temps t^* . La différence entre les moyennes de survie restreintes mesure le bénéfice moyen en survie jusqu'au temps de suivi t^* . Cette mesure présente donc un intérêt en termes de compréhension pour les patients, même si la notion d'un horizon fixe peut être difficile à interpréter.

Cette mesure peut être utilisée, que les taux instantanés d'événement soient proportionnels ou non [12]. Les moyennes de survie restreintes sont peu fréquemment rapportées, possiblement du fait de leur interprétation difficile lorsque la probabilité de survie au temps t^* est loin de 0.

Plutôt que la différence des moyennes de survie restreintes, il est possible d'utiliser le rapport de ces moyennes. Ce choix est illustratif de la différence entre une mesure absolue et relative d'un effet thérapeutique. Le rapport entre les moyennes de survie restreintes est égal au HR lorsque les temps de survie suivent une distribution exponentielle et proportionnelle.

La différence entre les moyennes de survie non restreintes

Comme précisé précédemment, estimer la moyenne de survie sans restreindre le temps de suivi nécessite que l'estimation des courbes de survie atteigne 0 %. Comme cela arrive rarement, aucune estimation non paramétrique de la moyenne de survie n'est habituellement rapportée. Néanmoins, il est possible de modéliser les distributions des temps de survie, permettant ainsi d'estimer les moyennes de survie non restreintes, sous l'hypothèse que le modèle utilisé est valide. Les modèles paramétriques adaptés aux données censurées sont largement utilisés, notamment en ingénierie, mais assez rarement en médecine où les méthodes non paramétriques ou semi-paramétriques (modèle de Cox) sont la règle. De plus, les modèles paramétriques peuvent être adaptés lorsque les *hazards* ne sont pas proportionnels, ce qui peut être une propriété intéressante. La limite de cette approche est évidemment que le résultat dépendra du modèle choisi. Même si des méthodes de diagnostic existent, permettant de sélectionner un modèle adapté aux données [13], il est souvent difficile de définir à l'avance le modèle le plus adapté. Cette dernière

problématique explique certainement la faible utilisation de ces modèles pour l'analyse des données d'essais cliniques. De plus, le modèle choisi sera toujours critiquable lorsque la proportion de patients présentant l'événement d'intérêt est faible (par exemple, traitement adjuvant du cancer du sein).

Le modèle de Cox dépendant du temps

Le modèle de Cox est le modèle de référence pour estimer le HR dans le domaine de la recherche clinique en oncologie. Une limite déjà évoquée du modèle de Cox à taux proportionnel est que le HR peut varier au cours du temps. Il est possible de modéliser la non-proportionnalité de l'effet du traitement en utilisant un modèle de Cox incluant un effet du traitement dépendant du temps [14]. Dans ce cas, un minimum de deux paramètres est estimé par le modèle : un paramètre correspond à l'effet de base du traitement, indépendamment du temps ; un second paramètre correspond à l'effet du traitement en fonction du temps. Le paramètre peut estimer un effet continu du temps ou un changement de l'effet du traitement après un certain temps de suivi. La modélisation peut être plus complexe et faire intervenir plus de deux paramètres. L'avantage de cette méthode est qu'elle permet une description fine de l'effet du traitement au cours du temps. La limite de cette approche dans le cadre de l'analyse des essais randomisés est qu'elle nécessite que le modèle soit défini avant d'avoir observé les données, et il est souvent difficile d'anticiper la fonction qui représentera de façon appropriée l'évolution de l'effet du traitement avec le temps. Une autre limite est la définition du test principal d'évaluation de l'effet thérapeutique. En effet, le fait que l'effet d'un traitement varie au cours du temps n'implique pas forcément qu'il soit efficace (par exemple, effet positif en début de traitement et effet délétère après un temps de suivi plus long). Enfin, l'interprétation finale peut être difficile lorsqu'effectivement l'effet du traitement varie au cours du temps, qui plus est si la variation est non linéaire.

