Surrogate Endpoints for Overall Survival in Cancer Randomized Controlled Trials

Marion Savina

université de BORDEAUX

Thèse
présentée pour obtenir le grade de
Docteur de l'Université de Bordeaux

Ecole Doctorale Sociétés, Politique, Santé Publique
Spécialité Santé Publique, option Biostatistique

Par Marion SAVINA

Critères de Substitution à la Survie Globale dans les Essais Cliniques Randomisés en Cancérologie

Surrogate Endpoints for Overall Survival in Cancer Randomized Controlled Trials

Sous la direction de Carine BELLERA

Soutenue le 14 décembre 2017 devant les membres du jury :

M. SALMI Rachid	Pr, INSERM U1219, Bordeaux	Président
Mme MOLLEVI Caroline	Dr, INSERM U1194, Montpellier	Rapporteuse
M. CHAMOREY Emmanuel	Dr, Université Nice-Sophia Antipolis	Rapporteur
Mme MATHOULIN-PELISSIER Simone	Pr, INSERM U1219, Bordeaux	Examinatrice
Mme MIGEOT Virginie	Pr, INSERM CIC1402, Poitiers	Examinatrice
Mme GOURGOU Sophie	MsC, Institut du Cancer de Montpellier	Invitée
M. PORCHER Raphaël	Dr, INSERM U1153, Paris	Invité
Mme BELLERA Carine	Dr, INSERM U1219, Bordeaux	Directrice de thèse

Abstract

In cancer randomized controlled trials (RCT), a surrogate endpoint is intended to substitute a clinically relevant endpoint, e.g. overall survival (OS), and it is supposed to predict treatment effect. Alternative endpoints, for example progression-free survival, are increasingly being used in place of OS as primary efficacy endpoints in RCTs. In practice however, the surrogate properties of these endpoints are not systematically assessed. We performed a systematic literature review to identify surrogate endpoints validated in oncology. We next conducted MAs to evaluate surrogate endpoints in two cancer settings: advanced soft-tissue sarcoma and adjuvant breast cancer. Results could not definitely validate surrogate endpoints in these indications. OS must remain the primary efficacy endpoint in these settings, even though alternative endpoints may provide valuable input in earlier phase studies (phase II trials, futility analyses). This work provides key information for the design of cancer RCTs, in particular for the choice of primary endpoints to assess treatment efficacy.

Résumé

Dans les essais cliniques randomisés (ECR) en cancérologie, un critère de substitution est une mesure biologique utilisée à la place d'un critère cliniquement pertinent pour le patient, par exemple la survie globale (SG), qui doit permettre de prédire l'effet attendu du traitement. Des critères alternatifs à la SG, par exemple la survie sans progression, sont de plus en plus fréquemment utilisés en tant que critère de jugement principal dans les ECR. En pratique cependant, les capacités de substitution à la SG de ces critères ne sont pas systématiquement évaluées. Nous avons dressé un état des lieux des critères de substitution validés en cancérologie à partir d'une revue systématique de la littérature. Par la suite, nous avons évalué par une approche méta-analytique des critères de substitution dans le contexte des sarcomes des tissus mous en situation avancée et du cancer du sein en situation adjuvante. Les résultats n'ont pas permis de définitivement valider de critères de substitution à la SG dans ces indications. La SG doit donc rester le critère de jugement principal des ECR, même si certains critères alternatifs restent informatifs dans des évaluations plus précoces (phase II, analyse de futilité), sous réserve que les données de survie continuent à être recueillies. Ce travail fournit des informations clés pour le développement des ECR en cancérologie afin notamment de sélectionner au mieux les critères de jugement de l'efficacité thérapeutique.

Acknowledgments

A mon président de jury, le Pr Louis-Rachid Salmi,

Merci de m'avoir fait l'honneur d'accepter de présider mon jury de thèse.

Au Pr Emmanuel Chamorey,

Au Pr Virginie Migeot,

Au Dr Raphaël Porcher,

Pour avoir accepté de faire partie de mon jury de thèse.

Au Dr Caroline Mollevi,

A Mme Sophie Gourgou,

Pour avoir accepté de faire partie de mon jury, mais également pour m'avoir donné l'opportunité de découvrir la recherche clinique et l'envie de continuer. Je vous remercie pour votre gentillesse, votre bienveillance et vos encouragements.

Au Pr Simone Mathoulin-Pélissier,

Pour m'avoir fait confiance et donné l'occasion de faire cette thèse. Je vous remercie de tout le soutien que vous m'avez apporté, vous compter parmi mon jury est un honneur.

A ma directrice de thèse, le Dr Carine Bellera,

Pour m'avoir donné l'occasion de faire cette thèse avec toi. Je te remercie de la confiance et du soutien que tu m'as apporté, dans le travail de thèse en lui-même mais également dans mes questionnements plus personnels. Je suis très fière d'être ta première thésarde.

Au Pr Virginie Rondeau,

Pour avoir accepté de faire partie de mon comité de suivi. Merci pour tes encouragements, ta disponibilité et ton aide méthodologique très précieuse durant ces trois années.

To Dr Saskia Litière,

For your amazing availability and support. It was a real pleasure to work with you.

To Dr Tomasz Burzykowski,

For agreeing to be part of my steering committee. Thank you for your availability and kindness, it was a great pleasure to work with you.

A toute l'unité de recherche clinique de l'Institut Bergonié,

Merci à Marina, Sandrine, Stéphanie, Antoine, Véronique, Derek et Maïté pour leur soutien, leur bonne humeur et les fous rires partagés.

A mes collègues de bureau et amis,

Merci à Lucile, Clément, Louis, et Hélène pour votre soutien et votre bonne humeur, nécessaires pour tenir jusqu'au bout.

A Angéline,

Merci pour… tellement de choses ! D'avoir atterri dans mon bureau tout d'abord, même si tu n'y étais pas pour grand-chose, et d'avoir su briser la glace si facilement. Merci pour ces 3 années de galères, de stress, de coups durs, de débats féministes et politiques, de sport (un peu), de repas (un peu plus), d'apéros (beaucoup), de création et de fous rires. Je n'aurais vraiment pas pu rêver meilleure partenaire de thèse (et autres) !

A mes amis,

A Julianne, toujours présente, toujours souriante malgré les moments difficiles. J'admirerai toujours ta force et ton courage. Je souhaite que ta nouvelle vie t'apporte tout ce que tu mérites.

A mes amis de M2, Mathilde, Anne et Yassine, même si on ne se voit pas aussi souvent que je le souhaiterais vous êtes toujours très importants pour moi.

A Charlotte, mon amie de toujours et pour toujours, je te remercie d'être là pour moi. Nos vies peuvent nous éloigner quelques temps, mais jamais très longtemps.

A Alexis, merci pour ton soutien indéfectible même dans les moments les plus difficiles. Je souhaite de tout mon cœur que tu réussisses à croire en toi autant que tu crois en moi.

A Justine, Emilie, Séverine, Antoine, Pierre et Vincent, à Dorian, et à mes équipiers de voile Didier, Fabrice et Stéphane, merci pour tous ces moments d'évasion. Mon cerveau de blonde n'aurait certainement pas tenu sans !

A Paulin,

Merci d'être qui tu es. Merci de ne pas t'être arrêté, perdu au milieu du supermarché, et d'être venu quand je t'ai invité à te réfugier dans ma cabane. Merci de croire et de me faire croire que tout est possible. Tu m'aides tous les jours à me rapprocher de la personne que je souhaite devenir.

A ma famille, enfin, qui m’a toujours encouragée et soutenue dans mes choix,

A ma mère, merci pour ta bienveillance, ta gentillesse et ton amour. Tu es notre force.

A Laurent, mon père. Merci de ta pudeur, de ta présence à mes côtés, discrète, sans jamais s’imposer.

A Claire, ma sœur, mon modèle de force et de persévérance. Je t’aime.

A Benjamin, mon Benjamin… J’aurais tellement de choses à te dire mais tu les connais déjà. Suis tes envies, quoiqu’il arrive, quoique les gens te disent. Tu y arriveras, j’ai entièrement confiance en toi.

Scientific valorization

Thesis-related communications

Publications in peer-reviewed journals

i. **Savina M**, Le Cesne A, Blay J-Y, Ray-Coquard I, Mir O, Toulmonde M, Cousin S, Terrier P, Ranchere-Vince P, Meeus P, Stoeckle E, Honoré C, Sargos P, Sunyach M-P, Le Péchoux C, Giraud A, Bellera C, Le Loarer F, Italiano A. Patterns of care and outcomes of patients with METAstatic soft tissue SARComa in a real-life setting: the METASARC observational study. *BMC Medicine* (2017) 15:78. (Impact factor: 8.005)

ii. **Savina M**, Gourgou S, Italiano A, Dinart D, Rondeau V, Pénel N, Mathoulin-Pélissier S, Bellera C (2017). Meta-analyses evaluating surrogate endpoints for overall survival in cancer randomized trials: A critical review. *In press, Critical Reviews in Oncology / Hematology*. (Impact factor: 5.039)

iii. **Savina M**, Litière S, Italiano A, Burzykowski T, Bonnetain F, Gourgou S, Rondeau V, Blay J-Y, Cousin S, Duffaud F, Gelderblom H, Gronchi A, Judson I, Le Cesne A, Lorigan P, Maurel J, van der Graaf W, Verweij J, Mathoulin-Pélissier S, Bellera C (2017). Surrogate endpoints in advanced sarcoma trials: a meta-analysis. *Submitted, Journal of Clinical Oncology.* (Impact factor : 24.00)

iv. **Savina M**, Bellera C, Jacot W, Burzykowski T, Bonnetain F, Kerbrat P, Roché H, Spielmann M, Goldhirsch A, Paridaens R, Mathoulin-Pélisser S, Gourgou S (2017). Surrogate endpoints for Overall Survival in Adjuvant trials of breast cancer: A Meta-Analysis. *In preparation.*

Oral communications

i. **Savina M.** Modalités de traitement et survie des patients atteints de sarcome des tissus mous métastatiques : l'étude METASARC. **EPICLIN 11 / 24èmes Journées des Statisticiens de CLCC**, May 2017, Saint-Etienne, France.

ii. **Savina M.** Surrogate endpoints for overall survival in adjuvant breast cancer: a meta-analysis of randomized controlled trials. **Journées Scientifiques du SIRIC BRIO (Bordeaux Recherche Intégrée en Oncologie)**, November 2016, Bordeaux, France.

iii. **Savina M**, Laghzali Y, Mathoulin-Pélissier S, Bellera C, Gourgou S. Critères de substitution à la Survie Globale dans les Essais Cliniques Randomisés dans le contexte du cancer du sein en situation adjuvante : une méta-analyse de 5 ECR. **EPICLIN 10 / 23èmes Journées de Statisticiens de CLCC**, May 2016, Strasbourg, France.

iv. **Savina M**, Pénel N, Litière S, Rondeau V, Toulmonde M, Italiano A, Mathoulin-Pélissier S, Bellera C. Evaluation de critères de substitution à la Survie Globale dans les Essais Cliniques Randomisés (ECR) portant sur les sarcomes en situation métastatique : une méta-analyse de 10 ECR. **Journées du Cancéropôle Grand Sud-Ouest**, November 2015, Bordeaux, France.

v. **Savina M**, Litière S, Pénel N, Rondeau V, Burzykowski T, Toulmonde M, Italiano A, Mathoulin-Pélissier S, Bellera C. Surrogate properties of survival endpoints in metastatic soft-tissue sarcoma: a meta-analysis. **36th Annual Conference of the International Society for Clinical Biostatistics (ISCB)**, August 2015, Utrecht, the Netherlands.

vi. **Savina M**, Pénel N, Litière S, Rondeau V, Toulmonde M, Italiano A, Mathoulin-Pélissier S, Bellera C. Evaluation de critères de substitution à la Survie Globale dans les Essais Cliniques Randomisés (ECR) portant sur les sarcomes en situation métastatique : une méta-analyse de 10 ECR. **EPICLIN 9 / 22èmes Journées de Statisticiens de CLCC**, May 2015, Montpellier, France.

vii. **Savina M**, Laghzali Y, Vernerey D, Bellera C, Gourgou-Bourgade S, Bonnetain F. Hiérarchisation et validation de critères de survie en tant que critères substitutifs à la survie globale. **20èmes Journées de Statisticiens de CLCC**, June 2013, Marseille, France.

Invited seminars

i. **Savina M**, Laghzali Y, Mathoulin-Pélissier S, Bellera C, Gourgou S. Surrogate endpoints in cancer randomized controlled trials: Application to sarcoma and breast cancer. **6èmes Journées Annuelles du Club Statistiques et Mathématiques Appliquées à la Cancérologie (SMAC)**, May 2017, Montpellier, France.

Posters

i. **Savina M**, Jacot W, Mathoulin-Pélissier S, Laghzali Y, Bellera C, Gourgou S. Surrogate endpoints for Overall Survival in Randomized Controlled Trials evaluating adjuvant treatment for breast cancer: a Meta-analysis. **European Society for Medical Oncology (ESMO) 2017 Congress**, September 2017, Madrid, Spain.

ii. **Savina M.** Evaluation of surrogate endpoints for overall survival in cancer randomized controlled trials. **Journées Scientifiques du SIRIC BRIO (Bordeaux Recherche Intégrée en Oncologie)**, November 2016, Bordeaux, France.

iii. **Savina M**, Litière S, Pénel N, Rondeau V, Burzykowski T, Toulmonde M, Italiano A, Mathoulin-Pélissier S, Bellera C. Surrogate properties of survival endpoints in metastatic soft-tissue sarcoma: a meta-analysis. **Annual Meeting of the American Society of Clinical Oncology (ASCO)**, May-June 2015, Chicago, United States of America.

Other communications

Publications in peer-reviewed journals

i. Anota A, Barbieri A, **Savina M**, Pam A, Gourgou-Bourgade S, Bonnetain F, Bascoul-Mollevi C. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. *Health and Quality of Life Outcomes* (2014) 12:192.

ii. Palussière J, Chomy F, **Savina M**, Deschamps F, Gaubert JY, Laurent F, Meunier C, Renault A, Bonnefoy O, Bellera C, Mathoulin-Pélissier S, de Baere T. Radiofrequency Ablation of Stage IA Non–Small Cell Lung Cancer in patients ineligible for surgery : results of a prospective multicenter phase II trial. Submitted to *Journal of Thoracic Oncology*.

iii. Mathoulin-Pélissier S, Bui BN, Chevreau C, Bauvin S, Cupissol D, Lebrun-Ly V, Grosclaude P, Lacourt A, Coureau G, Molinié F, Bompas E, **Savina M**, Bellera C. Intervention Program for Improvement of the Initial Management of Soft Tissue Sarcomas: results of a Before-After Controlled Study in five French areas (IPSSAR study). *In preparation.*

Oral communications

i. Anota A, Barbieri A, **Savina M**, Gourgou-Bourgade S, Bonnetain F, Bascoul-Mollevi C. Impact de l'effet Response Shift sur l'analyse longitudinale de la qualité de vie relative à la santé dans les essais cliniques en cancérologie: une étude de simulation. **EPICLIN 9 / 22èmes Journées des Statisticiens de CLCC**, May 2015, Montpellier, France.

ii. Anota A, Barbieri A, **Savina M**, Pam A, Gourgou-Bourgade S, Bonnetain F, Bascoul-Mollevi C. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. **EPICLIN 8 / 21èmes Journées des Statisticiens de CLCC**, May 2014, Bordeaux, France.

iii. Anota A, Barbieri A, **Savina M**, Pam A, Gourgou-Bourgade S, Bonnetain F, Bascoul-Mollevi C. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. **Workshop « Evaluation et analyse de la qualité de vie : nouveaux développements méthodologiques »**, April 2014, Montpellier, France.

iv. Anota A, **Savina M**, Bascoul-Mollevi C, Bonnetain F. QoLR : un package R pour l'analyse longitudinale de la qualité de vie en cancérologie. **20èmes Journées des Statisticiens de CLCC**, June 2013, Marseille, France.

v. Anota A, **Savina M**, Bascoul-Mollevi C, Bonnetain F. QoLR: un package R pour l'analyse longitudinale de la qualité de vie en cancérologie. **EPICLIN 7 (Epidémiologie et Recherche Clinique)**, May 2013, Paris, France.

Posters

i. **Savina M**, Bui BN, Chevreau C, Bauvin S, Cupissol D, Lebrun-Ly V, Grosclaude P, Lacourt A, Coureau G, Molinié F, Bompas E, Bellera C, Mathoulin-Pélissier S. Intervention Program for Improvement of the Initial Management of Soft Tissue Sarcomas: results of a Before-After Controlled Study in five French areas (IPSSAR

study). (**Student award winner**). **EPICLIN 10 / 23èmes Journées de Statisticiens de CLCC**, May 2016, Strasbourg, France.

ii. Anota A, Barbieri A, **Savina M**, Gourgou-Bourgade S, Bonnetain F, Bascoul-Mollevi C. Impact of Response Shift effect on the longitudinal analysis of Health-related quality of life in oncology clinical trials: a simulation study. **22th Annual Conference of the International Society for Quality of Life Research (ISOQOL)**, October 2015, Vancouver, Canada.

iii. Anota A, Barbieri A, **Savina M**, Pam A, Ychou M, Gourgou-Bourgade S, Bonnetain F, Bascoul-Mollevi C. Comparison of three longitudinal analysis models for the health-related quality of life in oncology: a simulation study. **21th Annual Conference of the International Society for Quality of Life Research (ISOQOL)**, October 2014, Berlin, Germany.

iv. Anota A, **Savina M**, Bascoul-Mollevi C, Bonnetain F. QoLR: an R package for the longitudinal analysis of Health-related quality of life in oncology. **20th Annual Conference of the International Society for Quality of Life Research (ISOQOL)**, October 2013, Miami, USA.

Résumé substantiel en français

Introduction

Processus de développement des traitements anticancéreux

Le développement d'un médicament repose sur un processus long et onéreux. Les différentes étapes de ce processus ont pour objectif d'apporter la preuve de l'efficacité, ou de l'inefficacité, du traitement tout en assurant la sécurité des sujets.

Dans un premier temps, la phase préclinique, le produit thérapeutique est testé sur des cellules de culture (*in vitro*) et des systèmes vivants non humains (*in vivo*). Ces études préliminaires permettent de collecter des informations pharmacologiques, pharmacocinétiques et sur la toxicité afin de déterminer la dose qui sera administrée à l'homme par la suite. Le produit passe ensuite par les trois phases cliniques (phase I à III) à l'issue desquelles une éventuelle autorisation de mise sur le marché (AMM) est délivrée. En oncologie, les essais cliniques incluent essentiellement des patients atteints d'un cancer.

L'objectif principal des essais de phase I est de tester la relation dose / toxicité du traitement afin de déterminer la dose maximale tolérée par le patient, c'est-à-dire, la dose maximale n'entraînant pas de toxicités excessives. La phase II vise à évaluer l'activité anticancéreuse du produit thérapeutique, afin d'éliminer les molécules inefficaces. Si l'essai de phase II conclut à l'efficacité du traitement, il est ensuite comparé au traitement standard, ou à un placebo, dans un essai de phase III. Ces derniers ont pour objectif d'apporter la confirmation du bénéfice clinique du nouveau traitement par rapport au traitement standard. Les résultats des essais de phase III seront utilisés par les autorités de santé lors de la décision de délivrer ou non une AMM.

Le processus de développement d'un médicament est particulièrement réglementé. Peu de médicaments parviennent au terme de ce processus et obtiennent une AMM. Entre 1998 et 2008, seulement 6% des traitements anticancéreux testés en phase I ont obtenu une AMM, et plus de 50% des produits testés en phase II ont échoué en phase III, principalement par manque d'efficacité (1). Si cette proportion élevée d'échecs s'explique en partie par une absence réelle d'efficacité des nouvelles molécules, le choix du design d'étude ou du critère principal d'évaluation de l'efficacité peuvent également en être la cause. Le critère principal se doit d'être objectif et reproductible. Il détermine le nombre de patients nécessaires dans l'essai et ainsi les conclusions de l'essai. Dans les essais de phase III en cancérologie, l'objectif étant d'améliorer la survie des patients, la survie globale (SG), définie par le délai entre les dates de randomisation et de décès, est le critère validé par les autorités de santé dans le contexte des essais cliniques randomisés (ECR) (2,3).

Critères de substitution - Définitions

Un critère clinique est une variable qui reflète le statut du patient (qualité de vie ou survie). Un biomarqueur est une variable objective, indicatrice d'un changement biologique, pathogénique ou pharmacologique lié à une intervention thérapeutique. On peut citer par exemple la survie sans progression (progression-free survival – PFS), définie par le délai entre les dates de randomisation et de progression ou décès. Un critère de substitution est un biomarqueur dont l'objectif est de remplacer un critère clinique(4). Selon l'International Conference on Harmonization Guidelines on Statistical Principles for Clinical Trials(4), un critère de substitution doit satisfaire trois conditions : (i) être biologiquement pertinent, (ii) au niveau du patient, être associé et permettre de prédire le critère final (association au niveau individuel), et (iii) au niveau de l'essai, l'effet traitement sur le critère doit être associé et permettre de prédire l'effet traitement sur le critère final, en l'occurrence la SG dans les ECR en cancérologie (association au niveau de l'essai).

Critères de substitution et Autorités de Santé

L'utilisation des critères de substitution est guidée par le besoin de réduire le nombre de patients, la durée des essais, et à terme les délais de mise sur le marché des nouveaux traitements et le coût global des ECR. Si la SG est la référence pour mesurer l'efficacité d'un traitement dans les ECR en oncologie (2,3), son utilisation comme critère principal présente certaines limites : effectifs de patients importants, problématique des « cross-over » et des lignes de traitements successives, inclusions des décès toutes causes, etc. Dans ce contexte, des critères alternatifs de mesure du bénéfice clinique sont fréquemment utilisés. Ces critères incluent des évènements cliniques et/ou biologiques, tels que la progression de la maladie ou la toxicité du traitement, observés plus fréquemment et plus précocement que le décès. Ces critères composites, communément utilisés dans les essais de phase II, sont de plus en plus utilisés comme critère principal dans les essais de phase III (5). Dans ce contexte, les autorités de santé ont adapté leurs recommandations sur l'évaluation des traitements permettant ainsi de délivrer des AMM sur la base de critères d'évaluation autres que la SG. L'Agence Européenne du Médicament (AEM) et la Food and Drug Administration (FDA) aux Etats-Unis proposent des processus d'approbation conditionnels, ou accélérés, basés sur des critères autres que la SG. Ces AMM conditionnelles ont pour objectif, dans les cas de maladies avec un pronostic sévère, d'accélérer la mise à disposition de traitements efficaces constituant un besoin médial non résolu ("unmet medical need"), c'est-à-dire, pour lesquels les options thérapeutiques sont particulièrement limitées, voire inexistantes. Un nombre important d'AMM de traitements anticancéreux sont ainsi délivrées sur la base de critères alternatifs alors qu'aucune étude n'a permis de démontrer que ceux-ci avaient été formellement validés en tant que critère substitutif à la SG (6). De plus, bien qu'une

confirmation du bénéfice sur la SG soit requise à la suite de ces AMM conditionnelles, en pratique, ces études confirmatoires sont rarement menées (7). L'utilisation inappropriée de critères de substitution peut mener à l'AMM de traitements inefficaces sur la SG, tel que le bevacizumab pour le traitement du cancer du sein métastatique. En 2008, la FDA a ainsi délivré une AMM conditionnelle suite à un ECR ayant démontré une amélioration de la PFS (8). Trois ans plus tard, deux ECR ont cependant conclu à un effet très modeste sur la PFS ainsi qu'à une absence d'effet sur la SG (9,10). Ces résultats ont conduit au retrait de l'AMM du bevacizumab dans cette indication (11). L'évaluation en amont des propriétés de substitution des critères d'évaluation de l'efficacité alternatifs à la SG est ainsi primordiale pour s'assurer de mesurer précisément le bénéfice des nouveaux traitements.

Evaluation statistique des critères de substitution

L'évaluation statistique des critères de substitution repose sur l'estimation de deux types d'association : (i) l'association au niveau individuel, à savoir, le critère de substitution permet de prédire le critère final au niveau du patient, et (ii) l'association au niveau de l'essai, à savoir, l'effet traitement sur le critère de substitution permet de prédire l'effet traitement sur le critère final au niveau de l'essai. Depuis la publication en 1989 par Prentice des quatre critères opérationnels nécessaires à la validation statistique d'un critère de substitution (12), de nombreux travaux statistiques concernant l'évaluation des critères de substitution (13) ont été publiés. Les méthodes statistiques peuvent être classifiées en deux catégories, selon que celles-ci soient basées sur l'analyse d'un ECR unique (l'approche « single-trial »), ou de plusieurs essais (l'approche méta-analytique). Les approches « single-trial » incluent les critères de Prentice et coll. (12), la proportion de traitement expliqué de Freedman (14), l'association ajustée et l'effet relatif développés par Buyse et Molenberghs (15). Bien qu'opérationnellement simples à mettre en œuvre, ces techniques permettent uniquement d'évaluer l'association individuelle entre le critère de substitution candidat et le critère d'intérêt final, tel que la SG.

L'approche méta-analytique permet d'estimer l'association entre les effets du traitement sur le critère de substitution candidat et la survie globale. Parmi les développements proposés, l'approche en deux étapes développée par Burzykowski, Buyse et Molenberghs est considérée comme la plus rigoureuse (13). Celle-ci repose sur la modélisation conjointe des deux critères, le critère de substitution candidat, et le critère final. Lorsque l'on s'intéresse à deux critères de survie, cette modélisation conjointe repose sur une fonction de copule. Dans un premier temps, la modélisation permet l'estimation de l'association entre les critères au niveau individuel. Pour chaque essai, les effets du traitement sur le critère de substitution et sur la SG sont estimés en tenant compte de leur association. Dans la seconde étape, un modèle mixte permet d'estimer l'association au niveau de l'essai, c'est-à-dire l'association

entre les effets du traitement sur le critère de substitution et sur le critère final, permettant de prendre en compte les erreurs d'estimation des effets traitements. Ce modèle requiert la disponibilité de données individuelles. Des modèles simplifiés à partir de données agrégées (telles que publiées dans la littérature) ont été proposés pour l'estimation de l'association au niveau de l'essai. Dans ce contexte, une régression linéaire pondérée de l'effet traitement sur le critère final sur l'effet traitement sur le critère de substitution permet l'estimation de l'association au niveau de l'essai.

Enfin, l'effet seuil du critère de substitution (« surrogate threshold effect » - STE) a par la suite été développé (16). Ce paramètre permet d'estimer l'effet du traitement minimal à observer sur le critère de substitution afin de prédire un effet du traitement significatif sur la SG. Le STE apporte une information directe sur l'utilité du critère de substitution dans la pratique courante.

Evaluation de la force de l'association

A ce jour, plusieurs méthodes statistiques permettant l'estimation de différents paramètres sont utilisées dans les études évaluant des critères de substitution. Afin de guider les chercheurs dans la conduite et l'interprétation de ces études, des grilles ont été développées afin d'évaluer la force des associations rapportées. La grille de Taylor et Elston repose sur un système hiérarchique du niveau de preuve (17) : (i) niveau 1 : preuve de l'association au niveau de l'essai niveau, (ii) niveau 2 : preuve de l'association au niveau individuel, (iii) niveau 3 : preuve biologique. Cette grille permet une classification simple des critères de substitution, mais ignore cependant les paramètres statistiques spécifiques couramment rapportés. L'Institute of Quality and Efficiency in Health Care (IQWiG) allemand a proposé un schéma d'évaluation reposant sur (i) la qualité de la méthodologie employée et (ii) la force de l'association au niveau de l'essai rapportée (18). Alors que certains des éléments inclus dans l'évaluation de la qualité de la méthodologie peuvent être subjectifs. L'évaluation de la force de l'association au niveau de l'essai repose cependant sur des seuils précis. Enfin, le Biomarker-Surrogate Evaluation Schema (BSES) évalue quatre domaines : le schéma de l'étude de validation, le critère final utilisé, la méthode statistique et l'extrapolation des résultats (19). Contrairement à la grille IQWiG, le BSES évalue la force de l'association globale en se basant non seulement sur la mesure de la force de l'association au niveau de l'essai, mais également sur celle au niveau individuel, ainsi que le STE. Le BSES ne recommande cependant aucune méthode statistique pour l'estimation de ces paramètres. L'évaluation de la qualité du design de l'étude et l'extrapolation des résultats peuvent cependant être subjectives.

Objectifs

Un nombre croissant d'études sont menées afin d'évaluer des critères de substitution dans les ECR en cancérologie. Avec la multiplicité des méthodes d'évaluation disponibles et la difficulté à généraliser la validation d'un critère à différentes situations thérapeutiques, il est nécessaire de dresser un état des connaissances récent et exhaustif des critères de substitution disponibles pour les ECR en cancérologie. **Le premier objectif de cette thèse était d'identifier les méta-analyses menées dans le cadre de l'évaluation de critères de substitution à la SG en cancérologie par le biais d'une revue systématique de la littérature et d'évaluer le niveau de preuve apporté par chaque méta-analyse.**

Suite à cette revue, nous avons identifié deux situations thérapeutiques pour lesquelles la mise à disposition d'un critère de substitution permettrait d'améliorer le développement de futurs ECR : les sarcomes des tissus mous (STM) en situation avancée et le cancer du sein en situation adjuvante.

Notre second objectif était donc d'évaluer les propriétés de substitution à la SG de différents critères de survie dans le cadre des STM en situation avancée, à partir d'une méta-analyse de 14 ECR. Nous nous sommes intéressés à trois critères candidats : la PFS, le temps jusqu'à progression et le temps jusqu'à échec du traitement. Parallèlement, nous avons étudié les propriétés d'un autre critère qui suscite beaucoup d'intérêt auprès des oncologues, le *temps jusqu'au traitement suivant,* à partir d'une cohorte prospective. En effet, nous ne pouvions pas reconstituer ce critère à partir des essais disponibles dans le cadre de notre méta-analyse.

Enfin, notre dernier objectif concernait l'évaluation des propriétés de substitution à la SG de différents critères de survie dans le cadre du cancer du sein en situation adjuvante, à partir d'une analyse groupée de cinq ECR. Nous nous sommes intéressés à quatre critères candidats : la survie sans rechute, la survie sans maladie invasive, la survie sans rechute locorégionale, et la survie sans maladie à distance.

Etat de l'art : Critères de substitution validés dans les essais cliniques randomisés en cancérologie

Contexte et objectif

En oncologie, des critères d'évaluation de l'efficacité alternatifs à la SG sont de plus en plus communément utilisés à la place de la SG dans les ECR (5). Ces critères incluent différents types d'évènements autres que le décès, et ont ainsi l'avantage d'être observés plus fréquemment et plus précocement. On peut citer par exemple la PFS, définie précédemment. Leur utilisation est motivée par la nécessité de réduire le nombre de patients

à inclure, la durée et le coût des ECR. Leur utilisation nécessite cependant une évaluation rigoureuse afin de les valider comme critères de substitution à la SG. La qualité d'un critère de substitution dépend de la maladie considérée ainsi que du mécanisme d'action du traitement étudié, leur évaluation ne peut ainsi se faire que dans une situation thérapeutique précise. Ce travail avait pour objectif de dresser un état des lieux des critères de substitution validés en cancérologie.

Méthodes

Nous avons conduit une revue systématique de la littérature à partir des bases de données MEDLINE et SCOPUS afin d'identifier les méta-analyses évaluant des critères de substitution à la SG dans les ECR en oncologie. Les publications pertinentes ont été sélectionnées en suivant un procédé en deux étapes (sélection sur résumé puis sur article complet), à partir d'une grille de lecture standardisée, et par deux lecteurs indépendants. La force de l'association au niveau de l'essai a été évaluée à partir des critères IQWiG et BSES.

Résultats

Un total de 53 publications présentant 164 méta-analyses a été inclus dans cette revue de la littérature. La majorité des méta-analyses portaient sur l'évaluation de critères de substitution en situation avancée, essentiellement dans le contexte du cancer colorectal, du poumon et du sein. Les méthodologies employées pour l'estimation des associations au niveau individuel ou de l'essai étaient hétérogènes, et 17% ont utilisé la méthode en deux étapes de Buyse et Burzykowski. De même, il existait une forte variabilité dans les paramètres d'association rapportés, et pour lesquels le degré de précision (intervalles de confiance) n'était pas systématiquement rapporté. Ce dernier point nous a limités lors de l'application des grilles IQWiG et BSES pour juger de la qualité des études publiées. En situation adjuvante, plusieurs méta-analyses suggéraient une forte association au niveau de l'essai entre la survie sans maladie (disease-free survival - DFS) et la SG dans le cancer du côlon, du cancer du poumon non-à-petites-cellules, du cancer gastrique et des cancers oto-rhino-laryngologiques (ORL). En situation métastatique, les associations au niveau de l'essai entre la PFS et la SG étaient élevées pour le cancer colorectal, le cancer du poumon et les cancers ORL.

Conclusion

Malgré un nombre croissant d'études portant sur l'évaluation de critères de substitution en oncologie, un nombre limité de méta-analyses reposant sur une méthodologie statistique rigoureuse présentent des associations suffisamment élevées pour conclure à la validité du critère de substitution. Un certain nombre d'informations ne sont cependant pas rapportées (intervalles de confiance) par les publications, limitant ainsi l'évaluation de la qualité de ces

études à partir des grilles actuelles. En conclusion, les données actuellement disponibles suggèrent un nombre limité de situations thérapeutiques avec un critère de substitution validé : DFS en situation adjuvante pour le cancer du côlon, du cancer du poumon non-à-petites-cellules, du cancer gastrique et des cancers ORL ; PFS en situation métastatique pour le cancer colorectal, le cancer du poumon et les cancers ORL. Les associations estimées dans d'autres cadres thérapeutiques et/ou pour les autres critères évalués étaient trop faibles ou imprécises pour conclure.

Ce travail a été accepté pour publication dans le journal *Critical Reviews in Oncology / Hematology*.

Critères de substitution dans les essais cliniques randomisés portant sur les sarcomes des tissus mous en situation avancée

Contexte et objectif

Très hétérogènes, les STM se composent de plus de 50 sous-types histologiques, ce qui les rend particulièrement complexes à diagnostiquer, à étudier, et à guérir. Les STM représentent environ 1% des cancers de l'adulte en France (20). L'hétérogénéité, associée à une faible incidence, limite le développement de nouveaux traitements thérapeutiques dans cette indication. La validation d'un critère de substitution qui serait disponible plus précocement que la SG, et sur de plus grands effectifs, permettrait de réduire les effectifs et la durée des ECRs. A ce jour cependant, aucune méta-analyse sur données individuelles évaluant des critères de substitution à la SG dans cette situation thérapeutique n'a été publiée. Notre objectif était d'évaluer, à partir d'une méta-analyse sur données individuelles, les propriétés de substitution de trois critères communément utilisés dans cette indication: la PFS, le temps jusqu'à progression (time-to-progression – TTP) et le temps jusqu'à échec du traitement (time-to treatment failure – TTF).

Méthodes

Les essais inclus dans la méta-analyse ont été identifiés à partir d'une revue systématique de la littérature, du registre des essais ClinicalTrials.gov et en contactant les principaux groupes promoteurs français et européens (EORTC et UNICANCER). Dans un premier temps, les données des différents ECR ont été uniformisées. Pour chaque patient, les critères de survie, SG, PFS, TTP, et TTF, ont été recalculés selon les recommandations internationales pour la définition des critères de survie pour les ECR portant sur la SG après un suivi de 18 mois.

Les associations au niveau individuel et au niveau de l'essai ont été estimées en suivant le modèle en deux étapes de Burzykowski, Buyse et Molenberghs (13) et le modèle de

régression linéaire pondérée. La force des associations a été évaluée à partir de la grille IQWiG (18). Deux analyses en sous-groupes ont été menées. Dans la première analyse, nous avons uniquement inclus les essais comparant des traitements systémiques à des chimiothérapies à base de doxorubicin ou ifosfamide en première ligne de traitement. Dans la seconde analyse, seuls les patients atteints d'un leiomyosarcome, un des sous-types histologiques de STM les plus courants, ont été inclus. Enfin, nous avons conduit différentes analyses de sensibilité afin d'évaluer la robustesse des modèles de régression linéaire pondérée (variation des règles de censure et de pondération).

Résultats

Nous avons recueilli les données individuelles de 14 ECR (2846 patients) évaluant des traitements systémiques pour des patients adultes atteints d'un STM métastatique. Au niveau individuel, les trois critères de substitution candidats étaient modérément associés à la SG. Au niveau de l'essai, les associations estimées étaient faibles et manquaient de précision. Elles ont ainsi été classées « moyennes » selon les critères IQWiG.

Conclusion

Les résultats de cette méta-analyse n'ont pas permis de valider un critère de substitution à la SG dans le cadre des STM en situation avancée. Notre étude présentait cependant certaines limites liées en partie à la nature même des STM (hétérogénéité des sous-types histologiques, faibles effectifs en termes d'ECR). Le critère principal des ECR dans l'évaluation des traitements dans le cadre des STM en situation métastatique doit rester la SG.

Ce travail a été soumis pour publication auprès du *Journal of Clinical Oncology*.

Temps jusqu'au prochain traitement: un critère alternatif pour les essais cliniques randomisés portant sur les sarcomes des tissus mous en situation avancée?

Contexte et objectif

A ce jour, aucun critère de substitution à la SG n'a été validé dans le contexte des STM en situation avancée. Alors que la PFS est cependant fréquemment utilisé, celle-ci présente certaines limites dans les ECR : la PFS repose sur l'imagerie radiologique, dont la lecture et l'interprétation peuvent être subjectives (22) ; les critères définissant les seuils de progression (e.g. RECIST (23,24)) sont en constante évolution, voire remis en question dans le contexte des nouvelles thérapies ; la date exacte de progression reste inconnue même si on se réfère généralement à la date de visite comme date d'approximation.

Le temps jusqu'au traitement suivant (time-to-next treatment – TNT) est un critère d'évaluation de l'efficacité utilisé dans certaines pathologies en hématologie et dans certains cancers. Défini comme le temps jusqu'à initiation d'un nouveau traitement après échec du traitement à l'étude, le TNT inclut toutes les causes possibles de changement de traitement. Ce critère reflèterait ainsi une altération de l'état du patient, quelle qu'elle soit, et serait donc un marqueur éventuel de la SG. Dans ce contexte, l'objectif de ce travail exploratoire était d'évaluer l'association, au niveau individuel, entre le TNT et la SG.

Méthodes

Les informations concernant les traitements donnés hors essai après échec du traitement à l'étude ne sont, à ce jour, pas recueillies de manière systématique dans le cadre d'ECR. Aussi, nous nous sommes basés sur une cohorte nationale prospective de patients atteints d'un STM métastatique. L'association au niveau individuel entre le TNT et la SG a été estimée pour différentes lignes de traitement. Un modèle de copule a été appliqué afin d'estimer un coefficient de rang de Spearman entre les deux critères de survie.

Résultats

Le TNT et la SG étaient fortement associés au niveau individuel, plus particulièrement en première ligne de traitement (R^2_{ind} = 0.76 IC95% [0.73 ; 0.78]).

Conclusion

Ce travail exploratoire a permis d'estimer l'association au niveau individuel. Ce travail fournit ainsi des résultats prometteurs. Le TNT mériterait une évaluation plus approfondie en tant que critère de substitution à la SG, sur la base d'une méta-analyse d'ECR, afin de permettre une estimation de l'association entre les effets du traitement sur le TNT et sur la SG.

Les résultats de ce travail ont fait l'objet d'une publication dans *BMC Medicine* (Savina et al. 2017).

Critères de substitution dans les essais cliniques randomisés portant sur le cancer du sein en situation adjuvante

Contexte et objectif

Malgré une incidence croissante, la survie des patientes atteintes du cancer du sein s'est significativement améliorée depuis 30 ans. En pratique, les ECR dans cette indication impliquent donc l'inclusion d'un nombre important de patients et un long suivi afin d'observer le bénéfice d'un nouveau traitement sur la SG. Un critère de substitution observé plus fréquemment et surtout plus précocement que la SG serait donc un atout important. L'objectif de ce travail était d'évaluer les propriétés de substitution à la SG de quatre critères de survie

dans le cancer du sein en situation adjuvante : la survie sans rechute (RFS), la survie sans maladie invasive (invasive disease-free survival - iDFS), la survie sans rechute locorégionale (locoregional relapse-free survival - LRFS) et la survie sans maladie à distance (distant disease-free survival - DDFS).

Méthodes

Nous avons mené une analyse groupée de cinq ECR évaluant des chimiothérapies, seules ou en combinaison avec une hormonothérapie, comme traitement adjuvant du cancer du sein. Dans un premier temps, les données des différents ECR ont été uniformisées. Pour chaque patient, les critères de survie, SG, RFS, iDFS, LRFS et DDFS, ont été recalculés selon les recommandations internationales pour la définition des critères de survie pour les ECR portant sur le cancer du sein (25,26). La RFS, iDFS, LRFS et DDFS ont été censurés après un suivi de cinq ans, et l'OS après un suivi de sept ans.

Les associations au niveau individuel et au niveau de l'essai, respectivement R^2_{ind} et R^2_{trial}, ont été estimées en suivant le modèle en deux étapes de Burzykowski, Buyse et Molenberghs (2SM) (13) et le modèle de régression linéaire pondérée par la taille de l'essai. Etant donné le nombre limité d'ECR, nous nous sommes basés sur le centre participant plutôt que sur l'ECR afin d'estimer l'association au niveau de l'essai. La force des associations a été évaluée à partir de la grille IQWiG (18). Nous avons testé la capacité de prédiction des modèles de régression linéaire par une méthode de validation croisée.

Résultats

Les critères étudiés ont révélé une forte association avec la SG, que ce soit au niveau individuel ou de l'essai. Sur la base du critère IQWiG, la LRFS à cinq ans a été classée fortement associée à la SG à sept ans. Pour les autres critères, les associations ont été classées « moyenne ». Les associations au niveau de l'essai avec des temps de suivi plus courts étaient modérées à élevées, avec des intervalles de confiance très larges.

Conclusion

Les résultats de cette analyse suggèrent que la LRFS à cinq ans est un critère de substitution raisonnable pour la SG à sept ans. Cette conclusion devrait cependant être modérée par les limites méthodologiques auxquelles nous avons été confrontés. En effet, l'utilisation du centre participant plutôt que l'essai comme unité d'analyse a probablement artificiellement augmenté la précision des estimations des associations au niveau essai. Cette analyse est la première menée sur données individuelles pour évaluer des critères de substitution dans le cancer du sein en situation adjuvante. Ces résultats prometteurs seront à confirmer à partir d'une méta-analyse conduite sur un effectif plus important d'ECR.

Un article présentant les résultats de cette analyse est en cours de relecture auprès des co-auteurs pour soumission au *Journal of Clinical Oncology*.

Conclusion générale et perspectives

L'utilisation des critères de substitution, disponibles plus rapidement et sur des effectifs plus faibles que la SG, est motivée par la nécessité d'accélérer le processus de développement des molécules. Malgré un nombre croissant de méta-analyses évaluant les critères de substitution à la SG, un faible nombre de critères ont cependant été formellement validés comme tel. Les données actuellement disponibles suggèrent un nombre limité de situation thérapeutique avec un critère de substitution validé : DFS en situation adjuvante pour le cancer du côlon, du cancer du poumon non-à-petites-cellules, du cancer gastrique et des cancers ORL ; PFS en en situation métastatique pour le cancer colorectal, le cancer du poumon et les cancers ORL.

Notre méta-analyse a démontré que le critère principal des ECR portant sur l'évaluation des traitements dans le cadre des STM en situation métastatique doit rester la SG, même si la PFS, le TTP et le TTF restent pertinents dans le cadre d'étude de futilité et/ou d'essais de phase II. Le TNT, que nous n'avons pu évaluer que via une cohorte prospective, présente une forte corrélation au niveau individuel avec la SG.

Dans le cadre du cancer du sein en situation adjuvante, les résultats de notre analyse groupée de cinq essais suggèrent de bonnes associations entre RFS, iDFS, LRFS et DDFS et la SG. Cependant, l'utilisation des centres participants plutôt que les essais comme unité d'analyse pour l'estimation de l'association au niveau de l'essai a probablement artificiellement réduit les intervalles de confiance associés à nos estimations. Ces résultats prometteurs seront à confirmer à partir d'une méta-analyse conduite sur un effectif plus important d'ECR.

Le nombre limité de critères de substitution validés en cancérologie peut s'expliquer par de multiples facteurs. Premièrement, la plupart des méta-analyses publiées repose sur un fragment des données disponibles, même lorsque les auteurs tentent d'inclure exhaustivement tous les ECR éligibles (27). Or les données publiées diffèrent significativement des données non publiées et cette absence d'exhaustivité est susceptible de biaiser les résultats. Deuxièmement, les méthodes statistiques pour l'évaluation des critères de substitution sont multiples, complexes et pour la plupart requièrent une quantité importante de données, tant en terme d'effectif de patients que d'ECR. Troisièmement, l'absence de consensus en termes (i) de paramètres à estimer pour mesurer la capacité de substitution, (ii) de méthodologie statistique pour le calcul de ces paramètres, et (iii) de seuil de validation pour ces paramètres, est une limite importante à la conduite de méta-analyses

rigoureuse et ainsi à la validation des critères de substitution. Enfin, la validation d'un critère de substitution reste spécifique à une certaine indication : validation au sein d'une population atteinte d'un « même » cancer et traité par des molécules au mécanisme d'action identique. Comme nous l'avons illustré avec les sarcomes, ceux-ci sont constitués de près de 50 sous-types histologiques. Il peut être particulièrement complexe d'établir avec certitude qu'une même molécule a le même mécanisme d'action quelque soit le sous-type histologique. Les méta-analyses nécessaires à la validation des marqueurs de substitution sont alors conduites sur des populations plus ou moins hétérogènes, « diluant » ainsi les associations étudiées. Cette remarque s'applique bien entendu également dans le cadre de l'hétérogénéité des traitements.

Les études basées sur des critères de substitution en cancérologie sont susceptibles d'être basées sur effectifs moindres, d'être plus rapides et moins coûteuses que des études basées sur la SG. De fait, l'utilisation de ces critères est indéniablement attrayante. Cette attractivité devrait augmenter au cours des prochaines années, notamment grâce aux avancées en biologie cellulaire et moléculaire. Celles-ci génèrent de nouveaux traitements nécessitant des tests ainsi que des nouveaux marqueurs qui pourraient servir de critères de substitution. Les études de validation sur les critères de substitution, même si celles-ci sont pour la plupart « négatives », restent des sources d'information importantes. En effet, certains critères, même si non validés en tant que critères de substitution à la SG, continuent à jouer un rôle légitime dans les études de phase II en permettant d'indiquer des premiers signes d'activité des traitements étudiés, ou dans des essais de phase III dans le cadre d'analyses intermédiaires.

Notations and abbreviations

Notations

- $f(.)$ denotes the density function, or density function, of a variable
- $f(.|.)$ denotes the density function of a variable conditional to one or more other variables
- $F(.)$ denotes the distribution function of a variable
- $P(.)$ denotes the probability of an event
- $\int .$ denotes the integral of a function
- $\sum .$ denotes a sum of values
- $var\ (.)$ denotes the variance of a variable
- $E[.]$ denotes the expectation of a variable
- $e^{\cdot}$ Denotes the exponential function
- .|. denotes a variable conditional to one or more other variables
- $\hat{.}$ denotes the estimation of a variable
- $\ddot{.}$ denotes the prediction of a variable
- $i = 1, \dots, n$ denotes the trial

  $j = 1, \dots, n_i$ denotes the patient in trial i
- T is the final – or true – outcome, with T_j the observed value for patient j and T_{ij} the observed value for patient j of trial i

  S is the surrogate endpoint, with S_j the observed value for patient j and S_{ij} the observed value for patient j of trial i

  X is a categorical variable identifying the treatment received, with X_j the observed value for patient j and X_{ij} the observed value for patient j of trial i
- R^2_{ind} is the individual-level association

  R^2_{trial} is the trial-level association
- μ is the intercept of the regression model of T on S

  μ_S is the intercept of the regression model of S on X and μ_{Si} the intercept specific to trial i

  μ_T is the intercept of the regression model of T on X and μ_{Ti} the intercept specific to trial i

  $\tilde{\mu}_T$ is the intercept of the regression model of T on X and S
- m_{Si} is the random intercept in the regression model of S on X for trial i

  m_{Ti} is the random intercept in the regression model of T on X for trial i
- γ is the effect of T on S

  α is the effect of X on S, with α_i the effect of X on S specific to trial i

  β is the effect of X on T, with β_i the effect of X on T specific to trial i

- a_i is the random effect of S on X for trial i
- b_i is the random effect of T on X for trial i
- β_S is the effect of X on T adjusted for S

  γ_X is the effect of S on T adjusted for X
- ε_j is the random error term of patient j in the regression model of T on S

  ε_{Sj} is the random error term of patient j in the regression model of S on X and ε_{Sij} is the random error term of patient j of trial i in the regression model of S on X

  ε_{Tj} is the random error term of patient j in the regression model of T on X and ε_{Tij} is the random error term of patient j of trial i in the regression model of T on X

  $\tilde{\varepsilon}_{Tj}$ is the random error term of patient j in the regression model of T on X and S
- Σ is the covariance matrix of the random vector $(\varepsilon_{Sj}, \varepsilon_{Tj})$ with
  - σ_{SS} the variance of ε_{Sj}
  - σ_{TT} the variance of ε_{Tj}
  - σ_{ST} the covariance between ε_{Sj} and ε_{Tj}
- $\tilde{\Sigma}$ is the covariance matrix of the random vector (a_i, b_i) with
  - d_{aa} the variance of a_i
  - d_{bb} the variance of b_i
  - d_{ab} the covariance of a_i and b_i
- D is the covariance matrix of the random vector $(m_{Si}, m_{Ti}, a_i, b_i)$ with
  - d_{SS} the variance of m_{Si}
  - d_{TT} the variance of m_{Ti}
  - d_{aa} the variance of a_i
  - d_{bb} the variance of b_i
  - d_{ST} the covariance of m_{Si} and m_{Ti}
  - d_{Sa} the covariance of m_{Si} and a_i
  - d_{Sb} the covariance of m_{Si} and b_i
  - d_{Ta} the covariance of m_{Ti} and a_i
  - d_{Tb} the covariance of m_{Ti} and b_i
  - d_{ab} the covariance of a_i and b_i
- $C_\theta(.)$ or $C_\theta\{.\}$ is a single-parameter copula function with θ the copula parameter
- λ_{Sij} is the hazard function of S for patient j of trial i with λ_{Si} the baseline hazard function for trial i

  λ_{Tij} is the hazard function of T for patient j of trial i with λ_{Ti} the baseline hazard function for trial i

Abbreviations

AE: Adverse Event

BC: Breast Cancer

BSES: Biomarker-Surrogate Evaluation Schema

CI: Confidence Interval

DDFS: Distant Disease-Free Survival

DFS: Disease-Free Survival

DLT: Dose-Limiting Toxicity

EMA: European Medicine Agency

FDA: Food and Drug Administration

HR: Hazard Ratio

HRA: Health Regulatory Authorities

iDFS: Invasive Disease-Free Survival

IPD: Individual-Patient Data

IQWiG: Institute for Quality and Efficiency in Health Care

LRFS: Locoregional Relapse-Free Survival

MA: Meta-Analysis

MTD: Maximum Tolerated Dose

NCI: National Cancer Institute

NCI CTC-AE: National Cancer Institute Common Terminology Criteria for Adverse Events

OS: Overall Survival

PFS: Progression-Free Survival

PTE: Proportion of Treatment Explained

QoL: Quality of Life

RCC: Renal-Cell Carcinoma

RCT: Randomized Controlled Trial

RE: Related Effect

RECIST: Response Evaluation Criteria in Solid Tumors

RFS: Relapse-Free Survival

RP2D: Recommended Phase II Dose

STE: Surrogate Threshold Effect

STEP: Surrogate Threshold Effect Proportion

STS: Soft-Tissue Sarcoma

TNT: Time-to-Next Treatment

TTP: Time-To Progression

USA: United States of America

Table of Contents

Acknowledgments 3

Scientific valorization 6

Résumé substantiel en français 10

Notations and abbreviations 22

List of Tables 28

List of Figures 28

1 Introduction 29

1.1 Development process for anti-cancer drugs 29

1.1.1 The different stages of drug development 29

1.1.2 Failures in the drug development process 31

1.2 Surrogate endpoints 32

1.2.1 Definitions 32

1.2.2 Surrogate endpoints in regulatory settings 33

1.2.2.1 The European Medicines Agency (EMA) 34

1.2.2.2 The Food and Drug Administration (FDA) 34

1.2.2.3 Increasing use of surrogate endpoints 36

1.2.2.4 Misuse of surrogate endpoints: an illustration 36

1.3 Statistical evaluation of surrogate endpoints 37

1.3.1 Single-trial methods for the assessment of surrogate endpoints 38

1.3.1.1 Prentice’s definition and operational criteria 38

1.3.1.2 Freedman’s proportion of treatment effect 40

1.3.1.3 Related effect and adjusted association 40

1.3.2 The meta-analytic approach 41

1.3.2.1 The two-stage model 42

1.3.2.2 The simplified models 45

1.3.2.3 The three-level model 48

1.3.2.4 The surrogate threshold effect 49

1.3.2.5 Summary of methods for the statistical evaluation of surrogate endpoints 50

1.3.3 Quality of validation studies on surrogate endpoints and strength of evidence 51
1.4 Synthesis and research objectives 54
2 Meta-analyses evaluating surrogate endpoints for overall survival in cancer randomized trials: a critical review 56
2.1 Introduction 56
2.2 Publication 56
3 Surrogate endpoints in metastatic soft-tissue sarcoma trials 92
3.1 Introduction 92
3.2 Publication 93
3.3 Supplementary analyses 114
3.3.1 Sensitivity analysis: Censoring process 114
3.3.2 Sensitivity analysis: Weighting process 115
4 Time-to-next treatment: an alternative endpoint in advanced sarcoma trials? 116
4.1 Introduction 116
4.2 Publication 116
5 Surrogate endpoints in adjuvant breast cancer trials 128
5.1 Introduction 128
5.2 Publication 128
6 General discussion 145
6.1 Conclusion on the thesis work 145
6.2 Critical insight and perspectives 147
References 149
Appendices 154
Appendix A: The German Institute of Quality and Efficiency in Health Care (IQWiG) decision tree for the overall conclusion on the surrogate validity (18) 154
Appendix B: Biomarker-Surrogacy (BioSurrogate) Evaluation Schema (19) 155

List of Tables

Table 1: Surrogate-based approvals for which subsequent trials report an OS benefit or a lack of survival benefit or for which no trials exist showing or refuting a survival benefit (7) ..35

Table 2: Individual- and trial-level associations with 2-year OS (N_{trial} = 15; $N_{patient}$ = 2846) – Data censored at a cut-off date 114

Table 3: Trial-level associations with 2-year OS estimated by meta-regression with different weighting approaches.....115

List of Figures

Figure 1: Drug development process.....29

Figure 2: Representation of the trial-level surrogacy.....33

Figure 3: Illustration of individual and trial-level associations (36).....38

Figure 4: Correlation between treatment effects on PFS (log(HR_{PFS}), X-axis) and on OS (log(HR_{OS}), Y-axis) in 18 trials 47

Figure 5: Correlation between treatment effects on progression-free and on overall survival in historical trials (circles), in irinotecan trials (squares), and in oxaliplatin trial (diamond). A logarithmic scale is used for both axes 50

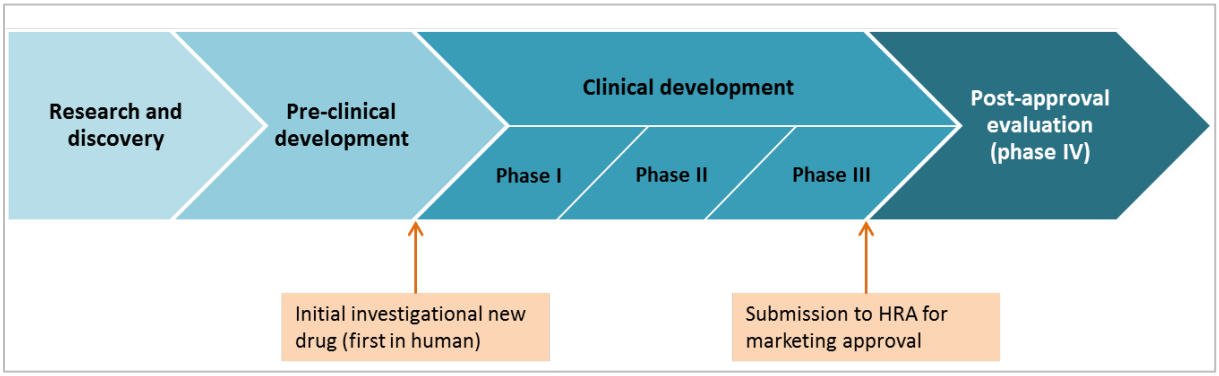
1 Introduction

1.1 Development process for anti-cancer drugs

1.1.1 The different stages of drug development

The development of a new therapeutic product is a long and expensive process that has to pass through multiple stages to provide sufficient evidence of the drug efficacy while insuring safety (28) (Figure 1).

Figure 1: Drug development process. HRA = health regulatory authorities.



The drug development process can be broadly classified as pre-clinical and clinical. Pre-clinical refers to experimentation that occurs before it is given to human subjects, and clinical refers to experimentation with humans.

Pre-clinical studies are the first steps of drug development. They involve testing for biological activity in laboratory (*in vitro*), and preliminary tests on animals (*in vivo*). These studies are essential to ascertain that the new drug or therapy is sufficiently promising to be introduced into humans. Preclinical studies provide pharmacology, pharmacokinetics and toxicity data that will help define the dose which will be administered to humans.

When authorized to be tested on humans, the therapeutic product goes through three main clinical phases (phase I to phase III) which will eventually lead to a marketing approval. In oncology, trials include volunteer patients with cancer for whom validated treatments have failed.

Phase I trials aim to establish the recommended dose and/or schedule of new drugs or drug combinations for phase II trials (29). Endpoints include toxicity endpoints, usually reported as per the National Cancer Institute NCI Common Terminology Criteria for Adverse Events (NCI CTC-AE), a descriptive terminology utilized for Adverse Event (AE) reporting and severity grading. A dose-limiting toxicity (DLT) is defined as an AE that is serious enough to prevent an increase in dose or level. DLTs are determined *a priori*, in order to subsequently identify

the maximum tolerated dose (MTD), defined as the highest dose that does not cause unacceptable rate of DLTs. The recommended phase II dose (RP2D) is defined based on the MTD, the overall safety profile of the drug, and its pharmacokinetic profile. To this extent, increasing doses of the drug are tested on different cohorts (one to three patients) until the highest dose with acceptable DLT rate is found. The guiding principle for dose escalation is to avoid unnecessary exposure of patients to sub-therapeutic doses of the drug while preserving safety and maintaining rapid accrual. Dose escalation methods for phase I cancer clinical trials fall into two broad classes: the rule-based designs, which include the traditional 3+3 design and its variations, and the model-based designs. The rule-based designs assign patients to dose levels according to pre-specified rules based on actual observations of DLTs from the clinical data. Typically, the MTD or recommended dose for phase II trials is determined by the pre-specified rules as well. On the other hand, the model-based designs assign patients to dose levels and define the MTD based on the estimation of the target toxicity level by a model depicting the dose–toxicity relationship. Phase I clinical trials in oncology are small (20 to 50 subjects), single-arm, open-label, sequential studies that usually include patients with a good performance status whom cancer has progressed despite standard treatments.

Phase II trials aim to assess preliminary signs of anti-tumor activity of a new drug, in order to screen out ineffective drugs and identify promising new drugs for further evaluation in phase III trials. Typical phase II endpoints include disease response or non-progression, defined, for example, as per the Response Evaluation Criteria in Solid Tumors (RECIST) in the case of solid tumors (24). The investigational drug is prescribed at the RP2D determined in preliminary phase I trials. For ethical reasons, studies of new agents in oncology usually are designed with two or more stages of accrual allowing early stopping due to inactivity of the agent. Randomization can also be employed in phase II trials. The primary aim in such case is not to formally compare the treatment arms as in subsequent phase III trials, but rather to collect efficacy data in a similar population treated with the standard strategy (in the absence of historical data), or to assess distinct administration schedules and/or routes of the drug. Sample size usually ranges from 30 to 60 patients, but can sometimes include more patients, specifically for randomized phase II trials.

Building on the data from phase I and II trials, phase III trials aim to provide confirmatory proof of the clinical benefit of a new treatment, by demonstrating that the new treatment is superior, non-inferior, or equivalent either to no treatment, placebo, or the best available therapy. Phase III trials are randomized to ensure that groups are alike in all important prognostic factors and only differ in the treatment each group receives, thus providing the basis for causality inference. While clinical benefit in phase III oncology trials is typically

measured using overall survival (OS), defined as the time between the date of randomization and the date of death from any cause, progression-free survival (PFS) or time-to-progression (TTP) are often selected as alternative endpoints. However, controversy exists regarding the use of these alternative endpoints to OS as primary endpoints, as it will be further discussed. Phase III trials are conducted in large patient population, usually hundreds of subjects. Results of phase III trials will guide health regulatory authorities (HRA) in the grant approval process.

Finally, phase IV trials are carried out once the drug has been approved by HRA. They aim to identify and evaluate the long-term effects of the new drug over a lengthy period for a greater number of patients than in phase III, usually thousands. New drugs can be tested continuously to uncover more information about efficacy, safety and side effects after being approved for marketing. Phase IV often test the drugs' effect on specific demographics. This can include pregnant women or people who are currently taking other medication to assess drug interactions.

1.1.2 Failures in the drug development process

The drug development process is a highly regulated path that defeats a large number of candidate products proving to be too toxic or not efficient enough. As a result, in the past several years, the frequency of drug approvals granted by HRA has been extremely low. For illustration, Kola et al. analyzed the success of drugs from first-in-man to registration by the Food and Drug Administration (FDA) in the United States of America (USA) over a ten-year period for ten major pharmaceutical companies (1,30). Between 1998 and 2008, only 5% of oncology compounds that initiated first-in-man studies were successfully registered by the FDA. In the late phases of development, more than 50% of oncology drugs that succeeded in phase II trials subsequently failed in phase III trials. Finally, of all oncology products that were presented for grant approval (and for which we can thus consider that at least one phase III trial led to significant efficacy findings), almost 30% failed registration. Subsequent studies investigated causes of failures of phase III trials and reported that 80% of failures of phase III cancer trials were attributable to lack of efficacy (1). Although these high failure rates can be attributable to a true absence of treatment efficacy, an inappropriate choice of the primary endpoint for the assessment of treatment efficacy may partly explain these results. Defining the primary efficacy endpoint is a key step when designing a phase III trial, as it will drive the estimated sample size, which is a function of the expected size of the treatment effect, the degree of variability in the measure of the treatment effect, as well as the type I and type II error rates. The endpoint should be robust and objectively defined, while properly accounting for the underlying disease mechanism as well as the mechanism of action of the investigational drug. While OS is considered the most reliable cancer endpoint and used by

the HRA for regular drug approval, alternative endpoints are used to assess clinical benefit of new treatments in phase III trials.

1.2 Surrogate endpoints

1.2.1 Definitions

A **clinical endpoint** is defined as a characteristic or variable that reflects how a patient feels (e.g. quality of life [QoL]), functions (e.g. QoL, patient reported outcomes), or survives (OS). A **biomarker** is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (e.g. PFS). The Biomarkers Definitions Working Group defines a **surrogate endpoint** as “a biomarker that is intended to substitute a clinical endpoint” (4). Surrogate endpoints and biomarkers provide valuable advantages in clinical research: those usually require shorter study duration and smaller sample size. As such they are ethically more attractive and they will incur less cost for the research. An extensive number of biomarkers are now available that could become potential surrogates due to the modern advances in biological and medical technologies: countless tissue, cellular, and hormonal factors; advanced imaging techniques; genomics, proteomic, metabolomics, etc. (31).

Surrogacy requires both clinical validation, that is a strong biological rationale, and statistical validation. According to the International Conference on Harmonization Guidelines on Statistical Principles for Clinical Trials, an endpoint must satisfy three conditions to be formally validated as a surrogate (Figure 2):

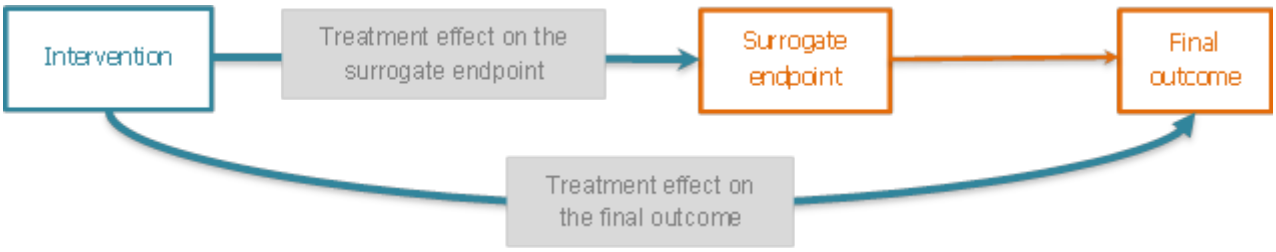
(i) the endpoint should be biologically relevant,

(ii) at the patient level, the endpoint must enable predicting the final outcome (individual-level association),

(iii) at the trial level, the treatment effect on the endpoint must enable predicting the treatment effect on the final outcome (trial-level association).

Clinical assessment, i.e. biological relevance (condition [i]), is directly related to the goal of the study, and as such to the disease and to the mechanism of action of the treatment investigated. This assessment does not rely on statistical measures but on biological knowledge. In oncology, in regards to the severity of the disease, it is reasonable to assume that disease progression is related to survival. Conversely, toxicity is usually directly related to the treatment efficacy, even though it might be lived as a deterioration of the patient status. The assessment of the individual and trial-level associations, that is conditions [ii] and [iii], rely on statistical methods that will be detailed in chapter 1.3.

Figure 2: Representation of the trial-level surrogacy



1.2.2 Surrogate endpoints in regulatory settings

OS presents multiple advantages in cancer randomized controlled trials (RCT): it is universally accepted as a measure of clinical benefit for the patient; it is objectively defined, both in terms of events and date of incidence; it is easily and precisely measured and thus reproducible; it can be exhaustively collected. As such, OS has been validated by HRAs (2,3). On the other hand, OS presents some limitations. Observing a benefit on OS may require a large number of patients and/or considerable time for patient follow-up, specifically in the adjuvant setting. As such, costs for trials may be increased, and there might be delays in the introduction of possible beneficial treatments for patients. The multiple lines of treatments, in particular in the advanced setting, may affect OS and thus bias the assessment of the true treatment effect (although this corresponds to standards and should be balanced across treatment arms). In this context, the development of alternative endpoints that could capture treatment benefit appropriately and be measurable earlier, is central for the evolution of clinical research in oncology.

Surrogate endpoints may not be inherently meaningful but aim to predict OS. They could therefore reduce the duration (and cost) of RCTs, limit patients' inclusions, answer clinical research questions in a shorter timeline, and potentially accelerate the drug approval process. These composite endpoints include biological and/or clinical events other than death, such as disease progression or treatment toxicity. These include, for instance, PFS, defined as the time from randomization to disease progression or death, for advanced diseases, or DFS, defined as the time from randomization to relapse in the adjuvant setting. Other types of events that do not directly reflect disease progression, can also be included in composite endpoints. Time-to-next treatment (TNT) is defined as the time from randomization to the initiation of a new therapy following treatment with the investigational drug. The use of TNT relies on the concept that a change in treatment usually occurs in response to a real change in the patient status, and so includes efficacy and toxicity components. These alternative endpoints commonly used in phase II trials, are increasingly replacing OS as primary endpoints in phase III trials (5). In that context, HRAs extended their

recommendations on the evaluation of anticancer therapeutics to allow the use of some endpoints other than OS.

1.2.2.1 The European Medicines Agency (EMA)

In Europe, the EMA states that “confirmatory trials should demonstrate that the investigational product provides clinical benefit”, defined in its guidelines as prolonged survival, PFS or DFS (3). Nonetheless, it specifies that OS remains the most persuasive outcome, so that proof of benefit on PFS or DFS should be accompanied with at least a trend of superiority, estimated with sufficient precision, for OS. Since 2006, the EMA added the possibility of a “conditional approval” for specific cases, including life-threatening diseases (3). This marketing authorization is valid for one year and subject to the following obligations: (i) the benefit-risk balance is positive, (ii) the applicant will likely be able to provide confirmatory data on the benefit-risk balance, (iii) the product fulfils an unmet medical need, and (iv) the benefits to public health of immediate availability of the medicinal product outweigh the risks inherent in the fact that additional data are still required.

In 2016, the EMA provided five conditional approvals and seven regular approvals for anticancer drugs. Among those 12 approvals, eight were based on an endpoint other than OS. Regarding conditional approvals, PFS was used to approve lenvatinib (mulikinase inhibitor), in combination with everolimus (inhibitor of mammalian target of rapamycin [mTOR]) for previously treated renal-cell carcinoma (RCC), alectinib (anaplastic lymphoma kinase [ALK] inhibitor) for previously treated ALK-positive non-small cell lung cancer and ixazomib (proteasome inhibitor) in combination with lenalidomide (immunomodulatory drug) and dexamethasone (corticosteroid) for the treatment of patients with multiple myeloma who received at least one prior chemotherapy treatment. Conditional approval of venetoclax (B-cell lymphoma-2 inhibitor) for chronic lymphotic leukaemia was granted based on response rate. The EMA granted regular approvals of two drugs for previously treated multiple myeloma based on PFS: daratumumab and elotuzumab, two monoclonal antibodies. Similarly, regular approvals were granted to palbociclib (selective inhibitor of the cyclin-dependent kinases CDK4 and CDK6) in the context of HR+/HER2 metastatic breast cancer (BC) based on PFS. Finally, chlormethine (cytotoxic agent based on mustard gas) was granted regular approval based on response rate for the treatment of cutaneous T-cell lymphoma. Interestingly, as it will be further discussed, none of the endpoints used for these approvals has been formally validated as a surrogate endpoint for OS.

1.2.2.2 The Food and Drug Administration (FDA)

In the USA, the FDA developed two types of approval processes: accelerated conditional approval and regular approval. Even though OS is the most reliable and preferred endpoint,

regular approvals can be granted based on “established surrogate endpoints” when assessing OS gives rise to difficulties, i.e. “long follow-up in large trials and subsequent cancer therapy potentially confounding survival analysis” (32). These alternative endpoints can be based on tumor assessment (e.g. PFS, DFS, or response rate) or symptom assessment. Accelerated approvals aim at reducing the delay in marketing of efficient drugs for serious or life-threatening diseases with no efficient therapeutic options by relieving the pre-grant evaluation process. The Accelerated Approval regulations, instituted by the FDA in 1992, allow drugs for serious conditions that filled an unmet medical need to be approved based on an endpoint “reasonably likely to predict clinical benefit” but not yet validated as surrogate for OS. To limit the risk of error, post-approval clinical trials are however required to confirm the benefit on OS.

In practice, confirmation studies to attest the benefit on OS, following conditional approval based on an alternative endpoint, are not systematically conducted. A recent review focused on all cancer drug approvals granted by the FDA based on a surrogate endpoint between 2009 and 2014 (7). The authors highlighted that of the 25 drugs which received accelerated approvals, three (12%) failed to show a benefit on OS in subsequent studies and 16 (64%) remained untested for OS at the time of the study. More alarming, 12 (40%) of the 30 products traditionally approved grants (37%) failed to demonstrate a benefit on OS in subsequent studies (table 1). Finally, recent works have shown that the term “unmet medical need” used in the context of FDA conditional accelerated approval, is imprecise and widely overused (27,33). In the context of oncology notably, almost 25% of the 237 cancer indications described as “unmet medical need” referred to indications with high incidence (more than 1000 annual cases), several existing regimens recommended by the National Comprehensive Cancer Network and a 5-year survival greater than 50% (33). For numerous contemporary FDA approvals of cancer drugs based on endpoints other than OS, the use of these alternative endpoints was thus neither justified nor justifiable.

Table 1: Surrogate-based approvals for which subsequent trials report an OS benefit or a lack of survival benefit or for which no trials exist showing or refuting a survival benefit (7)

	Approvals (N [%])		
Indication	Proven OS benefit	No OS benefit	OS benefit unknown
Total (N=55)	10 (18.2)	15 (27.3)	30 (54.5)
Accelerated approval (N=25)	6 (24.0)	3 (12.0)	16 (64.0)
Traditional approval (N=30)	4 (13.3)	12 (40.0)	14 (46.7)

1.2.2.3 Increasing use of surrogate endpoints

The use of surrogates for OS as primary endpoints in cancer RCTs and drug approvals has been increasing. While OS was the primary endpoint in 49% of trials between 1995 and 2004 (27), this proportion declined to 36% between 2005 and 2009. At the same time, the use of time-to-event endpoints other than OS, such as PFS, as primary endpoints increased from 26% to 43%. Consequently, drug approvals granted based on endpoints other than OS rose in the past decades. Indeed, between 2005 and 2007 only 23% of drug approvals were based on surrogate endpoints (34). This proportion reached 67% between 2008 and 2012 (6). This substantial use of alternative endpoints to OS in contemporary oncology RCTs and approvals raises the issue of the assessment of their surrogate properties and, as such, the appropriateness of using them as primary endpoint for evaluating the benefit of new therapies.

1.2.2.4 Misuse of surrogate endpoints: an illustration

Invalid surrogate endpoints can lead to the marketing of toxic drugs that do not improve OS, as illustrated with the recent experience with bevacizumab (anti-VEGF monoclonal antibody). In 2008, bevacizumab obtained FDA accelerated approval for administration with paclitaxel for untreated metastatic BC. This decision was based on the results of a phase III trial that highlighted a 5.9 month improvement in median PFS when bevacizumab was added to paclitaxel (8). In 2011, however, the approval was withdrawn as two randomized trials highlighted much more moderate results (9,10). The randomized AVADO (9) trial included 736 patients in three treatment arms: placebo plus docetaxel (placebo), bevacizumab 15 mg/kg plus docetaxel ($bevacizumab_{15}$), and bevacizumab 7.5 mg/kg plus docetaxel ($bevacizumab_{7.5}$). Even though the benefit of addition of bevacizumab to docetaxel on PFS was statistically significant ($bevacizumab_{15}$ *vs* placebo: HR [PFS] = 0.67, p-value = 0.001 with 95% confidence interval 95%CI = [0.54; 0.83]; $bevacizumab_{7.5}$ *vs* placebo: HR [PFS] = 0.80 with 95%CI = [0.65; 1.00]), the gain was very modest (median PFS: 8.1 months in the placebo arm, 10.0 months in the $bevacizumab_{15}$ arm and 9.0 months in the $bevacizumab_{7.5}$ arm). More importantly, OS was similar in all three treatment arms, with median values ranging from 30.2 months in the $bevacizumab_{15}$ arm to 31.9 months in the placebo arm, and no statistically significant gain in OS was detected ($bevacizumab_{15}$ *vs* placebo: HR [OS] = 1.03 with 95% CI = [0.7; 1.33]; $bevacizumab_{7.5}$ *vs* placebo: HR [OS] = 1.05 with 95% CI = [0.81; 1.36]). The RIBBON-1 trial was a randomized, double-blind, placebo-control, phase-III trial (10). The study enrolled 1237 patients to compare the efficacy and safety of bevacizumab when combined with different standard chemotherapy regimens versus those regimens alone for first-line treatment of patients with human epidermal growth factor receptor 2–negative metastatic BC: capecitabine in the CAPE cohort (N = 615) and taxane or

anthracycline in the Tax/Anthra cohort (N = 622). The two cohorts were independently powered and analyzed in parallel. As for the AVADO trial, the treatment effect on PFS was statistically significant and in favor of bevacizumab (CAPE cohort: HR [PFS] = 0.69 with 95% CI = [0.56; 0.84]; Tax/Anthra cohort: HR [PFS] = 0.64 with 95% CI = [0.52; 0.80]), however the gain in median PFS was short (5.7 months to 8.6 months in the CAPE cohort and 8.0 months to 9.2 months in the Tax/Anthra cohort). Additionally, the treatment effect on OS was not statistically significant in either the CAPE cohort (HR [OS] = 0.85 with 95% CI = [0.63; 1.14]) or the Tax/Anthra cohort (HR [OS] = 1.03 with 95% CI = [0.77; 1.38]). Following the publication of these results, the Oncologic Drugs Advisory Committee recommended the removal of bevacizumab approval in this indication (11).

1.3 Statistical evaluation of surrogate endpoints

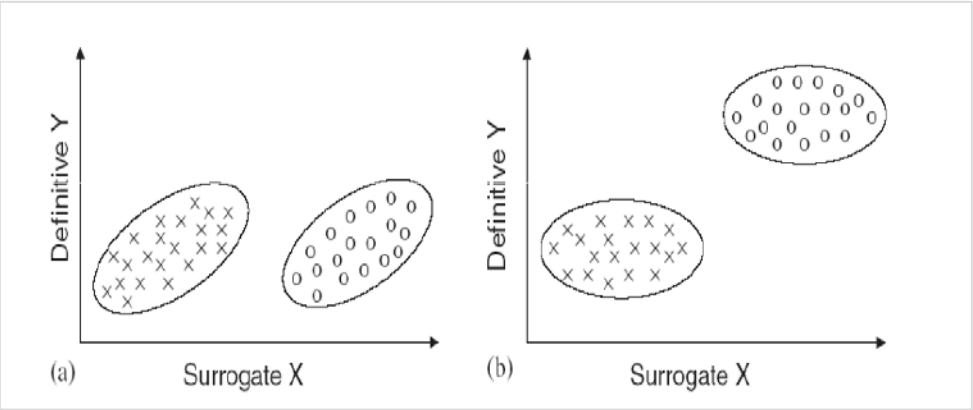
Surrogacy requires both clinical validation, that is a strong biological rationale, and statistical validation. Statistical validation requires assessing simultaneously:

(i) the individual-level association, that is, one must ensure that the candidate surrogate endpoint enables adequate prediction of the final endpoint,

(ii) the trial-level association, that is, one must ensure that treatment effect on the candidate surrogate endpoint enables adequate prediction of the treatment effect on the final endpoint.

Both conditions must be satisfied, as “a correlate does not a surrogate make” (35). As illustrated by Figure 3, correlation between an endpoint and a final endpoint is not a sufficient condition for validating a surrogate endpoint. On each graph, the final outcome (“definitive Y”, y-axis) is plotted against the candidate surrogate (“surrogate X”, x-axis) for individuals treated with two distinct treatments, represented either by bubbles or crosses. On the left-hand side, we observe a strong correlation between X and Y within each treatment group (an increase on the x-axis is associated with an increase on the y-axis, with a common slope). This suggests an association between the two endpoints. On the other hand, while there is an obvious difference in the mean value of X between the two treatment groups (x-axis), there is no such difference when one considers the endpoint Y (y-axis). A treatment effect on X cannot be translated into a treatment effect on Y. Conversely, on the right-hand side, there is no obvious association between X and Y within treatment group, suggesting an absence of correlation between X and Y. On the other hand, we observe a difference in the mean value of X between the two treatment groups, as well as in the mean value of Y between the two treatment groups. A treatment effect on X is associated with a treatment on Y, although X and Y are not correlated.

Figure 3: Illustration of individual and trial-level associations (36)



Since 1989 and the definition of the operational criteria for surrogacy introduced by Prentice (12), several statistical methods have been developed for the assessment of surrogate endpoints (13). These methods are classified into two main categories: the single-trial and the meta-analytic approaches. The first approach relies on data analysis from a single trial which is, per se, a key limitation for the estimation of the trial-level association. Meta-analytic approaches consist of data analysis from several RCTs, and as such allow for the estimation of the trial-level association.

1.3.1 Single-trial methods for the assessment of surrogate endpoints

1.3.1.1 Prentice's definition and operational criteria

In 1989, Prentice defined a surrogate endpoint as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint" (12). We consider the triplet (T, S, X) where T is the final – or true – outcome, S is the surrogate endpoint and X a categorical variable identifying the treatment received. The Prentice's definition can then be mathematically translated as:

$$f(S|X) = f(S) \leftrightarrow f(T|X) = f(T), \qquad (1.1)$$

where $f(S)$ and $f(T)$ represent respectively the probability distributions of the surrogate S and the true outcome T, and $f(S|X)$ and $f(T|X)$ represent respectively the distributions of the surrogate S and the final outcome T conditional to the treatment received X. Definition (1.1) implies that the validity of a surrogate is linked to the treatment under consideration. Since a direct verification of the equivalence raises practical issues in terms of repetition of experiment and data availability, Prentice introduces four operational criteria to assess if the triplet (T, S, X) fulfills this definition:

c1. $f(S|X) \neq f(S)$, i.e. the treatment has a significant effect on the surrogate endpoint,

c2. $f(T|X) \neq f(T)$, i.e. the treatment has a significant effect on the final outcome,

c3. $f(T|S) \neq f(T)$, i.e. the surrogate endpoint has a significant impact on the final outcome,

c4. $f(T|S,X) = f(T|S)$, i.e. the surrogate endpoint fully captures the treatment effect on the final outcome.

We first consider the case where S and T are two normally distributed endpoints, with respective means μ_S and μ_T. We denote by S_j and T_j, the values of the surrogate and the final endpoints for the j[th] patient. X_j is a covariate, here the treatment, attributed to the j[th] patient. The verification of c1 and c2 relies on the tests of significance for the parameters α and β, respectively, in the following models:

$$S_j = \mu_S + \alpha X_j + \varepsilon_{Sj}, \tag{1.2}$$

$$T_j = \mu_T + \beta X_j + \varepsilon_{Tj}, \tag{1.3}$$

where the error terms $(\varepsilon_{Sj}, \varepsilon_{Tj})$ are assumed to follow a joint zero-mean normal distribution with variance-covariance matrix Σ, defined as:

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}. \tag{1.4}$$

Criterion c3 can be verified by testing the non-nullity of the parameter γ in the linear model below:

$$T_j = \mu + \gamma S_j + \varepsilon_j, \tag{1.5}$$

where μ represents the fixed intercept of the linear model, i.e. the mean baseline value of the true endpoint and ε_j is a random error term for patient j.

Finally, to verify the fourth and last criterion, we consider the relationship between T, X and S, derived from (1.2)-(1.3):

$$T_j = \tilde{\mu}_T + \beta_S X_j + \gamma_X S_j + \tilde{\varepsilon}_{Tj}. \tag{1.6}$$

where $\tilde{\mu}_T$ is the fixed intercept of the model, and β_S and γ_X correspond respectively to the fixed effects of the treatment X on the true endpoint T adjusted for the surrogate S and S on T adjusted for X. The parameters β_S and γ_X are given by:

$$\beta_S = \beta - \sigma_{TS}\sigma_{SS}^{-1}\alpha, \tag{1.7}$$

$$\gamma_X = \sigma_{TS}\sigma_{SS}^{-1}, \tag{1.8}$$

The variance of $\tilde{\varepsilon}_{Tj}$ in (1.6) is defined by

$$var\left(\tilde{\varepsilon}_{Tj}\right) = \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}. \tag{1.9}$$

Criterion c4 implies that $\beta_S \equiv 0$ in model (1.6).

Although conceptually intuitive, these criteria present practical issues. Firstly, for c1 and c2 to be fulfilled, it is necessary to observe a significant treatment effect on both the surrogate and the final endpoints. In which case, one can argue that the validation of a surrogate endpoint might not be needed. Secondly, criterion c4, which assumes perfect surrogacy, relies on proving that a null hypothesis is true. This criterion can be useful to reject a poor surrogate endpoint, but cannot be proven.

1.3.1.2 Freedman's proportion of treatment effect

To get round these methodological impediments and address the issue of less than perfect surrogacy, Freedman et al. followed a quantitative approach for the validation of surrogate endpoints (14). Based on the work by Prentice, they proposed an operational measure of surrogacy that reflects the proportion of treatment effect (PTE) explained by the surrogate. To be a valid surrogate, an endpoint should explain a large proportion of the treatment effect. They define the PTE as:

$$PTE = 1 - \frac{\beta_S}{\beta} \qquad (2.1)$$

where β and β_S are the effects of treatment X on the true endpoint T respectively with and without adjustment on the surrogate endpoint S. In the case of censored time-to-event endpoints, β and β_S can be estimated using Cox proportional hazard models. As the PTE corresponds to a ratio of parameters a confidence interval (CI) can be calculated, based on the delta method or Fieller's theorem.

Freedman et al. pointed out that, in order to be precise, the PTE requires a strong treatment effect on the final outcome and a large number of observations, which might not be the case in most randomized clinical trials, thus leading to wide CIs. Additionally, despite its name, the PTE is not a proportion, and its value can be out of the range [0, 1], for instance when the unadjusted and adjusted treatment effects β and β_S are on opposite sides of 0 or when the unadjusted treatment effect β_S is very strong in comparison with β. In those cases, the interpretation of the PTE becomes complex. Additional measures of association were developed to address these limitations.

1.3.1.3 Related effect and adjusted association

For a surrogate endpoint to be useful in practice, one must be able to predict the treatment effect on the final outcome based on the observed treatment effect on the surrogate. Buyse and Molenberghs developed the relative effect (RE), which is the ratio of the treatment effect on the final outcome to the treatment effect on the surrogate endpoint (15). They formally define RE as:

$$RE = \frac{\beta}{\alpha} \quad (2.2)$$

where α and β are the treatment effects on the surrogate and true endpoints. As the PTE, RE is a ratio of parameters so that its CI can be estimated using the delta method or Fieller's theorem. When $RE = 1$, the treatment effects are equal, and as such indicate good surrogacy. When RE is inferior to one, the treatment effect on the true endpoint is weaker than the one on the surrogate. However, as long as the predicted treatment effect on the final endpoint remains clinically relevant, such endpoints may still be useful.

The RE quantifies the link between the treatment effects on the surrogate and the true endpoints. This value however eludes the individual-patient level. Buyse and Molenberghs proposed a second complementary measure called the adjusted association, ρ_X, that aims at reflecting the association between the two endpoints regardless of the treatment (13):

$$\rho_X = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}} \quad (2.3)$$

where σ_{ST}, σ_{SS} and σ_{TT} are the elements of the covariance matrix Σ defined in (1.4). This measure quantifies the association between the endpoints at the patient level.

In case of normally distributed endpoints, one can show that the RE and ρ_X are linked to the PTE by the following relationship:

$$PTE = \lambda\rho_X \frac{1}{RE},$$

where $\lambda = \sqrt{\sigma_{TT}\sigma_{SS}^{-1}}$ (37). The variance ratio λ is in fact a nuisance parameter, which reflects the precision of the estimations of the surrogate and the true endpoints. The PTE amalgamates the association between the surrogate and the true endpoints at the individual level, reflected by the adjusted association ρ_X, and the association between the treatment effects, estimated by the RE. This increases the difficulty in interpreting the PTE. However, the adjusted association ρ_X and the RE are not free from practical issues either. Firstly, the CIs for the RE may be wide when the sample sizes are not sufficiently large. More importantly, the estimation of the RE is based on one single trial, and as such relies on the strong assumption that the relation between the treatment effects on the surrogate and the true endpoints is multiplicative, which can only be verified using a set of trials. As a result, the RE was the basis of the meta-analytic approach developed subsequently.

1.3.2 The meta-analytic approach

Contrary to the single-trial approaches, meta-analytic schemas go beyond the evaluation of the association between the endpoints at the sole patient level (individual-level association).

Indeed, the key motivation for identifying a valid surrogate endpoint is to be able to predict the treatment effect on the final endpoint based on the treatment effect observed on the surrogate. This prediction requires assessing the association between the treatment effects on the two endpoints (trial-level association), which requires analyzing data from several RCTs.

In the following section, we note n the number of trials pooled in the analysis and n_i the number of patients included in each trial $i \in [1; n]$.

1.3.2.1 The two-stage model

Normally distributed endpoints

We first consider the case where the surrogate endpoint S and the final endpoint T are normally distributed. The two-stage approach developed by Burzykowski, Molenberghs and Buyse relies on a linear mixed-effect model (13). We consider the patient j in trial i. In the first stage, the distributions of S and T given the treatment X are modelled using fixed-effect linear models, as follows:

$$S_{ij}|X_{ij} = \mu_{Si} + \alpha_i X_{ij} + \varepsilon_{Sij} \quad (3.1)$$

$$T_{ij}|X_{ij} = \mu_{Ti} + \beta_i X_{ij} + \varepsilon_{Tij}, \quad (3.2)$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts, α_i and β_i are the trial-specific effects of the treatment X on the surrogate S and the final endpoint T in trial i. Finally, ε_{Sij} and ε_{Tij}, the error terms for the j^{th} patient of trial i, are assumed to be correlated and mean-zero normally distributed, with covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}. \quad (3.3)$$

The individual-level association corresponds to the association between the endpoints after adjustment for the treatment effect. Its estimation requires the construction of the distribution of T, given S and X, which is derived from (3.1) and (3.2) as follows:

$$T_{ij}|X_{ij}, S_{ij} \sim N\left\{\mu_{Ti} - \sigma_{ST}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{ST}\sigma_{SS}^{-1}\alpha_i)X_{ij} + \sigma_{ST}\sigma_{SS}^{-1}S_{ij};\ \sigma_{TT} - \sigma_{ST}^2\sigma_{SS}^{-1}\right\}. \quad (3.4)$$

The individual-level association, R_{ind}^2, is then reflected by the squared correlation between the two adjusted variables $S_{ij} - (\mu_{Si} + \alpha_i X_{ij})$ and $T_{ij} - (\mu_{Ti} + \beta_i X_{ij})$, defined as:

$$R_{ind}^2 = R_{\varepsilon_{Tij}|\varepsilon_{Sij}}^2 = \sigma_{ST}^2/\sigma_{SS}\sigma_{TT}. \quad (3.5)$$

At the second stage, the trial-specific intercepts, μ_{Si} and μ_{Ti}, and the treatment effects, α_i and β_i, are split up into a fixed component and a random component as follows:

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}. \quad (3.6)$$

The random part, $\begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}$, is assumed to follow a zero-mean normal distribution with dispersion matrix D given by:

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Tb} & d_{ab} & d_{bb} \end{pmatrix}. \quad (3.7)$$

To assess the trial-level association, we consider a new trial case $i = 0$ for which data are available for the surrogate endpoint but not for the final endpoint. By fitting (3.1) and (3.6) to the new trial's parameters, we obtain:

$$S_{0j}|X_{0j} = \mu_{S0} + \alpha_0 X_{0j} + \varepsilon_{S0j}, \quad (3.8)$$

with:

$$\hat{m}_{S0} = \hat{\mu}_{S0} - \hat{\mu}_S$$

$$\hat{a}_0 = \hat{\alpha}_0 - \hat{\alpha}.$$

We are interested in the estimation of the effect of X on T, i.e. the parameter β_0, knowing the observed effect of X on S, i.e. the parameter α_0. As b_0, m_{S0} and a_0 are normally distributed, then $(\beta + b_0|m_{S0}, a_0)$ is also normally distributed with mean and variance derived from (3.7) as follows:

$$E(\beta + b_0|m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix} \quad (3.9)$$

$$Var(\beta + b_0|m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \quad (3.10)$$

Theoretically, a perfect surrogate at the trial level would require the conditional variance (3.10) to be equal to zero. In practice, the **coefficient of determination, $\boldsymbol{R^2_{trial}}$,** is considered a measure of the strength of the trial-level association:

$$R^2_{trial} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \quad (3.11)$$

If the dispersion matrix D (3.7) is definite positive, then the parameter R^2_{trial} is unitless and comprised between 0, reflecting a very poor trial-level association, and 1, meaning a perfect trial-level association.