Les modèles de survie avec guérison

Les modèles de survie avec guérison sont des modèles utilisés lorsque l'on s'attend à ce que certains malades soient guéris de la maladie étudiée (par exemple, l'immunothérapie, les traitements adjuvants), et donc ne présentent jamais l'événement d'intérêt. Avec ce type de modèle, les courbes de survie ne descendent pas forcément jusqu'à 0 % en fin de suivi mais atteignent un plateau [15]. Dans les modèles classiques, dits de mélange, il est possible d'évaluer à la fois l'effet du traitement sur la proportion de patients guéris, et l'effet du traitement parmi les patients non guéris. La limite principale de cette approche, en recherche clinique, est qu'une guérison est définie théoriquement à un temps d'observation infini. Or, le plus souvent, le temps de suivi est relativement court lors de l'analyse

principale d'un essai clinique, ce qui limitera la possibilité d'utiliser ce type de modèle (problème de convergence et incertitude sur l'estimation de la proportion de patients guéris). En cas de suivi relativement court, la seule façon de pouvoir ajuster ce type de modèle est de faire l'hypothèse que les patients, qui n'ont pas présenté l'événement après un certain temps, sont guéris, mais c'est une hypothèse très forte, assez souvent considérée comme inacceptable.

Le bénéfice net

Le bénéfice net est une mesure relativement récente, qui n'est pas spécifique des analyses de survie, mais peut être utilisée pour comparer deux groupes de patients randomisés en fonction d'un critère de survie [16, 17].

« Relation directe entre le HR et le bénéfice net »

Il s'agit de la probabilité pour un patient pris au hasard dans le groupe traitement d'avoir une survie plus longue d'au moins m mois ou années qu'un patient pris au hasard dans le groupe témoin, moins la probabilité opposée. Le seuil m étant le seuil de bénéfice minimal jugé cliniquement pertinent. Le bénéfice net vaut donc 1 si tous les patients du groupe traitement vont mieux que tous les patients du groupe témoin, -1 si tous les patients du groupe traitement vont moins bien que tous les patients du groupe témoin, et 0 s'il n'y a pas de différence entre les groupes. Le bénéfice net peut prendre n'importe quelle valeur entre -1 et 1. Il existe une relation directe entre le HR et le bénéfice net en l'absence de censure : $\text{bénéfice net} = [1 - \text{HR}] / [1 + \text{HR}]$ [18]. En présence de censure, la relation peut être ajustée simplement sous l'hypothèse de proportionnalité. Si les *hazards* sont non proportionnels, il n'existe plus de relation directe entre le HR et le bénéfice net. Néanmoins, le bénéfice net reste alors interprétable, ce qui n'est plus le cas du HR. Le bénéfice net présente un intérêt pour communiquer les résultats d'un essai avec les cliniciens, et possiblement également avec les patients, puisqu'il pose la question du bénéfice sous l'angle d'un patient posant la question : « Quelle augmentation/diminution de probabilité ai-je de survivre plus longtemps (d'au moins m mois/années) avec le traitement qu'avec le contrôle ? ». À noter que le bénéfice net peut également être utilisé pour analyser simultanément plusieurs critères de jugement (par exemple, balance bénéfice/risque), mais cela ne sera pas abordé dans cet article.

Le win ratio

Le *win ratio* est une mesure relative de l'effet thérapeutique qui a été proposée en parallèle du bénéfice net, et qui suit une philosophie identique [19]. La différence est que le *win ratio* est le rapport (et non la différence) entre la probabilité pour un patient pris au hasard dans le groupe traitement d'avoir une survie plus longue qu'un

patient pris au hasard dans le groupe témoin, divisée par la probabilité opposée. Le *win ratio* a été proposé initialement pour analyser de façon conjointe plusieurs critères cardiovasculaires, mais tous les développements réalisés pour analyser la survie en utilisant le bénéfice net peuvent lui être appliqués.

Tests statistiques (tableau 3)

Le log-rank

Le test du log-rank est de loin le plus utilisé pour comparer deux distributions de survie. L'hypothèse nulle, à tester, est celle de l'identité des distributions de survie. Ce test correspond à l'attribution d'un poids identique à tous les événements, quel que soit le temps de leur survenue. Il revient à comparer le nombre d'événements observés dans le groupe témoin au nombre d'événements attendus sous l'hypothèse nulle. Sous l'hypothèse de proportionnalité définie dans le chapitre « Mesurer l'effet du traitement », le test du log-rank présente des propriétés optimales en termes de puissance. C'est-à-dire qu'aucun test ne sera plus puissant que lui pour rejeter l'hypothèse nulle. Une limite du test du log-rank est qu'il est basé sur une statistique qui ne présente pas d'interprétation évidente, et qui n'est, de ce fait, jamais rapportée en pratique.

Les log-rank pondérés

Lorsque le traitement présente uniquement un effet à court terme ou au contraire l'effet n'apparaît qu'au

bout d'un certain temps (par exemple après un an, deux ans, dix ans de suivi), l'hypothèse de proportionnalité ne tient plus et le test du log-rank est suboptimal.

Il existe une classe de tests, les tests du log-rank pondérés, dont le principe est de donner un poids différent aux événements en fonction de leur temps de survenue. Le test du log-rank standard est ainsi un membre de cette famille de tests et correspond à un poids identique quel que soit le temps de survenue des événements. Il est possible de donner plus de poids aux événements précoces (survenant alors que la probabilité de survie est élevée), ou à l'inverse plus de poids aux événements tardifs (survenant alors que la probabilité de survie est basse) [20]. Les tests du log-rank pondérés sont plus puissants que le test du log-rank standard lorsque les *hazards* ne sont pas proportionnels. Néanmoins, la limite de cette approche est que la statistique de test n'a toujours pas d'interprétation directe. De plus, le choix de la pondération est arbitraire et peut conduire à une perte de puissance si l'effet du traitement est proportionnel au cours du temps ou si l'effet dans le temps est l'opposé de celui qui était attendu.

Les comparaisons par paires généralisées

Le bénéfice net, évoqué dans la section « Tests statistiques », est estimé en utilisant les comparaisons par paires généralisées [17]. Ce paramètre peut servir à tester l'hypothèse nulle d'un bénéfice net égal à 0. Cette hypothèse nulle correspond à la situation où un patient du groupe traitement a la même probabilité d'avoir une

Tableau 3. Avantages et inconvénients des différents tests statistiques de l'effet thérapeutique.

Table 3. Advantages and disadvantages of various statistical tests for treatment effects.

Tests	Avantages	Inconvénients
Test du log-rank	<ul style="list-style-type: none"> • Presque toujours utilisé • Puissance optimale en cas de <i>hazards</i> proportionnels • Donne le même poids aux événements précoces et tardifs 	<ul style="list-style-type: none"> • Absence d'interprétation clinique de la statistique de test • Non optimal si l'effet du traitement est précoce ou retardé (non proportionnel)
Test du log-rank pondéré	<ul style="list-style-type: none"> • Meilleure puissance que le log-rank en cas de <i>hazards</i> non proportionnels si les poids sont attribués correctement 	<ul style="list-style-type: none"> • Rarement utilisé • Absence d'interprétation clinique de la statistique de test • Nécessite de définir artificiellement des poids différents en fonction du temps de survenue des événements • Non optimal si l'effet du traitement est proportionnel
Comparaisons par paires généralisées	<ul style="list-style-type: none"> • Permet de pré-spécifier le bénéfice minimal considéré cliniquement pertinent • Utilisable lorsque les <i>hazards</i> ne sont pas proportionnels • Test plus puissant que le test du log-rank si un seuil élevé est choisi et que l'effet du traitement est retardé 	<ul style="list-style-type: none"> • Nouvelle méthode donc peu d'expérience • Nécessite de pré-spécifier le bénéfice minimal cliniquement pertinent pour pouvoir l'utiliser comme analyse principale • Puissance non optimale si l'effet du traitement est proportionnel

Take home messages

- À part le *hazard ratio*, différentes mesures de l'effet d'un traitement peuvent être utilisées, incluant la différence entre les moyennes de survie restreintes ou non restreintes, le *win ratio*, et le bénéfice net.
- La comparaison des moyennes de survie, restreintes ou non, est cliniquement interprétable.
- Le bénéfice net est une mesure relativement récente, qui peut être utilisée pour comparer deux groupes de patients en fonction d'un critère de survie.
- Pour l'interprétation statistique et en fonction des situations, plusieurs alternatives au test du log-rank sont disponibles.

meilleure ou une moins bonne survie qu'un patient du groupe témoin. La notion de différence minimale cliniquement pertinente est intégrée dans la définition du bénéfice net. Si le bénéfice net est statistiquement différent de 0, et que le choix du critère de jugement et du seuil a été réalisé de façon adaptée, il est possible de conclure que le traitement expérimental est meilleur de façon statistiquement significative et cliniquement pertinente.

Tests basés sur des modèles

D'autres tests peuvent être utilisés, notamment basés sur les paramètres utilisés lors de la modélisation des courbes de survie ou la modélisation de l'effet du traitement au cours du temps (par exemple, modèles de guérison, Cox dépendant du temps). La limite de ce type d'approche, comme évoquée dans la section « Mesurer l'effet du traitement », est en général la difficulté de prédéfinir un paramètre ou un groupe de paramètres du modèle pouvant servir pour définir une hypothèse nulle et une hypothèse alternative cliniquement satisfaisante et compréhensible [13].