These equations can be simplified by making some reasonable assumptions. If one assumes independency between the random parameters related to the intercepts (m_{Si}, m_{Ti}) and the ones associated with the treatment effects (a_i, b_i), then the matrix D given in (3.7) adheres to the following structure:

$$D_0 = \begin{pmatrix} d_{SS} & d_{ST} & 0 & 0 \\ d_{ST} & d_{TT} & 0 & 0 \\ 0 & 0 & d_{aa} & d_{ab} \\ 0 & 0 & d_{ab} & d_{bb} \end{pmatrix},$$

The mean and variance of $(\beta + b_0|a_0)$ can then be reduced to:

$$E(\beta + b_0|a_0) = \beta + \frac{d_{ab}}{d_{aa}}(\alpha_0 - \alpha)$$

$$Var(\beta + b_0|a_0) = d_{bb} - \frac{d_{ab}^2}{d_{aa}}$$

with the corresponding coefficient of determination given by:

$$R^2_{trial} = R^2_{b_i|a_i} = \frac{d_{ab}^2}{d_{aa}d_{bb}}. \qquad (3.12)$$

Time-to-event endpoints

In oncology, the efficacy endpoints used in phase III trials are mainly time-to-event endpoints such as OS or PFS. In that context, Burzykowski et al. extended the two-stage model for normally distributed endpoints to time-to-event endpoints (38). Models (3.1) – (3.2) are thus replaced by the following copula model:

$$F(s,t) = P\big(S_{ij} \geq s, T_{ij} \geq t\big) = C_\theta\big\{F_{Sij}(s), F_{Tij}(t)\big\}, \qquad s,t \geq 0. \qquad (4.1)$$

where F_{Sij} and F_{Tij} are the marginal survival functions of the surrogate and final endpoints, and C_θ is a copula function. An *m*-dimensional copula function C_θ is defined from the unit m-cube $[0,1]^m$ to the unit interval $[0,1]$ and satisfies the following conditions (39):

1) $C_\theta(1, \dots, a_n, 1, \dots 1) = a_n$, for every $n \leq m$ and $a_n \in [0,1]$;
2) $C_\theta(a_1, \dots, a_m) = 0$ if $a_n = 0$ for any $n \leq m$;
3) C_θ is *m*-increasing,

where θ is a copula-specific parameter or vector of parameters, which measures the dependence between the marginal distributions. These conditions imply that the one-dimensional margins are uniform (40). Copula functions thus enable expressing joint distributions in terms of marginal distributions. Indeed, if we consider an *m*-variate density function F, the associated copula satisfies $F(y_1, \dots, y_m) = C_\theta\{F_1(y_1), F_m(y_m)\}$.

In theory, any bivariate copula function could be used in model (4.1), since the margins do not depend on the copula function. In practice however, we will select the copula function that suits the available data the most. The choice is guided by different elements, some being graphics (41), statistical tests (42) or fitting measures such as Akaike's information criterion (AIC) (43) or Schwarz's Bayesian criterion (44). In surrogate validation, applications mainly turned to three copula functions: the Clayton copula (45), the Hougaard copula (46) and the Plackett copula (47).

Depending on the copula function, the parameter θ will have different intervals of definition. In order to simplify its interpretation, **the individual-level association $\boldsymbol{R^2_{ind}}$** is estimated by a Spearman's rank correlation coefficient $\rho_{Spearman}$ calculated based on θ using the following equation:

$$R^2_{ind} = \rho_{Spearman} = 12 \int C_\theta(u,v)dudv - 3.$$

The parameter R^2_{ind} is then defined on the unit interval, with 1 reflecting a perfect association and 0, a null association.

As the second stage of the model remains unchanged, **the trial-level association $\boldsymbol{R^2_{trial}}$** is estimated following equation (3.11).

For illustration, Buyse et al. followed the two-stage model to evaluate PFS as a surrogate for OS in advanced colorectal cancer. They conducted a meta-analysis (MA) of ten RCTs (N = 3089) comparing fluouracil leucovorin with either fluouracil alone or with raltitrexed (48). PFS was highly correlated to OS at both the patient ($R^2_{ind} = 0.82;\ 95\%\ \mathrm{CI} = [0.82;\ 0.83]$) and the trial levels ($R^2_{trial} = 0.99;\ 95\%\ \mathrm{CI} = [0.94;\ 1.04]$). Authors concluded that PFS is an acceptable surrogate endpoint for OS in the context of advanced colorectal cancer treated by chemotherapy.

1.3.2.2 The simplified models

Individual patients' data versus aggregated data

The two-stage model described in the previous section assumes that individual patients' data (IPD) are available, that is, information with regards to the treatment, the surrogate endpoint, and the final endpoint are available on an individual basis. IPD limit the loss of information and ensure the homogeneity of the data in terms of endpoint definition and follow-up duration. As a result, both the individual and trial-level associations can be estimated. IPD however require extensive time and resources to overcome administrative (granting regulatory authorizations) and data management (standardization of the data, merging of databases, etc.) issues. In that context, simplified approaches have been proposed for the estimation of the trial-level association R^2_{trial} when only aggregated data are available, that

is, assuming summary measures of treatment effects are available only at the trial level. When IPD are available, the treatment effects are estimated separately for each trial *i*, and each endpoint. Otherwise, treatment effect estimations are extracted from the literature.

For time-to-event endpoints, the treatment effects are usually estimated by HRs from Cox proportional hazards models. Consider the j[th] patient in the i[th] trial, then the hazard functions for the surrogate endpoint, $\lambda_{Sij}(.)$, and the final endpoint, $\lambda_{Tij}(.)$, are given respectively by:

$$\lambda_{Sij}(s_{ij}) = \lambda_{Si}(s_{ij})e^{\alpha_i X_{ij}} \quad (4.2)$$

$$\lambda_{Tij}(t_{ij}) = \lambda_{Ti}(t_{ij})e^{\beta_i X_{ij}}, \quad (4.3)$$

where s_{ij} and t_{ij} are the values of S and T, respectively, for patient j of trial i. λ_{Si} and λ_{Ti} are the baseline hazard functions for S and T for trial i, and α_i and β_i are trial-specific fixed effects of treatment X_{ij} on S and T, respectively, for patient j of trial i.

Meta-regression with weighted fixed treatment effects

In this approach, treatment effects are considered as fixed. The association between the treatment effects is modelled through the linear regression of $\hat{\beta}_i$ on $\hat{\alpha}_i$:

$$\hat{\beta}_i = \mu + \hat{\alpha}_i \omega_i \gamma + \varepsilon_i. \quad (4.4)$$

where μ and γ are the fixed intercept and slope of the linear model, respectively and ε_i, the trial-specific error parameter for patient i, is assumed to follow a zero-mean normal distribution. As the treatment effects on the true and the surrogate endpoints, $\hat{\beta}_i$ and $\hat{\alpha}_i$ respectively, are not exact observations but estimations, trial-specific parameters ω_i are introduced in the model to weight the estimations as a function of their precision. The trial-level association is estimated by the determination coefficient R^2_{trial} of model (4.4) defined by:

$$R^2_{trial} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{\beta}_i - \ddot{\beta}_\iota\right)^2}{\sum_{i=1}^{n}\left(\hat{\beta}_i - E\left[\hat{\beta}_i\right]\right)^2}, \quad (4.5)$$

where $\hat{\beta}_i$ is the estimated treatment effect on T for trial i and $\ddot{\beta}_\iota$ is treatment effect on T for trial i predicted based on model (4.4).

In practice, for purpose of simplicity, the choice of the weights most usually turns to the trial size. However, the sample size may not adequately reflect the accuracy of the treatment effect estimations. For time-to-event endpoints, the precision of the estimations does not depend on the number of patients but rather on the number of observed events, which can thus be used as weights in equation (4.4). Though, the estimation errors of the treatment effects on both the surrogate and the final endpoints cannot be simultaneously introduced in

the model with this strategy. One alternative is to rely on the geometric mean of the variance of the HRs to account for the multi-directional error around the two HRs (49). The geometric mean estimates are calculated as:

$$\left(\left(HR[S]_{up} - HR[S]_{low}\right) \times \left(HR[T]_{up} - HR[T]_{low}\right)\right), \qquad (4.6)$$

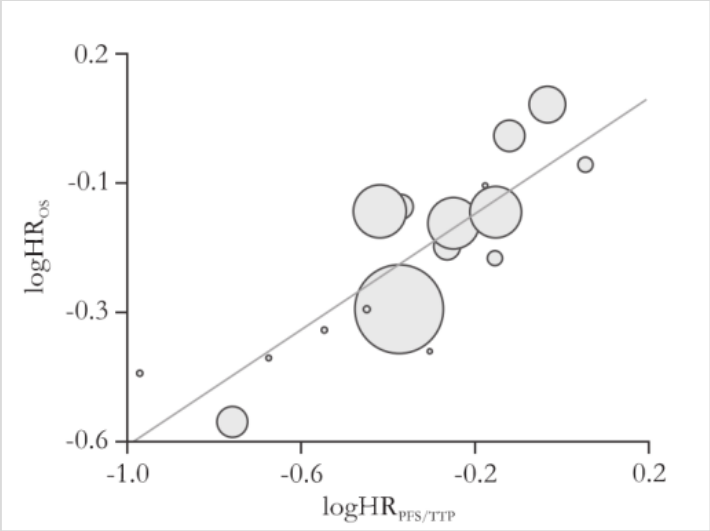
where $HR[S]_{up}$, $HR[S]_{low}$ are the upper and lower limits, respectively, of the 95% CI of the HR for the surrogate endpoint, and $HR[T]_{up}$ and $HR[T]_{low}$ are the upper and lower limits, respectively, of the 95% CI of the HR for the final endpoint.

For illustration, this approach was used in the context of advanced pancreatic cancer to evaluate the surrogate properties for OS of PFS (50). Based on data from the literature (18 trials), authors conducted a meta-regression to assess the association between the log transformation of the treatment effects on PFS and on OS, using weights proportional to the trial size. They estimated a high trial-level association between PFS and OS with a correlation coefficient R_{trial} of 0.78 (95% CI = [0.49; 0.91]) and a coefficient of determination R^2_{trial} of 0.69.

In figure 4, the treatment effects on OS (log[HR_{OS}]) for each trial are plotted as a function of the treatment effects on PFS (log[HR_{PFS}]), along with the regression line from equation (4.4). The size of the bubble reflects the weight associated with the trial, i.e. the size of the trial.

Authors concluded that PFS might be considered a surrogate endpoint for OS in this specific therapeutic situation.

Figure 4: Correlation between treatment effects on PFS (log(HR_{PFS}), X-axis) and on OS (log(HR_{OS}), Y-axis) in 18 trials



Meta-regression with random treatment effects

In this approach, after estimation of the treatment effects with models (4.2)-(4.3), we introduce trial-specific random effects instead of weights to account for estimation errors:

$$\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \begin{pmatrix}a_i\\ b_i\end{pmatrix}. \quad (4.6)$$

The random elements a_i and b_i are assumed to follow a zero-mean normal distribution with variance-covariance matrix given by:

$$\tilde{\Sigma} = \begin{pmatrix}d_{aa} & d_{ab}\\ d_{ab} & d_{bb}\end{pmatrix}, \quad (4.7)$$

The trial-level association is then estimated by the coefficient of determination R^2_{trial} defined in (3.12). In this modelling strategy, estimation errors for the treatment effects on both the candidate surrogate and the final endpoints are taken into account through the introduction of random effect associated with the trial. Contrary to the weighted meta-regression with fixed treatment effects, the expression of the estimation errors is not parametric.

The meta-regression with random treatment effects was used by Flaherty et al. to assess the trial-level association between PFS and OS in the context of metastatic melanoma (51). Based on aggregated data from 12 RCTs (N = 4416), they conducted a fixed- and random-effects statistical analysis. The random-effects approach estimated a high trial-level association (R^2_{trial} = 0.71). Despite the wide confidence interval (95% CI = [0.29; 0.90]), they concluded that PFS is a robust surrogate endpoint for OS based on the results of the meta-regression with weighted fixed treatment effects (R^2_{trial} = 0.89; 95% CI = [0.68; 0.97]).

1.3.2.3 The three-level model

In the previous sections, the unit used for the estimation of the association between the treatment effects was the trial. Two levels were then considered: the patient and the trial. In some cases however, the number of trials available is low and the unit of choice is, for instance, the center or the country. One can decide to keep the two-level strategy and simply replace the trial level by a center or country level. Another option is to consider three levels, with patients nested within centers (or countries) themselves nested within trials.

For each patient $k = 1, \dots, n_{ij}$ in center $j = 1, \dots, N_i$ within trial $i = 1, \dots, M$, let S_{ijk} and T_{ijk} be random variables denoting the surrogate and the true endpoints respectively, and X_{ijk} the binary variable indicating the treatment group. Then we have:

$$S_{ijk}|X_{ijk} = \mu_S + m_{Si} + m_{Sij} + (\alpha + a_i + a_{ij})X_{ijk} + \varepsilon_{Sijk} \quad (4.8)$$

$$T_{ijk}|X_{ijk} = \mu_T + m_{Ti} + m_{Tij} + (\beta + b_i + b_{ij})X_{ijk} + \varepsilon_{Tijk}. \quad (4.9)$$

where μ_S and μ_T are fixed intercepts, m_{Si} and m_{Ti} are random intercepts for trial i and m_{Sij} and m_{Tij} are random intercepts for center j in trial i. As previously, α and β are the fixed treatment effects on S and T respectively, and a_i and b_i are the random treatment effects on S and T respectively for trial i. a_{ij} and b_{ij} are random treatment effects on S and T respectively for center j in trial i. Finally, ε_{Sijk} and ε_{Tijk} are the individual-specific error terms, assumed to be zero-mean normally distributed with covariance matrix Σ given by (3.3).

The vector of random effects related to the trials (a_i, b_i) is still assumed to be zero-mean normally distributed with covariance matrix D given by (3.7). Similarly, we assume that the vector of random effects related to the centers (a_{ij}, b_{ij}) is zero-mean normally distributed with covariance matrix:

$$D' = \begin{pmatrix} d'_{SS} & d'_{ST} & d'_{Sa} & d'_{Sb} \\ d'_{ST} & d'_{TT} & d'_{Ta} & d'_{Tb} \\ d'_{Sa} & d'_{Ta} & d'_{aa} & d'_{ab} \\ d'_{Sb} & d'_{Tb} & d'_{ab} & d'_{bb} \end{pmatrix}. \qquad (4.10)$$

Compared to the two-level model, the significant increase in assumptions and number of parameters to be estimated might lead to computational issues. As for the two-level approach, simplifications can be considered by assuming independency between random effects, or even by suppressing some of the random effects.

1.3.2.4 The surrogate threshold effect

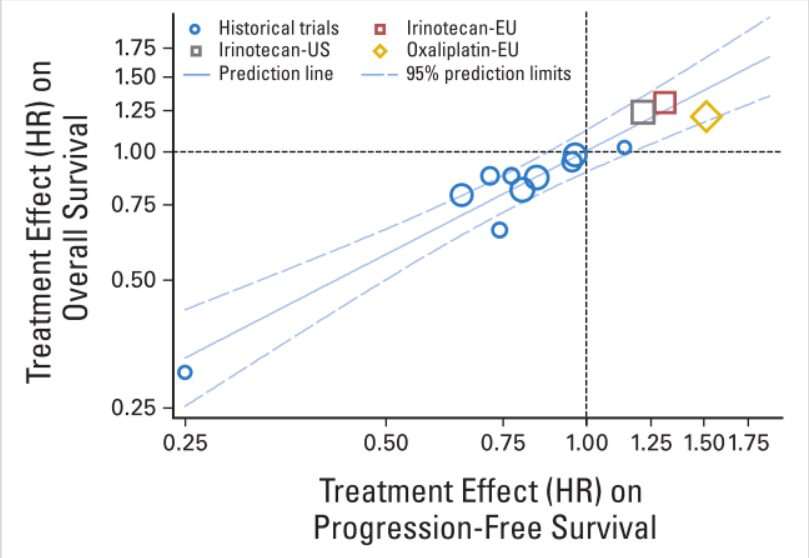
In 2006, Burzykowski and Buyse introduced the surrogate threshold effect (STE), defined as "the minimum treatment effect on the surrogate necessary to predict a non-zero effect on the true endpoint" (16). From a statistical point of view, it is calculated based on the 95% prediction bound around the regression line depicting the treatment effect on the true endpoint as a function of the treatment effect on the surrogate endpoint.

Let's denote $l()$ the lower prediction limit function of this 95% prediction bound, the STE corresponds to the value α_0 such as $l(\alpha_0) = 0$. In the context of normally distributed endpoints, the STE can be identified graphically by the value of α corresponding to the intersection point between the lower prediction limit and the horizontal line with equation (y = 0).

For illustration, Buyse et al. estimated the STE based on the prediction model developed using the two-stage model to evaluate PFS as surrogate for OS in advanced colorectal cancer as discussed in earlier chapter 1.3.2.2 (48). The STE estimated reached 0.86. In this particular case, the treatment effects were estimated by HR, so that the interpretation of the STE slightly differs, as in the case of treatment effect measured with HR, the null effect corresponds to a value of 1 instead of 0. In this example, a STE of 0.86 means that the

confidence interval around the estimated HR_{PFS} should fall below 0.86 in order to predict a significant benefit on OS (upper bound of the predicted HR_{OS} < 1). Figure 5 depicts the observed treatment effects on OS (HR_{OS}) as a function of the observed treatment effects on PFS (HR_{PFS}) with the 95% prediction interval, for the example of Buyse et al. The STE is the value of HR_{PFS} for which the horizontal line with equation (HR_{OS} = 1) crosses the upper bound of the prediction band.

Figure 5: Correlation between treatment effects on progression-free and on overall survival in historical trials (circles), in irinotecan trials (squares), and in oxaliplatin trial (diamond). A logarithmic scale is used for both axes



1.3.2.5 Summary of methods for the statistical evaluation of surrogate endpoints

In his landmark paper, Prentice proposed a formal definition of surrogate endpoints, outlined how they could be validated, and discussed intrinsic limitations in the surrogate marker validation quest (12). Much debate ensued, since many authors perceived a formal criteria-based approach as too stringent and not straightforward to verify (35). Freedman *et al.* took Prentice's approach one step further by introducing the 'proportion explained', which is the proportion of the treatment effect mediated by the surrogate (14). Buyse and Molenberghs discussed some problems with the proportion explained and proposed to replace it by two new measures (15). The first, defined at the population level and termed 'relative effect', is the ratio of the overall treatment effect on the true endpoint over that on the surrogate endpoint. The second is the individual-level association between both endpoints, after accounting for the effect of treatment, and referred to as 'adjusted association'.

These concepts were next extended to situations in which data are available from several RCTs (13,52). The individual-level association between the surrogate and final endpoints

carries over naturally, the only change required being an additional stratification to account for the presence of multiple experiments. The experimental unit can be the center in a multicenter trial, or the trial in a MA context. This latter situation has been extensively emphasized, because an informative validation of a surrogate endpoint will typically require large numbers of observations coming from several trials. Moreover, meta-analytic data usually carry a degree of heterogeneity not encountered in a single trial, caused by differences in patient population, study design, treatment regimens. These sources of heterogeneity increase one's confidence in the validity of a surrogate endpoint, when the relationship between the effects of treatment on the surrogate and the true endpoints tends to remain constant across such different situations. The notion of relative effect can then be extended to a trial-level measure of association between the effects of treatment on both endpoints. The two measures of association, one at the individual level, the other at the trial level, are proposed as an alternative way to assess the usefulness of a surrogate endpoint. This approach also naturally yields a prediction for the effect of treatment on the true endpoint, based on the observation of the effect of treatment on the surrogate endpoint. Finally, the surrogate threshold effect brings information regarding the practical use of a surrogate endpoint. Indeed, it indicates the minimum treatment effect required on the surrogate to predict a non-null treatment effect on the final outcome.

As of today, the two-stage model is recognized as the most statistically rigorous approach (53,54). This approach makes the maximum use of all IPD and allows for the estimation of both levels of association. It takes into account the estimation errors in the calculation of the trial-level association. Aside from the administrative issues related to obtaining IPD from multiple RCTs, the main restriction of the two-stage approach is related to numerical concerns as numerical convergence may be difficult to reach if the variability within or between trials is not sufficient. A large number of RCTs with different treatment effects is preferable, even if in theory, this approach allows one to validate a surrogacy endpoint in the absence of treatment effect, simply because of random sampling variability.

1.3.3 Quality of validation studies on surrogate endpoints and strength of evidence

Validation of a surrogate endpoint is a complex process for which several statistical approaches are available that can lead to the estimation of multiple parameters, specifically the individual-level and/or the trial-level association. As a result, judging the quality of published studies assessing surrogate endpoints, and the strength of the reported associations can be overly complex. Consequently, validation grids have been developed to guide researchers with the interpretation of results of surrogacy studies.

In 2009, Elston and Taylor proposed a three-level hierarchy to assess the level of evidence of a surrogate endpoint (17). They distinguished three levels of evidence: level 1, the highest evidence demonstrating that the treatment effects on the candidate surrogate and the final outcome are correlated at the trial level; level 2, evidence demonstrating that the candidate surrogate and the final outcome are correlated at the patient level; and level 3, evidence of biological plausibility of an association between the candidate surrogate and the final outcome. As per the Elston and Taylor's criterion, one endpoint shall meet a trial-level association R^2_{trial} , estimated from (3.11) or (3.12), superior to 0.7 to be ranked as level-1 evidence of surrogacy. Of note, this ranking depends essentially on the point estimate, and does not account for the variability in the estimation process.

In 2011, the German Institute of Quality and Efficiency in Health Care (IQWiG), an independent health technology assessment agency that assesses the benefits and harms of drug and non-drug technologies on behalf of the German Federal Joint Committee and the Federal Ministry of Health, developed recommendations for the evaluation of surrogate endpoints (18). They consider two evaluation stages, the first one, rather subjective, rates the reliability of evidence and the second one, more objective, focuses on the strength of evidence. They classify the methodological reliability of the study into four categories: low, moderate, limited or high reliability. This classification is based on the following elements: (i) application of a recognized approach described in the specialized statistical literature, (ii) conduct of analyses to test the robustness and generalizability of results, (iii) systematic compilation of data, (iv) sufficient restriction of indications or degrees of disease severity and of interventions, and (v) clear definitions of the endpoints investigated. The evaluation of the strength of evidence relies on objective conditions on the trial-level association estimated by a correlation coefficient R_{trial}. If we consider an unweighted model and no covariates R_{trial} can be approximated by the squared root of R^2_{trial} estimated from (3.11) or (3.12) (54). The correlation is considered high when the lower limit of the 95% CI of the coefficient correlation R_{trial} is superior to 0.85, low when the upper limit of the 95% CI of R_{trial} is inferior to 0.7. Otherwise, the strength of the correlation is ranked as medium, meaning that no conclusion can be drawn based on the results. The authors then provide a decision tree for the overall conclusion on the surrogate validity based on both the reliability and the strength of the association (Appendix A).

The first version of the Biomarker-Surrogate Evaluation Schema (BSES) was proposed in 2007 and then upgraded to a revised version published in 2010 (19). The BSES includes four domains: study design, target outcome, statistical evaluation and generalizability. Each of the four domains is associated with a four-level rank (0 to 3) leading to a global score ranging from 0 to 12, calculated based on specific guidelines (Appendix B). The study domain

assesses the reliability of the study based on the sources of data (MA of RCTs or observational cohort, number and quality of RCTs). The target outcome domain focuses on the type of endpoints evaluated. The statistical evaluation domain ranks the strength of the overall association between the candidate surrogate and OS, based on three surrogacy measures: the individual-level association R^2_{ind}, the trial-level association R^2_{trial} estimated by (3.11) or (3.12) and the STE proportion (STEP), a transformation of the STE estimated by (5.1). No recommendation regarding methods for the calculation of R^2_{ind}, R^2_{trial} and STEP are provided. Finally, the generalizability domain focuses on the clinical and pharmacological evidence supporting surrogacy, and the generalizability across populations and drug-class mechanisms. The overall level of evidence is finally classified as proof of surrogacy, high probability of surrogacy, hint of surrogacy, no proof of surrogacy, and proof of low correlation.

To our knowledge, three validation scales have been introduced to evaluate the level of evidence of studies assessing surrogate endpoints. The hierarchy scale introduced by Elston and Taylor (17) is easy to use and interpret, however, it eludes the differences between the studies in terms of methodological quality and reliability. The IQWiG guidelines attempt to fill that void by including an evaluation of the reliability of evidence relying on multiple parameters such as study design, statistical methodology and sensibility analyses. Although these elements are relevant, the authors did not develop a formal grading scale, so that appreciation of reliability may be subject to bias. The evaluation of the strength of evidence relies on objective and robust conditions on the trial-level association and is the most conservative, even though it eludes the statistical method question. Finally, the BSES schema evaluates multiple aspects of the MA and the strength of the association is not only assessed through the trial-level association but also considers the individual-level association and the STE. However, it does not provide any recommendations regarding the statistical approach for the estimation of these surrogacy measures, and more particularly, the STEP. Additionally, even though more detailed than the IQWiG criteria, a significant subjectivity remains, notably when assessing the quality of the study design and the generalizability of the endpoint. The assessment of the generalizability is also controversial, since it emphasizes consistency across different drug classes. As the validity of a surrogate endpoint depends on the mechanisms of action of both the disease and the treatment, the recommendations so far were to look at drugs class by class as a surrogate may be relevant for one drug class but not for the other. Depending on the diseases and the treatments, this item may not be pertinent.

1.4 Synthesis and research objectives

Surrogate endpoints are increasingly being used to replace OS as primary endpoints in cancer RCTs, even though their capacity to substitute for and predict OS has not been systematically assessed. Several studies evaluating surrogate endpoints in cancer RCTs have been conducted in the past 20 years, focusing on multiple therapeutic settings, populations and differing in terms of statistical methodology. To the best of our knowledge, three reviews of MAs evaluating surrogate endpoints in cancer trials have been published (55–57). Two reviews were restricted to specific types of tumors (e.g. solid tumors only (56)) or settings (e.g. advanced setting only (56)), and the last one did not clearly report the search strategy (55), which neither allows the reader to reproduce the results nor to assess the exhaustiveness of the search. With the increasing number of studies evaluating surrogate endpoints, an update of current knowledge, with a formal evaluation of the level of evidence, is necessary. **The first objective of this thesis was to identify MAs that evaluated surrogate endpoints for OS in oncology and assess the strength of evidence provided by these studies.** We performed a systematic review to identify MAs of cancer RCTs assessing surrogate endpoints for OS, and evaluated the strength of the reported associations based on (i) the German Institute of Quality and Efficiency in Health Care guidelines and (ii) the Biomarker-Surrogate Evaluation Schema. This work is presented in chapter 2, and is currently in press in *Critical Reviews in Oncology and Hematology*.

The review highlighted certain therapeutic settings where the surrogacy question has only been seldom, if at all, addressed, such as in advanced soft-tissue sarcoma (STS) and in adjuvant breast cancer. Due to their extremely low incidence, conducting large RCTs to evaluate the benefit of new treatment for metastatic STS is complex. The identification of valid surrogate endpoints for OS that would be observed sooner and more frequently than OS, thereby reducing the number of included patients, would be of a great advantage for clinical research. As highlighted in our critical review, only one MA evaluating response rate and PFS as surrogates for OS in metastatic STS was conducted, but it was limited to the analysis of aggregated data (58). **The second objective of this thesis was thus to assess the surrogate properties for OS of three commonly used time-to-event endpoints in advanced STS: progression-free survival (PFS), time-to-progression (TTP) and time-to-treatment failure (TTF).** This work is presented in chapter 3. The manuscript is currently under review in the *Journal of Clinical Oncology*.

Some other composite endpoints such as time-to-next treatment (TNT), even though not as widespread as progression-based endpoints, might be interesting candidate surrogate endpoint. TNT is defined as the time from baseline (randomization, inclusion or initiation of the treatment) to initiation of a new treatment after failure of the previous one. By definition, it

includes all possible reasons for switching treatment, so that it might more accurately reflect a change in the patient status. TNT however is not systematically collected in experimental studies. **The third objective of this thesis was to conduct a preliminary investigation of TNT as a potential candidate surrogate endpoint for OS, and thus to assess the individual-level correlation between TNT and OS,** based on a multicenter cohort study of STS patients. This work is presented in chapter 4 and the manuscript has been published in *BMC Medicine*.

Surrogate endpoints are particularly relevant in the context of rare disease, such as STS, but also in settings with extended OS, thus requiring longer trial duration, such as in adjuvant BC trials. While several MAs have been conducted in the metastatic setting (57), only one MA evaluated the surrogate properties of various endpoints for adjuvant BC trials (59), as highlighted in our critical review. The study was however limited to the analysis of aggregated data. **The fourth objective of this thesis was thus to assess the surrogate properties for OS of four time-to-event endpoints commonly used in this setting: RFS, invasive disease-free survival (iDFS), distant disease-free survival (DDFS) and locoregional relapse-free survival (LRFS).** This work is presented in chapter 5. The manuscript is currently under review by the co-authors.

2 Meta-analyses evaluating surrogate endpoints for overall survival in cancer randomized trials: a critical review

2.1 Introduction

In cancer RCTs, alternative endpoints are increasingly being used in place of OS to reduce sample size, duration and cost of trials. It is necessary to ensure that these endpoints are valid surrogates for OS. The objective of this preliminary work was to (i) identify MAs that evaluated surrogate endpoints for OS and (ii) assess the strength of evidence for each MA.

We performed a systematic review through a computerized search of MEDLINE and SCOPUS databases to identify MAs that evaluated potential surrogate endpoints for OS in cancer RCTs. Relevant publications were selected based on a two-step process and using a standardized data extraction grid. We also assessed the strength of the associations reported in each MA based on (i) the German Institute of Quality and Efficiency in Health Care guidelines and (ii) the Biomarker-Surrogate Evaluation Schema.

We retrieved 164 MAs from 53 publications. Most MAs focused on three cancer localizations: colorectal, lung and breast cancer. Overall, data suggests that DFS had reasonable surrogate properties for OS in adjuvant treatment for colon cancer, non-small-cell lung cancer, gastric cancer, and head and neck cancer. In the advanced setting, PFS may be an appropriate surrogate endpoint for OS colorectal cancer, lung, and head and neck cancers. This work highlighted an important heterogeneity in terms of statistical methods used for surrogacy assessment, as well as in terms of reported parameters.

Consensual frameworks for the assessment of surrogacy evidence as well as for the reporting of summary parameters are required for improved assessment of published MAs on surrogate endpoints.

This work is currently in press in *Critical Reviews in Oncology / Hematology*.

2.2 Publication

Meta-analyses evaluating surrogate endpoints for overall survival in cancer randomized trials: A critical review

Marion Savina[1,2], Sophie Gourgou[3], Antoine Italiano[4,5], Derek Dinart[6], Virginie Rondeau[2], Nicolas Penel[7], Simone Mathoulin-Pélissier[1,2], Carine Bellera[1,2]

[1] Clinical and Epidemiological Research Unit, Institut Bergonié, Comprehensive Cancer Center, Bordeaux, France
[2] University of Bordeaux, ISPED, Centre INSERM U1219 Bordeaux Population Health, Epicene Team, Bordeaux, France
[3] Biometrics unit, Institut du Cancer de Montpellier, Comprehensive Cancer Center, Montpellier, France
[4] Department of Medical Oncology, Institut Bergonié, Comprehensive Cancer Center, Bordeaux, France
[5] Oncogenesis and Therapeutic Targeting of the Sarcoma Cell, ACTION, INSERM U1218, University of Bordeaux, France
[6] Clinical and Epidemiological Research Unit, Institut Bergonié, Comprehensive Cancer Center, Bordeaux, France
[7] Department of Medical Oncology, Centre Oscar Lambret, Comprehensive Cancer Center, Lille, France

Abstract

Background: In cancer randomized controlled trials (RCT), alternative endpoints are increasingly being used in place of overall survival (OS) to reduce sample size, duration and cost of trials. It is necessary to ensure that these endpoints are valid surrogates for OS. Our aim was to identify meta-analyses that evaluated surrogate endpoints for OS and assess the strength of evidence for each meta-analysis (MA).

Materials and methods: We performed a systematic review to identify MA of cancer RCTs assessing surrogate endpoints for OS. We evaluated the strength of the association between the endpoints based on (i) the German Institute of Quality and Efficiency in Health Care guidelines and (ii) the Biomarker-Surrogate Evaluation Schema.

Results: Fifty-three publications reported on 164 MA, with heterogeneous statistical methods. Disease-free survival (DFS) and progression-free survival (PFS) showed good surrogacy properties for OS in colorectal, lung and head and neck cancers. DFS was highly correlated to OS in gastric cancer.

Conclusion(s): The statistical methodology used to evaluate surrogate endpoints requires consistency in order to facilitate the accurate interpretation of the results. Despite the limited number of clinical settings with validated surrogate endpoints for OS, there is evidence of good surrogacy for DFS and PFS in tumor types that account for a large proportion of cancer cases.

Key words: surrogate endpoint, cancer, overall survival, meta-analysis, randomized controlled trial, systematic review.

Introduction

In cancer randomized controlled trials (RCT), overall survival (OS) is the gold standard primary efficacy endpoint. However, observing a benefit on OS may require a significant number of patients, considerable time for patient follow-up and as such, substantial trial costs and delays in the introduction of possibly beneficial treatments. Additionally, the added lines of treatment and the ethics of withholding a potentially useful treatment drive researchers to include a possibility for crossover in trial designs, especially in the metastatic setting. In such case, crossover from the control (or standard of care) treatment arm towards the experimental strategy is allowed, which might lead to a biased estimation of the OS benefit. The development of alternative endpoints, that could capture treatment benefit appropriately and be measurable earlier, is central to clinical oncology research progress.

Alternative survival endpoints, such as progression-free survival (PFS) in trials of metastatic diseases or disease-free survival (DFS) in the adjuvant setting, are increasingly replacing OS in phase III trials (1). In the United States of America, a large proportion of new cancer drug approvals granted by the Food and Drug Administration (FDA) are based on such alternative endpoints (2-5). Moreover, clinical practice guidelines often expand recommendations for approved drugs based on studies assessing surrogates. For illustration, carfilzomib received FDA approval in 2012 based on response rate (RR) for relapse and/or refractory multiple myeloma (6). In 2013, the National Comprehensive Cancer Network (NCCN) then added untreated myeloma to the FDA approval in its guidelines based on the results of two small uncontrolled I/II trials highlighting a benefit on RR. Since Medicare and most major private insurers follow the NCCN guidelines for the establishment of their coverage policy, this expansion impacted the patient's access to carfilzomib (7). Recent data have shown that almost half of cancer drug approvals are based on endpoints that have never been formally evaluated as surrogates for OS (5). Out of the 36 drugs approved by the FDA between 2008 and 2012 based on an endpoint other than OS, only 5 were shown to improve OS in subsequent randomized studies. Of the 31 remaining drugs, 18 failed to show a benefit on OS and 13 remained untested for OS in August 2015 (4).

The use of these alternative endpoints relies on the hypothesis that they can adequately replace, i.e. be valid surrogates of OS, otherwise this might lead to the marketing of drugs that do not improve OS. This issue is well illustrated with the example of bevacizumab which obtained FDA accelerated approval in 2008 for metastatic breast cancer based on PFS improvement (8). In 2011, approval was withdrawn as multiple randomized trials highlighted that the PFS gain was more modest than predicted by earlier trials, and importantly that there was no improvement in OS (9).

The Biomarkers Definitions Working Group defines a surrogate endpoint as "a biomarker that is intended to substitute a clinical endpoint" (10). Validating surrogacy requires both clinical

and statistical validations. These include (i) a strong biological rationale, (ii) a strong correlation between the candidate surrogate and the final endpoint (individual-level association), and (iii) a strong correlation between the treatment effects observed on the candidate surrogate and the final endpoint (trial-level association) (11). These correlations can be ascertained by performing a meta-analysis of multiple randomized trials.

Over the last two decades, several meta-analyses evaluating surrogate endpoints for different cancer localizations such as resectable colorectal (12), advanced lung (13) and metastatic breast cancer (14) have been published. In this context, we aim to provide an up-to-date review of meta-analyses evaluating surrogate endpoints for OS in cancer RCTs.

Methods

The systematic review of meta-analyses involved three steps: selection of meta-analyses, data extraction, and scoring of the strength of the trial-level association reported between the candidate surrogate endpoint and OS.

Selection

We performed a systematic review through a computerized search of MEDLINE and SCOPUS databases to identify meta-analyses of RCTs assessing surrogate endpoints for OS in human trials. The wide search algorithm included the following key words: neoplasm, cancer, oncology, surrogacy, surrogate endpoints, correlation, association and prediction (Additional file 1). We included all studies published up to 18 July 2016 and available in English or French. We selected relevant publications based on a two-step process using a standardized data extraction grid (Additional file 2), designed and validated by two readers who independently checked both steps of the selection process. In the first step, general information was retrieved based on the abstract. Publications were ineligible if the abstract presented at least one of the following characteristics: letter/comment to the editor, conference abstract, not conducted in humans, not related to cancer only, study led on healthy patients, explicitly unrelated to surrogate endpoints, not a validation study of surrogate endpoints. In the second step of the selection process, we read the full manuscripts for the selected abstracts. Publications were ineligible if they included at least one of the following characteristics: did not consider OS as the final endpoint, did not follow a meta-analytical framework, did not report the trial-level association, or reported on a simulation study. When several publications on the exact same dataset were available (e.g. different methodologies to address the same objective), we only considered the original study. When studies led to multiple updates, we only included the last publication. We have presented the results of the selection process following the PRISMA guidelines for the reporting of systematic reviews and meta-analyses (15).

Data extraction

For full manuscripts that met eligibility criteria, we collected information regarding general

characteristics of the meta-analyses: cancer localization, disease setting (adjuvant, neoadjuvant, advanced), number and design (phase II or III) of trials, number of patients, and period of inclusion. We collected the sources of the review used to identify the trials included in each MA: published articles and/or congress abstracts and/or trial registries. When the analysis was based on a dataset collected previously for another purpose, and thus no review was undertaken, we assigned the label “convenience sample” to the data source. We collected information regarding the type of data – aggregated data or individual-patient data (IPD) – and retrieved the statistical methods for the assessment of the individual-level association, and the trial-level association (16). The individual-level surrogacy reflects the association between the candidate surrogate and the final endpoint (OS) regardless of the treatment. It can be estimated by a linear correlation coefficient (correlation of Pearson) for normal endpoints. In the context of time-to-event endpoints, this estimation is more complex. The two endpoints are jointly modelled, using for instance a copula function or a frailty model, to estimate a correlation parameter. When IPD are not available, the individual-level association is usually estimated using aggregated data from each treatment arm (such as response rate or median survival time) with linear models. Two main methods are available for the estimation of the trial-level association, i.e. the association between the treatment effects on the candidate surrogate and on the final endpoint. In the first approach, the treatment effects are estimated independently by hazard ratios (HR) or odds ratios (OR). The association between the treatment effects is then assessed by the coefficient of determination R^2 of a linear regression model, usually weighted by the trial size. The second method follows the two-stage model introduced by Burzykowski and Buyse (17, 18). In the first stage, the treatment effects are estimated simultaneously using a bivariate model. The association between the treatment effects is then estimated using an error-in variable model, to adjust for estimation errors.

When available, the minimum treatment effect on the surrogate necessary to predict a non-zero effect on the true endpoint, or surrogate threshold effect (STE) (19), was collected as well as the surrogate threshold effect proportion (STEP).

Scoring the strength of the association

We relied on two frameworks for the assessment of the strength of the association (Additional file 3). Both require the estimation of the trial-level association. As mentioned above, this association can be estimated by the coefficient of determination (R^2) or the coefficient of correlation (r) of a regression model. The strength of the trial-level association was scored as per the German Institute of Quality and Efficiency in Health Care (IQWiG) guidelines (20): high correlation (lower limit of the 95% confidence interval of $r \geq 0.85$), low correlation (upper limit of the 95% confidence interval of $r \leq 0.7$), or medium in any other case, meaning that the validity of the surrogate remains unclear. If the coefficient of determination R^2 was provided instead of the correlation coefficient r, then r was calculated

by taking the square root. When no confidence interval was reported, we considered the strength of the trial-level association as medium. To distinguish those studies truly classified as “medium” based on the reported point estimate and its confidence interval from studies that reported high correlations without reporting on the precision or with wide confidence intervals, we assigned the level “medium +” to the latter ones. We also relied on the Biomarker-Surrogate Evaluation Schema (BSES) (21), which requires the estimation of three parameters: the individual-level association (R^2_{ind}), the trial-level association (R^2_{trial}) and the surrogate threshold effect proportion (STEP). The BSES classifies the association between the surrogate and OS as excellent ($R^2_{trial} \geq 0.60$ and STEP ≥ 0.3 and $R^2_{ind} \geq 0.60$, where R^2_{ind} is the patient-level association and R^2_{trial} is the trial-level association), good ($R^2_{trial} \geq 0.4$ and STEP ≥ 0.2 and $R^2_{ind} \geq 0.4$), fair ($R^2_{trial} \geq 0.2$ and STEP ≥ 0.1 and $R^2_{ind} \geq 0.2$) or otherwise poor. If the correlation coefficient was provided instead of the coefficient of determination, then R^2 was calculated by taking the square of r. Since most studies did not report the STEP, we also calculated an adapted score (BSES2), that considered the BSES classification algorithm but ignoring the STEP. We considered the studies that did not use formal hazard ratios to estimate the trial-level association as not evaluable with none of the three scales. Finally, to limit subjectivity and errors when ranking the strength of association, two assessors independently ranked the IQWiG and the BSES scores. In case of discordance, a third reader assessed the strength of association. If the discordance remained, the three readers met to reach an agreement (MS, DD, CB).

Results

Throughout the manuscript, we refer to the notation N_P and N_{MA}, to denote number of publications or number of MA respectively.

Selection

The algorithm initially retrieved a total of 4222 publications. At the end of selection process, we retained 53 articles (Figure 1).

Characteristics of the meta-analyses

The 53 publications (12-14, 22-71) reported on 164 meta-analyses (table 2). Some cancer settings were investigated in multiple publications: advanced colorectal cancer (N_{MA} = 37 [23% of eligible MA] and N_P = 11 [21% of eligible publications]), metastatic breast cancer (N_{MA} = 25 [15%] and N_P = 9 [17%]), advanced lung cancer (N_{MA} = 11 [7%] and N_P = 6 [11%]), metastatic renal-cell carcinoma (N_{MA} = 8 [5%] and N_P = 4 [8%]), advanced gastric cancer (N_{MA} = 3 [2%] and N_P = 3 [6%]) and advanced pancreatic cancer (N_{MA} = 23 [14%] and N_P = 3 [6%]). Characteristics of the 164 meta-analyses are described in table 1. A large majority of meta-analyses relied on aggregated data (N_{MA} = 125, 76.2%). Most meta-analyses described the literature search process to identify relevant trials (N_{MA} = 158, 96.3%). Identification was essentially based on published articles and abstracts (N_{MA} = 53; 32.3%) or published manu-

scripts only (N_{MA} = 49, 29.9%). In addition to the trial-level association, 57 studies (34.8%) reported the individual-level association, among which 51 reported the associated confidence intervals (31.1%). Commonly used statistical methods for the estimation of the individual-level association included the two-step model (N_{MA} = 28, 17.1%), non-parametric methods to estimate a rank correlation (N_{MA} = 19, 11.6%) and weighted linear regression models using survival rates or median survival times (N_{MA} = 21, 12.8%). Weighted or adjusted linear regression models were most often applied for estimation of the trial-level association (N_{MA} = 109, 66.5%). Among those 109 MA however, 42 (39%) did not use hazard ratios but differences or ratios of median survival times as estimators of the treatment effects. A total of 17 meta-analyses (10.4%) reported the STE, and none reported the STEP.

Scoring the strength of the association

The individual-level association, trial-level association and surrogate threshold effect reported are available in table 2. Out of the 164 meta-analyses, a total of 95 (57.9%) were based on a methodology consistent with the IQWiG guidelines. The trial-level association of these meta-analyses was classified as high (N_{MA} = 17; 17.9%), medium + (N_{MA} = 19; 20%), medium (N_{MA} = 52; 54.7%) or low (N_{MA} = 7; 7.4%) (Table 2). According to the BSES, the association between the endpoints and OS was not evaluable for all meta-analyses, since none estimated the STEP. Only 37 meta-analyses (22.6%) estimated all parameters required to apply the BSES2. Among these 37 studies, the association was classified as excellent (N_{MA} = 15), good (N_{MA} = 7), fair (N_{MA} = 5) or poor (N_{MA} = 10) (Table 2). For 12 meta-analyses, the trial-level association was considered high according to IQWIG and excellent as per BSES2 (Table 3).

Discussion

We provided an overview of current evidence of surrogate endpoints for OS in cancer RCTs. Contrary to previous reviews (6, 72, 73), we did not restrict our review to specific types of cancers or therapeutic settings, nor to specific endpoints and relied on a large research algorithm to be as exhaustive as possible. We identified 20 additional publications (22, 23, 25, 27, 28, 38, 39, 41, 48, 53, 55, 60, 62-66, 68, 70, 71), among which nine focused on therapeutic settings not reported in previous reviews (41, 62-66, 70, 71) such as glioblastoma (68) or soft-tissue sarcoma (71). This highlights the evolving literature on surrogacy and the importance of regular updates. To assess the strength of the association between the candidate surrogate survival endpoint and OS, we relied on two frameworks, which, although they have not been validated, can be considered complementary: the German IQWiG guidelines and an adaptation of the BSES. If the IQWiG guidelines are the most conservative, the adapted BSES grid includes the individual-level association in the evaluation of the level of proof of surrogacy.

We retrieved 164 meta-analyses from 53 publications. There is an important variability in

terms of number of meta-analyses reported by publication. About 43% of the 53 publications reported one single meta-analysis. For some cancer types however, all available meta-analyses were reported by one publication (e.g. head and neck cancer). On the other hand, 19 of the 23 meta-analyses conducted in pancreatic cancer came from the same publication.

Most meta-analyses focused on three cancer localizations: colorectal, lung and breast cancer. One explanation is the high incidence of these cancers, and thus the potentially large populations that could be affected by the validation of a surrogate in these settings. The urgency of new treatments due to poor survival outcomes also justifies that more meta-analyses have been conducted for cancer types with poor prognosis (in the advanced setting in general, but also pancreatic or head and neck cancer). Finally, related to the incidence issue, robust statistical validation of surrogacy requires a large dataset of large trials. As such, it is expected that more meta-analyses will be conducted for those cancer types where more data are available. In these settings, where several MA are available, one may be interested in pooling these data to come up with some “grand” estimate of a surrogacy measure. Although tempting, this approach should be avoided. Some trials may have been included in distinct meta-analyses, among which heterogeneity in terms of disease setting and/or mechanism of action of the treatment is likely to be an issue. Finally, accounting for the estimation errors associated with the correlation coefficient reported by each individual MA would be overly complex.

In some therapeutic settings the surrogacy question has only seldom been treated, if at all, although the availability of a surrogate would be particularly relevant, e.g. for those disease with longer life expectancy or low incidence. For instance, in adjuvant breast cancer (10-year OS: 78%), a surrogate endpoint such as DFS would be a great asset to reduce the duration of trial. However, only one meta-analysis based on aggregated data was identified (43). Conducting large RCTs, and thus meta-analyses in rare tumors is overly complex, as illustrated with the case of metastatic sarcoma, for which we could only retrieve one meta-analysis based on aggregated data (71).

Publications did not systematically report the sources of the trials analyzed, and when reported, authors usually relied on published reports only. Meta-analyses conducted on non-exhaustive data, specifically excluding unpublished trials, might lead to a selection bias (6) that could impact the surrogacy measures. Trials with negative or non-significant findings, even if very large, are more likely to remain unpublished and thus excluded in the calculation of the surrogacy measures (74-77). It is however difficult to infer the impact of these inclusions on the estimated surrogacy measures. The search for unpublished trials and the use of trial registries is however complex and does not guarantee the availability of the required parameters.

We observed a noteworthy heterogeneity in terms of statistical methods. Most meta-analyses identified relied on aggregated data for the estimation of the surrogacy metrics. Historically,

meta-analyses originate from the quantitative synthesis of findings of studies with the trial as unit of measurement, so that IPD is not required. Additionally, the use of aggregated data enables to free oneself from the time-consuming challenges that are the collection (granting of the regulatory authorizations) and data-management of IPD. Consequently, the most exhaustive meta-analyses are based on aggregated data from the literature. In the context of surrogate endpoints however, there are arguments in favor of IPD. In particular, the use of IPD limits the loss of information, especially for the estimation of the patient-level association, enables the standardization of the endpoint definition and the follow-up duration, as well as the application of the most robust statistical method for surrogacy assessment, based on IPD (17, 18). On the other hand, the cost of IPD cannot be ignored. The logistics, feasibility and human resources required to obtain all necessary authorizations, to standardize, merge and analyze IPD is tremendous.

Most meta-analyses relied on a methodology that did not enable us to evaluate the strength of the association irrespective of the scale considered. Regarding the BSES score, since STEP was never reported, the quality of the meta-analysis was not evaluable. When ignoring the STEP and thus relying on the BSES2 score, some studies attained both a high IQWiG score and an excellent BSES2 score, which indicates good surrogate properties for OS. In the adjuvant setting, our review suggests that DFS is a good surrogate endpoint for OS in colon cancer treated with fluoropyrimidine-based chemotherapy, operable and locally advanced non-small-cell lung cancer treated with chemotherapy and/or radiotherapy, curatively resected gastric cancer and locally advanced head and neck cancer treated with chemotherapy (12, 13, 24, 57, 61). Similarly, PFS showed very strong surrogate properties for OS in the context of advanced colorectal cancer treated with fluorouracil- or leucovorin-based chemotherapy, locally advanced lung cancer and locally advanced head and neck cancer treated with radiotherapy (13, 34, 61). Conversely, the following endpoints showed very poor surrogate properties for OS: response rate and a four-category response criteria in advanced colorectal cancer (28, 38); response rate, complete response and objective response in metastatic breast cancer (47, 52); pathologic complete response in breast cancer treated by neoadjuvant chemotherapy or targeted therapy (44). We also identified promising endpoints as surrogates for OS. Indeed, further evaluation of PFS as a surrogate endpoint for OS should be considered in metastatic colorectal cancer treated with targeted therapy in 1st line setting (30) and in glioblastoma (68). Similarly, DFS showed promising results in locally advanced head and neck cancer treated with chemotherapy (61).

It is expected that good surrogacy findings could be a proxy of innovation in the field and vice-versa. The availability of valid surrogates is indeed likely to drive researchers to use these surrogates as primary endpoints, leading to faster trials results and thus accelerates the innovation process. Conversely, innovative fields are likely to lead to more trials, more meta-analyses, and thus potentially more valid surrogate endpoints.

Our study presents some limitations. The two frameworks we relied on to assess of the strength of the association reported in the meta-analyses have not been formally validated. In addition, these frameworks investigate additional domains to the strength of the statistical association that we focused on. For instance, the IQWiG framework also investigates the reliability of the study to be rated as low, moderate, limited or high. Although the authors describe the study characteristics to consider, they did not develop a formal grading scale, so that appreciation of reliability may be subject to bias. Similarly, the BSES involves assessing the generalizability of the surrogate domain which emphasizes consistency across different drug classes. The recommendations so far were to look to drug class by class as a surrogate may be relevant for one drug class and not for the other. Difference in the results of different classes of drugs may be explained by different biological mechanisms. Depending of the diseases and the treatments available, this item may not be pertinent. As for the IQWiG framework, we thus limited the BSES evaluation to the statistical evaluation of the surrogacy measures. In addition, we relied on two independent assessments to limit subjectivity when ranking the strength of association. In case of discordance, an agreement was reached after discussion with a third independent reader (MS, DD, CB).

Interestingly, only two publications used formal validation grids to interpret their results: Mauguen et al. relied on the BSES grid (13) and Zer et al. (71) used a criterion proposed by Burnand et al. (78). Additionally, five publications assessed the strength of the association using criteria defined a priori (42, 43, 50, 61, 65). The subjectivity of these indicators, the absence of recommendation concerning their application, and more importantly the absence of reporting of some of the parameters required to apply these frameworks, are major limitations to their use and interpretation. For instance, the BSES score did not allow validating any of the endpoints evaluated, partly because of the absence of the STEP parameter, which was not reported in the meta-analyses.

Some studies lacked statistical rigor, which prevented us from accurately evaluating the strength of the associations. Indeed, a rather sizeable number of meta-analyses did not report the confidence intervals for correlation coefficients, even for the most recent publications. On the 69 meta-analyses reported in the 21 papers published since 2014, less than half (N_{MA} = 28; 40.6%) reported the confidence intervals or standard errors associated with the surrogacy measures, i.e. coefficient of correlation or coefficient of determination. In addition, only five (7.2%) relied on IPD and 10 (14.5%) reported the STE.

Possible explanations for the absence of strong evidence of surrogate endpoints in cancer RCTs are multiple: absence of standardized statistical methods and measures for the evaluation of surrogacy, heterogeneity of quality indicators for the assessment of the strength of evidence, use of non-exhaustive sets of trials, and difficulties in gathering IPD. In addition to these ‘technical’ explanations, clinical and biological rationale should also be considered (79). First, contrary to OS, alternative endpoints such as PFS are prone to biases such as

measurement errors, x-ray reader and clinician interpretation, or imprecise date of event. The change in tumor size required to be considered as a progression might also be too small to have an impact on time to death. In addition, with the introduction of immunotherapeutic agents, the definition of progression has to be adapted and the issue of surrogacy might be impacted. Discrepancies between treatment effect on OS and PFS has been observed in different cancer types treated with immunotherapy. For instance the phase III RCT evaluating eribulin mesylate versus dacarbazine in patients with advanced soft-tissue sarcoma highlighted a significant effect on OS (HR 0.77; 95% CI 0.62-0.95) but not on PFS (HR 0.88; 95% CI 0.71-1.09) (80). As of today however, there are not enough MA data to properly address the surrogacy issue in the context of immunotherapeutic agents. Finally, RCTs with crossover designs, especially unidirectional crossover, might complicate the surrogacy evaluation.

From a statistical point of view, the use of surrogates in place of OS for drug approvals 5, justifies their assessment. From a clinical perspective however, one can wonder if improving OS remains the primary objective in certain therapeutic situations. In metastatic cancer for instance, with the multiplication of lines of treatment, the change in OS is not only impacted by the experimental treatment but also by the sequence of post-progression treatments (79). One can then wonder if the treatment still aims at improving OS or focuses on controlling disease progression. If so, the correlation link between the alternative endpoints and OS might not be as central as in the adjuvant setting, and progression-based endpoints such as PFS or TTP might still be relevant even though they do not prove to be good surrogates for OS.

Conclusion

The literature on surrogate endpoints in cancer RCTs is evolving quickly. We provided a summary of evidence on alternative endpoints to OS to be used as primary efficacy endpoints in cancer trials for various therapeutic situations, solid and non-solid tumors, adjuvant and advanced settings. Overall, data suggests that DFS has adequate surrogate properties for OS in the context of adjuvant treatment for colon cancer, non-small-cell lung cancer, gastric cancer, and head and neck cancer. In advanced settings, PFS may be an appropriate surrogate endpoint for OS in the context of metastatic colorectal cancer, lung cancer, and head and neck cancer. Consistency in statistical methods is required for surrogacy validation, frameworks for assessment of surrogacy evidence, and reporting of summary parameters (including their precision), for improved assessment of published meta-analyses on surrogate endpoints.

References

1. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival End Point Reporting in Randomized Cancer Clinical Trials: A Review of Major Journals. J Clin Oncol. 2008 Aug 1;26(22):3721–6.
2. Sridhara R, Johnson JR, Justice R, Keegan P, Chakravarty A, Pazdur R. Review of oncology and hematology drug product approvals at the US Food and Drug Administration between July 2005 and December 2007. J Natl Cancer Inst. 2010 Feb 24;102(4):230–43.
3. Johnson JR, Ning Y-M, Farrell A, Justice R, Keegan P, Pazdur R. Accelerated approval of oncology products: the food and drug administration experience. J Natl Cancer Inst. 2011 Apr 20;103(8):636–44.
4. Kim C, Prasad V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. JAMA Intern Med. 2015 Dec 1;175(12):1992.
5. Kim C, Prasad V. Strength of Validation for Surrogate End Points Used in the US Food and Drug Administration's Approval of Oncology Drugs. Mayo Clin Proc. 2016 Jun;91(6):713–25.
6. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. JAMA Intern Med. 2015 Aug;175(8):1389–98.
7. McGivney WT. NCCN Guidelines and Their Impact on Coverage Policy. J Natl Compr Canc Netw. 2010 Jun 1;8(6):625–625.
8. Miller K, Wang M, Gralow J, Dickler M, Cobleigh M, Perez EA, et al. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. N Engl J Med. 2007 Dec 27;357(26):2666–76.
9. Carpenter D, Kesselheim AS, Joffe S. Reputation and precedent in the bevacizumab decision. N Engl J Med. 2011 Jul 14;365(2):e3.
10. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar;69(3):89–95.
11. Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports. Health Technol Assess Winch Engl. 2009 Jan;13(8):iii, ix–xi, 1-50.
12. Buyse M, Burzykowski T, Michiels S, Carroll K. Individual- and trial-level surrogacy in colorectal cancer. Stat Methods Med Res. 2008 Oct;17(5):467–75.
13. Mauguen A, Pignon J-P, Burdett S, Domerg C, Fisher D, Paulus R, et al. Surrogate endpoints for overall survival in chemotherapy and radiotherapy trials in operable and locally advanced lung cancer: a re-analysis of meta-analyses of individual patients' data. Lancet Oncol. 2013 Jun;14(7):619–26.
14. Burzykowski T, Buyse M, Piccart-Gebhart MJ, Sledge G, Carmichael J, Lück H-J, et al. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2008 Apr

20;26(12):1987–92.

15. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7):e1000097.
16. The Evaluation of Surrogate Endpoints | Tomasz Burzykowski | Springer [Internet]. [cited 2017 Sep 22]. Available from: http://www.springer.com/la/book/9780387202778
17. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostat Oxf Engl. 2000 Mar;1(1):49–67.
18. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. J R Stat Soc Ser C Appl Stat. 2001 Nov;50(4):405–22.
19. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat. 2006 Jul;5(3):173–86.
20. Validity of surrogate endpoints in oncology: Executive summary of rapid report A10-05, Version 1.1. In: Institute for Quality and Efficiency in Health Care: Executive Summaries [Internet]. Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2005. Available from: http://www.ncbi.nlm.nih.gov/books/NBK198799/
21. Lassere MN, Johnson KR, Schiff M, Rees D. Is blood pressure reduction a valid surrogate endpoint for stroke prevention? an analysis incorporating a systematic review of randomised controlled trials, a by-trial weighted errors-in-variables regression, the surrogate threshold effect (STE) and the biomarker-surrogacy (BioSurrogate) evaluation schema (BSES). BMC Med Res Methodol [Internet]. 2012 Dec [cited 2016 Dec 27];12(1). Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-27
22. Sargent D, Shi Q, Yothers G, Van Cutsem E, Cassidy J, Saltz L, et al. Two or three year disease-free survival (DFS) as a primary end-point in stage III adjuvant colon cancer trials with fluoropyrimidines with or without oxaliplatin or irinotecan: data from 12,676 patients from MOSAIC, X-ACT, PETACC-3, C-06, C-07 and C89803. Eur J Cancer Oxf Engl 1990. 2011 May;47(7):990–6.
23. Alonso A, Molenberghs G. Evaluating time to cancer recurrence as a surrogate marker for survival from an information theory perspective. Stat Methods Med Res. 2008 Oct;17(5):497–504.
24. Green E, Yothers G, Sargent DJ. Surrogate endpoint validation: statistical elegance versus clinical relevance. Stat Methods Med Res. 2008 Oct;17(5):477–86.
25. Sargent DJ, Patiyil S, Yothers G, Haller DG, Gray R, Benedetti J, et al. End points for colon cancer adjuvant trials: observations and recommendations based on individual patient data from 20,898 patients enrolled onto 18 randomized trials from the ACCENT Group. J Clin Oncol Off J Am Soc Clin Oncol. 2007 Oct 10;25(29):4569–74.
26. Sargent DJ, Wieand HS, Haller DG, Gray R, Benedetti JK, Buyse M, et al. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol Off J Am Soc Clin Oncol. 2005 Dec 1;23(34):8664–70.

27. Ghosh D, Taylor JMG, Sargent DJ. Meta-analysis for surrogacy: accelerated failure time models and semicompeting risks modeling. Biometrics. 2012 Mar;68(1):226–32.
28. Ciani O, Buyse M, Garside R, Peters J, Saad ED, Stein K, et al. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. J Clin Epidemiol. 2015 Jul;68(7):833–42.
29. Shi Q, de Gramont A, Grothey A, Zalcberg J, Chibaudel B, Schmoll H-J, et al. Individual patient data analysis of progression-free survival versus overall survival as a first-line end point for metastatic colorectal cancer in modern randomized trials: findings from the analysis and research in cancers of the digestive system database. J Clin Oncol Off J Am Soc Clin Oncol. 2015 Jan 1;33(1):22–8.
30. Sidhu R, Rong A, Dahlberg S. Evaluation of progression-free survival as a surrogate endpoint for survival in chemotherapy and targeted agent metastatic colorectal cancer trials. Clin Cancer Res Off J Am Assoc Cancer Res. 2013 Mar 1;19(5):969–76.
31. Giessen C, Laubender RP, Ankerst DP, Stintzing S, Modest DP, Mansmann U, et al. Progression-free survival as a surrogate endpoint for median overall survival in metastatic colorectal cancer: literature-based analysis from 50 randomized first-line trials. Clin Cancer Res Off J Am Assoc Cancer Res. 2013 Jan 1;19(1):225–35.
32. Chirila C, Odom D, Devercelli G, Khan S, Sherif BN, Kaye JA, et al. Meta-analysis of the association between progression-free survival and overall survival in metastatic colorectal cancer. Int J Colorectal Dis. 2012 May;27(5):623–34.
33. Wilkerson J, Fojo T. Progression-free survival is simply a measure of a drug's effect while administered and is not a surrogate for overall survival. Cancer J Sudbury Mass. 2009 Oct;15(5):379–85.
34. Buyse M, Burzykowski T, Carroll K, Michiels S, Sargent DJ, Miller LL, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2007 Nov 20;25(33):5218–24.
35. Tang PA, Bentzen SM, Chen EX, Siu LL. Surrogate end points for median overall survival in metastatic colorectal cancer: literature-based analysis from 39 randomized controlled trials of first-line chemotherapy. J Clin Oncol Off J Am Soc Clin Oncol. 2007 Oct 10;25(29):4562–8.
36. Johnson KR, Ringland C, Stokes BJ, Anthony DM, Freemantle N, Irs A, et al. Response rate or time to progression as predictors of survival in trials of metastatic colorectal cancer or non-small-cell lung cancer: a meta-analysis. Lancet Oncol. 2006 Sep;7(9):741–6.
37. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-Analysis Group in Cancer. Lancet Lond Engl. 2000 Jul 29;356(9227):373–8.
38. Burzykowski T, Molenberghs G, Buyse M. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. J R Stat Soc Ser A Stat Soc. 2004 Feb;167(1):103–24.
39. Hotta K, Kato Y, Leighl N, Takigawa N, Gaafar RM, Kayatani H, et al. Magnitude of the benefit of

progression-free survival as a potential surrogate marker in phase 3 trials assessing targeted agents in molecularly selected patients with advanced non-small cell lung cancer: systematic review. PloS One. 2015;10(3):e0121211.

40. Hayashi H, Okamoto I, Taguri M, Morita S, Nakagawa K. Postprogression survival in patients with advanced non-small-cell lung cancer who receive second-line or third-line chemotherapy. Clin Lung Cancer. 2013 May;14(3):261–6.
41. Foster NR, Renfro LA, Schild SE, Redman MW, Wang XF, Dahlberg SE, et al. Multitrial Evaluation of Progression-Free Survival as a Surrogate End Point for Overall Survival in First-Line Extensive-Stage Small-Cell Lung Cancer. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer. 2015 Jul;10(7):1099–106.
42. Hotta K, Fujiwara Y, Matsuo K, Kiura K, Takigawa N, Tabata M, et al. Time to progression as a surrogate marker for overall survival in patients with advanced non-small cell lung cancer. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer. 2009 Mar;4(3):311–7.
43. Ng R, Pond GR, Tang PA, MacIntosh PW, Siu LL, Chen EX. Correlation of changes between 2-year disease-free survival and 5-year overall survival in adjuvant breast cancer trials from 1966 to 2006. Ann Oncol Off J Eur Soc Med Oncol. 2008 Mar;19(3):481–6.
44. Berruti A, Amoroso V, Gallo F, Bertaglia V, Simoncini E, Pedersini R, et al. Pathologic complete response as a potential surrogate for the clinical outcome in patients with breast cancer after neoadjuvant therapy: a meta-regression of 29 randomized prospective studies. J Clin Oncol Off J Am Soc Clin Oncol. 2014 Dec 1;32(34):3883–91.
45. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet Lond Engl. 2014 Jul 12;384(9938):164–72.
46. Miksad RA, Zietemann V, Gothe R, Schwarzer R, Conrads-Frank A, Schnell-Inderst P, et al. Progression-free survival as a surrogate endpoint in advanced breast cancer. Int J Technol Assess Health Care. 2008;24(4):371–83.
47. Hackshaw A, Knight A, Barrett-Lee P, Leonard R. Surrogate markers and survival in women receiving first-line combination anthracycline chemotherapy for advanced breast cancer. Br J Cancer. 2005 Nov 28;93(11):1215–21.
48. Adunlin G, Cyrus JWW, Dranitsaris G. Correlation between progression-free survival and overall survival in metastatic breast cancer patients receiving anthracyclines, taxanes, or targeted therapies: a trial-level meta-analysis. Breast Cancer Res Treat. 2015 Dec;154(3):591–608.
49. Petrelli F, Barni S. Surrogate endpoints in metastatic breast cancer treated with targeted therapies: an analysis of the first-line phase III trials. Med Oncol Northwood Lond Engl. 2014 Jan;31(1):776.
50. Beauchemin C, Cooper D, Lapierre M-È, Yelle L, Lachaine J. Progression-free survival as a potential surrogate for overall survival in metastatic breast cancer. OncoTargets Ther. 2014;7:1101–10.
51. Sherrill B, Amonkar M, Wu Y, Hirst C, Stein S, Walker M, et al. Relationship between effects on time-to-disease progression and overall survival in studies of metastatic breast cancer. Br J

Cancer. 2008 Nov 18;99(10):1572–8.

52. Bruzzi P, Del Mastro L, Sormani MP, Bastholt L, Danova M, Focan C, et al. Objective response to chemotherapy as a potential surrogate end point of survival in metastatic breast cancer patients. J Clin Oncol Off J Am Soc Clin Oncol. 2005 Aug 1;23(22):5117–25.
53. Bria E, Massari F, Maines F, Pilotto S, Bonomi M, Porta C, et al. Progression-free survival as primary endpoint in randomized clinical trials of targeted agents for advanced renal cell carcinoma. Correlation with overall survival, benchmarking and power analysis. Crit Rev Oncol Hematol. 2015 Jan;93(1):50–9.
54. Johnson KR, Liauw W, Lassere MND. Evaluating surrogacy metrics and investigating approval decisions of progression-free survival (PFS) in metastatic renal cell cancer: a systematic review. Ann Oncol Off J Eur Soc Med Oncol. 2015 Mar;26(3):485–96.
55. Petrelli F, Barni S. Surrogate end points and postprogression survival in renal cell carcinoma: an analysis of first-line trials with targeted therapies. Clin Genitourin Cancer. 2013 Dec;11(4):385–9.
56. Delea TE, Khuu A, Heng DYC, Haas T, Soulières D. Association between treatment effects on disease progression end points and overall survival in clinical studies of patients with metastatic renal cell carcinoma. Br J Cancer. 2012 Sep 25;107(7):1059–68.
57. Oba K, Paoletti X, Alberts S, Bang Y-J, Benedetti J, Bleiberg H, et al. Disease-free survival as a surrogate for overall survival in adjuvant trials of gastric cancer: a meta-analysis. J Natl Cancer Inst. 2013 Nov 6;105(21):1600–7.
58. Paoletti X, Oba K, Bang Y-J, Bleiberg H, Boku N, Bouché O, et al. Progression-free survival as a surrogate for overall survival in advanced/recurrent gastric cancer trials: a meta-analysis. J Natl Cancer Inst. 2013 Nov 6;105(21):1667–70.
59. Shitara K, Matsuo K, Muro K, Doi T, Ohtsu A. Correlation between overall survival and other endpoints in clinical trials of second-line chemotherapy for patients with advanced gastric cancer. Gastric Cancer Off J Int Gastric Cancer Assoc Jpn Gastric Cancer Assoc. 2014 Apr;17(2):362–70.
60. Shitara K, Ikeda J, Yokota T, Takahari D, Ura T, Muro K, et al. Progression-free survival and time to progression as surrogate markers of overall survival in patients with advanced gastric cancer: analysis of 36 randomized trials. Invest New Drugs. 2012 Jun;30(3):1224–31.
61. Michiels S, Le Maître A, Buyse M, Burzykowski T, Maillard E, Bogaerts J, et al. Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta-analyses of individual patient data. Lancet Oncol. 2009 Apr;10(4):341–50.
62. Petrelli F, Coinu A, Borgonovo K, Cabiddu M, Barni S. Progression-free survival as surrogate endpoint in advanced pancreatic cancer: meta-analysis of 30 randomized first-line trials. Hepatobiliary Pancreat Dis Int HBPD INT. 2015 Apr;14(2):124–31.
63. Colloca G, Venturino A, Guarneri D. Analysis of Response-Related and Time-to-event Endpoints in Randomized Trials of Gemcitabine-Based Treatment Versus Gemcitabine Alone as First-Line Treatment of Patients With Advanced Pancreatic Cancer. Clin Colorectal Cancer. 2016 Sep;15(3):264–76.
64. Colloca G, Venturino A, Guarneri D. Analysis of Clinical End Points of Randomised Trials Including

Bevacizumab and Chemotherapy versus Chemotherapy as First-line Treatment of Metastatic Colorectal Cancer. Clin Oncol R Coll Radiol G B. 2016 Oct;28(10):e155-164.

65. Chen Y-P, Sun Y, Chen L, Mao Y-P, Tang L-L, Li W-F, et al. Surrogate endpoints for overall survival in combined chemotherapy and radiotherapy trials in nasopharyngeal carcinoma: Meta-analysis of randomised controlled trials. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2015 Aug;116(2):157–66.
66. Cartier S, Zhang B, Rosen VM, Zarotsky V, Bartlett JB, Mukhopadhyay P, et al. Relationship between treatment effects on progression-free survival and overall survival in multiple myeloma: a systematic review and meta-analysis of published clinical trial data. Oncol Res Treat. 2015;38(3):88–94.
67. Flaherty KT, Hennig M, Lee SJ, Ascierto PA, Dummer R, Eggermont AMM, et al. Surrogate endpoints for overall survival in metastatic melanoma: a meta-analysis of randomised controlled trials. Lancet Oncol. 2014 Mar;15(3):297–304.
68. Han K, Ren M, Wick W, Abrey L, Das A, Jin J, et al. Progression-free survival as a surrogate endpoint for overall survival in glioblastoma: a literature-based meta-analysis from 91 trials. Neuro-Oncol. 2014 May 1;16(5):696–706.
69. Lee L, Wang L, Crump M. Identification of potential surrogate end points in randomized clinical trials of aggressive and indolent non-Hodgkin's lymphoma: correlation of complete response, time-to-event and overall survival end points. Ann Oncol Off J Eur Soc Med Oncol. 2011 Jun;22(6):1392–403.
70. Moriwaki T, Yamamoto Y, Gosho M, Kobayashi M, Sugaya A, Yamada T, et al. Correlations of survival with progression-free survival, response rate, and disease control rate in advanced biliary tract cancer: a meta-analysis of randomised trials of first-line chemotherapy. Br J Cancer. 2016 Apr 12;114(8):881–8.
71. Zer A, Prince RM, Amir E, Abdul Razak A. Evolution of Randomized Trials in Advanced/Metastatic Soft Tissue Sarcoma: End Point Selection, Surrogacy, and Quality of Reporting. J Clin Oncol Off J Am Soc Clin Oncol. 2016 May 1;34(13):1469–75.
72. Sherrill B, Kaye JA, Sandin R, Cappelleri JC, Chen C. Review of meta-analyses evaluating surrogate endpoints for overall survival in oncology. OncoTargets Ther. 2012;5:287–96.
73. Ciani O, Davis S, Tappenden P, Garside R, Stein K, Cantrell A, et al. Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. Int J Technol Assess Health Care. 2014 Jul;30(3):312–24.
74. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA. 1998 Jan 28;279(4):281–6.
75. Krzyzanowska MK, Pintilie M, Tannock IF. Factors associated with failure to publish large randomized trials presented at an oncology meeting. JAMA. 2003 Jul 23;290(4):495–501.
76. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A-W, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PloS One. 2008 Aug 28;3(8):e3081.

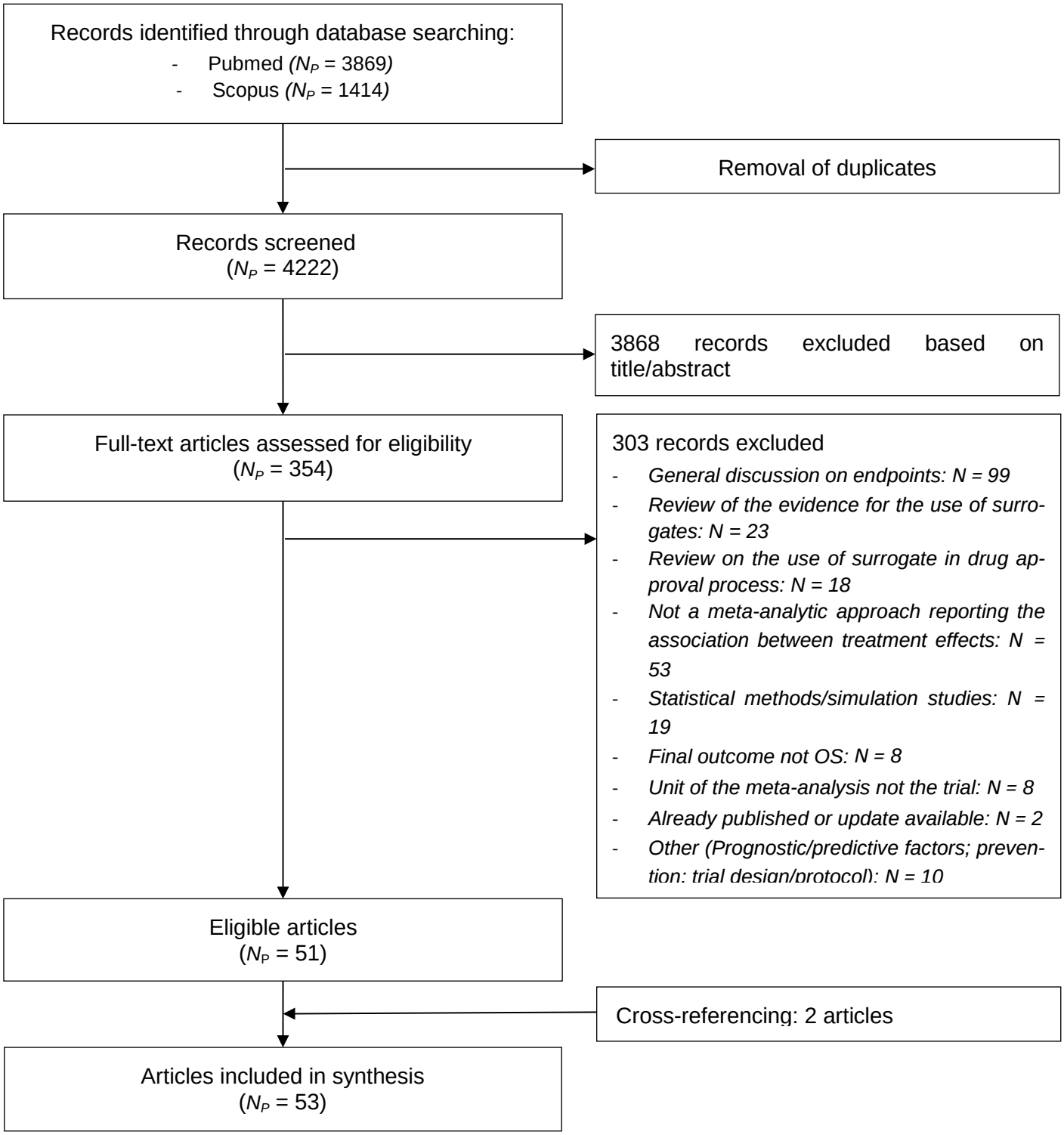
77. Dwan K, Gamble C, Williamson PR, Kirkham JJ, Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PloS One. 2013;8(7):e66844.
78. Burnand B, Kernan WN, Feinstein AR. Indexes and boundaries for “quantitative significance” in statistical decisions. J Clin Epidemiol. 1990;43(12):1273–84.
79. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? J Clin Oncol Off J Am Soc Clin Oncol. 2012 Apr 1;30(10):1030–3.
80. Schöffski P, Chawla S, Maki RG, Italiano A, Gelderblom H, Choy E, et al. Eribulin versus dacarbazine in previously treated patients with advanced liposarcoma or leiomyosarcoma: a randomised, open-label, multicentre, phase 3 trial. Lancet Lond Engl. 2016 Apr 16;387(10028):1629–37.
81. Thirion P, Michiels S, Pignon JP, Buyse M, Braud AC, Carlson RW, et al. Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: an updated meta-analysis. J Clin Oncol Off J Am Soc Clin Oncol. 2004 Sep 15;22(18):3766–75.
82. Advanced Colorectal Cancer Meta-Analysis Project. Modulation of 5-fluorouracil by leucovorin in patients with advanced colorectal cancer: evidence in terms of response rate. J Clin Oncol 1992;10:896–903.
83. Advanced Colorectal Cancer Meta-Analysis Project. Meta-analysis of randomized trials testing the biochemical modulation of 5-fluorouracil by methotrexate in metastatic colorectal cancer. J Clin Oncol 1994;12:960–69.
84. Meta-Analysis Group In Cancer. Efficacy of intravenous continuous infusion of 5-fluorouracil compared with bolus administration in patients with advanced colorectal cancer. J Clin Oncol 1998;16:301–08.
85. Meta-Analysis Group In Cancer. Reappraisal of hepatic arterial infusion in the treatment of nonresectable liver metastases from colorectal cancer. J Natl Cancer Inst 1996;88:252–58.
86. Non-small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. Non-small Cell Lung Cancer Collaborative Group. BMJ. 1995 Oct 7;311(7010):899–909.
87. Aupérin A, Le Péchoux C, Rolland E, Curran WJ, Furuse K, Fournel P, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2010 May 1;28(13):2181–90.
88. NSCLC Meta-analyses Collaborative Group, Arriagada R, Auperin A, Burdett S, Higgins JP, Johnson DH, et al. Adjuvant chemotherapy, with or without postoperative radiotherapy, in operable non-small-cell lung cancer: two meta-analyses of individual patient data. Lancet Lond Engl. 2010 Apr 10;375(9722):1267–77.
89. Mauguen A, Le Péchoux C, Saunders MI, Schild SE, Turrisi AT, Baumann M, et al. Hyperfractionated or accelerated radiotherapy in lung cancer: an individual patient data meta-analysis. J Clin Oncol Off J Am Soc Clin Oncol. 2012 Aug 1;30(22):2788–97.
90. Bourhis J, Le Maître A, Baujat B, Audry H, Pignon J-P, Meta-Analysis of Chemotherapy in Head,

Neck Cancer Collaborative Group, et al. Individual patients' data meta-analyses in head and neck cancer. Curr Opin Oncol. 2007 May;19(3):188–94.

91. Pignon JP, Bourhis J, Domenge C, Designé L. Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. MACH-NC Collaborative Group. Meta-Analysis of Chemotherapy on Head and Neck Cancer. Lancet Lond Engl. 2000 Mar 18;355(9208):949–55.
92. Pignon J-P, le Maître A, Bourhis J, MACH-NC Collaborative Group. Meta-Analyses of Chemotherapy in Head and Neck Cancer (MACH-NC): an update. Int J Radiat Oncol Biol Phys. 2007;69(2 Suppl):S112-114.
93. Bourhis J, Overgaard J, Audry H, Ang KK, Saunders M, Bernier J, et al. Hyperfractionated or accelerated radiotherapy in head and neck cancer: a meta-analysis. Lancet Lond Engl. 2006 Sep 2;368(9538):843–54

Figures

Figure 1: Flow of information through the different phases of the systematic review of studies reporting on the validation of surrogate endpoints for overall survival in randomized cancer trials, as per PRISMA guidelines (15)



Tables

Table 1: Characteristics of the meta-analyses reporting on surrogate time-to-event endpoints for overall survival in randomized cancer trials (NMA = 164)

	N_{MA}	%
Tumor type (number of publications investigating the tumor type)		
Colorectal (N_P = 18)	46	28
Breast (N_P = 12)	35	21.3
Lung (N_P = 6)	17	10.4
Pancreas (N_P = 3)	23	14
Biliary tract cancer (N_P = 1)	9	5.5
Renal Cell Carcinoma (N_P = 4)	8	4.9
Head and Neck (N_P = 1)	8	4.9
Gastric (N_P = 4)	4	2.4
Non-hodgkin lymphoma (N_P = 1)	4	2.4
Nasopharyngeal carcinoma (N_P = 1)	2	1.2
Soft-tissue sarcoma (N_P = 1)	2	1.2
Ovarian (N_P = 1)	1	0.6
Glioblastoma (N_P = 1)	1	0.6
Melanoma (N_P = 1)	1	0.6
Multiple myeloma (N_P = 1)	1	0.6
All solid tumors (N_P = 1)	2	1.2
Investigational treatment setting		
Neoadjuvant	5	3
Adjuvant	18	11
Locally advanced	13	7.9
Advanced/metastatic	122	74.4
All settings	6	3.7
Candidate surrogate endpoint by setting		
Neoadjuvant		
Pathologic complete response (pCR)	5	100
Adjuvant		
Disease-free survival (DFS)	17	94.4
Time-to-recurrence (TTR)	1	5.6
Locally advanced		
Progression-free survival (PFS)	8	61.5
Time-to progression (TTP)	1	7.7
Duration of locoregional control (DLRC)	4	30.8
Advanced / Metastatic		
Progression-free survival (PFS)	45	36.9
Time-to progression (TTP)	11	9
PFS or TTP (PFS/TTP)	23	18.9
Response rate (RR)	24	19.7
Time-to failure (TTF)	1	0.8
Other	18	14.8
All settings		
Progression-free survival (PFS)	2	33.3
PFS/TTP	2	33.3
Complete response (CR)	2	33.3
Data source		
Published articles	49	29.9

	N_{MA}	%
Published articles and abstracts	53	32.3
Published articles and trial registries	4	2.4
Published articles, abstracts and trial registries	27	16.5
Convenience sample	25	15.6
Not well described	6	3.7
Type of data		
Individual patient data	39	23.8
Aggregated data	125	76.2
Surrogacy measures and statistical methods		
Individual-level surrogacy*		
Non-parametric rank correlation measure	19	11.6
Weighted linear regression on rates or median times	21	12.8
2-step model	28	17.1
Other (new methodology)	4	2.4
Trial-level surrogacy*		
Non-parametric rank correlation measure	42	25.6
Weighted linear regression or error-in-measurement model on treatment effects estimated by HR	67	40.9
Weighted linear regression model on treatment effects estimated by ratios or differences of medians	42	25.6
2-step model	36	22
Other (new methodology)	4	2.4
Surrogate threshold effect (STE)	17	10.4
Surrogate threshold effect proportion (STEP)	0	0

* Does not add-up to 100% since several methods can be used in each meta-analysis

Table 2: Meta-analyses assessing surrogate endpoints for OS in cancer randomized trials: characteristics, results, and strength of the association (164 meta-analyses published in 53 manuscripts)

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
Colon cancer														
DFS	(22)	Stage II or III patients	NA	Adjuvant, fluoropyrimidines with or without oxaliplatin or irinotecan	1997 to 2002	Convenience sample (26)	IPD	6	12 676	--	R^2_{HR} = 0.58 [0.02; 1] R^2_{2SM} = 0.37 [0; 0.98]	--	Medium	NE
		Stage III patients	NA	Adjuvant, fluoropyrimidines with or without oxaliplatin or irinotecan	1997 to 2002	Convenience sample (26)	IPD	6	9 395	--	R^2_{HR} = 0.91 [0.54; 1] R^2_{2SM} = 0.86 [0.64; 1]	--	Medium +	NE
	(12)	Stage II or III patients	NA	Adjuvant, fluoropyrimidines alone or in combination (ACCENT dataset)	1977 to 1999	Convenience sample (26)	IPD	10	10 255	ρ_{2SM} = 0.96 [0.95; 0.97]	R^2_{2SM} = 0.94 [0.87; 1.01]	0.93	High	Excellent
	(24)	Stage II or III patients	NA	Adjuvant, fluoropyrimidines alone or in combination (ACCENT dataset)	1977 to 1999	Convenience sample (26)	IPD	13	>10 000	τ_{2SM} = 0.85 [0.72; 0.99] R^2 = 0.88; ρ = 0.93	R^2_{2SM} = 0.90 [0.89; 0.90] R^2_{HR} = 0.80; ρ = 0.85	--	High	Excellent
	(25)	Stage II patients	NA	Adjuvant, fluoropyrimidines alone or in combination (ACCENT dataset)	1977 to 1999	Convenience sample (26)	IPD	18	6 966	--	R^2_{2SM} = 0.70 [0.47; 0.93] ρ = 0.70 [0.44; 0.80]	--	Medium	NE
		Stage III patients	NA	Adjuvant, fluoropyrimidines alone or in combination (ACCENT dataset)	1977 to 1999	Convenience sample (26)	IPD	18	13 932	--	R^2_{2SM} = 0.88 [0.78; 0.98] ρ = 0.92 [0.83; 0.95]	--	High	NE
	(26)	None	NA	Adjuvant, fluoropyrimidines alone or in combination (ACCENT dataset)	1977 to 1999	Not well described	IPD	18	20 898	τ_{2SM} = 0.87 [0.87; 0.88]	R^2_{2SM} = 0.78 [0.60; 0.96]	--	Medium +	Excellent
TTR	(27)	Stage II or III patients	NA	Adjuvant, fluoropyrimidines alone or in combination (ACCENT dataset)	1977 to 1999	Convenience sample (26)	IPD	12	13 977	τ_{2SM} = 0.86	R^2_{2SM} = 0.96 [0.93; 0.99]	--	High	NE
	(23)	Stage II or III patients	NA	Adjuvant, fluoropyrimidines alone or in combination (ACCENT dataset)	1977 to 1999	Convenience sample (26)	IPD	10	10 255	R^2_h = 0.84 [0.83; 0.85]	R^2_{2SM} = 0.82 [0.44; 0.95] R^2_h = 0.85 [0.53; 0.96]	--	Medium +	Good
Colorectal cancer														
PFS	(28)	Advanced or metastatic disease	All lines	Pharmacologic therapies	2003 to 2013*	Articles, abstracts and trial registries	AD	36	NS	--	R^2_{HR} = 0.34 [0.10; 0.59] ρ = 0.75	0.8	Medium	NE
	(29)	Metastatic disease	1st line	Biologic and non-biologic agents	1997 to 2006	Not well described	IPD	22	16 762	R^2 = 0.69 [0.58; 0.79] ρ_{2SM} = 0.51 [0.50; 0.52]	R^2_{HR} = 0.54 [0.33; 0.75] R^2_{2SM} = 0.46 [0.24; 0.68]	0.57	Medium	Fair
		Metastatic disease	1st line	Non-biologic agents only	1997 to 2006	Not well described	IPD	22	9 439	R^2 = 0.59 [0.39; 0.79] ρ_{2SM} = 0.47 [0.46; 0.49]	R^2_{HR} = 0.59 [0.31; 0.87] R^2_{2SM} = 0.35 [0; 0.71]	0,59	Medium	Poor
		Metastatic disease	1st line	Biologic agents in at least one treatment arm	1997 to 2006	Not well described	IPD	22	7 323	R^2 = 0.69 [0.48; 0.91] ρ_{2SM} = 0.55 [0.54; 0.56]	R^2_{HR} = 0.52 [0.24; 0.80] R^2_{2SM} = 0.45 [0.16; 0.75]	0,45	Medium	Poor
	(30)	Metastatic disease	All lines	Fluoropyrimidines alone or in combination	2000 to 2011*	Articles and abstracts	AD	24	20 438	--	R^2_{HR} = 0.73 [0.53; 0.85] r_{HR} = 0.86 [0.73; 0.92]	0.90	Medium +	NE
		Metastatic disease	1st line	Fluoropyrimidines alone or in combination	2000 to 2011*	Articles and abstracts	AD	18	14 124	--	R^2_{HR} = 0.80 [0.60; 0.90] r_{HR} = 0.90 [0.70; 0.95]	0.91	Medium +	NE
		Metastatic disease	All lines	Fluoropyrimidines alone or in combination and targeted therapies in at least 1 arm	2000 to 2011*	Articles and abstracts	AD	12	12 060	--	R^2_{HR} = 0.80 [0.55; 0.91] r_{HR} = 0.89 [0.74; 0.95]	0.87	Medium +	NE
		Metastatic disease	1st line	Fluoropyrimidines alone or in combination and targeted therapies in at least 1 arm	2000 to 2011*	Articles and abstracts	AD	8	7309	--	R^2_{HR} = 0.91 [0.73; 0.96] r_{HR} = 0.95 [0.86; 0.98]	0.90	High	NE
		Metastatic disease	All lines	Fluoropyrimidines alone or in combination and anti-EGFR antibody therapies in at least 1 arm	2000 to 2011*	Articles and abstracts	AD	9	7 792	--	R^2_{HR} = 0.68 [0.32; 0.87] r_{HR} = 0.83 [0.56; 0.93]	0.77	Medium	NE
		Patients with advanced or metastatic wild-type *KRAS* tumors	All lines	Fluoropyrimidines alone or in combination and anti-EGFR antibodies in at least 1 arm	2000 to 2011*	Articles and abstracts	AD	9	NS	--	R^2_{HR} = 0.72 [0.24; 0.91] r_{HR} = 0.85 [0.49; 0.95]	0.72	Medium +	NE