Conclusion

Lors de l'analyse d'un essai clinique, l'évaluation de l'effet thérapeutique sur un critère de type temps jusqu'à événement peut être réalisée de multiples façons. Les méthodes les plus utilisées (Kaplan-Meier, modèle de Cox, test du log-rank) sont le plus souvent adaptées et présentent l'avantage notable d'être bien connues et évaluées. L'écueil majeur du test du log-rank est que le test n'est pas basé sur une mesure interprétable de l'ampleur de l'effet thérapeutique. Néanmoins, les autres méthodes présentées dans cet article peuvent être utilisées en addition, voire dans certains cas en remplacement, des méthodes habituelles. Ainsi, l'effet du traitement peut être décrit sous une perspective différente, ce qui peut s'avérer utile chaque fois que les décisions prises sont basées sur

la comparaison de données de survie dans un essai clinique. Bien qu'aucune des mesures existantes ne soit capable de prédire l'effet thérapeutique pour un patient individuel, les mesures absolues de l'effet du traitement peuvent s'avérer plus informatives pour les patients que les mesures relatives comme le HR. Nous encourageons les investigateurs et les statisticiens à intégrer les méthodes présentées dans cet article lors du design et de l'analyse des essais cliniques randomisés, et notamment les moyennes de survie restreintes et le bénéfice net.

RÉFÉRENCES

1. Punt CJA, Buyse M, Kohne C-H, et al. Endpoints in adjuvant treatment trials: a systematic review of the literature in colon cancer and proposed definitions for future trials. *J Natl Cancer Inst* 2007 ; 99 (13) : 998-1003.
2. Saad ED, Buyse M. Overall survival: patient outcome, therapeutic objective, clinical trial end point, or public health measure? *J Clin Oncol* 2012 ; 30 (15) : 1750-4.
3. Subramanian J, Madadi AR, Dandona M, Williams K, Morgensztern D, Govindan R. Review of ongoing clinical trials in non-small cell lung cancer: a status report for 2009 from the ClinicalTrials.gov website. *J Thorac Oncol* 2010 ; 5 (8) : 1116-9.
4. Chao C, Studts JL, Abell T, et al. Adjuvant chemotherapy for breast cancer: how presentation of recurrence risk influences decision-making. *J Clin Oncol* 2003 ; 21 (23) : 4299-305.
5. Bobbio M, Demichelis B, Giustetto G. Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet* 1994 ; 343 (8907) : 1209-11.
6. Com-Nougé C, Guérin S, Rey A. Assessment of risks associated with multiple events. *Rev Epidemiol Sante Publique* 1999 ; 47 (1) : 75-85.
7. Seppä K, Hakulinen T, Pokhrel A. Choosing the net survival method for cancer survival estimation. *Eur J Cancer* 2015 ; 51 (9) : 1123-9.
8. Zipkin DA, Umscheid CA, Keating NL, et al. Evidence-based risk communication. *Ann Intern Med* 2014 ; 161 (4) : 270.
9. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014 ; 32 (22) : 2380-5.
10. Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010 ; 21 (1) : 13-5.
11. Alexander BM, Schoenfeld JD, Trippa L. Hazards of hazard ratios — deviations from model assumptions in immunotherapy. *N Engl J Med* 2018 ; 378 (12) : 1158-9.
12. 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* 2013 ; 13 : 152.
13. Cleall S. Modelling survival data in medical research (second edition). *Pharm Stat* 2004 ; 3 (1) : 69-169.
14. Bellera C, MacGrogan G, Debled M, de Lara CT, Brouste V, Mathoulin-Pelissier S. 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* 2010 ; 16 : 10-20.
15. Othus M, Mitchell A, Barlogie B, Morgan G, Crowley J. Cure-rate survival models and their application to cancer clinical trials. In: *Frontiers of biostatistical methods and applications in clinical oncology*. Singapore : Springer Singapore, 2017. p. 165-78.
16. Péron J, Roy P, Ozenne B, Roche L, Buyse M. The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. *JAMA Oncol* 2016 ; 2 (7) : 901-5.
17. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med* 2010 ; 29 (30) : 3245-57.
18. Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials* 2008 ; 5 (6) : 641-2.
19. 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.
20. Fleming TRHDP. A class of hypothesis tests for one and two samples of censored survival data. *Commun Stat* 1984 ; 1 (10) : 763-94.