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
	(31)	Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	None	2000 to 2012*	Articles and abstracts	AD	50	22 736	R^2 = 0.86 [0.79; 0.91]	R^2_{HR} = 0.87 [0.67; 0.93]	--	Medium +	Excellent
		Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	Chemotherapy only	2000 to 2012*	Articles and abstracts	AD	40	17 887	R^2 = 0.81 [0.71; 0.88]	R^2_{HR} = 0.93 [0.49; 0.97]	--	Medium +	Good
		Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	Oxaliplatin-based chemotherapy	2000 to 2012*	Articles and abstracts	AD	31	10 060	R^2 = 0.69 [0.36; 0.87]	R^2_{HR} = 0.68 [0.41; 0.85]	--	Medium	Fair
		Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	Irinotecan-based chemotherapy	2000 to 2012*	Articles and abstracts	AD	24	7 301	R^2 = 0.74 [0.59; 0.86]	R^2_{HR} = 0.82 [0.52; 0.95]	--	Medium +	Good
		Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	Chemotherapy + antibodies	2000 to 2012*	Articles and abstracts	AD	19	4 849	R^2 = 0.52 [0.09; 0.88]	R^2_{HR} = 0.47 [0.05; 0.72]	--	Medium -	Poor
		Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	Chemotherapy + bevacizumab	2000 to 2012*	Articles and abstracts	AD	11	3 310	R^2 = 0.45 [0.00; 0.84]	R^2_{HR} = 0.84 [0.05; 0.94]	--	Medium +	Poor
		Advanced (no locally advanced/ unresectable) or metastatic disease; patients with wild-type *KRAS* tumors only	1st line	Chemotherapy + anti-EGFR antibodies	2000 to 2012*	Articles and abstracts	AD	7	1 335	R^2 = 0.96 [-0.76; 1]	R^2_{HR} = 0.28 [0; 0.92]	--	Medium	Poor
	(32)	Metastatic disease	1st, 2nd or 3rd line	None	< 2009*	Articles and abstracts	AD	35	NS	r = 0.89 [0.83; 0.93] ρ = 0.78 [0.66; 0.85]	R^2_{med} = 0.59	--	NE	NE
	(33)	Advanced or metastatic disease	NS	None	NS	Articles	AD	NS		--	R^2_{HR} = 0.52	--	Medium	NE
	(34)	Advanced disease	NS	At least 1 arm with fluorouracil + leucovorin (c)	1981 to 1990	Not well described	IPD	10	3 089	ρ_{2SM} = 0.82 [0.82; 0.83]	R^2_{2SM} = 0.99 [0.94;1.04]	0.86	High	Excellent
	(35)	Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	None	1999 to 2005*	Articles, abstracts and trial registries	AD	39	18 668	ρ = 0.79 [0.65; 0.87]	R^2_{med} = 0.65 ρ = 0.74 [0.47; 0.88]	--	NE	NE
TTP	(28)	Advanced or metastatic disease	All lines	Pharmacologic therapies	2003 to 2013	Articles, abstracts and trial registries	AD	9	NS	--	R^2_{HR} = 0.65 [0.09; 0.92] ρ = 0.80	0.61	Medium	NE
	(33)	Metastatic disease	1st, 2nd or 3rd line	None	< 2009*	Articles and abstracts	AD	27	NS	r = 0.75 [0.59; 0.84] ρ = 0.59 [0.37; 0.74]	R^2_{med} = 0.32	--	NE	NE
	(34)	Advanced or metastatic disease	NS	At least 1 arm with fluouracil + leucovorin (c)	1981 to 1990	Convenience sample (81)	IPD	10	3 089	R^2_h = 0.84 [0.82; 0.85]	R^2_h = 0.82 [0.40; 0.95] R^2_{2SM} = 0.88 [0.52; 0.97]	--	Medium +	Good
	(36)	Advanced or metastatic disease	1st line	Chemotherapy	< 2005*	Articles and abstracts	AD	146	35 337	--	R^2_{med} = 0.33	--	NE	NE
	(35)	Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	None	1999 to 2005*	Articles, abstracts and trial registries	AD	39	18 668	ρ = 0.24 [0.13; 0.55]	ρ = 0.52 [0; 0.81]	--	NE	NE
PFS/TTP	(32)	Metastatic disease	1st line	None	< 2009*	Articles and abstracts	AD	62	23 527	r = 0.87 [0.82; 0.91] ρ = 0.76 [0.67; 0.82]	R^2_{med} = 0.48	--	NE	NE
	(32)	Metastatic disease	2nd line	None	< 2009*	Articles and abstracts	AD	48	NS	--	R^2_{med} = 0.54	--	NE	NE
	(32)	Metastatic disease	1st, 2nd or 3rd line	None	< 2009*	Articles and abstracts	AD	13	NS	--	R^2_{med} = 0.37	--	NE	NE

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
	(32)	Metastatic disease	1st, 2nd or 3rd line	None	< 2009*	Articles and abstracts	AD	20	NS	--	R^2_{HR} = 0.69	--	NE	NE
RR	(28)	Advanced or metastatic disease	All lines	Pharmacologic therapies	2003 to 2013*	Articles, abstracts and trial registries	AD	32	NS	--	R^2_{HR} = 0.06 [0.01; 0.29] ρ = 0.53	< 0.28	Low	NE
	(30)	Metastatic disease	All lines	Fluoropyrimidines alone or in combination and targeted therapies in at least 1 arm, phase-III trials only	2000 to 2011*	Articles and abstracts	AD	12	12 060	--	r_{HR} = 0.50	--	Medium	NE
	(30)	Patients with advanced or metastatic wild-type *KRAS* tumors	All lines	Fluoropyrimidines alone or in combination and anti-EGFR antibodies in at least 1 arm	2000 to 2011*	Articles and abstracts	AD	9	NS	--	r_{HR} = 0.68	--	Medium	NE
	(36)	Advanced or metastatic disease	1st line	Chemotherapy	< 2005*	Articles and abstracts	AD	146	NS	--	R^2_{med} = 0.10	--	NE	NE
	(35)	Advanced (no locally advanced/ unresectable) or metastatic disease	1st line	None	1999 to 2005*	Articles, abstracts and trial registries	AD	39	18 668	ρ = 0,59 [0.42; 0.72]	ρ = 0.39 [0.08; 0.63]	--	NE	NE
	(37)	Advanced disease	1st line	Fluoropyrimidine with fluorouracil or floxuridine	< 1991*	Convenience sample (82-85)	IPD	25	3 791	--	R^2_{HR} = 0.38 [0.09; 0.68]	--	Medium	NE
Response	(38)	Advanced disease	1st line	Fluoropyrimidine with fluorouracil or floxuridine	1990 to 1996	Convenience sample (82-85)	IPD	27	4 010	θ_{2SM} = 6.78 [6.01; 7.55]	R^2_{2SM} = 0.16 [0; 0.42]	--	Low	Poor
Lung cancer														
DFS	(13)	Operable and locally advanced NSCLC	NA	Adjuvant chemotherapy *vs* no chemotherapy	NS	Convenience sample (86-89)	IPD	17	5 319	τ_{2SM} = 0.83 [0.83; 0.83]	R^2_{2SM} = 0.92 [0.88; 0.95]	--	High	Excellent
	(13)	Operable and locally advanced NSCLC	NA	Adjuvant radiotherapy + chemotherapy *vs* no chemotherapy	NS	Convenience sample (86-89)	IPD	7	2 247	τ_{2SM} = 0.87 [0.87; 0.87]	R^2_{2SM} = 0.99 [0.98; 1]	--	High	Excellent
PFS	(39)	Advanced NSCLC with or without molecular selection	NS	Molecular targeted agents alone (not in combination with other treatment modalities)	2003 to 2014*	Articles, abstracts and trial registries	AD	18	7 633	--	R^2_{HR} = 0.23	--	Medium	NE
	(39)	Advanced NSCLC with molecular selection	NS	Molecular targeted agents alone (not in combination with other treatment modalities)	2003 to 2014*	Articles, abstracts and trial registries	AD	8	NS	--	R^2_{HR} = 0	--	Medium	NE
	(39)	Advanced NSCLC without any molecular selection	NS	Molecular targeted agents alone (not in combination with other treatment modalities)	2003 to 2014*	Articles, abstracts and trial registries	AD	10	NS	--	R^2_{HR} = 0.41	--	Medium	NE
	(41)	Extensive-stage SCLC	1st line	None	1982 to 2007	Not well described	IPD	10	2 855	τ_{2SM} = 0.58	R^2_{HR} = 0.83 [0.43; 0.95] R^2_{2SM} = 0.90 (SE = 0.27)	0.67	Medium +	NE
	(13)	Locally advanced NSCLC	NA	Radiotherapy + sequential chemotherapy *vs* radiotherapy alone	NS	Convenience sample (86-89)	IPD	8	1 458	τ_{2SM} = 0.77 [0.77; 0.77]	R^2_{2SM} = 0.96 [0.93; 0.99]	--	High	Excellent
	(13)	Locally advanced NSCLC	NA	Radiotherapy + concurrent chemotherapy *vs* radiotherapy alone	NS	Convenience sample (86-89)	IPD	15	2 552	τ_{2SM} = 0.85 [0.85; 0.85]	R^2_{2SM} = 0.97 [0.96; 0.66]	--	High	Excellent
	(13)	Locally advanced NSCLC	NA	Radiotherapy + sequential chemotherapy *vs* radiotherapy + concurrent chemotherapy	NS	Convenience sample (86-89)	IPD	6	1 201	τ_{2SM} = 0.83 [0.83; 0.83]	R^2_{2SM} = 0.89 [0.81; 0.97]	--	High	Excellent
	(13)	Locally advanced SCLC or NSCL	NA	Modified radiotherapy *vs* standard radiotherapy	NS	Convenience sample (86-89)	IPD	12	2 685	τ_{2SM} = 0.81 [0.81; 0.81]	R^2_{2SM} = 0.96 [0.93; 0.98]	--	High	Excellent
	(40)	Advanced NSCLC	NS	Chemotherapy alone or in combination with molecularly targeted agents	2000 to 2011*	Articles and abstracts	AD	18	11 310	ρ = 0.51	ρ = 0.29	--	NE	NE
TTP	(39)	Advanced NSCLC	1st line	Systemic chemotherapy *vs* cytotoxic or molecular-targeted agents	1994 to 2006*	Articles, abstracts and trial registries	AD	54	23 157	--	R^2_{med} = 0.33	--	NE	NE

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
	(36)	Advanced or metastatic NSCLC	1st line	Chemotherapy	< 2005*	Articles and abstracts	AD	191	44 125	--	R^2_{med} = 0.19	--	NE	NE
RR	(39)	Advanced NSCLC with or without molecular selection	NS	Molecular targeted agents alone (not in combination with other treatment modalities)	2003 to 2014*	Articles, abstracts and trial registries	AD	18	7 633	--	R^2_{HR} = 0.10	--	Medium	NE
	(39)	Advanced NSCLC with molecular selection	NS	Molecular targeted agents alone (not in combination with other treatment modalities)	2003 to 2014*	Articles, abstracts and trial registries	AD	8	NS	--	R^2_{HR} = 0.04	--	Medium	NE
	(39)	Advanced NSCLC without any molecular selection	NS	Molecular targeted agents alone (not in combination with other treatment modalities)	2003 to 2014*	Articles, abstracts and trial registries	AD	10	NS	--	R^2_{HR} = 0.43	--	Medium	NE
	(36)	Advanced or metastatic NSCLC	1st line	Chemotherapy	< 2005*	Articles and abstracts	AD	146	35 337	--	R^2_{med} = 0.16	--	NE	NE
Breast cancer														
DFS	(43)	No locally advanced disease	NA	Adjuvant systemic treatment	1970 to 2002	Articles	AD	126	NS	--	R^2_{med} = 0.38	--	NE	NE
	(43)	No locally advanced disease, node-positive patients	NA	Adjuvant systemic treatment	1970 to 2002	Articles	AD	79	NS	--	R^2_{med} = 0.39	--	NE	NE
	(43)	No locally advanced disease, node-negative patients	NA	Adjuvant systemic treatment	1970 to 2002	Articles	AD	20	NS	--	R^2_{med} = 0.39	--	NE	NE
	(43)	No locally advanced disease	NA	Adjuvant chemotherapy	1970 to 2002	Articles	AD	79	NS	--	R^2_{med} = 0.43	--	NE	NE
	(43)	No locally advanced disease	NA	Adjuvant hormonotherapy	1970 to 2002	Articles	AD	47	NS	--	R^2_{med} = 0.37	--	NE	NE
pCR	(44)	None	NA	Neoadjuvant chemotherapy or neoadjuvant anti-HER2 targeted therapy and cytotoxic therapy	1990 to 2009	Articles and abstracts	AD	29	14 641	--	R^2_{HR} = 0.09 [0.01; 0.41]	--	Low	NE
	(44)	None	NA	Neoadjuvant anthracycline- and taxane-based *vs* anthracycline-based regimens	1990 to 2009	Articles and abstracts	AD	9	4 894	--	R^2_{HR} = 0.03 [0; 0.82]	--	Medium	NE
	(44)	None	NA	Neoadjuvant chemotherapy, intensified *vs* standard-dose regimens	1990 to 2009	Articles and abstracts	AD	8	2 862	--	R^2_{HR} = 0.57 [0.19; 0.93]	--	Medium	NE
	(44)	None	NA	Neoadjuvant capecitabine-containing *vs* standard regimens	1990 to 2009	Articles and abstracts	AD	7	3 678	--	R^2_{HR} = 0.15 [0.03; 0.91]	--	Medium	NE
	(45)	None	NA	Neoadjuvant chemotherapy	1990 to 2011*	Articles	IPD	12	9 440	--	R^2_{HR} = 0.24 [0.00; 0.70]	--	Medium	NE
PFS	(33)	Advanced or metastatic disease	NS	None	NS	Articles	AD	NS	NS	--	R^2_{HR} = 0.78	--	Medium +	NE
	(46)	Metastatic disease	NS	Anthracycline- based regimens	1980 to 2002	Articles and abstracts	AD	16	4 323	--	R^2_{HR} = 0.43	--	Medium	NE
	(46)	Metastatic disease	NS	Taxane- based regimens	1980 to 2002	Articles and abstracts	AD	15	5 893	--	R^2_{HR} = 0.35	--	Medium	NE
	(14)	Metastatic disease	1st line	Anthracycline- and/or taxane-based regimens	NS	Articles and abstracts	IPD	11	3 953	ρ_{2SM} = 0.69 [0.69; 0.69]	r_{2SM} = 0.48 [0; 1]	--	Medium	Poor
TTP	(14)	Metastatic disease	1st line	Anthracycline- and/or taxane-based regimens	NS	Articles and abstracts	IPD	11	3 953	ρ_{2SM} = 0.68 [0.68; 0.68	r_{2SM} = 0.49 [0; 1]	--	Medium	Poor

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
	(47)	Metastatic disease	1st line	Anthracycline-based regimens	1966 to 2005*	Articles	AD	42	9 163	--	R^2_{HR} = 0.56 (SE = 0.0928)	--	Medium	NE
PFS/TTP	(48)	Metastatic disease	All lines	Anthracycline, taxane or targeted therapy in at least one arm	1990 to 2015*	Articles, abstracts and trial registries	AD	84	NS	--	R^2_{HR} = 0.31	--	Medium	NE
	(48)	Metastatic disease	All lines	Anthracycline, taxane or targeted therapy in at least one arm	2004 to 2015*	Articles, abstracts and trial registries	AD	40	NS	--	R^2_{HR} = 0.31	--	Medium	NE
	(48)	Metastatic disease	1st line	Anthracycline, taxane or targeted therapy in at least one arm	1990 to 2015*	Articles, abstracts and trial registries	AD	48	NS	--	R^2_{HR} = 0.30	--	Medium	NE
	(48)	Metastatic disease	≥ 2nd line	Anthracycline, taxane or targeted therapy in at least one arm	1990 to 2015*	Articles, abstracts and trial registries	AD	27	NS	--	R^2_{HR} = 0.55	--	Medium	NE
	(48)	Metastatic disease		Anthracycline, taxane or targeted therapy in at least one arm, cross-over allowed	1990 to 2015*	Articles, abstracts and trial registries	AD	20	NS	--	R^2_{HR} = 0.49	--	Medium	NE
	(49)	Advanced or metastatic disease	1st line	Chemotherapy + targeted agents	2000 to 2012*	Articles	AD	20	10 138	R^2 = 0.61 ρ = 0.81 [0.58; 0.92]	R^2_{HR} = 0.73 ρ = 0.7 [0.39; 0.87];	--	Medium +	Fair
	(50)	Advanced or metastatic disease	NS	Chemotherapy only	< 2010*	Articles, abstracts and trial registries	AD	144	43 459	ρ = 0.428	R^2_{med} = 0.86; ρ = 0.427	--	NE	NE
	(51)	Metastatic disease, no locally advanced disease	NS	None	1994 to 2008*	Articles	AD	67	NS	r = 0.38	R^2_{med} = 0.30	--	NE	NE
	(51)	Metastatic disease, no locally advanced disease	NS	Anthracycline- based regimens	1994 to 2008*	Articles	AD	36	NS	--	R^2_{med} = 0.43	--	NE	NE
	(51)	Metastatic disease, no locally advanced disease	NS	Hormone- based regimens	1994 to 2008*	Articles	AD	12	NS	--	R^2_{med} = 0.24	--	NE	NE
	(51)	Metastatic disease, no locally advanced disease, HER2+ patients	NS	None	1994 to 2008*	Articles	AD	4	NS	--	R^2_{med} = 0.93	--	NE	NE
	(51)	Metastatic disease, no locally advanced disease	1st line	None	1994 to 2008*	Articles	AD	46	NS	--	R^2_{med} = 0.28	--	NE	NE
	(51)	Metastatic disease, no locally advanced disease	Not 1st line	None	1994 to 2008*	Articles	AD	21	NS	--	R^2_{med} = 0.32	--	NE	NE
RR	(14)	Metastatic disease	1st line	Anthracycline- and/or taxane-based regimens	NS	Articles and abstracts	IPD	11	3 953	--	r_{2SM} = 0.57 [0; 1]	--	Medium	NE
	(52)	Metastatic disease	All lines	Standard *vs* intensified Epirubicin-containing chemotherapy	NS	Articles	IPD	10	2 126	--	R^2_{HR} = 0.10 [0.00; 0.43]	--	Low	NE
	(47)	Metastatic disease	1st line	Anthracycline-based regimens	1966 to 2005*	Articles	AD	42	9 163	--	R^2_{HR} = 0.34 (SE = 0.0590)	--	Low	NE
CR	(47)	Metastatic disease	1st line	Anthracycline-based regimens	1966 to 2005*	Articles	AD	42	9 163	--	R^2_{HR} = 0.12 (SE = 0.0521)	--	Low	NE
OR	(47)	Metastatic disease	1st line	Anthracycline-based regimens	1966 to 2005*	Articles	AD	42	9 163	--	R^2_{HR} = 0.38 (SE = 0.0380)	--	Low	NE
DCR	(14)	Metastatic disease	1st line	Anthracycline- and/or taxane-based regimens	NS	Articles and abstracts	IPD	11	3 953	--	r_{2SM} = 0.47 [0; 1]	--	Medium	NE
Renal cell carcinoma														
PFS	(53)	Metastatic disease	NS	Targeted therapy alone	NS	Articles and abstracts	AD	10	7 236	r = 0.85 [0.61–0.95] R^2 = 0.73 ρ = 0.69 [0.28–0.89] τ = 0.55 [0.13–0.87]	r_{HR} = 0.45; R^2_{HR} = 0.20; ρ = 0.78; τ = 0.34	--	Medium	NE

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
	(53)	Metastatic disease	NS	Immunotherapy alone	NS	Articles and abstracts	AD	9	2 829	r = 0.84 [0.63–0.93] R^2 = 0.71 ρ = 0.85 [0.64–0.94] τ = 0.69 [0.49–0.88]	r_{HR} = 0.63; R^2_{HR} = 0.66; ρ = 0.80; τ = 0.66	--	Medium	NE
	(54)	Metastatic disease	NS	Chemotherapy, immunotherapy or targeted therapy	1988 to 2008	Articles and trial registries	AD	30	NS	--	R^2_{med} = 0.49	3.65	NE	NE
	(54)	Metastatic disease	NS	Targeted therapy	1988 to 2008	Articles and trial registries	AD	11	NS	--	R^2_{med} = 0.44	--	NE	NE
	(55)	Metastatic disease	1st line	Targeted therapy	< 2011*	Articles	AD	6	3 188	ρ = 0.87; R^2 = 0.97	R^2_{med} = 0.07; ρ = 0.36; r_{med} = 0.26	--	NE	NE
	(56)	Metastatic disease	NS	Interleukin-2, interferon, axitinib, lapatinib, pazopanib, sunitinib, sorafenib, bevacizumab, everolimus, or temsirolimus	1997 to 2010*	Articles and abstracts	AD	31	10 943	--	R^2_{HR} = 0.63; r_{HR} = 0.80	--	Medium	NE
RR	(55)	Metastatic disease	1st line	Targeted therapy	< 2011*	Articles	AD	6	3 188	ρ = 0.96	R^2_{med} = 0.27; ρ = 0.49; r_{med} = 0.52;	--	NE	NE
DCR	(55)	Metastatic disease	1st line	Targeted therapy	< 2011*	Articles	AD	6	3 188	ρ = 1	R^2_{med} = 0.95; ρ = 1; r_{med} = 0.97;	--	NE	NE
Gastric cancer														
DFS	(57)	Curatively resected gastric cancer	NA	Adjuvant chemotherapy *vs* surgery alone	< 2004	Articles and trial registries	IPD	14	3 288	τ_{2SM} = 0.97 [0.97; 0.98]	R^2_{2SM} = 0.96 [0.93; 1]	--	High	Excellent
PFS	(58)	Advanced or recurrent gastric cancer	NS	Chemotherapy	< 2006	Articles	IPD	20	4 069	τ_{2SM} = 0.85 [0.85; 0.85]	R^2_{2SM} = 0.61 [0.04; 1]	--	Medium	Poor
PFS/TTP	(59)	Advanced or metastatic disease	2nd line	Chemotherapy	2002 to 2013*	Articles and abstracts	AD	10	4 286	ρ = 0.56 [0.34; 0.74]	ρ = 0.36 [0; 1]	--	NE	NE
	(60)	Advanced or metastatic disease	NS	Chemotherapy	1966 to 2010*	Articles and abstracts	AD	36	10 484	ρ = 0.70 [0.59; 0.82]	ρ = 0.80 [0.68; 0.92]	--	NE	NE
Ovarian cancer														
PFS	(33)	Advanced or metastatic disease	NS	None	NS	Articles	AD	NS	NS	--	R^2_{HR} = 0.73	--	Medium +	NE
Head & neck cancer														
DFS	(61)	Locally advanced disease	NS	Adjuvant chemotherapy	1965 to 2000*	Convenience sample (90-92)	IPD	9	2068	ρ_{2SM} = 0.82 [0.82; 0.82]	r_{2SM} = 0.93 [0.85; 1.01]	--	High	Excellent
PFS	(61)	Locally advanced disease	NS	Radiotherapy	1965 to 2000*	Convenience sample (93)	IPD	15	6515	ρ_{2SM} = 0.86 [0.86; 0.86]	r_{2SM} = 0.98 [0.97; 1]	--	High	Excellent
	(61)	Locally advanced disease	NS	Concomitant chemotherapy	1965 to 2000*	Convenience sample (90-92)	IPD	56	9530	ρ_{2SM} = 0.86 [0.86; 0.86]	r_{2SM} = 0.86 [0.79; 0.93]	--	Medium +	Excellent
	(61)	Locally advanced disease	NS	Induction chemotherapy	1965 to 2000*	Convenience sample (90-92)	IPD	32	4631	ρ_{2SM} = 0.90 [0.90; 0.90]	r_{2SM} = 0.79 [0.66; 0.92]	--	Medium	Good
DLRC	(61)	Locally advanced disease	NS	Adjuvant chemotherapy	1965 to 2000*	Convenience sample (90-92)	IPD	9	2068	ρ_{2SM} = 0.65 [0.64; 0.65]	r_{2SM} = 0.84 [0.67; 1.01)	--	Medium	Good
	(61)	Locally advanced disease	NS	Radiotherapy	1965 to 2000*	Convenience sample (93)	IPD	15	6515	ρ_{2SM} = 0.76 [0.76; 0.76]	r_{2SM} = 0.94 [0.89; 1]	--	High	Good
	(61)	Locally advanced disease	NS	Concomitant chemotherapy	1965 to 2000*	Convenience sample (90-92)	IPD	56	9530	ρ_{2SM} = 0.76 [0.76; 0.77]	r_{2SM} = 0.72 [0.60; 0.85]	--	Medium	Fair
	(61)	Locally advanced disease	NS	Induction chemotherapy	1965 to 2000*	Convenience sample (90-92)	IPD	32	4631	ρ_{2SM} = 0.53 [0.28; 0.78]	r_{2SM} = 0.59 [0.36; 0.81]	--	Medium	Poor

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
Pancreatic cancer														
PFS/TTP	(62)	Advanced or metastatic disease	1st line	Gemcitabine alone *vs* polychemotherapy	2002 to 2013*	Articles, abstracts and trial registries	AD	30	8 467	$R^2 = 0.6$ r = 0.75 [0.62; 0.85]	$R^2_{HR} = 0.69$ $R^2_{HR} = 0.64$; $r_{HR} = 0.78$ [0.49; 0.91]	--	Medium	Fair
PFS	(63)	Advanced or metastatic disease	1st line	Gemcitabine in combination *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	39	NS	--	$\rho = 0.67$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	23	NS	--	$\rho = 0.71$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine + targeted therapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	16	NS	--	$\rho = 0.66$	--	NE	NE
	(64)	Advanced or metastatic disease	1st line	Bevacizumab and chemotherapy *vs* chemotherapy	2000 to 2014*	Articles	AD	9	NS	--	$R^2_{med} = 0.71$	--	NE	NE
TTP	(63)	Advanced or metastatic disease	1st line	Gemcitabine in combination *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	4	NS	--	$\rho = 0.63$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	3	NS	--	$\rho = 0$	--	NE	NE
TTF	(63)	Advanced or metastatic disease	1st line	Gemcitabine in combination *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	3	NS	--	$\rho = 0.50$	--	NE	NE
DOR	(63)	Advanced or metastatic disease	1st line	Gemcitabine in combination *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	7	NS	--	$\rho = 0.76$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	3	NS	--	$\rho = 0.50$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine + targeted therapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	4	NS	--	$\rho = 0.40$	--	NE	NE
RR	(63)	Advanced or metastatic disease	1st line	Gemcitabine in combination *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	41	NS	--	$R^2_{med} = 0.15$; $\rho = 0.29$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	26	NS	--	$\rho = 0.23$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine + targeted therapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	15	NS	--	$\rho = 0.54$	--	NE	NE
	(64)	Advanced or metastatic disease	1st line	Bevacizumab and chemotherapy *vs* chemotherapy	2000 to 2014*	Articles	AD	11	NS	--	$R^2_{med} = 0.58$	--	NE	NE
DCR	(63)	Advanced or metastatic disease	1st line	Gemcitabine in combination *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	33	NS	--	$R^2_{med} = 0.56$; $\rho = 0.61$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	18	NS	--	$\rho = 0.71$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine + targeted therapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	15	NS	--	$\rho = 0.53$	--	NE	NE
	(64)	Advanced or metastatic disease	1st line	Bevacizumab and chemotherapy *vs* chemotherapy	2000 to 2014*	Articles	AD	6	NS	--	$R^2_{med} = 0.56$	--	NE	NE
CBR	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	7	NS	--	$\rho = 0.78$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	6	NS	--	$\rho = 0.66$	--	NE	NE
CA19-9	(63)	Advanced or metastatic disease	1st line	Gemcitabine in combination *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	6	NS	--	$\rho = 0$	--	NE	NE
	(63)	Advanced or metastatic disease	1st line	Gemcitabine-based chemotherapy *vs* Gemcitabine alone	1997 to 2014*	Articles	AD	5	NS	--	$\rho = 0$	--	NE	NE

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
Nasopharyngeal Carcinoma														
PFS	(65)	Non-metastatic disease	NS	Combined chemotherapy and radiotherapy	1979 to 2010	Articles, abstracts and trial registries	AD	15	3 760	--	R^2_{HR} = 0.99	≤ 0.84	Medium +	NE
TTP	(65)	Non-metastatic disease	NS	Combined chemotherapy and radiotherapy	1979 to 2010	Articles, abstracts and trial registries	AD	9	2 422	--	R^2_{HR} = 0.88	≤ 0.83	Medium +	NE
Multiple myeloma														
PFS	(66)	None	NS	None	2002 to 2013*	Articles and trial registries	AD	21	12 048	--	R^2_{HR} = 0.63 [0.43; 0.84] r_{HR} = 0.80 [0.58; 0.90]	--	Medium	NE
Melanoma – Metastatic														
PFS	(67)	Non-resectable or metastatic melanoma	NS	Dacarbazine *vs* any systemic therapy	NS	Articles, abstracts and trial registries	AD	12	4 416	--	r_{HR} = 0.71 [0.29; 0.90]	--	Medium	NE
Glioblastoma														
PFS	(68)	None	NS	None	1991 to 2012*	Articles and abstracts	AD	10	7 125	--	R^2_{HR} = 0.92 [0.72; 0.99]	--	High	NE
Indolent NHL														
PFS/TTP	(69)	None	1st line	Chemotherapy	1990 to 2009*	Articles, abstracts and trial registries	AD	20	5 128	ρ = 0.56 [0.2; 0.78]	ρ = 0.26 [0; 0.72]	--	NE	NE
CR	(69)	None	1st line	Chemotherapy	1990 to 2009*	Articles, abstracts and trial registries	AD	20	5 128	--	ρ = 0.21 [0; 0.50]	--	NE	NE
Aggressive NHL														
PFS/TTP	(69)	None	1st line	Chemotherapy	1990 to 2009*	Articles, abstracts and trial registries	AD	38	16 103	ρ = 0.85 [0.71; 0.92]	R^2_{med} = 0.66 ρ = 0.90 [0.73; 0.96]	--	NE	NE
CR	(69)	None	1st line	Chemotherapy	1990 to 2009*	Articles, abstracts and trial registries	AD	38	16 103	--	ρ = 0.50 [0.23; 0.74]	--	NE	NE
Biliary tract cancer														
PFS/TTP	(70)	Advanced disease	1st line	Chemotherapy	< 2015*	Articles and abstracts	AD	19	2 148	--	R^2_{med} = 0.66 [0.32; 0.85]	--	NE	NE
	(70)	Advanced disease	1st line	Gemcitabine-based regimens only	< 2015*	Articles and abstracts	AD	15	2 148	--	R^2_{med} = 0.78 [0.46; 0.92]	--	NE	NE
	(70)	Advanced disease	1st line	Targeted therapy only	< 2015*	Articles and abstracts	AD	7	2 148	--	R^2_{med} = 0.78 [0.14; 0.96]	--	NE	NE
RR	(70)	Advanced disease	1st line	Chemotherapy	< 2015*	Articles and abstracts	AD	17	2 148	--	R^2_{med} = 0.29 [0.01; 0.65]	--	NE	NE
	(70)	Advanced disease	1st line	Gemcitabine-based regimens only	< 2015*	Articles and abstracts	AD	14	2 148	--	R^2_{med} = 0.39 [0.02; 0.75]	--	NE	NE
	(70)	Advanced disease	1st line	Targeted therapy only	< 2015*	Articles and abstracts	AD	7	2 148	--	R^2_{med} = 0.43 [0.03; 0.89]	--	NE	NE
DCR	(70)	Advanced disease	1st line	Chemotherapy	< 2015*	Articles and abstracts	AD	17	2 148	--	R^2_{med} = 0.34 [0.02; 0.69]	--	NE	NE
	(70)	Advanced disease	1st line	Gemcitabine-based regimens only	< 2015*	Articles and abstracts	AD	14	2 148	--	R^2_{med} = 0.60 [0.17; 0.86]	--	NE	NE
	(70)	Advanced disease	1st line	Targeted therapy only	< 2015*	Articles and abstracts	AD	7	2 148	--	R^2_{med} = 0.44 [0.03; 0.89]	--	NE	NE

Endpoint	Ref	Disease specifications	Line of treatment	Treatment specification	Inclusion or publication (*) period	Data source	Type of data	*N* trials	*N* patients	Individual-level association	Trial-level association	STE	Strength of association as per… IQWiG	BSES2
Soft tissue sarcoma														
PFS	(71)	Advanced or metastatic disease	All lines	Systemic therapy in at least one arm	1974 to 2014*	Articles and abstracts	AD	14	NS	--	r_{HR} = 0.61	--	Medium	NE
RR	(71)	Advanced or metastatic disease	All lines	Systemic therapy in at least one arm	1974 to 2014*	Articles and abstracts	AD	11	NS	--	r_{HR} = 0.51	--	Medium	NE
All solid tumors														
PFS	(33)	Metastatic disease	NS	None	NS	Articles	AD	66	NS	--	R^2_{HR} = 0.62	--	Medium	NE
RR	(33)	Metastatic disease	NS	None	NS	Articles	AD	66	NS	--	R^2_{HR} = 0.37	--	Medium	NE

DFS = disease-free survival; pCR = pathologic complete response; TTR = time-to recurrence; PFS = progression-free survival; TTP = time-to progression; RR = response rate; TTF = time-to treatment failure; CR = complete response; OR = objective response (partial or complete response); DCR = disease control rate; EFS = event-free survival; DLRC = duration of locoregional control; DCR = disease control rate; GIST = gastro intestinal stromal tumor; DOR = duration of response; CBR = clinical benefit response; CA19-9 = cancer antigen 19-9 response-related criteria

ρ = non-parametric Spearman rank correlation coefficient; r = Pearson correlation coefficient; R^2 = proportion of variation estimated by weighted linear regression model; ρ_{2SM} = Spearman rank coefficient estimated by the two-step model; τ_{2SM} = Kendal's tau estimated by the two-step model; r_{HR} = Pearson correlation coefficient between treatment effects estimated by hazard ratios; R^2_{HR} = proportion of variation estimated using treatment effects estimated by hazard ratios; ; r_{med} = Pearson correlation coefficient between treatment effects estimated by difference or ratios of medians; R^2_{med} = proportion of variation estimated using treatment effects estimated by differences or ratios of medians; r_{2SM} = Pearson correlation coefficient between treatment effects estimated using the 2-step model; R^2_{2SM} = proportion of variation between treatment effects estimated using the 2-step model

Table 3: Meta-analyses presenting a trial-level correlation ranked as 'High' as per the IQWIG framework and an association ranked as 'Excellent' as per the adapted BSES2 framework.

Endpoint	Cancer localization	Disease specifications	Treatment specifications
DFS	Colon cancer	Stage II or III patients	Adjuvant setting, fluoropyrimidines alone or in combination (12,22)
	Lung cancer	Operable and locally advanced NSCLC	Adjuvant treatment by chemotherapy and/or radiotherapy (13)
	Gastric cancer	Curatively resected gastric cancer	Adjuvant chemotherapy (57)
	Head & neck cancer	Locally advanced disease	Adjuvant chemotherapy (61)
PFS	Colorectal cancer	Advanced / Metastatic disease	Fluorouracil- and leucovorin-based chemotherapy (32)
	Lung cancer	Locally advanced NSCLC	Radiotherapy alone or in combination with chemotherapy (13)
	Lung cancer	Locally advanced SCLC or NSCLC	Radiotherapy (13)
	Head & neck cancer	Locally advanced disease	Radiotherapy (61)

DFS = disease-free survival; PFS = progression-free survival; SCLC = small-cell lung cancer; NSCLC = non-small-cell lung cancer.

Additional files

Additional file 1: Number of events observed for progression-free survival, time-to-progression and time-to-treatment failure after 6 and 12 months of follow-up; All trials (N_{trial} = 15; $N_{patient}$ = 2846) [a]

The following search algorithm was launched on PUBMED (last update: 18 July 2016):
(neoplasms[mh] OR cancer [Title/Abstract] OR oncology [Title/Abstract] OR tumor [Title/Abstract] OR tumour [Title/Abstract] OR lymphoma [Title/Abstract] OR sarcoma [Title/Abstract] OR melanoma [Title/Abstract] OR myeloma [Title/Abstract] OR carcinoma [Title/Abstract]) AND (surrogate[Title/Abstract] OR surrogacy[Title/Abstract] OR correlation[Title/Abstract] OR association[Title/Abstract] OR prediction[Title/Abstract]) AND (endpoint [Title/Abstract] OR “end point” [Title/Abstract] OR endpoints [Title/Abstract] OR “end points” [Title/Abstract] OR "end-point" [Title/Abstract] OR "end-points" [Title/Abstract])
Limits were as follows: Article, English or French language, Human only

The following search algorithm was launched on SCOPUS (last update: 18 July 2016):

TITLE-ABS-KEY (neoplasm OR cancer OR oncology OR tumor OR tumour OR lymphoma OR sarcoma OR melanoma OR myeloma OR carcinoma) AND TITLE-ABS-KEY (surrogate OR surrogacy OR correlation OR association) AND TITLE-ABS-KEY (endpoint OR "end point" OR endpoints OR "end points" OR "end-point" OR "end-points") AND (LIMIT-TO (EXACTKEYWORD , "Human") OR LIMIT-TO (EXACTKEYWORD, ‘Humans”)

Additional file 2: Data extraction grid

Publication identification	
Journal: 1st author: ☐ Clinical journal	Title: Publication date: ☐ Statistical journal
Title/Abstract selection	
Human only: Cancer only: Type of publication: Healthy patients or patients in remission: Explicitly unrelated to surrogate end-points: Theme of the publication:	☐YES ☐NO ☐YES ☐NO ☐ General article ☐ Comment ☐ Letter to editor ☐ Conference abstract ☐ YES ☐NO ☐To be verified ☐ YES ☐NO Comment: ☐ Validation study of surrogate endpoints ☐ Review ☐ Guidelines
Publication selected: ☐ YES ☐ NO	
Full-paper selection	
Publication available in French or English	☐YES: ☐French ☐English ☐NO
Type of analysis: Type of trials included: Type of data: Final endpoint: Endpoints evaluated as surrogates: Cancer localization: Setting:	☐Single-trial analysis ☐Meta-analysis: trials included ☐Phase II ☐Phase III Randomized : ☐YES ☐NO Comparative : ☐YES ☐NO ☐Aggregate data ☐Individual data ☐OS ☐Other: ☐PFS ☐TTP ☐DFS ☐Response rate ☐Other: ☐Breast ☐Colorectal ☐Lung ☐Stomach ☐Prostate ☐Gastric ☐Other: ☐Adjuvant ☐Neoadjuvant ☐Advanced/Metastatic
Statistic methods employed:	Individual-level association: ☐KAPPA coefficient ☐% of agreement ☐Non-parametric correlation coefficient ☐Weighted linear regression on: ☐Joint modelling Trial-level association: ☐KAPPA coefficient ☐% of agreement ☐Non-parametric correlation coefficient ☐Weighted linear regression on HR estimated independently ☐Weighted linear regression on HR estimated simultaneously (2-step model) ☐Weighted linear regression on estimator other than HR

	☐Error-in-variable model on HR estimated independently ☐Error-in-variable model on HR estimated simultaneously (2-step model)
Meta-analytic unit:	☐Trial ☐Center ☐Country ☐Other:
Number of patients:	
Individual-level association estimated:	
Trial-level association estimated:	
Surrogate Threshold Effect (STE):	
STE proportion (STEP):	
Authors conclusions:	
BSES score (Statistical evaluation domain)	☐ Poor ☐ Fair ☐ Good ☐ Excellent ☐ Not evaluable
Adapted BSES score	☐ Poor ☐ Fair ☐ Good ☐ Excellent ☐ Not evaluable
IQWiG score	☐ Low ☐ Medium ☐ Medium + ☐ High ☐ Not evaluable
Publication selected:	**☐ YES ☐ NO**

Additional file 3: Frameworks for the assessment of the strength of the association

We relied on two frameworks for the assessment of the strength of evidence of the validation studies on surrogate endpoints.

The Biomarker-Surrogate Evaluation Schema (BSES) is aimed at assessing the level of surrogacy evidence based on the surrogacy measures and characteristics of the study (16). The level of proof of surrogacy of each meta-analysis is classified as proof of surrogacy, high probability of surrogacy, hint of surrogacy, no proof of surrogacy, and proof of low correlation. The BSES includes four domains: study design, target outcome, statistical evaluation and generalizability (16). Each of the four domains is associated with a four-level rank (0 to 3) leading to a global score ranging from 0 to 12, calculated based on the BSES guidelines. We specifically focused on the "Statistical Evaluation" domain, for which estimation of the individual-level and trial-level associations, as well as the STEP are required. The trial-level association R^2 is classified as excellent ($R^2 \geq 0.60$ and STEP ≥ 0.3 and $R^2_{ind} \geq 0.60$, where R^2_{ind} is the patient-level association), good ($R^2 \geq 0.4$ and STEP ≥ 0.2 and $R^2_{ind} \geq 0.4$), fair ($R^2 \geq 0.2$ and STEP ≥ 0.1 and $R^2_{ind} \geq 0.2$) or poor otherwise.

The German Institute of Quality and Efficiency in Health Care (IQWiG) guidelines classify the trial-level association as a function of the correlation coefficient r as high (lower limit of the 95% confidence interval of $r \geq 0.85$), low correlation (upper limit of the 95% confidence interval of $r \leq 0.7$), or medium in any other case, meaning that the validity of the surrogate remains unclear(15). If the coefficient of determination (R^2) is provided instead of the correlation coefficient r, then r is calculated by taking the square root.

3 Surrogate endpoints in metastatic soft-tissue sarcoma trials

3.1 Introduction

Soft-tissue sarcoma (STS) develop from soft tissues like fat, muscle, nerves, fibrous tissues, blood vessels or deep skin tissues. With more than 50 different histological subtypes, such as leiomyosarcoma or liposarcoma, and increasing number of molecular subtypes, STS are very heterogeneous cancers, which make them overly complex to diagnose, study and consequently to cure. These rare tumors account for 1% of all malignancies in adults (20). In France, 4000 new cases of STS are diagnosed each year.

Despite adequate locoregional treatment, up to 40% of patients with STS develop metastases. When metastases are detected, the standard of care is based on palliative chemotherapy with a limited efficacy. The median OS of patients with metastatic STS is shorter than two years.

The rarity and heterogeneity of STS increase the difficulty of conducting large and homogeneous trials to evaluate new treatment. A surrogate endpoint that would be observed more frequently than OS would help reduce the number of patients to include and would be a great asset for medical research on STS treatment. Nonetheless, as highlighted in the previous section, to date, there is no MA based on IPD evaluating surrogate endpoints in advanced STS.

We conducted a MA on IPD from 14 RCTs evaluating systemic treatment and/or targeted therapy for patients with advanced STS to assess the surrogate properties of three commonly used progression-based endpoints: progression-free survival (PFS), time-to progression (TTP), and time-to treatment failure (TTF). We first performed a systematic review of RCTs in advanced STS through a computerized search on MEDLINE. To limit publication bias, we also examined trial registries and contacted European sponsoring groups. Formal conventions were drafted to set the terms of the transfer regarding the nature and format of the data and the project valorization. An important work of data-management predated the analysis. This stage aimed at homogenizing the data in terms of computing format, definition of the time-to-event endpoints and censoring process. Each endpoint, candidate surrogates and OS, was recalculated to ensure an identical definition and follow-up across trials. Following the two-stage approach (chapter 1.3.2.1) and the simplified meta-regression with weighted fixed treatment effects (chapter 1.3.2.2), we estimated the individual- and trial-level associations with OS of PFS, TTP and TTF.

The MA did not lead to significant evidence to validate the candidate endpoints as surrogates for OS when assessing systemic treatment in advanced STS. OS should therefore remain the primary endpoint in RCTs conducted in this setting.

Trial design for advanced STS is particularly challenging due to the low incidence and the heterogeneity of the disease and treatments, which may have contributed to weaken the observed correlations between candidate surrogates and OS. It is however reasonable to assume that an effect on OS can only be achieved if there is an effect also on disease progression. As such, alternative endpoints, e.g. PFS, remain useful in testing new treatments in earlier drug development stages, such as in phase II trials or phase III futility assessment, provided OS data are collected throughout the trial. This work is currently under review by the *Journal of Clinical Oncology*.

In supplementary analyses, we explored the robustness of the models by conducting additional sensitivity analyses. In the first analysis, different approaches for data censoring were investigated. In the second analysis, we investigated different weighting procedures for the meta-regression model. The results of these supplementary analyses are presented in section 3.3.

3.2 Publication

Surrogate endpoints in advanced sarcoma trials: a meta-analysis

Marion Savina[1,2], Saskia Litière[3], Antoine Italiano[4], Tomasz Burzykowski[5], Franck Bonnetain[6], Sophie Gourgou[7], Virginie Rondeau[2], Jean-Yves Blay[8], Sophie Cousin[4], Florence Duffaud[9], Hans Gelderblom[10], Alessandro Gronchi[11], Ian Judson[12], Axel Le Cesne[13], Paul Lorigan[14], Joan Maurel[15], Winette van der Graaf[12,16,17], Jaap Verweij[18], Simone Mathoulin-Pélissier[1,2], Carine Bellera[1,2]

[1] Clinical and Epidemiological Research Unit, Institut Bergonié, Comprehensive Cancer Center, Bordeaux, France; [2] University of Bordeaux, ISPED, Centre INSERM U1219 Bordeaux Population Health, Epicene Team, Bordeaux, France; [3] European Organisation for Research and Treatment of Cancer (EORTC), Brussels, Belgium; [4] Medical Oncology unit, Institut Bergonié, Regional Comprehensive Cancer Center, Bordeaux, France; [5] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium; [6] Methodology and Quality of life in Oncology Unit, EA3181.CHU Besançon, France; [7] Biometrics unit, Institut du Cancer de Montpellier, France; [8] Comprehensive Cancer Center Léon Bérard Lyon and University Claude Bernard Lyon I, Lyon, France; [9] Medical Oncology unit, University Hospital La Timone and University of Aix-Marseille, Marseille, France; [10] Department of Clinical Oncology, Leiden University Medical Center, Leiden, The Netherlands; [11] Fondazione Istituto di Ricovero e Cura a Carattere Scientifico, Istituto Nazionale dei Tumori, Milano, Italy; [12] Royal Marsden NHS Foundation Trust, London, United Kingdom; [13] Medicine Department, Comprehensive Cancer Center Institut Gustave Roussy, Villejuif, France; [14] University of Manchester/The Christie NHS Foundation Trust, Manchester, UK; [15] Department of Medical Oncology, Hospital Clinic, CIBERehd, Translational Genomics and Targeted Therapeutics in Solid Tumors (IDIBAPS), Barcelona, Spain; [16] The Institute of Cancer Research, London, United Kingdom; [17] Radboud University Medical Centre, Department of Medical Oncology, Nijmegen, Netherlands; [18] Department of Medical Oncology, Erasmus University Medical Center, Rotterdam, The Netherlands

Abstract

Introduction: Alternative endpoints to overall survival (OS) are frequently used to assess treatment efficacy in randomized controlled trials (RCT). Their properties in terms of surrogate outcomes for OS need to be assessed. We evaluated the surrogate properties of progression-free survival (PFS), time-to-progression (TTP) and time-to-treatment failure (TTF) in advanced soft tissue sarcomas (STS).

Methods: We performed a meta-analysis using individual-patient data (IPD). Trials were identified by searches of MEDLINE and ClinicalTrials.gov and by contacting European sponsoring groups. European phase II/III RCTs evaluating therapies for adults with advanced STS were eligible. Statistical methods included weighted linear regression and the two-stage model introduced by Buyse and Burzykowski. The strength of the trial-level association was ranked according to the German Institute for Quality and Efficiency in Health Care (IQWiG) guidelines.

Results: IPD from 14 RCTs (N=2846) were analyzed. Individual-level associations were moderate (highest for 12-month PFS: $\rho_{Spearman}$=0.66; 95%CI [0.63; 0.68]). Trial-level associations were ranked as low for the three endpoints as per the IQWiG criterion.

Conclusion: Our results do not support strong surrogate properties of PFS, TTP and TTF for OS in advanced STS.

Key words: surrogate endpoint, soft-tissue sarcoma, overall survival, meta-analysis, randomized controlled trial

Introduction

The choice of the primary endpoint is a key step when designing a randomized controlled trial (RCT). In oncology, the most commonly used endpoint to assess the efficacy of a new treatment in RCT is overall survival (OS) which is easily measurable, objectively defined as the time from randomization to death and validated by health regulatory authorities [1]. Alternative time-to-event endpoints are commonly used in practice in phase II trials and increasingly being used instead of OS in phase III trials [2]. These composite endpoints include death as well as biological and clinical events, such as disease progression or treatment toxicity. Such endpoints are developed due to the need to reduce the number of patients included, the trial duration, the delay to reach trials' conclusions and ultimately the cost of the trials. However, their use in practice does not guarantee their validity as surrogates for OS. It is therefore essential to rigorously assess their surrogate properties for OS, and as such whether or not they can be used as primary endpoints for assessing the benefit of new therapies. This approach does not preclude their intrinsic value as parameters of patient benefit of a treatment.

The International Conference on Harmonization (ICH) E9 Harmonized Tripartite guidelines - approved by the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) - do not provide any recommendations on the use of specific statistical methods for the validation of surrogate endpoints. However, the meta-analytic surrogacy evaluation scheme proposed by Buyse and Burzykowski et al. [3,4] has been widely used and is considered as the most statistically rigorous [5,6]. This approach requires individual-patient data (IPD) from multiple RCTs with similar design and treatment to address surrogacy from a multi-level framework. At the patient level, the surrogate endpoint should be correlated and predictive of the final endpoint regardless of the treatment (individual-level association). At the trial level, the treatment effect on the surrogate endpoint should be correlated and predictive of the treatment effect on the final endpoint (trial-level association).

Soft-tissue sarcomas (STS) are a heterogeneous group of diseases that account for 1% of all malignancies in adults [7]. Despite adequate locoregional treatment, up to 40% of patients with STS develop metastatic disease [7]. When metastases are detected, the standard of care is palliative chemotherapy. Due to their rarity, conducting large RCTs to evaluate the benefit of new treatment for metastatic STS is complex. The identification of valid surrogate endpoints for OS that would be observed sooner and more frequently than OS, thereby reducing the number of included patients, would be of a great advantage for clinical research. To our knowledge, only one meta-analysis evaluating response rate and PFS as surrogates for OS in metastatic STS was conducted [8]. The study was however limited to the analysis of aggregated data.

We performed a meta-analysis of phase II / III RCTs using IPD to assess the surrogate properties for OS of three commonly used time-to-event endpoints in advanced STS: progression-free survival (PFS), time-to-progression (TTP) and time-to-treatment failure (TTF). This manuscript follows the international recommendations of the PRISMA guidelines for reporting meta-analysis [9].

Methods

This study is registered on the clinical trial registry clinicaltrials.gov (identifier: NCT02873923).

Study selection

We identified trials by using a computerized search on MEDLINE with the following search algorithm: “sarcoma"[MeSH] AND "randomized controlled trial"[Text Word] AND trial[Text Word]. We limited our research to trials published before April the 7th, 2016. We also searched for trials on ClinicalTrials.gov and by contacting European sponsoring groups (EORTC, UNICANCER). Trials were eligible if they met the following criteria: (i) phase II or III randomized trials on humans, (ii) evaluating therapies for adults with advanced (i.e. locally advanced or metastatic) STS, (iii) at least one time-to-event endpoint other than OS as outcome, (iv) published or soon to be published in French or English, (v) signed agreement from the principal investigator and the sponsor, and (vi) available IPD.

Patients, data and outcomes

For all patients, we gathered clinical and histological data at baseline, date of randomization, data related to treatment allocation, disease evaluations during trial, date of last follow-up or death, survival status, cause of death (if applicable), along with any randomization variable. We assessed the surrogate properties of PFS, TTP and TTF evaluated at six and twelve months for 18-month OS. Outcomes were defined following the international DATECAN guidelines [10]. OS was defined as the time from randomization to all-cause death. When death was not observed, 18-month OS was censored at the date of last contact with the patient or at 18 months whichever came first. PFS was defined as the time from randomization to progression or all-cause death, whichever came first. TTP was defined as the time to progression or cancer-related death. Finally, TTF was defined as the time to progression or cancer-related death or treatment-related toxic death, whichever came first. When none of the events included in the definition was observed, 6- and 12-month PFS, TTP and TTF were censored at the date of last follow-up or 6 months, respectively 12 months, of follow-up whichever came first.

Surrogacy measures

The individual-level surrogacy was assessed following a copula-based approach, with OS

and the surrogate endpoints jointly modelled using a one-parameter copula. The individual-level associations were estimated by the Spearman rank correlation coefficient ($\rho_{Spearman}$) calculated from the copula parameter [4].

The trial-level surrogacy - the association between the treatment effects - was evaluated with two frameworks. In the weighted regression model (WLR) approach, treatment effects on OS and PFS/TTP/TTF were estimated separately for each trial, based on the logarithm of the hazard ratios (log(HR)) using Cox proportional hazard models. We assessed the association between the treatment effects using the coefficient of determination (R^2_{WLR}) of a linear regression model weighted by the trial size. The second method follows the two-stage model (2SM) adapted to time-to-event endpoints introduced by Burzykowski et al. [4]. First, we simultaneously estimated the treatment effects on OS and on the candidate surrogate endpoints in each trial using a bivariate survival model based on the one-parameter Clayton copula. This approach enables taking into account the correlation between the endpoints in the estimation of the HR. We then estimated the association between the treatment effects (Weibull-distribution-based log(HR)) using an error-in-variable model, a regression model that allows taking into account the estimation errors. We assessed the trial-level association using the coefficient of determination (R^2_{2SM}).

All analyses were made on an intention-to-treat basis. We reported confidence intervals for a two-sided confidence-level of 95% (95% CI). All analyses were performed using SAS software v9.3 following Burzykowski et al. [4].

Strength of association

The strength of the trial-level association was ranked according to the Institute for Quality and Efficiency in Health Care (IQWiG) guidelines [11]: high association (lower limit of the 95% CI of $R^2 \geq 0.72$), low association (higher limit of the 95% CI of $R^2 \leq 0.49$) or medium association (neither low nor high), meaning that the validity of the surrogate remains unclear.

Subgroup analyses

To control trials' heterogeneity, we performed two additional subgroup analyses. In the 1st subgroup analysis, we retained only trials focusing on first line treatment that included doxorubicin- or ifosfamide-based therapies in the control arm. In the second analysis, all trials were eligible, but only patients with leiomyosarcomas were included.

Results

Data

After screening 231 abstracts, we identified 21 trials eligible for inclusion and obtained the trial sponsor's agreement for 19 RCTs (Figure 1). IPD were available for a total of 14 RCTs [12-25]. Trials characteristics are presented in table 1. Three trials had two experimental arms

evaluating different administration schedules for the same drug [12,17,25]. We combined the two experimental arms into one for these studies. One trial was designed as two parallel randomized comparative studies [16], and it was included in the meta-analysis as two distinct trials so that we considered a total of 15 trials. Aside from one trial [20], RCTs evaluated chemotherapy-based regimens. Most trials compared an experimental chemotherapy to a doxorubicin or ifosfamide-based chemotherapy regimen as first-line treatment (table 1). One trial focused on leiomyosarcomas [22] and one trial excluded liposarcomas [25], all 13 other trials presented similar histological subtypes inclusion criteria.

IPD from the 2846 patients included in the trials were analyzed. Median follow-up duration ranged from 9.4 to 93 months (median: 35.5 months). Figure 2 displays forest plots for the treatment effects estimated by hazard ratios (HR) on two-year OS, and one-year PFS, TTP and TTF for each trial.

Among the 2165 patients who died during the total follow-up, 1704 (78.7%) died during the first 18 months. During the first six months of follow-up, 1526 patients progressed and 634 patients died: 570 from cancer, 17 from treatment toxicity and 47 from other causes. During the first year of follow-up, 2042 patients progressed and 1311 patients died: 1234 from cancer, 19 from treatment toxicity and 58 from another cause. For each of the three candidate surrogate endpoints, the number of events observed at 6 and 12 months is provided in additional file 1.

Correlation between the candidate surrogate endpoints and OS (individual-level surrogacy)

We relied on a one-parameter Clayton copula model, considered the best fitting model compared to Plackett or Hougaard copula. Considering a six-month follow-up for the surrogate endpoints, the individual-level correlations with 18-month OS for the three endpoints evaluated were modest, with PFS showing the highest correlation (0.62; 95% CI [0.59; 0.65]) (Table 2). Correlations obtained when using a one-year follow-up for the surrogate endpoints were slightly higher (PFS: 0.66; 95% CI [0.63; 0.68]).

Correlation between treatment effects on the candidate surrogate endpoints and treatment effect on OS (trial-level surrogacy)

A total of 15 pairs of log(HR) were compared for each endpoint. When considering a six-month follow-up for the surrogates, the trial-level associations R^2_{WLR} and R^2_{2SM}, estimated with the WLR approach and the two-stage model respectively, were low ($R^2_{WLR} \leq 0.60$; $R^2_{2SM} \leq 0.60$) (Table 2). When considering a one-year follow-up, the association measures remained low ($R^2_{WLR} \leq 0.60$; $R^2_{2SM} \leq 0.05$). Regression curves calculated based on the WLR models are shown in figure 3. As per IQWiG guidelines, all trial-level associations estimated were ranked as medium.

Subgroup analyses

The first subgroup analysis focused on trials comparing systemic therapy to doxorubicin- or ifosfamide-based chemotherapies in the first-line setting (N_{trial}=11; $N_{patient}$=2243). When considering a six-month follow-up, the three endpoints were moderately associated with 18-month OS at the patient level ($0.56 \leq \rho_{Spearman} \leq 0.67$). At the trial level, the association between the candidate surrogates and 18-month OS was low ($R^2_{WLR} \leq 0.60$; $R^2_{2SM} \leq 0.11$) (Table 2). When considering a 12-month follow-up, the individual-level associations were slightly higher ($0.61 \leq \rho_{Spearman} \leq 0.70$). At the trial level, the associations between the candidate surrogates and OS increased, particularly when estimated with the two-stage model. For the second subgroup analysis focusing on leiomyosarcomas, the treatment effects on OS and on the candidate surrogates could not be computed for one trial due to lack of events, it was then excluded from the subgroup analysis. Individual-level correlations slightly decreased compared to the primary analysis. At the trial level however, the correlations significantly increased (Table 2).

Discussion

We pooled IPD data from 2846 patients included in 14 RCTs to evaluate the surrogate properties of PFS, TTP and TTF for OS in advanced STS. At the individual-level, associations between the three endpoints and OS were moderate, with the highest correlation observed for PFS. At the trial level, associations between the treatment effects on three endpoints and treatment effect on OS were low with wide confidence intervals. The strength of the trial-level association was quantified as medium as per the IQWiG criteria, indicating that the validity of the endpoints as surrogates for OS remains unclear.

Several statistical methods are available to assess surrogacy. We relied on the two-stage approach developed by Buyse and Burzykowski based on IPD [4], considered the most rigorous statistical approach for surrogacy assessment [5,6]. Similarly, several criteria have been proposed to assess the validity of surrogate endpoints [11,26,27]. Although they present differences, they all require a lower limit of the 95% CI for the trial-level correlation coefficient at least higher than 0.6 to definitely validate a surrogate endpoint. As such, they all corroborate the absence of surrogacy evidence.

To the best of our knowledge, this is the second meta-analysis conducted in advanced STS patients, and the first on IPD. In the first meta-analysis, conducted on aggregated data, the authors reported a 0.61 trial-level association when assessing the surrogate properties of PFS, and concluded that PFS was an appropriate surrogate for OS [8]. However, we feel that data are lacking to conclude strongly. No confidence interval for the trial-level association was reported, a key element to quantify the validity of a surrogate endpoint using appropriate criteria [11,26,27]. As the correlation estimate reported was derived from a smaller set of trials

that in our study, it is likely that the precision was also poor.

Trial design for advanced sarcoma is particularly challenging due to the rarity and the heterogeneity of the disease and treatments, which may contribute to weaken the observed correlations between candidate surrogates and OS [28]. Most STS trials include different clinical phenotypes to increase their statistical power, even though specific RCTs would be required [28-30]. In the present study, the distribution of sarcoma subtypes across trials was highly variable, with proportions ranging from 0% [22] to 18% [18] for liposarcoma, 18% [21] to 100% [22] for leiomyosarcoma and 0% [22] to 14% [23] for synovial sarcoma. Locally advanced and metastatic patients have different prognoses, yet they are often conflated in trials as "advanced" sarcomas. Heterogeneity, in terms of treatment settings, remains between the trials included in our study, which could also have weakened the association between the candidate surrogates and OS. In the present study, 11 trials included only 1st line treatment (79% of all patients), one trial included 1st and 2nd line treatments (3% of patients), one trial included 2nd line treatment only (3% of patients) and one trial included 2nd to 5th line treatments (13% of patients). Central review at study entry is also likely to interfere. Patients with inappropriate histologies or grades could be included and thus dilute the overall association. In our study, 11 out of the 14 trials reported that radiological central review was used at study entry and two indicated that histologies were reviewed locally or in a specialized center. Results from our sensitivity analysis on doxorubicin- or ifosfamide-based therapies as first-line metastatic treatment did not significantly differ from our main analysis. Results from the subgroup analysis on patients with leiomyosarcoma seem promising, however one should interpret these results with caution due to the limited number of patients included. These factors, however, should be accounted for when interpreting the statistically non-significant correlations observed in this meta-analysis. A balance between restrictions to homogeneous trials to limit the dilution of the correlation estimates, while maximizing the number of trials to ensure sufficient precision, is thus a complex exercise.

Finally, absence of surrogacy could also be explained if indeed the candidate endpoints (PFS, TTP, TTF) do not adequately predict OS. This may be an argument for surrogates such as pathological response which may not relate to OS because of micrometastases disease outside the resection areas responsible for OS, not measured by the surrogate, or if an intervention has offsite target effects that are independent of the disease process [31]. Such argument however seems less likely for endpoints such as PFS which encompasses both local, distant events, and deaths. There may be however some potential biologic explanations for why survival endpoints encompassing progression may be truly increased without a survival impact. Booth and Eisenhauer, for example, have questioned the mechanisms of actions of some agents, especially those targeting cell signaling and angiogenesis, and whether with chronic administration, they could delay progression for a

time but lead to evolutionary changes in tumors, producing a more aggressive phenotype after treatment, thus offsetting the earlier delay in progression [32].

One should also consider that the absence of surrogacy evidence might also be related to an absence of the treatment effect on OS. Alternative survival endpoints to OS, such as progression-free survival (PFS) in trials of metastatic diseases or disease-free survival (DFS) in the adjuvant setting, are increasingly replacing OS in phase III trials [2]. In the United States of America, a large proportion of new cancer drug approvals granted by the Food and Drug Administration (FDA) are based on such alternative endpoints [1,33,34]. In advanced STS, FDA granted approval for pazopanib in 2012 based on proof of benefit for PFS [20], even though at the time no study had assessed trial-level association between PFS and OS. The Accelerated Approval regulations, instituted by the FDA in 1992, allowed drugs for serious conditions that filled an unmet medical need to be approved based on a surrogate endpoint. Using a surrogate endpoint enabled the FDA to approve these drugs faster. As a result, an increasing number of anticancer drug product approvals by the FDA are made based on endpoints other than OS [1,33,34], some with no sufficient proof of their surrogate validity for OS [34]. This issue is well illustrated with the example of bevacizumab in metastatic breast cancer based [35,36]. In the context of accelerated approval, the FDA guidance on the term unmet medical need is imprecise. While this is not an issue in advanced sarcoma, recent data have shown that this term can often be overused [37], and as such the use of surrogates through pathways such as accelerated approvals, may be far greater than conditions with true unmet needs [31].

We could not include all the trials retrieved by our literature search, although trial-level meta-analyses for surrogacy assessment should be based on all the available evidence. Since data that is easily located and included in meta-analysis can have different correlations that unavailable or unreported data, attempt to validate surrogate endpoints can be biased. However, to date, no example of a surrogate validation study based on all relevant evidence exists [31].

Several conditions have to be met to ensure adequate validation of a surrogate endpoint: (i) a significant quantity of data, both in terms of trials and patients, (ii) homogeneity, in terms of disease, settings, and mechanisms of action of the drugs, and (iii) strong statistical thresholds. Although our meta-analysis did not lead to the validation of a surrogate in advanced sarcoma, endpoints that do not achieve the high bar of validated surrogate continue to be useful in testing new treatments [38]. In disease and treatment cases in which it is reasonable to assume that an effect on OS can only be achieved if there is also an effect on PFS, lack of an effect on PFS could be used as a phase II futility assessment (or early phase III futility assessment), assuming that the phase III end point is OS [39,40].

Conclusion

Our meta-analysis did not lead to significant evidence to validate PFS, TTP or TTF as surrogate markers for OS when assessing systemic treatment in advanced STS. OS should therefore remain the primary endpoint in a randomized phase III trial. One should however acknowledge that trial design for advanced soft tissue sarcoma is particularly challenging due to the rarity and the heterogeneity of the disease and treatments, which may have contributed to weaken the observed correlations between candidate surrogates and OS. In addition, it is reasonable to assume that an effect on OS can only be achieved if there is also an effect on disease progression. As such, alternative endpoints, e.g. PFS, remain useful in testing new treatments in earlier drug development stages, such as in phase II trials or phase III futility assessment, provided OS data are collected throughout the trial.

References

1. Sridhara R, Johnson JR, Justice R, Keegan P, Chakravarty A, Pazdur R. Review of oncology and hematology drug product approvals at the US Food and Drug Administration between July 2005 and December 2007. J Natl Cancer Inst. 2010 Feb 24;102(4):230–43.
2. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival end point reporting in randomized cancer clinical trials: a review of major journals. J Clin Oncol Off J Am Soc Clin Oncol. 2008 Aug 1;26(22):3721–6.
3. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostat Oxf Engl. 2000 Mar;1(1):49–67.
4. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. J R Stat Soc Ser C Appl Stat. 2001 Jan 1;50(4):405–22.
5. Green E, Yothers G, Sargent DJ. Surrogate endpoint validation: statistical elegance versus clinical relevance. Stat Methods Med Res. 2008 Oct;17(5):477–86.
6. Renfro LA, Shi Q, Sargent DJ. Surrogate End Points in Soft Tissue Sarcoma: Methodologic Challenges. J Clin Oncol Off J Am Soc Clin Oncol. 2016 Aug 22;
7. Coindre JM, Terrier P, Guillou L, et al. Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: a study of 1240 patients from the French Federation of Cancer Centers Sarcoma Group. Cancer. 2001 May 15;91(10):1914–26.
8. Zer A, Prince RM, Amir E, Abdul Razak A. Evolution of Randomized Trials in Advanced/Metastatic Soft Tissue Sarcoma: End Point Selection, Surrogacy, and Quality of Reporting. J Clin Oncol. 2016 May 1;34(13):1469–75.
9. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009 Jul 21;6(7):e1000097.
10. Bellera CA, Penel N, Ouali M, et al. Guidelines for time-to-event end point definitions in sarcomas and gastrointestinal stromal tumors (GIST) trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials)†. Ann Oncol Off J Eur Soc Med Oncol. 2015 May;26(5):865–72.
11. Institute for Quality and Efficiency in Health Care (IQWiG). Validity of surrogate endpoints in oncology: Executive summary of rapid report A10-05, Version 1.1. In: Institute for Quality and Efficiency in Health Care: Executive Summaries [Internet]. Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2005 [cited 2016 Dec 30]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK198799/
12. Nielsen OS, Dombernowsky P, Mouridsen H, et al. High-dose epirubicin is not an alternative to standard-dose doxorubicin in the treatment of advanced soft tissue sarcomas. A study of the EORTC soft tissue and bone sarcoma group. Br J Cancer. 1998 Dec;78(12):1634–9.
13. Verweij J, Lee SM, Ruka W, et al. Randomized phase II study of docetaxel versus doxorubicin in first- and second-line chemotherapy for locally advanced or metastatic soft tissue sarcomas in

adults: a study of the european organization for research and treatment of cancer soft tissue and bone sarcoma group. J Clin Oncol Off J Am Soc Clin Oncol. 2000 May;18(10):2081–6.

14. Le Cesne A, Judson I, Crowther D, et al. Randomized phase III study comparing conventional-dose doxorubicin plus ifosfamide versus high-dose doxorubicin plus ifosfamide plus recombinant human granulocyte-macrophage colony-stimulating factor in advanced soft tissue sarcomas: A trial of the European Organization for Research and Treatment of Cancer/Soft Tissue and Bone Sarcoma Group. J Clin Oncol Off J Am Soc Clin Oncol. 2000 Jul;18(14):2676–84.
15. Judson I, Radford JA, Harris M, et al. Randomised phase II trial of pegylated liposomal doxorubicin (DOXIL/CAELYX) versus doxorubicin in the treatment of advanced or metastatic soft tissue sarcoma: a study by the EORTC Soft Tissue and Bone Sarcoma Group. Eur J Cancer Oxf Engl 1990. 2001 May;37(7):870–7.
16. van Oosterom AT, Mouridsen HT, Nielsen OS, et al. Results of randomised studies of the EORTC Soft Tissue and Bone Sarcoma Group (STBSG) with two different ifosfamide regimens in first- and second-line chemotherapy in advanced soft tissue sarcoma patients. Eur J Cancer Oxf Engl 1990. 2002 Dec;38(18):2397–406.
17. Lorigan P, Verweij J, Papai Z, et al. Phase III Trial of Two Investigational Schedules of Ifosfamide Compared With Standard-Dose Doxorubicin in Advanced or Metastatic Soft Tissue Sarcoma: A European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group Study. J Clin Oncol. 2007 Jul 20;25(21):3144–50.
18. Maurel J, Lopez-Pousa A, de las Penas R, et al. Efficacy of Sequential High-Dose Doxorubicin and Ifosfamide Compared With Standard-Dose Doxorubicin in Patients With Advanced Soft Tissue Sarcoma: An Open-Label Randomized Phase II Study of the Spanish Group for Research on Sarcomas. J Clin Oncol. 2009 Apr 10;27(11):1893–8.
19. Fayette J, Penel N, Chevreau C, et al. Phase III trial of standard versus dose-intensified doxorubicin, ifosfamide and dacarbazine (MAID) in the first-line treatment of metastatic and locally advanced soft tissue sarcoma. Invest New Drugs. 2009 Oct;27(5):482–9.
20. van der Graaf WT, Blay J-Y, Chawla SP, et al. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): a randomised, double-blind, placebo-controlled phase 3 trial. The Lancet. 2012 May;379(9829):1879–86.
21. Bui-Nguyen B, Ray-Coquard I, Chevreau C, et al. High-dose chemotherapy consolidation for chemosensitive advanced soft tissue sarcoma patients: an open-label, randomized controlled trial. Ann Oncol Off J Eur Soc Med Oncol. 2012 Mar;23(3):777–84.
22. Pautier P, Floquet A, Penel N, et al. Randomized Multicenter and Stratified Phase II Study of Gemcitabine Alone Versus Gemcitabine and Docetaxel in Patients with Metastatic or Relapsed Leiomyosarcomas: A Federation Nationale des Centres de Lutte Contre le Cancer (FNCLCC) French Sarcoma Group Study (TAXOGEM study). The Oncologist. 2012 Sep 1;17(9):1213–20.
23. Judson I, Verweij J, Gelderblom H, et al. Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: a randomised controlled phase 3 trial. Lancet Oncol. 2014 Apr;15(4):415–23.

24. Gelderblom H, Blay JY, Seddon BM, et al. Brostallicin versus doxorubicin as first-line chemotherapy in patients with advanced or metastatic soft tissue sarcoma: an European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group randomised phase II and pharmacogenetic study. Eur J Cancer Oxf Engl 1990. 2014 Jan;50(2):388–96.

25. Bui-Nguyen B, Butrynski JE, Penel N, et al. A phase IIb multicentre study comparing the efficacy of trabectedin to doxorubicin in patients with advanced or metastatic untreated soft tissue sarcoma: the TRUSTS trial. Eur J Cancer Oxf Engl 1990. 2015 Jul;51(10):1312–20.

26. Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports. Health Technol Assess Winch Engl. 2009 Jan;13(8):iii, ix–xi, 1-50.

27. Lassere MN, Johnson KR, Schiff M, Rees D. Is blood pressure reduction a valid surrogate endpoint for stroke prevention? An analysis incorporating a systematic review of randomised controlled trials, a by-trial weighted errors-in-variables regression, the surrogate threshold effect (STE) and the Biomarker-Surrogacy (BioSurrogate) Evaluation Schema (BSES). BMC Med Res Methodol. 2012 Mar 12;12:27.

28. Constantinidou A, van der Graaf WTA. The fate of new fosfamides in phase III studies in advanced soft tissue sarcoma. Eur J Cancer Oxf Engl 1990. 2017 Oct;84:257–61.

29. Kummar S, Allen D, Monks A, et al. Cediranib for metastatic alveolar soft part sarcoma. J Clin Oncol Off J Am Soc Clin Oncol. 2013 Jun 20;31(18):2296–302.

30. Constantinidou A, Miah A, Pollack S, Jones RL. New drugs and clinical trial design in advanced sarcoma: have we made any progress? Future Oncol Lond Engl. 2013 Oct;9(10):1409–11.

31. Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? BMC Med. 2017 Jul 21;15(1):134.

32. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? J Clin Oncol Off J Am Soc Clin Oncol. 2012 Apr 1;30(10):1030–3.

33. Kim C, Prasad V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. JAMA Intern Med. 2015 Dec 1;175(12):1992.

34. Kim C, Prasad V. Strength of Validation for Surrogate End Points Used in the US Food and Drug Administration's Approval of Oncology Drugs. Mayo Clin Proc. 2016 Jun;91(6):713–25.

35. Miller K, Wang M, Gralow J, et al. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. N Engl J Med. 2007 Dec 27;357(26):2666–76.

36. Carpenter D, Kesselheim AS, Joffe S. Reputation and precedent in the bevacizumab decision. N Engl J Med. 2011 Jul 14;365(2):e3.

37. Lu E, Shatzel J, Shin F, Prasad V. What constitutes an "unmet medical need" in oncology? An empirical evaluation of author usage in the biomedical literature. Semin Oncol. 2017 Feb;44(1):8–12.

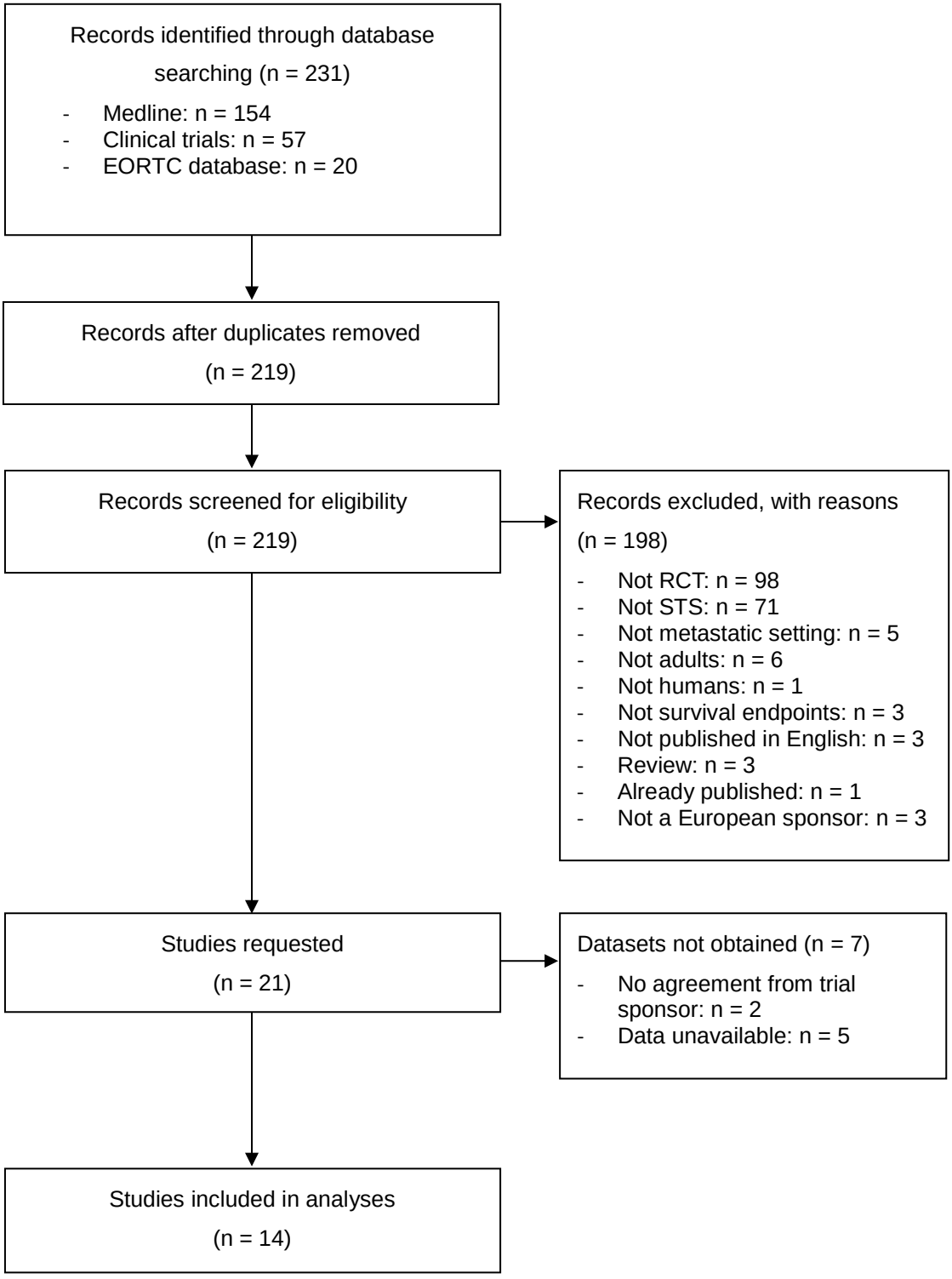
38. LeBlanc M, Tangen C. Surrogates for Survival or Other End Points in Oncology. JAMA Oncol. 2016 Feb;2(2):263–4.

39. Goldman B, LeBlanc M, Crowley J. Interim futility analysis with intermediate endpoints. Clin Trials Lond Engl. 2008;5(1):14–22.

40. Redman MW, Goldman BH, LeBlanc M, Schott A, Baker LH. Modeling the relationship between progression-free survival and overall survival: the phase II/III trial. Clin Cancer Res Off J Am Assoc Cancer Res. 2013 May 15;19(10):2646–56.

Figures

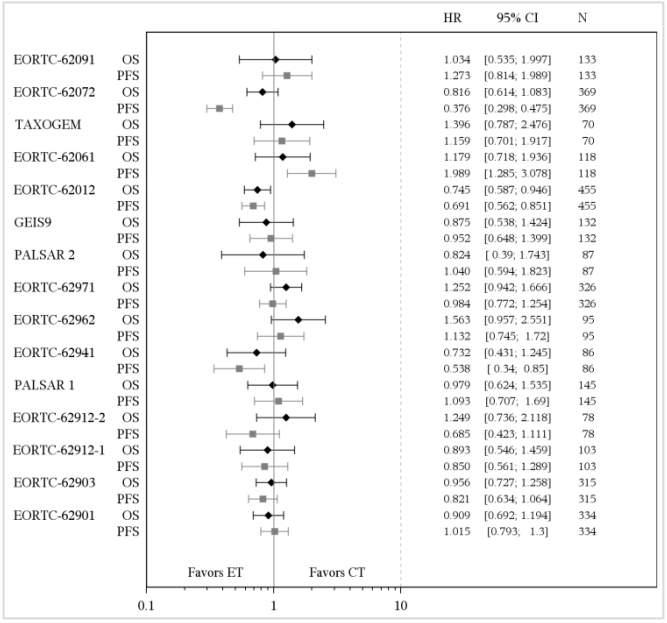
Figure 1 : Flow of information through the different phases of the study selection, as per PRISMA guidelines [9, a]



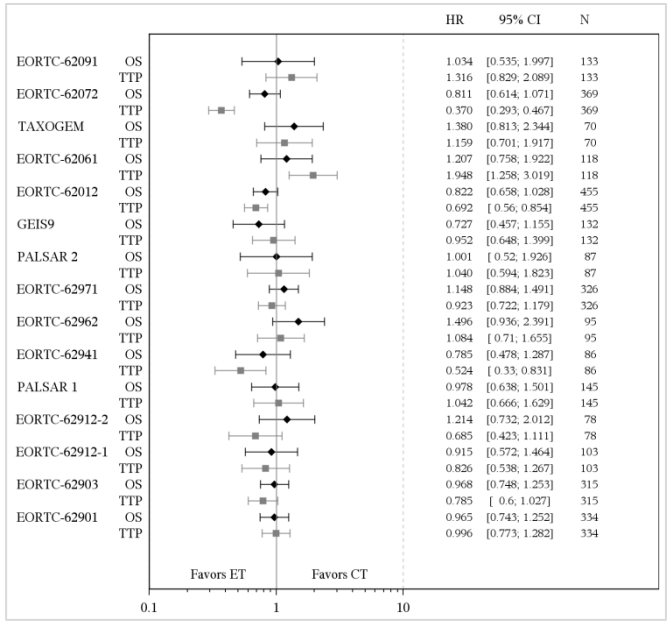
[a] EORTC = European Organization for Research and Treatment of Cancer; RCT = randomized control trial; STS = soft-tissue sarcoma

Figure 2: Forest plots of treatment effects (hazard ratios - HR) on 12-month progression-free survival (A), time-to progression (B) and time-to-treatment failure (C) and on 18-month overall survival (OS) estimated using separate Cox models. The first row for each trial shows the result for OS, and the second row shows the result for the candidate surrogate. The diamonds and squares represent the point estimates for OS and the candidate surrogate, respectively. The horizontal error bars show the 95% confidence interval (CI) of each hazard ratio (15 trials, 2846 patients). [a]

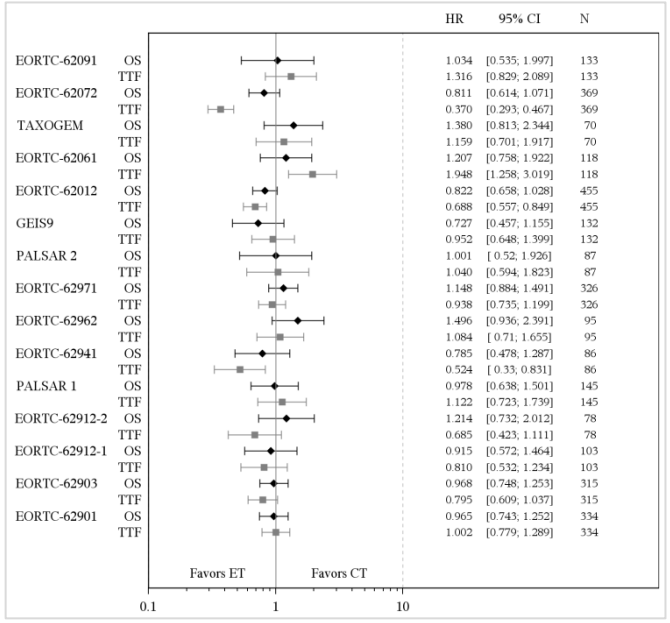
A: Progression-free survival (PFS)



B: Time-to progression (TTP)

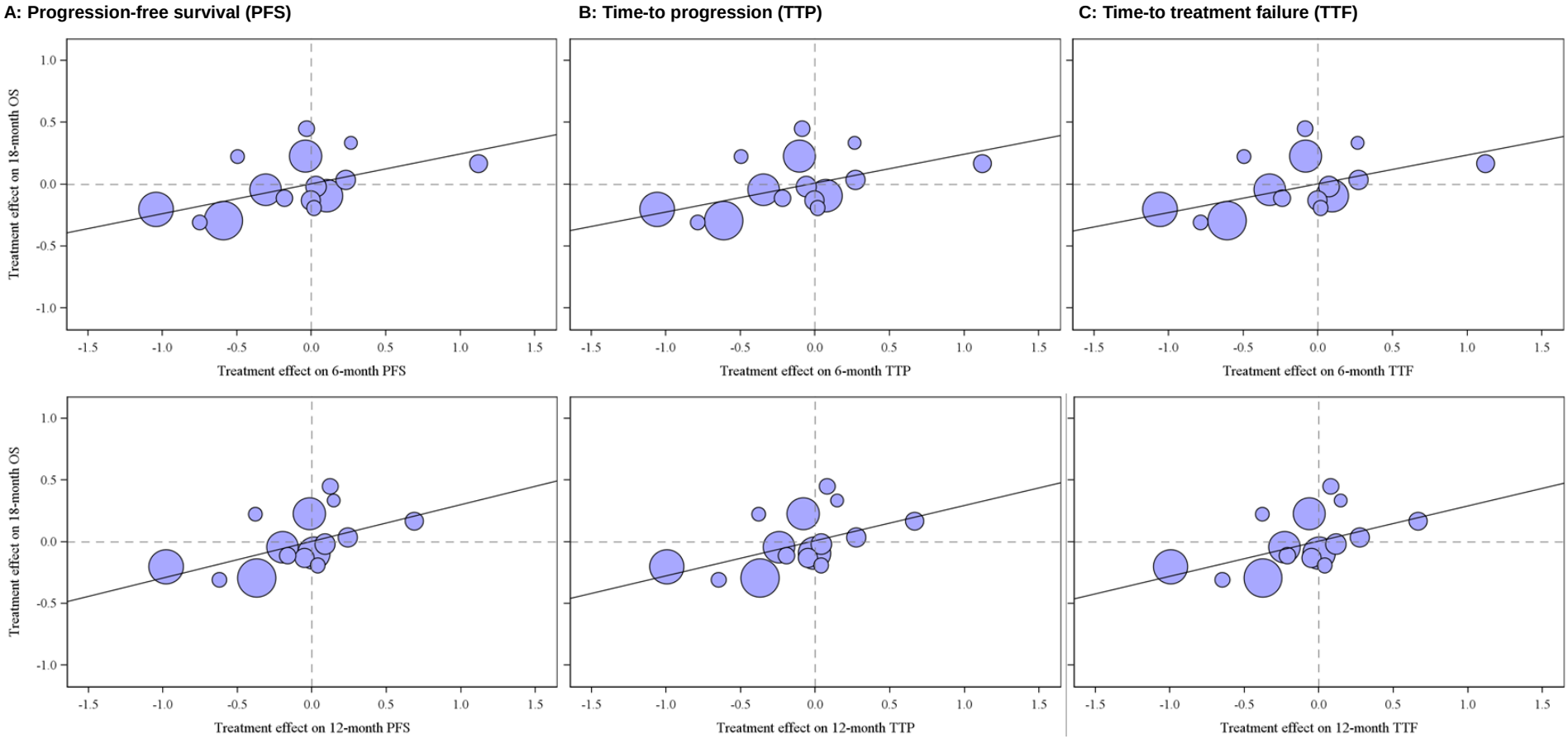


C: Time-to-treatment failure (TTF)



[a] CT = control treatment; ET = experimental treatment.

Figure 3: Trial-level association between treatment effects (log(HR)) on 18-month overall survival (OS) and (A) progression-free survival, (B) time-to progression and (C) time-to treatment failure evaluated at 6 months and 12 months estimated by the weighted linear regression approach. Each circle represents a trial, and the surface area of the circle is proportional to the size of the corresponding trial (15 trials) [a]



[a] HR = hazard ratio; PFS = progression-free survival; TTF = time-to-treatment failure; TTP = time-to-progression; OS = overall survival

Tables

Table 1: Key characteristics of the included trials [a]

Study	Phase	Inclusion period	N	Treatment line	Control arm	Experimental arm	Median follow-up	
							All patients	Patients alive
62091[25]	IIb/III	≥ 2011	133	1st line	Doxorubicin	Trabectedin	9.4 months	8.6 months
62072[20]	III	2008 - 2010	369	2nd to 5th line	Placebo	Pazopanib	14.6 months	12.2 months
Taxogem[22,b]	II	2006 - 2008	70	2nd line	Gemcitabine	Gemcitabine + Docetaxel + Lenograstime	32.5 months	24.9 months
62061[24]	II	2006 - 2008	118	1st line	Doxorubicin	Brostallicin	21.3 months	19.3 months
62012[23]	III	2003 - 2010	455	1st line	Doxorubicin	Intensified Doxorubicin + Ifosfamide	56.4 months	30.7 months
GEIS9[18]	II	2003 - 2007	132	1st line	Doxorubicin	Intensified Doxorubicin + Ifosfamide	22.5 months	15.4 months
Palsar2[21]	III	2000 - 2008	87	1st line	MAID	MAID + MICE	22.3 months	21.4 months
62971[17]	III	1998 - 2001	326	1st line	Doxorubicin	Ifosfamide	51.7 months	43.1 months
62962[15]	II	≥ 1997	95	1st line	Doxorubicin	Doxorubicin pegylated liposomal	35.2 months	14.5 months
62941[13]	II	≥ 1995	86	1st and 2nd line	Doxorubicin	Docetaxel	35.9 months	10.9 months
Palsar1[19]	III	1994 - 1997	145	1st line	MAID	Intensified MAID	93.0 months	89.7 months
62912[16]	II	1992 - 1994	78	2nd-line	Ifosfamide5 g/m²/1 day	Ifosfamide3 g/m²/3 days	30.6 months	16.2 months
		1994 - 1996	103	1st line	Ifosfamide5 g/m²/1 day	Ifosfamide3 g/m²/3 days	35.5 months	6.6 months
62903[14]	III	1992 - 1995	315	1st line	Doxorubicin + Ifosfamide	Doxorubicin + Ifosfamide + GM-GSF	91.4 months	40.1 months
62901[12]	II	≥ 1991	334	1st line	Doxorubicin	Epirubicin	50.2 months	13.3 months

[a] GM-GSF: Recombinant human granulocyte-macrophage colony-stimulating factor; MAID = Doxorubicin, Ifosfamide and Dacarbazine; MICE: Mesna, Ifosfamide, Carboplatin and Etoposide; Ref = reference.

[b] Only patients with leiomyosarcoma included.

Table 2: Individual- and trial-level associations between 6-month and 12-month progression-free survival, time-to-progression, time-to-treatment failure and 18-month overall survival [a]

		Individual-level association	Trial-level association	
Folow-up	**Endpoint**	**$\rho_{Spearman}$ [b] [95%CI]**	**R^2_{WLR} [c] [95%CI]**	**R^2_{2SM} [d] [95%CI]**
All trials (N_{trial} = 15; $N_{patient}$ = 2846)				
6 months	PFS	0.62 [0.59; 0.65]	0.33 [0.00; 0.60]	0.04 [0.00; 0.43]
	TTP	0.59 [0.56; 0.63]	0.32 [0.00; 0.58]	0.07 [0.00; 0.60]
	TTF	0.60 [0.57; 0.63]	0.32 [0.00; 0.58]	0.06 [0.00; 0.57]
12 months	PFS	0.66 [0.63; 0.68]	0.33 [0.00; 0.60]	0.00 [0.00; 0.05]
	TTP	0.63 [0.60; 0.66]	0.30 [0.00; 0.57]	0.00 [0.00; 0.02]
	TTF	0.64 [0.61; 0.67]	0.31 [0.00; 0.58]	0.00 [0.00; 0.01]
Doxorubicin- or ifosfamide-based treatment, first-line setting (N_{trial} = 11; $N_{patient}$ = 2243)				
6 months	PFS	0.63 [0.60; 0.67]	0.30 [0.00; 0.60]	0.00 [0.00; 0.08]
	TTP	0.60 [0.56; 0.64]	0.26 [0.00; 0.58]	0.00 [0.00; 0.11]
	TTF	0.61 [0.57; 0.65]	0.27 [0.00; 0.58]	0.00 [0.00; 0.06]
12 months	PFS	0.67 [0.64; 0.70]	0.39 [0.00; 0.66]	0.08 [0.00; 0.86]
	TTP	0.64 [0.61; 0.68]	0.31 [0.00; 0.61]	0.12 [0.00; 1.00]
	TTF	0.65 [0.62; 0.68]	0.32 [0.00; 0.62]	0.10 [0.00; 1.00]
Leiomyosarcomas (N_{trial} = 14; $N_{patient}$ = 1025)				
6 months	PFS	0.57 [0.51; 0.62]	0.59 [0.15; 0.76]	0.91 [0.00; 1.00]
	TTP	0.55 [0.49; 0.60]	0.58 [0.13; 0.75]	0.97 [0.00; 1.00]
	TTF	0.53 [0.48; 0.58]	0.59 [0.14; 0.76]	0.91 [0.00; 1.00]
12 months	PFS	0.59 [0.54; 0.64]	0.59 [0.16; 0.75]	0.91 [0.00; 1.00]
	TTP	0.52 [0.47; 0.58]	0.58 [0.15; 0.75]	0.97 [0.00; 1.00]
	TTF	0.53 [0.48; 0.58]	0.58 [0.15; 0.75]	0.91 [0.00; 1.00]

[a] CI = confidence interval; PFS = progression-free survival; TTF = time-to-treatment failure; TTP = time-to-progression.

[b] $\rho_{Spearman}$ represents the Spearman rank correlation coefficient between the candidate surrogates and overall survival.

[c] R^2_{WLR} represents the coefficient of determination between treatment effect on the candidate surrogates and overall survival based on weighted linear regression models.

[d] R^2_{2SM} represents the coefficient of determination between treatment effect on the candidate surrogates and overall survival based on the two-stage model [4].

Additional files

Additional file 1: Number of events observed for progression-free survival, time-to-progression and time-to-treatment failure after 6 and 12 months of follow-up; All trials (N_{trial} = 15; $N_{patient}$ = 2846) [a]

Folow-up	Endpoint	Number of events
6 months	TTP	1629
	TTF	1646
	PFS	1693
12 months	TTP	2226
	TTF	2245
	PFS	2303

[a] PFS = progression-free survival; TTF = time-to-treatment failure; TTP = time-to-progression.

Additional file 2: Number of events observed for progression-free survival, time-to-progression and time-to-treatment failure after 6 and 12 months of follow-up; Trials with doxorubicin- or ifosfamide-based control arm, first-line setting (N_{trial} = 11; $N_{patient}$ = 2243) [a]

Folow-up	Endpoint	Number of events
6 months	TTP	1203
	TTF	1220
	PFS	1264
12 months	TTP	1695
	TTF	1714
	PFS	1768

[a] PFS = progression-free survival; TTF = time-to-treatment failure; TTP = time-to-progression.

Additional file 3: Individual- and trial-level associations between 6-month and 12-month progression-free survival, time-to-progression, time-to-treatment failure and 24-month overall survival [a]

		Individual-level association	Trial-level association	
Folow-up	**Endpoint**	**$\rho_{Spearman}$ [b] [95%CI]**	**R^2_{WLR} [c] [95%CI]**	**R^2_{2SM} [d] [95%CI]**
All trials (N_{trial} = 15; $N_{patient}$ = 2846)				
6 months	PFS	0.60 [0.56; 0.63]	0.38 [0.02; 0.62]	0.18 [0.00; 1.00]
	TTP	0.57 [0.54; 0.60]	0.36 [0.01; 0.61]	0.35 [0.00; 1.00]
	TTF	0.58 [0.54; 0.61]	0.36 [0.00; 0.61]	0.16 [0.00; 1.00]
12 months	PFS	0.62 [0.59; 0.65]	0.40 [0.03; 0.64]	0.02 [0.00; 0.53]
	TTP	0.60 [0.57; 0.63]	0.37 [0.02; 0.62]	0.10 [0.00; 1.00]
	TTF	0.61 [0.58; 0.63]	0.38 [0.02; 0.63]	0.08 [0.00; 1.00]
Doxorubicin- or ifosfamide-based treatment, first-line setting (N_{trial} = 11; $N_{patient}$ = 2243)				
6 months	PFS	0.60 [0.57; 0.64]	0.29 [0.00; 0.59]	0.00 [0.00; 0.16]
	TTP	0.57 [0.53; 0.61]	0.26 [0.00; 0.57]	0.01 [0.00; 0.63]
	TTF	0.58 [0.54; 0.62]	0.26 [0.00; 0.57]	0.01 [0.00; 0.42]
12 months	PFS	0.64 [0.61; 0.67]	0.39 [0.00; 0.66]	0.19 [0.00; 1.00]
	TTP	0.61 [0.58; 0.64]	0.32 [0.00; 0.62]	0.71 [0.00; 1.00]
	TTF	0.62 [0.59; 0.65]	0.33 [0.00; 0.62]	0.44 [0.00; 1.00]
Leiomyosarcomas (N_{trial} = 14; $N_{patient}$ = 1025)				
6 months	PFS	0.54 [0.49; 0.60]	0.68 [0.26; 0.82]	NC
	TTP	0.52 [0.47; 0.58]	0.66 [0.23; 0.80]	NC
	TTF	0.53 [0.47; 0.58]	0.68 [0.26; 0.81]	NC
12 months	PFS	0.54 [0.49; 0.60]	0.74 [0.38; 0.85]	NC
	TTP	0.53 [0.47; 0.58]	0.73 [0.36; 0.84]	NC
	TTF	0.53 [0.48; 0.58]	0.74 [0.38; 0.85]	NC

[a] CI = confidence interval; NC = not computed; PFS = progression-free survival; TTF = time-to-treatment failure; TTP = time-to-progression.

[b] $\rho_{Spearman}$ represents the Spearman rank correlation coefficient between the candidate surrogates and overall survival.

[c] R^2_{WLR} represents the coefficient of determination between treatment effect on the candidate surrogates and overall survival based on weighted linear regression models.

[d] R^2_{2SM} represents the coefficient of determination between treatment effect on the candidate surrogates and overall survival based on the two-stage model (4).

3.3 Supplementary analyses

3.3.1 Sensitivity analysis: Censoring process

In the analyses presented in the manuscript, data were censored after a certain follow-up duration for each patient: two years for OS and six or twelve months for PFS, TTP and TTF. As a result, each patient benefited from the same follow-up duration for OS and the candidate surrogates in the MA. In order to be more realistic, we led a sensitivity analysis considering a common cut-off date that is, assuming all patients are followed-up until the end of follow-up for the last patient included.

For each trial, we identified the last patient included and the dates corresponding to a follow-up of (1) two years, (2) twelve months, and (3) six months for this patient. All other patients included in the trial were then censored at date (1) for the evaluation of two-year OS, date (2) for the evaluation of the 12-month candidate surrogates, and date (3) for the evaluation of the 6-months candidate surrogates.

The individual- and trial-level associations between the candidate surrogates and 2-year OS are presented in table 4.1.

Table 2: Individual- and trial-level associations with 2-year OS (N_{trial} = 15; $N_{patient}$ = 2846) – Data censored at a cut-off date

		Individual-level association	Trial-level association	
Follow-up	Endpoint	R^2_{ind} [95%CI]	R^2_{trial} (WLR)* [95%CI]	R^2_{trial} (2SM)** [95%CI]
6 months	PFS	0.62 [0.59; 0.65]	0.38 [0.02; 0.63]	0.01 [-0.81; 0.84]
	TTP	0.59 [0.56; 0.62]	0.35 [0.00; 0.60]	0.04*** [-0.15; 0.22]
	TTF	0.60 [0.57; 0.63]	0.35 [0.01; 0.61]	0.04*** [-0.15; 0.22]
12 months	PFS	0.63 [0.60; 0.65]	0.39 [0.02; 0.64]	0.03*** [-0.13; 0.18]
	TTP	0.60 [0.57; 0.63]	0.36 [0.01; 0.61]	0.03*** [-0.15; 0.22]
	TTF	0.61 [0.58; 0.64]	0.36 [0.01; 0.62]	0.03*** [-0.15; 0.22]

* R^2_{trial} (WLR): Trial-level coefficient correlation based on linear regression models weighted by the trial size (equation 4.5)

** R^2_{trial} (2SM): Trial-level coefficient correlation based on the two-stage model (equation 3.12);

*** R^2_{trial} (2SM) adjusted on estimation errors could not be computed, we report here unadjusted R^2_{2SM} (equation 4.5)

At the individual-level, the associations with OS were very close to those estimated in the primary analysis. At the trial-level, the associations estimated using the meta-regression models were also very similar whatever the censoring approach. The trial-level associations estimated using the two-stage models were lower than the ones from the primary analysis.

Note that due to convergence issues, the R^2_{2SM} adjusted on estimation errors could not be computed in this sensitivity analysis, so that the associations reported in table 4.1 are unadjusted.

3.3.2 Sensitivity analysis: Weighting process

When assessing the trial-level association by conducting a meta-regression, different weights can be used for the fixed treatment effects (chapter 1.3.2.2). If the sample size is the most common choice, it might not be the most relevant. The statistical power of an RCT is not directly determined by the number of patients but by the number of events observed during the trial. One approach might be to use the number of events observed as weights in the meta-regression.

To evaluate the impact of the choice of the weights on the estimation of the trial-level association, we performed sensitivity analyses using this alternative weighting procedures. We carried out a meta-regression with fixed treatment effects weighted by the number of events common to OS and the candidate surrogate observed. For PFS, all deaths were considered as events in the weighting process, whereas only deaths from cancer were considered as events for TTP. Regarding TTF, deaths from cancer and deaths from treatment toxicity were included in the weighting process. As the follow-up duration was not similar for OS and the candidate surrogate, we chose the shorter duration, i.e. six or twelve months, to define the weight of each trial. The trial-level associations assessed with the two weighting approaches – sample size and number of events – are presented in table 4.2.

Table 3: Trial-level associations with 2-year OS estimated by meta-regression with different weighting approaches

		R^2_{trial} (WLR) [95%CI] with weighting by…	
Follow-up	Endpoint	Sample size	Number of events
6 months	PFS	0.38 [0.02; 0.62]	0.38 [0.02; 0.62]
	TTP	0.36 [0.01; 0.61]	0.36 [0.01; 0.62]
	TTF	0.36 [0.00; 0.61]	0.24 [0.01; 0.62]
12 months	PFS	0.40 [0.03; 0.64]	0.43 [0.04; 0.66]
	TTP	0.37 [0.02; 0.62]	0.40 [0.03; 0.64]
	TTF	0.38 [0.02; 0.63]	0.41 [0.03; 0.64]

* R^2_{trial} (WLR): Trial-level coefficient correlation based on linear regression models weighted by trial size (equation 4.5)

Trial-level correlations estimated using the meta-regression weighted by the sample size and the numbers of deaths were very consistent, in terms of both point estimates and precision of the estimations.

4 Time-to-next treatment: an alternative endpoint in advanced sarcoma trials?

4.1 Introduction

Progression-based endpoints such as PFS or TTP have not been formally validated as surrogate endpoints for OS in the context of advanced STS (chapter 3). Other composite endpoints, not relying solely on progression evaluation, could then be worth investigating. Time-to-next treatment (TNT) is an established endpoint that is mostly applied in hematological malignancies and has recently been used in breast, colon, and prostate cancer (60–62). It is defined as the time from baseline (randomization, inclusion or initiation of the treatment) to initiation of a new treatment after failure of the previous one. By definition, it includes all possible reasons for switching treatment, so that it might more accurately reflect a change in the patient status. As information regarding subsequent treatments was not systematically available in the trials collected for the MA presented in the previous chapter, we relied on IPD from a prospective cohort of 1575 patients to investigate this hypothesis. We assessed prognostic factors of TNT and OS, and estimated the individual-level association between TNT and OS at different lines of treatment using a copula-based model.

This study highlighted a strong correlation between TNT and OS at the patient level, especially in the first-line setting. Further investigation of TNT as a surrogate for OS in a formal MA of RCTs would be worth undertaking. We therefore advise to collect the relevant information for the definition of TNT in future trials.

This work has been published in *BMC Medicine* (Savina et al. 2017).

4.2 Publication

Savina et al. BMC Medicine (2017) 15:78
DOI 10.1186/s12916-017-0831-7

BMC Medicine

RESEARCH ARTICLE Open Access

Patterns of care and outcomes of patients with METAstatic soft tissue SARComa in a real-life setting: the METASARC observational study

Marion Savina[1,2], Axel Le Cesne[3], Jean-Yves Blay[4], Isabelle Ray-Coquard[4], Olivier Mir[3], Maud Toulmonde[5], Sophie Cousin[5], Philippe Terrier[6], Dominique Ranchere-Vince[7], Pierre Meeus[8], Eberhard Stoeckle[9], Charles Honoré[10], Paul Sargos[11], Marie-Pierre Sunyach[12], Cécile Le Péchoux[13], Antoine Giraud[1], Carine Bellera[1,2], François Le Loarer[14] and Antoine Italiano[5,15*]

Abstract

Background: Well-designed observational studies of individuals with rare tumors are needed to improve patient care, clinical investigations, and the education of healthcare professionals.

Methods: The patterns of care, outcomes, and prognostic factors of a cohort of 2225 patients with metastatic soft tissue sarcomas who were diagnosed between 1990 and 2013 and documented in the prospectively maintained database of the French Sarcoma Group were analyzed.

Results: The median number of systemic treatments was 3 (range, 1–6); 27% of the patients did not receive any systemic treatment and 1054 (49%) patients underwent locoregional treatment of the metastasis. Half of the patients who underwent chemotherapy (n = 810) received an off-label drug. Leiomyosarcoma was associated with a significantly better outcome than the other histological subtypes. With the exception of leiomyosarcomas, the benefit of a greater than third-line regimen was very limited, with a median time to next treatment (TNT) and overall survival (OS) ranging between 2.3 and 3.7 months and 5.4 and 8.5 months, respectively. The TNT was highly correlated with OS. Female sex, leiomyosarcoma histology, locoregional treatment of metastases, inclusion in a clinical trial, and treatment with first-line polychemotherapy were significantly associated with improved OS in the multivariate analysis.

Conclusions: The combination of doxorubicin with a second drug, such as ifosfamide, represents a valid option, particularly when tumor shrinkage is expected to provide clinical benefits. After failure of the second-line therapy, best supportive care should be considered, particularly in patients with non-leiomyosarcoma histology who are not eligible to participate in a clinical trial. Locoregional treatment of metastasis should always be included in the therapeutic strategy when feasible. TNT may represent a useful surrogate endpoint for OS in clinical studies.

Keywords: Sarcoma, Metastases, Outcome, Patterns of care, Chemotherapy, Surgery

* Correspondence: a.italiano@bordeaux.unicancer.fr
[5]Department of Medicine, Institut Bergonié, Bordeaux, France
[15]Early Phase Trials and Sarcoma Units, Institut Bergonié, 229 Cours de l'Argonne, Bordeaux, France
Full list of author information is available at the end of the article

Background

Soft-tissue sarcomas (STSs) represent a heterogeneous group of diseases that account for 1% of all malignancies in adults [1]. Despite adequate locoregional treatment, up to 40% of patients with STSs will develop metastatic disease [1, 2]. When metastases are detected, the standard of care is based on palliative chemotherapy. Due to their rarity, no specific data on the comprehensive management and outcomes of metastatic STS patients are available.

A national network of care coordinated by three national reference centres has been set up through the support of the French National Cancer Institute for the management of STS patients. All suspected or diagnosed STS cases are reviewed by an accredited pathologist who is an expert in the field, and the cases are included in a national database. The aim of this study was to use this unique set of data to assess the modalities of treatment of patients with metastatic STS in a real-life setting, to evaluate their impact on the outcome according to the histological subtype, and to identify prognostic factors.

Methods

Patients

From 1990 to 2013, patients ≥ 18 years old with a diagnosis of metastatic STS (excluding gastrointestinal stromal tumors, visceral sarcomas, and Ewing tumors) who were evaluated at one of the three national reference centres designated by the French National Cancer Institute for the management of STS (Centre Léon Bérard, Lyon; Institut Bergonié, Bordeaux; and Institut Gustave Roussy, Villejuif) were included in the prospectively maintained database of the French Sarcoma Group. A histological review of all patients was performed by the members of the pathological sub-committee of the French Sarcoma Group. The histological diagnosis and grading was established according to the World Health Organization Classification of Tumours and to the French grading system [2, 3].

Outcomes

Time to next treatment (TNT) was defined as the time from the systemic treatment onset to the next treatment or death due to any cause, whichever came first. When neither death nor new systemic therapy was observed, TNT was censored at the date of last patient contact. Overall survival (OS) was defined as the interval between the diagnosis of metastatic disease or the first-line systemic therapy onset and the time of death. When death was not observed, OS was censored at the date of last patient contact.

Statistical analysis

The statistical analysis of the baseline demographics and clinical outcomes was based on all data available up to the cut-off date of December 31, 2015. Descriptive statistics were used to show the distribution of variables in the population. Multivariate logistic regression models were used to identify biological and clinical factors associated with the type of treatment received and with the probability of survival 5 years after the diagnosis of metastases. Follow-up times were described as median values based on the inverse Kaplan–Meier estimator [4].

Prognostic factors of TNT and OS were identified using Cox proportional hazard models. The variables included in the univariate and multivariate analyses are detailed in Additional file 1.

The correlation between TNT and OS was evaluated at each of the four first-lines of metastatic chemotherapy by a Spearman rank correlation coefficient and was expressed as a value between 0 (no association) and 1 (perfect association). We used a reviewed copula-based approach that introduced an iterative multiple imputation method [5] for the estimation of the correlation coefficient. The data were analyzed using the SAS v9.3 and R v3.3 software packages.

Results

Patients

A total of 2165 patients were included in this study. Their characteristics are presented in Table 1. The median follow-up duration was 61 months (range, 1–300). The five most frequently detected histological subtypes were leiomyosarcoma (LMS), undifferentiated pleomorphic sarcoma (UPS), synovial sarcoma (SS), dedifferentiated liposarcoma (DLPS), and malignant peripheral nerve sheath tumors (MPNST).

General treatment patterns

The general treatment patterns are described in Table 2. Patients over 75 years of age ($P < 0.0001$) and with MPNST ($P = 0.0136$) had a lower probability of receiving any systemic treatment, whereas presence of liver, lung, peritoneal, bone, pleural, skin, or lymphatic metastases was associated with a higher probability of receiving chemotherapy. Being over 75 years ($P < 0.0001$), DLPS ($P = 0.0031$), a grade 3 ($P = 0.0188$), and the presence of more than one metastatic site ($P < 0.0001$) were associated with a lower probability of receiving a locoregional treatment, whereas being a woman ($P = 0.0012$), SS ($P = 0.0026$), and the presence of lymphatic, brain, bone, skin, soft tissue, or peritoneal metastases were associated with an increased probability of locoregional treatment. Locoregional metastasis treatment was the sole treatment for 250 patients (11.55%). The metastasis localization was the only factor associated with the probability of receiving only locoregional treatment. Indeed, the presence of liver ($P < 0.0001$), lung ($P <$

Table 1 Patient characteristics according to the study population

	All patients (n = 2165)		Patients alive at 5 years (n = 224)		Patients treated with metastatic chemotherapy (n = 1575)	
	n	%	n	%	n	%
Sex						
Male	1055	48.73	92	41.07	754	47.87
Female	1110	51.27	132	58.93	821	52.13
Age at first metastasis						
< 75 years old	1886	87.11	216	96.43	1429	90.73
≥ 75 years old	279	12.89	8	3.57	146	9.27
Histology						
Leiomyosarcoma	502	23.19	60	26.79	396	25.14
UPS	203	9.38	9	4.02	141	8.95
DLPS	172	7.94	12	5.36	112	7.11
Synovial sarcoma	188	8.68	16	7.14	150	9.52
MPNST	80	3.70	11	4.91	50	3.17
Other	1020	47.11	116	51.79	726	46.10
Grade						
1	138	6.37	48	21.43	94	5.97
2	590	27.25	74	33.04	440	27.94
3	1083	50.02	63	28.13	765	48.57
Not available	354	16.35	39	17.41	276	17.52
Number of metastatic sites						
1	1780	82.22	199	88.84	1248	79.24
> 1	385	17.78	25	11.16	327	20.76
Metastatic sites						
Lung	1399	64.62	149	66.52	1075	68.25
Liver	410	18.94	34	15.18	352	22.35
Peritoneum	396	18.29	60	26.79	319	20.25
Bone	370	17.09	29	12.95	305	19.37
Lymph node	304	14.04	35	15.63	236	14.98
Skin	172	7.94	25	11.16	136	8.63
Soft tissue	173	7.99	36	16.07	135	8.57
Pleura	163	7.53	11	4.91	140	8.89
Brain	113	5.22	5	2.23	89	5.65
Bone marrow	12	0.55	0	0.00	10	0.63
Other	228	10.53	32	14.29	166	10.54

UPS undifferentiated pleomorphic sarcoma, *DLPS* dedifferentiated liposarcoma, *MPNST* malignant peripheral nerve sheath tumors

0.0001), pleural ($P = 0.0005$), and peritoneal ($P = 0.0087$) metastases was associated with a lower probability of locoregional treatment alone, whereas patients with soft-tissue metastases ($P = 0.0031$) were more likely to receive only a locoregional treatment. Best supportive care alone was more likely to be proposed to patients over 75 years ($P < 0.0001$), with a grade 3 tumor ($P = 0.0306$), or with multiple metastatic sites ($P = 0.0201$).

Systemic treatment patterns (Table 2)

The median number of systemic treatments received by the patients was 3 (min = 1 and max = 6) and did not significantly differ across the histological subtypes. Patients < 75 years old ($P < 0.0001$) and those with lymph node involvement ($P = 0.0001$) were more likely to receive polychemotherapy in the first-line setting. The most frequently prescribed off-label drug was gemcitabine. Female sex ($P = 0.0313$) and age ≥ 75 years ($P =$

Table 2 General patterns of treatment according to study population

	All patients (n = 2165)		Patients alive at 5 years (n = 224)		Patients treated with chemotherapy (n = 1575)	
	n	%	n	%		
Metastatic treatment received						
Best supportive care only	340	15.70	13	5.80	0	0.00
Locoregional treatment	1054	48.68	187	83.48	804	51.05
Surgery	408	38.71	82	43.85	282	35.07
Radiotherapy	254	24.10	12	6.42	213	26.49
Radiofrequency	42	3.98	9	4.81	33	4.10
Other	30	2.85	3	1.60	19	2.36
Combination	320	30.36	81	43.32	257	31.97
None	1111	51.32	37	16.52	771	48.95
Chemotherapy	1575	72.75	156	69.64	1575	100
None	590	27.25	68	30.36	–	–
1 line	489	22.59	54	34.62	489	31.05
2 lines	293	13.53	24	15.38	293	18.60
3 lines	240	11.09	21	13.46	240	15.24
4 lines	157	7.25	11	7.05	157	9.97
> 4 lines	396	17.27	46	29.49	396	25.15
Anthracycline received						
Yes	–	–	109	69.87	951	60.38
No	–	–	47	30.13	624	39.62
Anthracycline received as first line						
Yes	–	–	98	62.82	852	54.10
No	–	–	58	37.18	723	45.90
Polychemotherapy received as first line						
Yes	–	–	95	60.90	716	45.46
No	–	–	61	39.10	859	54.54
Inclusion in a clinical trial						
Yes:	–	–	55	35.26	332	21.08
Line 1	–	–	10	6.41	122	7.75
Line 2	–	–	17	16.67	107	9.85
Line 3	–	–	10	12.82	56	7.06
Line 4	–	–	7	12.28	30	5.42
Other lines	–	–	11	23.91	17	4.29
No	–	–	101	64.74	1243	78.92
Off-label drugs						
Yes:	–	–	99	63.46	810	51.43
Line 1	–	–	21	13.46	194	12.32
Line 2	–	–	22	21.57	203	18.69
Line 3	–	–	14	17.95	169	21.31
Line 4	–	–	21	36.84	142	25.68
Other lines	–	–	21	45.65	102	25.76
No	–	–	57	36.54	765	48.57

0.0003) were factors associated with a lower probability of being part of a clinical trial. On the contrary, patients with LMS or SS ($P = 0.0217$) and patients with liver ($P = 0.0072$), skin ($P = 0.0013$) or peritoneal ($P = 0.0036$) metastases were more likely to be included in a clinical trial during the course of their treatment.

Time to next treatment and overall survival

The median TNT and OS according to the treatment line setting for the five most frequent histological subtypes are described in Table 3. Patients with metastatic LMS had the longest median survival, whereas patients with UPS had the shortest. The benefit of systemic therapy beyond the second line setting was limited, with a median TNT ranging between 2.3 and 3.5 months except for LMS (>4 months). The correlation estimated between TNT and OS was similar and high regardless of the considered chemotherapy line (rho > 0.65); the highest value was observed in the first line setting (rho = 0.76; 95% CI, 0.73–0.78) (Table 4).

Prognostic factors for time to next treatment

We evaluated the prognostic TNT value calculated from the first line systemic therapy of the main biological, histological, and clinical factors for the 1575 patients who received at least one systemic treatment (Table 5).

Regarding the multivariate analysis, the following factors remained associated with an increased TNT: female sex, locoregional treatment of metastases, and administration of polychemotherapy in the first line of metastatic treatment (Table 5, Fig. 1). Only a grade 3 tumor at diagnosis remained associated with a decreased TNT (Table 5, Fig. 1).

Prognostic factors for OS

We evaluated the prognostic OS values of the main biological, histological, and clinical factors for the 1575 patients who received at least one systemic treatment (Table 6).

Table 3 Median time to next treatment (TNT) and overall survival (OS) according to the histological subtype and treatment setting

	Median TNT/OS (months)			
	TNT1/OS1[a]	TNT2/OS2[b]	TNT3/OS3[c]	TNT4/OS4[d]
LMS	8.0/24.9	5.6/17.3	4.6/12.3	4.4/9.2
UPS	4.8/11.0	3.5/7.9	2.3/3.7	3.5/6.2
DLPS	4.4/11.8	5.1/8.8	2.4/6.0	3.2/8.5
SS	8.7/19.7	5.7/11.7	3.4/7.8	2.3/6.0
MPNST	4.1/12.5	2.8/7.0	3.6/8.0	3.7/5.4

[a]Calculated from the date of first-line treatment onset
[b]Calculated from the date of second-line treatment onset
[c]Calculated from the date of third-line treatment onset
[d]Calculated from the date of fourth-line treatment onset
DLPS dedifferentiated liposarcomas, *LMS* leiomyosarcomas, *MPNST* malignant peripheral nerve sheath sarcomas, *SS* synovial sarcomas, *UPS* undifferentiated pleomorphic sarcomas

Table 4 Correlation between time to next treatment (TNT) and overall survival (OS)

	Spearman's rho	95% CI
TNT1/OS1[a]	0.76	0.73–0.78
TNT2/OS2[b]	0.70	0.67–0.73
TNT3/OS3[c]	0.68	0.65–0.72
TNT4/OS4[d]	0.73	0.70–0.76

[a]Calculated from the date of first-line treatment onset
[b]Calculated from the date of second-line treatment onset
[c]Calculated from the date of third-line treatment onset
[d]Calculated from the date of fourth-line treatment onset

The following factors remained associated with an increased OS in the multivariate analysis: female sex, LMS, locoregional treatment of metastases, inclusion in a clinical trial, and administration of polychemotherapy in the first line of metastatic treatment (Table 6, Fig. 2). A grade 3 tumor at diagnosis remained associated with a decreased OS (Table 6, Fig. 2).

Parameters correlated with 5-year survival

To evaluate the parameters associated with a long survival, we excluded patients alive and with a follow-up inferior to 5 years, leading to the inclusion of 1619 patients in this analysis. A total of 224 patients were alive 5 years after the diagnosis of metastasis. The characteristics and patterns of this population are described in Tables 1 and 2, respectively.

The odds ratios and confidence intervals estimated by the logistic regression model for the factors significantly associated with the probability of 5-year survival are presented in Fig. 3. The factors associated with a higher probability of 5-year survival were locoregional treatment of metastases (OR = 7.41; 95% CI, 4.42–12.41) and inclusion in a clinical trial (OR = 1.59; 95% CI, 1.04–2.42). A grade 3 tumor at the time of diagnosis of metastasis was associated with a lower probability of 5-year survival (OR = 0.32; 95% CI, 0.21–0.48).

To observe the impact of the locoregional treatment modality on the probability of 5-year survival, we replaced the binary variable "locoregional treatment: yes/no" by a categorical variable detailing the type of locoregional treatment received (surgery, radiotherapy, radiofrequency, other, combination, or none). The following locoregional treatment modalities were particularly and significantly associated with a higher probability of 5-year survival: surgery (OR = 11.20; 95% CI, 6.19–20.26), radiofrequency (OR = 15.62; 95% CI, 5.04–48.41), and combination of modalities (OR = 9.60; 95% CI, 5.38–17.14). Other types of treatment, such as radiotherapy, were also correlated with a better probability of long survival; however, the effect was not significant.

Table 5 Prognostic factors for time to next treatment

	Univariate analysis		Multivariate analysis	
Covariate	*P*	HR (95% CI)	*P*	HR (95% CI)
Sex (ref: Male)	0.0014	0.835 (0.747–0.933)	0.0013	0.825 (0.733–0.928)
Age (ref: < 75 years old)	0.0023	1.374 (1.120–1.686)	–	–
Histotype (ref: Other)				
LMS	0.5114	0.955 (0.831–1.097)	–	–
DLPS	0.0068	1.357 (1.088–1.692)	–	–
MPNST	0.3703	1.154 (0.843–1.580)	–	–
SS	0.8580	0.983 (0.811–1.191)	–	–
UPS	0.0375	1.243 (1.013–1.525)	–	–
Grade (ref: < 3)	< 0.0001	1.417 (1.258–1.596)	< 0.0001	1.372 (1.218–1.546)
Number of metastatic sites (ref: 1)	0.1175	1.118 (0.972–1.285)	–	–
Liver metastasis (ref: no)	0.1436	1.103 (0.967–1.259)	–	–
Locoregional treatment (ref: no)	< 0.0001	0.496 (0.442–0.556)	< 0.0001	0.487 (0.432–0.550)
Clinical trial in first line (ref: no)	0.6453	1.048 (0.859–1.277)	–	–
Anthracycline in first line (ref: no)	< 0.0001	0.756 (0.674–0.847)	–	–
Polychemotherapy in first line (ref: no)	< 0.0001	0.729 (0.651–0.815)	< 0.0001	0.743 (0.660–0.836)

DLPS dedifferentiated liposarcomas, *LMS* leiomyosarcomas, *MPNST* malignant peripheral nerve sheath sarcomas, *SS* synovial sarcomas, *UPS* undifferentiated pleomorphic sarcomas

Discussion

The heterogeneity of STS has rarely been taken into account in the design of clinical trials to investigate systemic therapies in STS patients. Our results indicated that LMS clearly represented a distinct STS subgroup with a significantly better outcome in the advanced setting. Previous studies have shown worse outcomes for LMS than the results obtained in our current analysis. The largest study published to date was a retrospective analysis of 2185 patients with advanced STS treated in

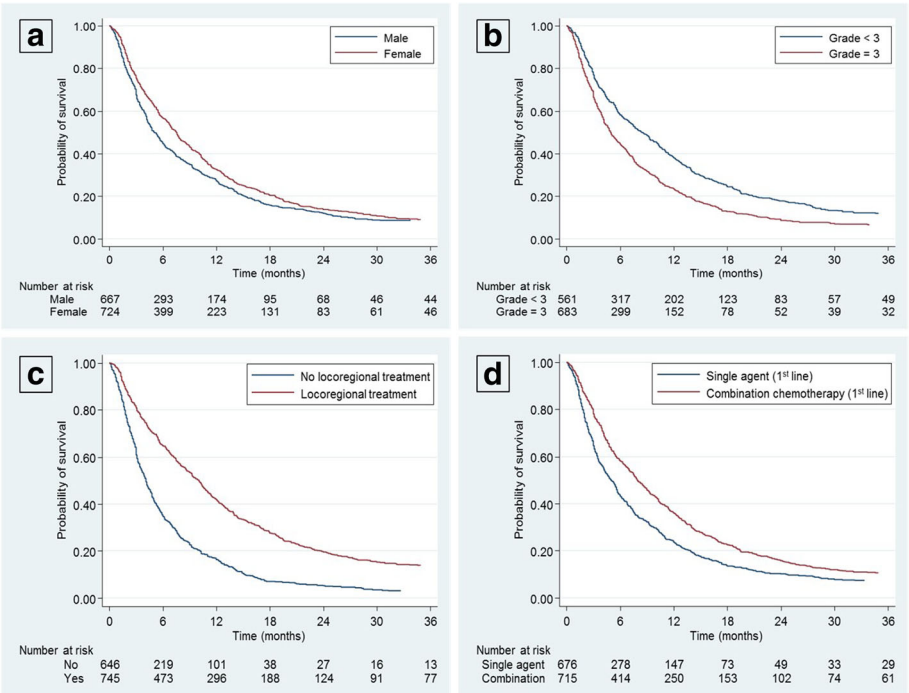


Fig. 1 Prognostic factors of time to next treatment – Kaplan–Meier curves. Kaplan-Meier Curves of time to next treatment according to (**a**) gender, (**b**) grade, (**c**) locoregional treatment of metastases, and (**d**) type of systemic treatment

Table 6 Prognostic factors for overall survival

	Univariate analysis		Multivariate analysis	
Covariate	*P*	HR (95% CI)	*P*	HR (95% CI)
Sex (ref: Male)	0.0002	0.801 (0.713–0.899)	0.0003	0.792 (0.698–0.900)
Age (ref: < 75 years old)	0.0024	1.389 (1.123–1.717)	–	–
Histotype (ref: Other)				
LMS	0.0004	0.765 (0.659–0.888)	0.0010	0.765 (0.652–0.897)
DLPS	0.0269	1.291 (1.030–1.619)	0.2034	1.171 (0.918–1.492)
MPNST	0.1368	1.273 (0.926–1.751)	0.2183	1.234 (0.883–1.726)
SS	0.4738	1.074 (0.883–1.307)	0.0764	1.206 (0.980–1.485)
UPS	0.0061	1.347 (1.089–1.668)	0.1839	1.168 (0.929–1.469)
Grade (ref: < 3)	< 0.0001	1.692 (1.491–1.920)	< 0.0001	1.687 (1.483–1.919)
Number of metastatic sites (ref: 1)	0.0136	1.200 (1.038–1.387)	0.0009	1.305 (1.115–1.528)
Liver metastasis (ref: no)	0.1056	0.891 (0.774–1.025)	–	–
Locoregional treatment (ref: no)	< 0.0001	0.412 (0.365–0.465)	< 0.0001	0.400 (0.351–0.455)
Clinical trial (ref: no)	< 0.0001	0.750 (0.653–0.862)	0.0002	0.755 (0.651–0.877)
Off-label drugs (ref: no)	< 0.0001	0.791 (0.703–0.890)	–	–
Anthracycline (ref: no)	0.0046	0.838 (0.741–0.947)	–	–
Anthracycline in first line (ref: no)	0.0127	0.861 (0.765–0.968)	–	–
Polychemotherapy in first line (ref: no)	0.0003	0.804 (0.715–0.902)	0.0023	0.822 (0.724–0.932)

DLPS dedifferentiated liposarcomas, *LMS* leiomyosarcomas, *MPNST* malignant peripheral nerve sheath sarcomas, *SS* synovial sarcomas, *UPS* undifferentiated pleomorphic sarcomas

the first-line studies of EORTC-STBSG; these patients showed no significant differences in terms of OS between LMS (492 cases) and the other histological subtypes, with a median OS of approximately 12 months [6]. However, this study, which focused only on first-line treatment, included patients diagnosed before the identification of the KIT mutation in gastrointestinal stromal tumors [4]. Therefore, a significant proportion of gastrointestinal stromal tumors, which are chemorefractory, were likely included in the LMS group. The better outcome of LMS may be explained by a specific biology but also by the potentially higher sensitivity to some anticancer agents such as gemcitabine, dacarbazine, or trabectedin. For instance, in a recent phase II randomized trial, patients with leiomyosarcomas of any origin benefited significantly from the combination of gemcitabine with dacarbazine, achieving a median progression-free survival (PFS) and OS of 4.9 and 13.8 months, respectively, versus 2.1 and 7.8 months, respectively, for the non-leiomyosarcoma subtypes [7]. Moreover, a large worldwide expanded access program for trabectedin showed a median OS of 16.2 months in 321 heavily pretreated leiomyosarcoma patients versus a median survival time of 11.9 months for the whole cohort of 903 patients [8].

We report here the first study assessing the outcomes of patients with advanced UPS. Some past reports included patients with malignant fibrous histiocytomas (MFHs). However, a significant subset of tumors initially diagnosed as MFH showed a specific line of differentiation (lipogenic, neurogenic, myogenic, or non-sarcomatous) [9–12]. "MFH" is now considered an obsolete terminology and has been replaced by the term UPS, which is a diagnosis of exclusion. We found that patients with advanced UPS had the worst outcome with the shortest TNT and a median OS of only 11 months. These results illustrate the particular resistance to chemotherapy of this histological subset and an intrinsically more aggressive biology. Further investigations are needed to better understand the mechanisms of their tumorigenesis and to define more appropriate therapeutic strategies.

Approximately 45% of the 1575 patients who underwent systemic therapy received a combination chemotherapy regimen in the first-line setting. The first-line chemotherapy for advanced, metastatic, or non-resectable STS is typically based on single-agent doxorubicin [13]. Indeed, the majority of clinical studies comparing single agents with combinations failed to show an OS advantage but consistently showed improvement in the response rates and PFS [14, 15]. Interestingly, our analysis showed a significant impact of the use of combination chemotherapy on OS, with a hazard ratio of 0.822 (0.724–0.932) and $P = 0.0003$. Judson et al. [14] recently published the results of a randomized clinical trial evaluating doxorubicin as a single agent in the control

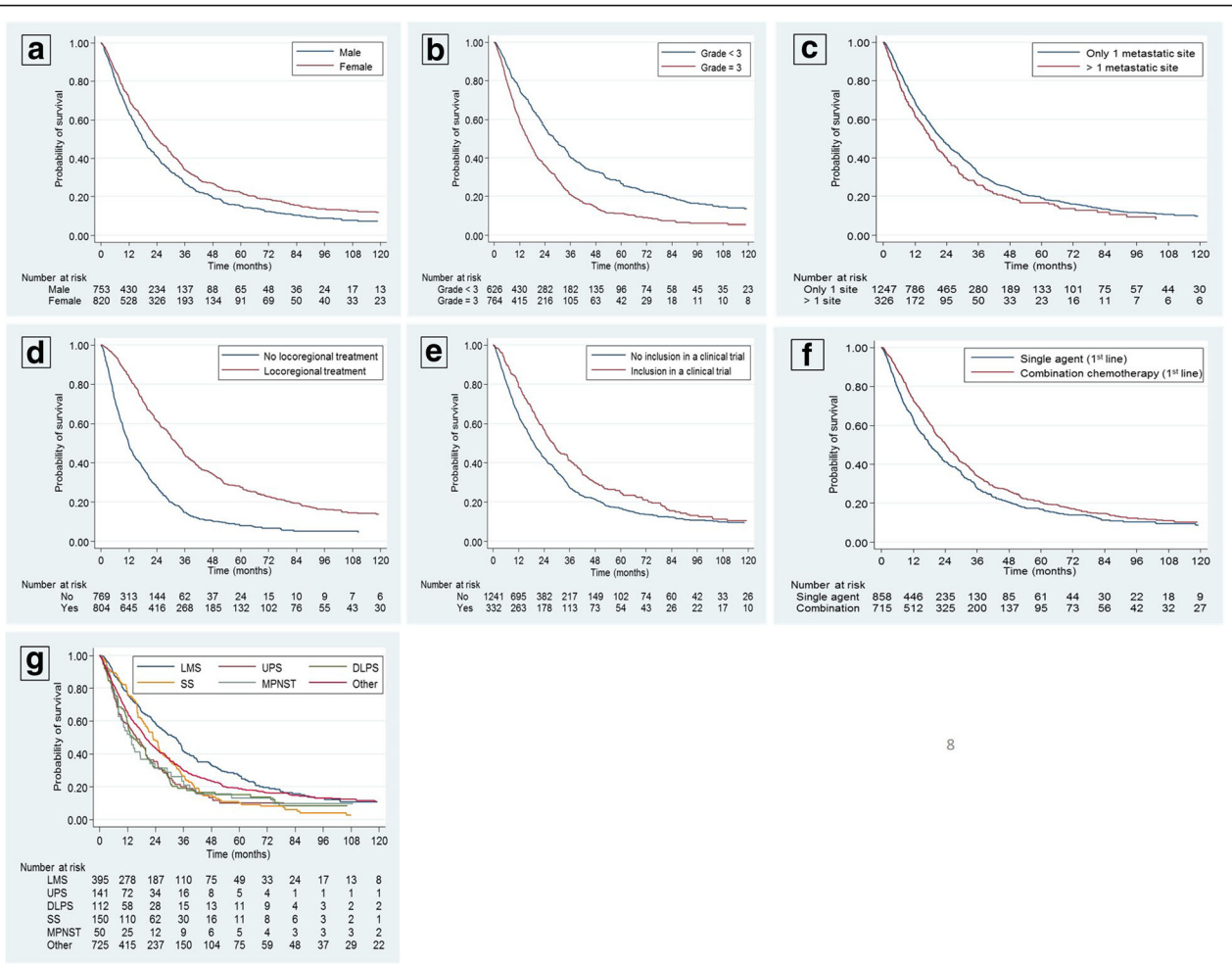


Fig. 2 Prognostic factors of overall survival – Kaplan–Meier curves. Kaplan-Meier curves of Overall survival according to (**a**) gender, (**b**) grade, (**c**) number of metastatic sites, (**d**) locoregional treatment of metastases, (**e**) inclusion in a clinical trial, (**f**) type of systemic treatment, (**g**) histological subtype

arm versus doxorubicin-ifosfamide in the experimental arm as a first-line treatment for advanced or metastatic STS. Although the Kaplan–Meier curves presented in the publication highlighted a difference between the two treatment arms in favor of polychemotherapy, the trial failed to detect a significant effect of polychemotherapy on OS, which was in contrast to our results. Our results suggest that the negative outcome of this study may simply be due to a lack of power as already suggested by Benjamin and Lee [16]. Indeed, by including 450 patients

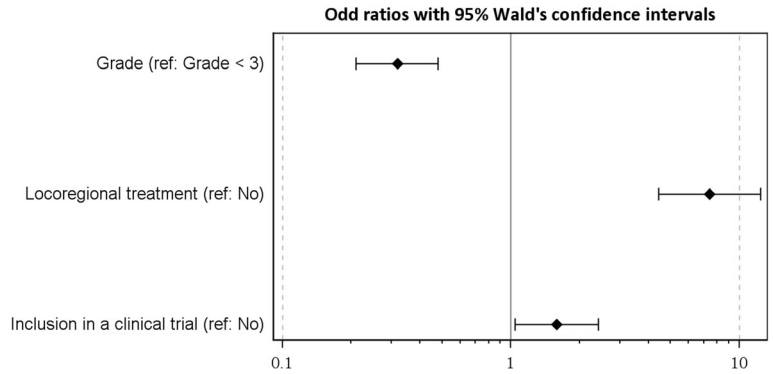


Fig. 3 Prognostic factors for 5-year survival – Odd ratios with 95% Wald's confidence intervals

and observing at least 366 events, the trial was designed to detect a maximum HR of 0.737. Due to the large size of our dataset, we were able to observe an HR of 0.822. Based on their hypotheses, a total of 827 events would be required to detect a similar treatment effect in a randomized clinical trial. Although our study suggests a benefit in terms of OS, clinicians should also be aware that randomized trials have clearly demonstrated that combination chemotherapy is more toxic than single-agent doxorubicin with a potential significant impact on the quality of life [14, 15]. Therefore, a combination of doxorubicin with a second drug such as ifosfamide should be used only after a careful discussion with the patient on the benefit/risk ratio of this approach, particularly when tumor shrinkage is expected to improve the symptoms or clinical benefits.

A high proportion of patients received more than two lines of systemic treatment. With the exception of leiomyosarcomas, our results indicate that the benefit of a greater than third-line regimen is very limited, with the median TNT and OS ranging between 2.3 and 3.7 months and 5.4 and 8.5 months, respectively. This result is consistent with the data from the PALETTE study, which led to the approval of pazopanib in advanced STS [17]. In that study, the number of previous lines of chemotherapy was a significant prognostic factor in the multivariate analysis for PFS with a significantly worse outcome in patients receiving pazopanib in the third- or fourth-line settings versus the first- or second-line settings. Given the potential toxicity and the moderate benefit of systemic therapy after failure of the second-line treatment, best supportive care should be considered as a reasonable option, particularly in patients with non-leiomyosarcoma histology and a poor performance status or patients who were not eligible to participate in a clinical trial. Notably, 50% of patients received an off-label drug during their treatment disease course. This result reflects the increasing evidence for the use of other drugs besides doxorubicin and ifosfamide in the sarcoma field. The most frequently prescribed off-label drug in this study was gemcitabine. Indeed, gemcitabine with or without docetaxel is commonly used in some specific sarcoma subsets, particularly in leiomyosarcomas and angiosarcomas [18–21], although neither of these drugs is approved for this indication. Another not yet approved drug that is frequently used in the sarcoma field is paclitaxel, which shows activity particularly in angiosarcomas [22, 23].

A significant proportion of patients with metastatic STS (27%) did not receive any systemic therapy. An age > 75 years was significantly associated with a lower probability of receiving any systemic treatment. Aging is associated with progressive functional declines, an increased prevalence of comorbidities, and a higher risk of cardiac and hematological toxicities related to anthracyclines [24–26]. These data may explain the reluctance of oncologists to use chemotherapy in elderly patients with STS and raises the question of the development of adapted chemotherapy regimens for elderly patients with advanced STS, such as low-dose cyclophosphamide [27] or liposomal doxorubicin [28].

A total of 49% of the patients received a loco-regional treatment of the metastasis, the most frequent of which were surgery followed by radiotherapy and radiofrequency ablation. The majority of these patients (71%) had lung metastases. The published evidence on the role of locoregional treatments, such as pulmonary metastasectomy, is derived from a small number of studies with limited sample sizes [29]. Primary bone sarcomas, which may represent a distinct disease, are often included in these analyses. Our present study differed from previous publications because we used a larger database cohort, which increased the power of the multivariate analysis; additionally, we focused on STS exclusively to enhance the homogeneity of the study population. As suggested by previous studies, patients who underwent a locoregional metastasis treatment had improved survival in the multivariate analysis. Arguments have suggested that an observational study may not provide evidence that a difference in survival is attributable to the locoregional treatment and that only a randomized trial can answer the question. However, we observed that more than 80% of metastatic patients alive 5 years after the diagnosis of metastasis had received a locoregional treatment, versus 50% in the general population, and this parameter was most significantly associated with the probability of being alive at 5 years in the logistic regression model. Precisely, the descriptive analyses of the patients alive after 5 years suggest that surgery, radiofrequency, and a combination of different modalities are particularly beneficial in terms of survival. This hypothesis was confirmed by our sensibility analysis, since we found that the positive effect on the probability of 5-year survival was significant for these three treatment modalities only.

No data are available from randomized clinical trials to define how best to integrate the locoregional treatment of metastases in the management of patients with advanced disease. The most recent attempts were made by the European Organisation for Research and Treatment of Cancer (EORTC-Protocol 62933) with a randomized multicenter trial to assess metastasectomy alone versus induction chemotherapy followed by metastasectomy in a targeted sample size of 340 patients. Started in 1996, this trial was closed due to poor accrual in November of 2000. Notably, we report here the first large series of patients who received non-surgical locoregional treatment of metastases, including 254 patients treated with radiotherapy, 42 with radiofrequency ablation, and 320 with a

combination of surgery plus radiotherapy or surgery plus radiofrequency ablation of metastases.

The gold standard endpoint in randomized clinical trials in oncology is OS. However, the use of a surrogate endpoint at an earlier stage in clinical trials would speed up the assessment of treatments and might reduce the cost of drug development. Studies that assess the use of alternative outcome measures, such as the response rate or PFS, as surrogate endpoints for OS in sarcoma patients showed only a modest if any correlation with PFS and OS [30, 31]. This issue was recently illustrated with the pivotal trial that led to eribulin approval in patients with liposarcomas that showed a benefit in OS but not in PFS [32]. TNT is an established endpoint that is mostly applied in hematological malignancies and has recently been used in breast, colon, and prostate cancer [33–35]. The use of this parameter is predicated on the concept that a change in treatment usually occurs in response to a real change in the patient status by integrating the efficacy and toxicity components. In our study, we found a strong correlation between TNT and OS. The prospective validation of this endpoint as a surrogate for OS should be done in future studies.

Conclusions

This study reports the most comprehensive information related to the patterns of care and outcome of STS with advanced disease managed in the real-life setting. Limitations include its observational nature, which provides a lower level of evidence than a conventional clinical trial, the lack of data related to visceral sarcomas and GIST, and to the safety of therapeutic interventions. However, there are several lines of evidence indicating that observational studies usually do provide valid information and could be used to exploit well-designed databases [36].

Additional file

Additional file 1: Supplementary Methods. (DOCX 11 kb)

Abbreviations

DLPS: Dedifferentiated liposarcoma; LMS: Leiomyosarcoma; MFHs: Malignant fibrous histiocytomas; MPNST: Malignant peripheral nerve sheath tumors; OS: Overall survival; PFS: Progression-free survival; SS: Synovial sarcoma; STSs: Soft-tissue sarcomas; TNT: Time to next treatment; UPS: Undifferentiated pleomorphic sarcoma

Acknowledgments

The authors are grateful to Jean-Baptiste Courreges, Myriam Jean-Denis, and Nouria Mesli for their contribution to the management of the French Sarcoma Group database.

Funding

The present study has been funded by the French National Cancer Institute.

Availability of data and materials

The datasets supporting the conclusions of this article cannot be shared for confidentiality reasons. The METASARC database contains the most comprehensive information related to the outcome of STS with advanced disease and will be continuously updated to help clinicians identify the best therapeutic options for their patients. Queries related to factors such as the activity of a drug in a specific histological subtype can be sent by email (metasarc@bordeaux.unicancer.fr). Similarly, our results will also be available for investigators who need a reference for the response and outcome in the assessment of an investigational drug in a specific setting.

Authors' contributions

Study concepts and design: AI. Acquisition, analysis or interpretation of data: MS, ALC, JYB, IRC, OM, MT, SC, PT, DRV, PM, ES, CH, PS, MPS, CLP, AG, CB, FLL, AI. Drafting of the manuscript: MS, ALC, JYB, IRC, OM, MT, SC, PT, DRV, PM, ES, CH, PS, MPS, CLP, AG, CB, FLL, AI. Critical revision of the manuscript for important intellectual content: MS, ALC, JYB, IRC, OM, MT, SC, PT, DRV, PM, ES, CH, PS, MPS, CLP, AG, CB, FLL, AI. All authors have given final approval of the version to be published. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

This study was approved by the ethics committee of the Comprehensive Cancer Center Institut Bergonié (Bordeaux, France).

Publisher's Note

Author details

[1]Clinical and Epidemiological Research Unit, Institut Bergonié, Bordeaux, France. [2]ISPED, INSERM U1219 Bordeaux Population Health Center, Epicene Team, Bordeaux, France. [3]Department of Medicine, Institut Gustave Roussy, Villejui, France. [4]Department of Medicine, Centre Leon Berard, Lyon, France. [5]Department of Medicine, Institut Bergonié, Bordeaux, France. [6]Department of Pathology, Institut Gustave Roussy, Villejuif, France. [7]Department of Pathology, Centre Leon Berard, Lyon, France. [8]Department of Surgery, Centre Leon Berard, Lyon, France. [9]Department of Surgery, Institut Bergonié, Bordeaux, France. [10]Department of Surgery, Institut Gustave Roussy, Villejuif, France. [11]Department of Radiotherapy, Institut Bergonié, Bordeaux, France. [12]Department of Radiotherapy, Centre Leon Berard, Lyon, France. [13]Department of Radiotherapy, Institut Gustave Roussy, Villejuif, France. [14]Department of Pathology, Institut Bergonié, Bordeaux, France. [15]Early Phase Trials and Sarcoma Units, Institut Bergonié, 229 Cours de l'Argonne, Bordeaux, France.

Received: 3 December 2016 Accepted: 3 March 2017
Published online: 10 April 2017

References

1. Coindre JM, Terrier P, Guillou L, et al. Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: a study of 1240 patients from the French Federation of Cancer Centers Sarcoma Group. Cancer. 2001;91:1914–26.
2. Fletcher C, Unni K, Mertens F, editors. World Health Organization Classification of Tumours Pathology and Genetics of Tumours of Soft Tissue and Bone. Lyon: IARC Press; 2013.
3. Trojani M, Contesso G, Coindre JM, et al. Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. Int J Cancer. 1984;33:37–42.
4. Hirota S, Isozaki K, Moriyama Y, et al. Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. Science. 1998;279:577–80.
5. Schemper M, Kaider A, Wakounig S, Heinze G. Estimating the correlation of bivariate failure times under censoring. Statist Med. 2013;32:4781–90.

6. Van Glabbeke M, van Oosterom AT, Oosterhuis JW, et al. Prognostic factors for the outcome of chemotherapy in advanced soft tissue sarcoma: an analysis of 2,185 patients treated with anthracycline-containing first-line regimens–a European Organization for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group Study. J Clin Oncol. 1999;17:150–7.
7. García-Del-Muro X, López-Pousa A, Maurel J, et al. Randomized phase II study comparing gemcitabine plus dacarbazine versus dacarbazine alone in patients with previously treated soft tissue sarcoma: a Spanish Group for Research on Sarcomas study. J Clin Oncol. 2011;29:2528–33.
8. Samuels BL, Chawla S, Patel S, et al. Clinical outcomes and safety with trabectedin therapy in patients with advanced soft tissue sarcomas following failure of prior chemotherapy: results of a worldwide expanded access program study. Ann Oncol. 2013;24:1703–9.
9. Coindre JM, Mariani O, Chibon F, et al. Most malignant fibrous histiocytomas developed in the retroperitoneum are dedifferentiated liposarcomas: a review of 25 cases initially diagnosed as malignant fibrous histiocytoma. Mod Pathol. 2003;16:256–62.
10. Fletcher CD. Pleomorphic malignant fibrous histiocytoma: fact or fiction? A critical reappraisal based on 159 tumors diagnosed as pleomorphic sarcoma. Am J Surg Pathol. 1992;16:213–28.
11. Fletcher CD, Gustafson P, Rydholm A, Willén H, Akerman M. Clinicopathologic re-evaluation of 100 malignant fibrous histiocytomas: prognostic relevance of subclassification. J Clin Oncol. 2001;19:3045–50.
12. Oda Y, Tamiya S, Oshiro Y, et al. Reassessment and clinicopathological prognostic factors of malignant fibrous histiocytoma of soft parts. Pathol Int. 2002;52:595–606.
13. ESMO/European Sarcoma Network Working Group. Soft tissue and visceral sarcomas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2014;25 Suppl 3:iii102–12.
14. Judson I, Verweij J, Gelderblom H, et al. Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: a randomised controlled phase 3 trial. Lancet Oncol. 2014;15:415–23.
15. Santoro A, Tursz T, Mouridsen H, et al. Doxorubicin versus CYVADIC versus doxorubicin plus ifosfamide in first-line treatment of advanced soft tissue sarcomas: a randomized study of the European Organization for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group. J Clin Oncol. 1995;13:1537–45.
16. Benjamin RS, Lee JJ. One step forward, two steps back. Lancet Oncol. 2014; 15:366–7.
17. van der Graaf WT, Blay JY, Chawla SP, et al. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): a randomised, double-blind, placebo-controlled phase 3 trial. Lancet. 2012;379:1879–86.
18. Ducoulombier A, Cousin S, Kotecki N, Penel N. Gemcitabine-based chemotherapy in sarcomas: a systematic review of published trials. Crit Rev Oncol Hematol. 2016;98:73–80.
19. Stacchiotti S, Palassini E, Sanfilippo R, et al. Gemcitabine in advanced angiosarcoma: a retrospective case series analysis from the Italian Rare Cancer Network. Ann Oncol. 2012;23:501–8.
20. von Mehren M, Randall RL, Benjamin RS, et al. Soft Tissue Sarcoma, Version 2.2016, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2016;14:758–86.
21. Duffaud F, Pautier P, Bui B, et al. A pooled analysis of the final results of the two randomized phase II studies comparing gemcitabine (G) vs. gemcitabine + docetaxel (G + D) in patients (pts) with metastatic/relapsed leiomyosarcoma (LMS). Ann Oncol. 2010;21 Suppl 8:viii408–16. doi:10.1093/annonc/mdq536.
22. Italiano A, Cioffi A, Penel N, et al. Comparison of doxorubicin and weekly paclitaxel efficacy in metastatic angiosarcomas. Cancer. 2012;118:3330–6.
23. Penel N, Bui BN, Bay JO, et al. Phase II trial of weekly paclitaxel for unresectable angiosarcoma: the ANGIOTAX Study. J Clin Oncol. 2008;26: 5269–74.
24. Swain SM, Whaley FS, Ewer MS. Congestive heart failure in patients treated with doxorubicin: a retrospective analysis of three trials. Cancer. 2003;97: 2869–79.
25. Fumoleau P, Roché H, Kerbrat P, et al. Long-term cardiac toxicity after adjuvant epirubicin-based chemotherapy in early breast cancer: French Adjuvant Study Group results. Ann Oncol. 2006;17:85–92.
26. Shayne M, Culakova E, Poniewierski MS, et al. Dose intensity and hematologic toxicity in older cancer patients receiving systemic chemotherapy. Cancer. 2007;110:1611–20.
27. Mir O, Domont J, Cioffi A, et al. Feasibility of metronomic oral cyclophosphamide plus prednisolone in elderly patients with inoperable or metastatic soft tissue sarcoma. Eur J Cancer. 2011;47:515–9.
28. Judson I, Radford JA, Harris M, et al. Randomised phase II trial of pegylated liposomal doxorubicin (DOXIL/CAELYX) versus doxorubicin in the treatment of advanced or metastatic soft tissue sarcoma: a study by the EORTC Soft Tissue and Bone Sarcoma Group. Eur J Cancer. 2001;37:870–7.
29. Treasure T, Fiorentino F, Scarci M, Møller H, Utley M. Pulmonary metastasectomy for sarcoma: a systematic review of reported outcomes in the context of Thames Cancer Registry data. BMJ Open. 2012;2: e001736.
30. Savina M, Litière S, Penel N, et al. Surrogate properties of survival endpoints in metastatic soft-tissue sarcoma: a meta-analysis. J Clin Oncol. 2015; 33(Suppl):abstr: 10547.
31. Zer A, Prince RM, Amir E, Abdul RA. Evolution of randomized trials in advanced/metastatic soft tissue sarcoma: end point selection, surrogacy, and quality of reporting. J Clin Oncol. 2016;34:1469–75.
32. Schöffski P, Chawla S, Maki RG, et al. Eribulin versus dacarbazine in previously treated patients with advanced liposarcoma or leiomyosarcoma: a randomised, open-label, multicentre, phase 3 trial. Lancet. 2016;387:1629–37.
33. Chudley L, McCann K, Mander A, et al. DNA fusion-gene vaccination in patients with prostate cancer induces high-frequency CD8(+) T-cell responses and increases PSA doubling time. Cancer Immunol Immunother. 2012;61:2161–70.
34. Liang C, Li L, Fraser CD, et al. The treatment patterns, efficacy, and safety of nab (®)-paclitaxel for the treatment of metastatic breast cancer in the United States: results from health insurance claims analysis. BMC Cancer. 2015;15:1019.
35. Teng CL, Wang CY, Chen YH, Lin CH, Hwang WL. Optimal sequence of irinotecan and oxaliplatin-based regimens in metastatic colorectal cancer: a population-based observational study. PLoS One. 2015;10:e0135673.
36. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000;342:1878–86.

5 Surrogate endpoints in adjuvant breast cancer trials

5.1 Introduction

Breast cancer (BC) is the most common cancer and the fifth more frequent cause of death from cancer in women worldwide. Despite its increasing incidence, we observed a significant decline in BC mortality in 30 years and the lengthening of survival. This improvement in survival can be explained by different factors, such as the increase of cancer screening, the enhancement of diagnosis methods or the development of adjuvant therapies. Conducting RCTs to assess BC treatments, especially in the adjuvant setting, thus implies including a large number of patients and a long follow-up to observe an OS benefit for a new drug. Therefore, the validation of surrogate endpoints for OS in the context of BC, and particularly in the adjuvant setting, is a key issue in clinical research. As highlighted in the review presented in chapter 2, rigorous MAs using IPD to assess surrogate endpoints are lacking in this setting.

We conducted a pooled analysis on IPD from five RCTs evaluating adjuvant chemotherapy for patient with BC. We assessed the surrogate properties of four time-to-event endpoints: relapse-free survival (RFS), invasive disease-free survival (iDFS), locoregional relapse-free survival (LRFS) and distant disease-free survival (DDFS). Each endpoint, candidate surrogates and OS, were recalculated to ensure identical definition and follow-up across trials. The individual- and trial-level associations were estimated following the two-stage model (chapter 1.3.2.3) and the meta-regression model with weighted fixed treatment effects (chapter 1.3.2.3). Based on the regression equation from a weighted fixed treatment effects model, we also estimated the STE. As commonly performed when the number of available trials is limited, we used the including centers instead of the trials as the analysis unit for the estimation of the trial-level associations.

This analysis showed good correlations between the candidate surrogates and OS, both at the individual- and the trial-level. Based on these results, the trial-level association between LRFS and OS was ranked high as per the IQWiG criteria. For the three other endpoints, the trial-level associations were ranked as medium. One should however acknowledge that the use of including centers instead of the trials may have artificially narrowed the reported confidence intervals. Further investigation based on a larger dataset of RCTs would be required to confirm these preliminary findings.

The manuscript is currently under review with the co-authors.

5.2 Publication

Surrogate Endpoints for Overall Survival in Adjuvant Breast Cancer Trials: a Pooled Analysis of 5 Randomized Controlled Trials

M.Savina, C. Bellera, W. Jacot, T. Burzykowski, F. Bonnetain, P. Kerbrat, H. Roché, M. Spielmann, A. Goldhirsch, R. Paridaens, S. Mathoulin-Pélisser, S. Gourgou.

Abstract

Introduction: Alternative endpoints to overall survival (OS) such as diease-free survival (DFS) are increasingly used to assess treatment efficacy in randomized controlled trials (RCT) to reduce the number of patients included and the trial duration. Their properties, in terms of surrogate markers, need to be assessed to ensure that they are adequate replacements for OS. We evaluated the surrogate properties of four time-to-event in adjuvant breast cancer.

Methods: We relied on a meta-analytical framework using individual-patient data to estimate individual-level association (association between the endpoints) and trial-level association (association between the treatment effects). Statistical methods included weighted linear regression (WLR) and the two-step model (2SM) from Burzykowski et al. The strength of the trial-level association was ranked according to the IQWiG guidelines. The prediction capacity was assessed using internal validation following a leave-one-out approach.

Results: Individual data from 5 RCTs (N=11676) were analyzed. We evaluated 5-year relapse-free survival (RFS), invasive DFS (iDFS), locoregional RFS (LRFS) and distant DFS (DDFS) as surrogate endpoints for 7-year OS. All four endpoints were highly associated with OS at the individual level. The trial-level association between LRFS and OS was ranked as high as per the IQWiG criteria and the model fitted showed good prediction properties. The three other candidate surrogates showed high trial-level association with OS however the estimations lacked precision.

Conclusion: The four endpoints were highly associated with seven-year OS at the individual level. At the trial level, only LRFS was highly associated with OS as per the IQWiG guidelines. These results suggest that LRFS is an interesting candidate surrogate for OS. Further evaluation on a larger set of trials is required to confirm these results and improve the precision of our estimations.

Key words: surrogate endpoint, breast cancer, overall survival, meta-analysis, randomized controlled trial

Introduction

When designing a randomized controlled trial (RCT), the choice of the primary efficacy endpoint is a key step. This endpoint should be measurable, sensitive to the treatment effect and clinically relevant. In oncology, the most commonly used endpoint to assess the efficacy of a new treatment is overall survival (OS), easily measurable, objectively defined as the time from randomization to death, and validated by health regulatory authorities. With a five-year OS rate close to 90% in 2012, the use of OS as primary endpoint in breast cancer trials implies delays in the evaluation of potentially usefull therapies, specifically in the adjuvant setting. Alternative endpoints commonly used in phase II trials, such as relapse-free survival (RFS) or progression-free survival (PFS), are increasingly used as primary endpoints in phase III RCT (1). These composite endpoints include not only death but also biological and clinical events, such as disease progression or toxicity, and can help reducing the number of patients, the duration and ultimately the cost of the trials.

The use of alternative endpoints in practice does not guarantee their validity as surrogates for OS. In the metastatic setting for instance, trials have shown a positive impact of everolimus and exemestane on progression-free survival (PFS) compared to exemestane alone in patients with hormone-receptor-positive advanced breast cancer (2). However, no significant treatment effect was subsequently observed on long-term OS (3), which might be explained either by a lack of power or due to the absence of a validated surrogate endpoint in this setting. It is essential to properly and rigorously assess the surrogate properties for OS of such alternative endpoints, to identify adequate primary endpoints for the benefit assessment of new therapies.

The International Conference on Harmonization (ICH) E9 Harmonized Tripartite guidelines - approved by the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) - does not provide any recommendations on the use of specific statistical methods for the validation of surrogate endpoints. However the meta-analytic surrogacy evaluation schema proposed by Buyse et al. and Burzykowski et al. (4,5) has been widely used and is recognized as the most statistically rigorous (6,7). This approach requires individual-patient data (IPD) from multiple RCTs with similar design and treatment to address surrogacy from a multi-level framework. At the patient level, the surrogate endpoint is supposed to be correlated and predictive of the final endpoint regardless of the treatment (individual-level association). At the trial level, the treatment effect (summarized by the hazard ratio [HR]) on the surrogate endpoint should be correlated and predictive of the treatment effect on the final endpoint (trial-level association).

In the absence of validated surrogate endpoints for OS in adjuvant beast cancer trials [CROH], we performed a pooled analysis of five RCTs to assess the surrogate properties of four time-to-event endpoints commonly used in this setting: RFS, invasive disease-free survival (iDFS), distant disease-free survival (DDFS) and locoregional relapse-free survival

(LRFS). To the best of our knowledge, no such analysis based on IPD has been conducted in the context of adjuvant breast cancer trials.

Methods

This study is registered on the clinical trial registry clinicaltrials.gov (identifier: NCT02873923).

Study selection

Trials were selected through contacts with acamedic groups (European Organisation for Research and Treatment [EORTC], the National Federation of French Cancer Center [UNICANCER]), and pharmaceutical groups. Trials were eligible if they suited the following criteria: (i) phase III trials, (ii) evaluating the efficacy of adjuvant chemotherapy and/or targeted therapy for adults with early breast cancer, (iii) with at least one time-to-event endpoint other than OS as outcome, (iv) published or presented in congress, (v) agreement from the principal investigator and the sponsor, and (vi) available IPD.

Data and Outcomes

We collected individual biological and histological data at baseline, date of randomization, data related to treatment allocation, disease evaluations during trial, toxicity, date of last follow-up or death, survival status, cause of death (if applicable), along with randomization variables. OS was defined as the time from randomization to death, whatever the cause. RFS, iDFS, DDFS and LRFS were defined according to the international DATECAN breast cancer guidelines (8,9). Events included as failures in the definition of each endpoint are listed in table 1. We assessed the surrogate properties for seven-year OS of the four endpoints evaluated after five years of follow-up.

Surrogacy measures

We assessed the surrogate properties for OS of the candidate surrogates by investigating their association with OS (individual-level surrogacy) as well as the association between the treatment effect (HR) on the candidate surrogate and the treatment effect on OS (trial-level surrogacy). We assessed the individual-level surrogacy following a copula-based approach (5). We jointly modelled the candidate surrogate and OS using a one-parameter Clayton copula function. We estimated the individual-level associations by the Spearmann rank correlation coefficient ($\rho_{Spearmann}$) calculated from the copula parameter.

We next estimated the trial-level surrogacy. In meta-analyses assessing surrogate endpoints, the number of trials might not be sufficient to properly estimate the trial-level association. In such case, each large trial might be subdivided into smaller groups of patients, based for instance on the including centers or country (4,10). We thus subdivided each of the five trials into smaller pseudo-trials based on including centers, that we will call trial-unit. Each trial-unit was designed to involve at least 400 patients (200 in each treatment arm) and at least four deaths (two in each treatment arm). Based on this subdivision, we followed two frameworks

to assess the trial-level surrogacy. First, we estimated the treatment effects on the candidate surrogate and OS for each trial-unit based on the logarithm of the HR (log[HR]) following Cox proportional hazards models. We modelled the association between the treatment effects on the two endpoints using a linear regression model weighted by the trial-unit size (WLR). To acount for the dependancy between trial-units from the same trial, we included a random effect related to the trial. We then estimated the trial-level association by the proportion of variation explained by the treatment effect on the candidate surrogate (R^2_{WLR}). The second method follows the two-step model (2SM) adapted for time-to-event endpoints introduced by Burzykowski et al. (5). First, we simultaneously estimated the treatment effects on the candidate surrogate and on OS using a bivariate survival model based on a one-parameter Clayton copula function. This approach enables taking into account the correlation between OS and the candidate surrogate in the estimations. To acount for the dependancy between trial-units from the same trial, we assumed different treatment effects for each trial-unit with constant baseline hazard within trials. To assess the association between the treatment effects (log(HR)), we estimated the coefficient of determination of an error-in-variable model (R^2_{2SM}), or of a classical linear mixed-effect regression model (unadjusted R^2_{2SM}) in the absence of convergence of the error-in-variable model. Finally, we estimated the surrogate threshold effect (STE) based on the WLR model. The STE corresponds to the minimum treatment effect on the surrogate endpoint necessary to predict a benefit on OS, i.e. to predict a HR for OS with a prediction interval strictly inferior to one (11).

We conducted all analyses on an intention-to-treat basis. Confidence intervals (CI) were calculated for a two-sided probability coverage of 95%. All analyses were performed using SAS software v9.3.

Strength of association

We ranked the strength of the trial-level association according to the Institute for Quality and Efficiency in Health Care (IQWiG) guidelines (12): high association (lower limit of the 95% confidence interval for $R^2 \geq 0.72$), low association (higher limit of the 95% confidence interval for $R^2 \leq 0.49$) or medium association (neither low nor high), meaning that the validity of the surrogate remains unclear.

Prediction capacity

To test prediction capacity of the endpoints, we performed internal validation following a leave-one-out cross-validation approach (13). We repeatedly fit the regression model using data from all trial-units except one, then used the model to predict the treatment effect on OS based on the treatment effects for the trial-unit that was excluded. We compared the observed treatment effect on OS (HR_{OS}) to the 95% prediction interval computed for each excluded trial-unit. We measured the accuracy of the prediction by estimating the mean squared error prediction (MSEP) defined as the squared mean of the difference between the

predicted treatment effects on OS based on RFS, iDFS, LRFS and DDFS, noted $HR_{OS/RFS}$, $HR_{OS/RFS}$, $HR_{OS/RFS}$, $HR_{OS/RFS}$ respectively, and the observed tratment effect HR_{OS}.

Sensitivity analyses

We conducted two sensitivity analyses. Firstly, we conducted sensitivity analyses to assess the impact of different follow-up durations on the individual- and the trial-level associations. We investigated the properties of the RFS, iDFS, LRFS and DDFS evaluated after three years of follow-up as surrogates for 5-year OS and 7-years OS. Secondly, to control for trials' heterogeneity in terms of patient selection, we performed a subgroup analysis excluding one trial that focused only on HER2-positive patients.

Results

Data

We retained five RCTs that included 11676 patients (14–18) (table 2). Each trial compared two treatment arms. Follow-up duration ranged from 53.4 to 95.8 months (median 74 months). A total of 1560 patients died during follow-up (13.4%). The number of observed events for RFS, iDFS, LRFS, DDFS and OS at various time points is presented in table 3. Forest plots presenting the treatment effects on seven-year OS and five-year RFS, iDFS, LRFS and DDFS estimated by Cox proportional hazards models for each trial-unit are presented in Figure 1.

Correlation between surrogate endpoints and OS (individual-level surrogacy)

Measures of the individual-level associations as estimated using the 2SM method are presented for seven-year OS and the candidate surrogates evaluated at five years (Table 4). Individual-level associations were high ($\rho_{Spearman} \geq 0.98$) for all candidate surrogates.

Correlation between treatment effects on the surrogate endpoints and treatment effect on OS (trial-level surrogacy)

Measures of the trial-level association as estimated using the WLR and the 2SM methods are presented for seven-year OS and the alternative endpoints evaluated after five years of follow-up (Table 4). Associations between treatment effect on the various endpoints and on OS are presented in figure 2 (WLR method). Based on the WLR method, point estimate for correlation coefficient ranged from 0.90 (iDFS) to 0.96 (RFS). Based on the 2SM method, point estimate for correlation coefficient ranged from 0.48 (iDFS) to 0.89 (LRFS). As per the IQWiG recommendations, five-year LRFS was considered highly associated with seven-year OS at the trial level (WLR or 2SM).

Surrogate threshold effect (STE)

The STE was estimated for each endpoint based on the weighted linear model on independently estimated log(HR) results (table 4). The highest STE was oserved for LRFS with a value of 0.79, that is, one should observe a treatment effect on LRFS inferior to 0.79

(HR_{LRFS}<0.79), in order to predict a treatment effect on OS inferior to 1 (HR_{OS}<1) with 95% probability.

Leave-one out cross-validation

The results of the leave-one-out cross-validation for the fitted weighted regression models for within-trial HR are presented in table 5. For RFS, the observed HR_{OS} fell within the 95% prediction interval in 15 of the 21 trial-units (71%). Similar results were observed for iDFS (65%), LRFS (67%) and DDFS (71%). The predicted HR_{OS} fell on the same side of the equivalence line (HR_{OS} = 1) as the observed HR_{OS} in 86%, 88%, 86% and 100% of the cases for respectively RFS, iDFS, LRFS and DDFS. The MSEP was inferior to 0.001 for the four prediction models.

Sensitivity analyses

With reduced follow-up duration, the individual-level associations were not impacted by the reduction of the follow-up duration while trial-level decreased (additional file 1). After exclusion of the HERA trial which included only HER2-positive patients, individual-level and trial-level associations were similar to that reported for the primary analysis (additional file 2).

Discussion

We pooled IPD from 11676 patients included in five RCTs to evaluate the surrogate properties of five-year RFS, iDFS, LRFS and DDFS for seven-year OS in ajuvant breast cancer. The four endpoints showed high individual-level associations with OS. At the trial level, LRFS was highly associated with OS as per the IQWiG criteria. The other three endpoints showed high trial-level association with OS but the estimations lacked precision, as illustrated by the large 95%CI. The fitted regression models provided good predictions of the treatment effect on OS based on the observed treatment effect on LRFS. Additionally, the STE estimated was high, which means that LRFS would be practically usefull.

Several statistical methods are available to assess surrogacy. We relied on the two-stage approach developed by Buyse and Burzykowski based on IPD (4), considered the most rigorous statistical approach for surrogacy assessment (5,6). Similarly, several criteria have been proposed to assess the validity of surrogate endpoints (11,26,27). Although they present differences, the IQWiG criteria is the most conservative. As such, they all corroborate the high level of evidence of surrogacy.

While several meta-analyses have been conducted in the metastatic setting (19), to the best of our knowledge this the first meta-analysis conducted on IPD to evaluate surrogate endpoints in adjuvant breast cancer. Previously, only one meta-analysis based on aggregated data evaluated the surrogate properties of various endpoints for adjuvant breast cancer trials (19). This study, however, did not lead to the validation of a surrogate endpoint. In the neo-adjuvant setting, pathological complete response (pCR) was also investigated (18). The analysis, conducted on IPD, did not lead to the validation of pCR as a surrogate

marker for OS; no additional time-to-event endpoint was investigated.

Some limitations to our study are worth noticing. Firstly, we relied on a convenient sample of RCTs for which IPD were available, rather than on an exhaustive set of all available adjuvant breast cancer trials. Since data that is easily located and included in meta-analysis can have different correlations than unavailable or unreported data, attempt to validate surrogate endpoints can be biased. However, to date, no example of a surrogate validation study based on all relevant evidence exists (28). Secondly, as commonly performed when the number of trials available is limited, we subdivided each of the five trials into groups of including centers for the assessment of the trial-level association. This approach definitely affects estimations, and specifically the reported 95%CI of the trial-level associations which are likely to be artificially narrow. Finally, some heterogeneity in terms of population remains between the trials, although the sensitivity analysis conducted after exclusion of the trial conducted on HER2-positive patients only did not reveal different results from the primary analysis (additional file 2).

The Accelerated Approval regulations, instituted by the FDA in 1992, allows drugs for serious conditions that fill an unmet medical need to be approved based on an end-point “likely to predict” the clinical outcome. As a result, an increasing number of anticancer drug product approvals by the FDA are made based on endpoints other than OS (1,33,34), some with no sufficient proof of their surrogate validity for OS (34). In the context of breast cancer, the FDA granted two accelerated approvals since 2009 based on response rate and PFS in the context of neoadjuvant (Pertuzumab) and metastatic (Lapatinib) breast cancer respectively. Two traditional approvals granted by the FDA (Everolimus and Pertuzumab) in the metastatic setting were based on PFS, even though there is no proof of its validity as surrogate for OS. The use of invalid surrogate endpoints can however lead to the marketing of drugs without a significant clinical benefit. This issue is well illustrated with the example of bevacizumab in metastatic breast cancer which was initially granted accelerated approval based on PFS data, but subsequently withdrawn following publication of OS results (20,21).

Several conditions have to be met to ensure adequate validation of a surrogate endpoint: (i) a significant quantity of data, both in terms of trials and patients, (ii) homogeneity, in terms of disease, settings, and mechanisms of action of the drugs, and (iii) strong statistical thresholds. Although our study was limited by the use of a convenient set of trials and a small number of trials, we observed high associations with OS for the four candidate endpoints investigated at both the individual and the trial level, and LRFS qualified as highly associated with OS as per the IQWiG criteria. The estimated trial-level associations between OS and RFS, iDFS ans DDFS were ranked as medium. Endpoints that do not achieve the high bar of validated surrogate continue to be useful in testing new treatments (22). Specifically, in settings in which it is reasonable to assume that an effect on OS can only be achieved if there is also an effect on disease progression, lack of an effect on the surrogate could be

used as a phase II futility assessment (or early phase III futility assessment), assuming that the phase III endpoint is OS (23,24).

Conclusion

To the best of our knowledge, this pooled analysis remains the first one conducted based on IPD in the context of adjuvant breast cancer. Our results suggest that LRFS may be a reasonnable surrogate endpoint for OS when assessing adjuvant systemic treatment in breast cancer. One should however acknowledge that this analysis was limited by the number of available trials, and the use of the including centers instead of the trials for the assessment of the trial-level associations might have artifically led to narrower confidence intervals. OS should remain the primary endpoint in RCT in this setting until meta-analyses conducted on a larger set of trials confirm our preliminary findngs. As it is reasonable to assume that an effect on OS can only be achieved if there is also an effect on disease progression, alternative endpoints such as LRFS remain useful in testing new treatments, provided that OS data are collected throughout the trial.

References

1. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival end point reporting in randomized cancer clinical trials: a review of major journals. J Clin Oncol Off J Am Soc Clin Oncol. 2008 Aug 1;26(22):3721–6.

2. Baselga J, Campone M, Piccart M, Burris HA, Rugo HS, Sahmoud T, et al. Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. N Engl J Med. 2012 Feb 9;366(6):520–9.

3. Piccart M, Hortobagyi GN, Campone M, Pritchard KI, Lebrun F, Ito Y, et al. Everolimus plus exemestane for hormone-receptor-positive, human epidermal growth factor receptor-2-negative advanced breast cancer: overall survival results from BOLERO-2†. Ann Oncol Off J Eur Soc Med Oncol. 2014 Dec;25(12):2357–62.

4. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostat Oxf Engl. 2000 Mar;1(1):49–67.

5. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. J R Stat Soc Ser C Appl Stat. 2001 Jan 1;50(4):405–22.

6. Green E, Yothers G, Sargent DJ. Surrogate endpoint validation: statistical elegance versus clinical relevance. Stat Methods Med Res. 2008 Oct;17(5):477–86.

7. Renfro LA, Shi Q, Sargent DJ. Surrogate End Points in Soft Tissue Sarcoma: Methodologic Challenges. J Clin Oncol Off J Am Soc Clin Oncol. 2016 Aug 22;

8. Gourgou-Bourgade S, Cameron D, Poortmans P, Asselain B, Azria D, Cardoso F, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials)†. Ann Oncol Off J Eur Soc Med Oncol. 2015 May;26(5):873–9.

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

10. Lassere M, Johnson K, Hughes M, Altman D, Buyse M, Galbraith S, et al. Simulation studies of surrogate endpoint validation using single trial and multitrial statistical approaches. J Rheumatol. 2007 Mar;34(3):616–9.

11. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat. 2006 Sep;5(3):173–86.

12. Institute for Quality and Efficiency in Health Care (IQWiG). Validity of surrogate endpoints in oncology: Executive summary of rapid report A10-05, Version 1.1. In: Institute for Quality and Efficiency in Health Care: Executive Summaries [Internet]. Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2005 [cited 2016 Dec 30]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK198799/

13. Efron B, Gong G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. Am Stat. 1983 Feb;37(1):36.

14. Bramwell VHC, Pritchard KI, Tu D, Tonkin K, Vachhrajani H, Vandenberg TA, et al. A randomized placebo-controlled study of tamoxifen after adjuvant chemotherapy in premenopausal women with early breast cancer (National Cancer Institute of Canada—Clinical Trials Group Trial, MA.12). Ann Oncol. 2010 Feb;21(2):283–90.

15. Roché H, Fumoleau P, Spielmann M, Canon J-L, Delozier T, Serin D, et al. Sequential adjuvant epirubicin-based and docetaxel chemotherapy for node-positive breast cancer patients: the FNCLCC PACS 01 Trial. J Clin Oncol Off J Am Soc Clin Oncol. 2006 Dec 20;24(36):5664–71.

16. Spielmann M, Roché H, Delozier T, Canon J-L, Romieu G, Bourgeois H, et al. Trastuzumab for patients with axillary-node-positive breast cancer: results of the FNCLCC-PACS 04 trial. J Clin Oncol Off J Am Soc Clin Oncol. 2009 Dec 20;27(36):6129–34.

17. Goldhirsch A, Gelber RD, Piccart-Gebhart MJ, de Azambuja E, Procter M, Suter TM, et al. 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): an open-label, randomised controlled trial. Lancet Lond Engl. 2013 Sep 21;382(9897):1021–8.

18. Kerbrat P, Desmoulins I, Roca L, Levy C, Lortholary A, Marre A, et al. Optimal duration of adjuvant chemotherapy for high-risk node-negative (N-) breast cancer patients: 6-year results of the prospective randomised multicentre phase III UNICANCER-PACS 05 trial (UCBG-0106). Eur J Cancer Oxf Engl 1990. 2017 Jul;79:166–75.

19. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. JAMA Intern Med. 2015 Aug;175(8):1389–98.

20. Miller K, Wang M, Gralow J, Dickler M, Cobleigh M, Perez EA, et al. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. N Engl J Med. 2007 Dec 27;357(26):2666–76.

21. Carpenter D, Kesselheim AS, Joffe S. Reputation and precedent in the bevacizumab decision. N Engl J Med. 2011 Jul 14;365(2):e3.

22. LeBlanc M, Tangen C. Surrogates for Survival or Other End Points in Oncology. JAMA Oncol. 2016 Feb;2(2):263–4.

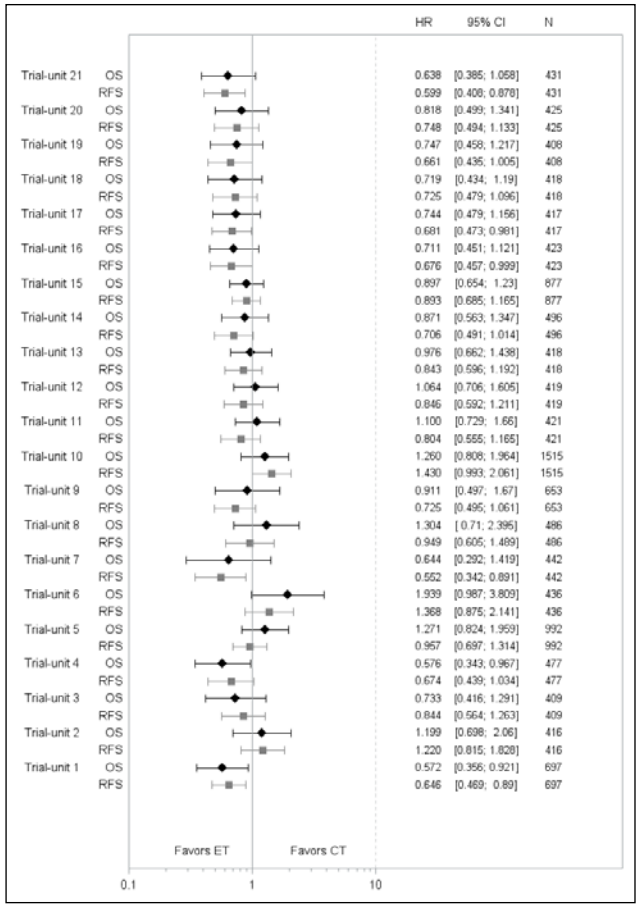
23. Goldman B, LeBlanc M, Crowley J. Interim futility analysis with intermediate endpoints. Clin Trials Lond Engl. 2008;5(1):14–22.

24. Redman MW, Goldman BH, LeBlanc M, Schott A, Baker LH. Modeling the relationship between progression-free survival and overall survival: the phase II/III trial. Clin Cancer Res Off J Am Assoc Cancer Res. 2013 May 15;19(10):2646–56.

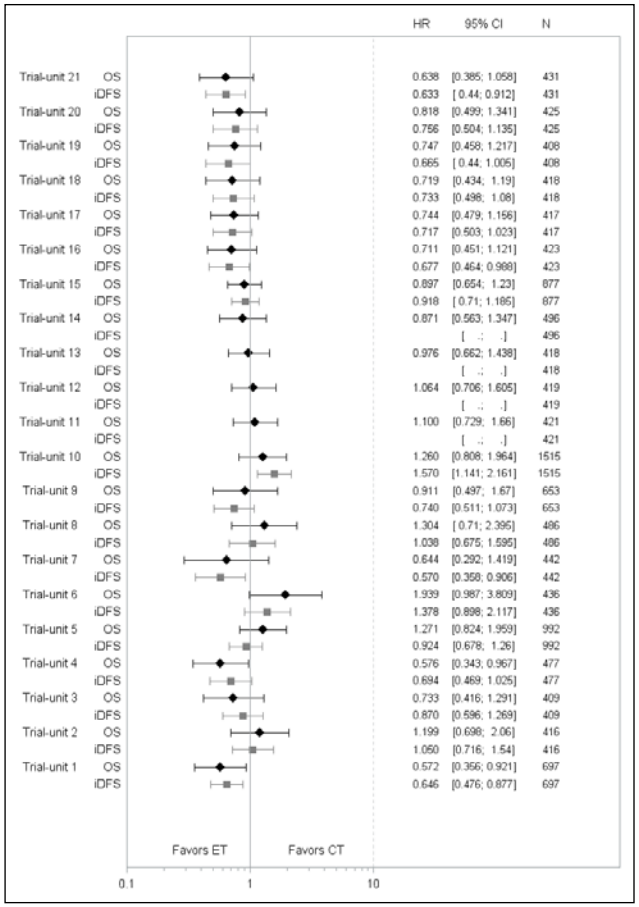
Figures

Figure 1. Forest plots of treatment effects on 7-year overall survival (OS) and on 5-year relapse-free survival (A), invasive disease-free survival (B), locoregional relapse-free survival (C) and distant disease-free survival (D) estimated using independent Cox models. The point estimates for the hazard ratio (HR) are presented for each trial-unit, for overall survival (OS, diamonds) and the candidate surrogate (squares), with the 95% confidence intervals (95%CI) and the number of patients per trial-unit.

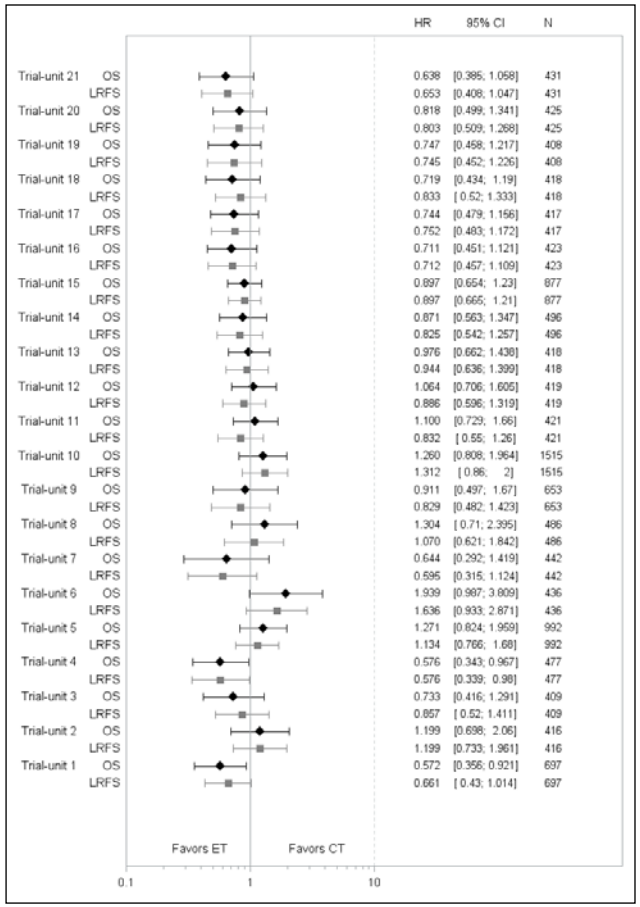
A: Relapse-free survival (RFS) – N = 11 676



B: Invasive disease-free survival (iDFS) – N = 9 922



C: Locoregional relapse-free survival (LRFS) – N = 11 676



D: Distant disease-free survival (DDFS) – N = 11 676

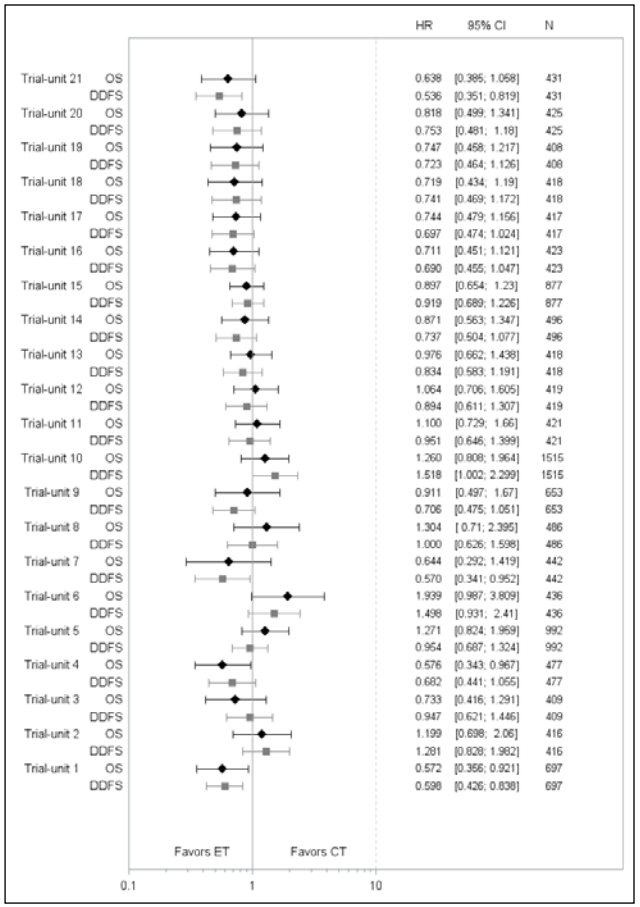
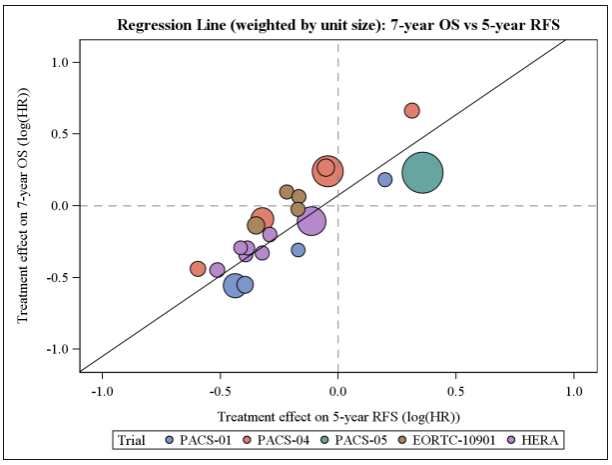


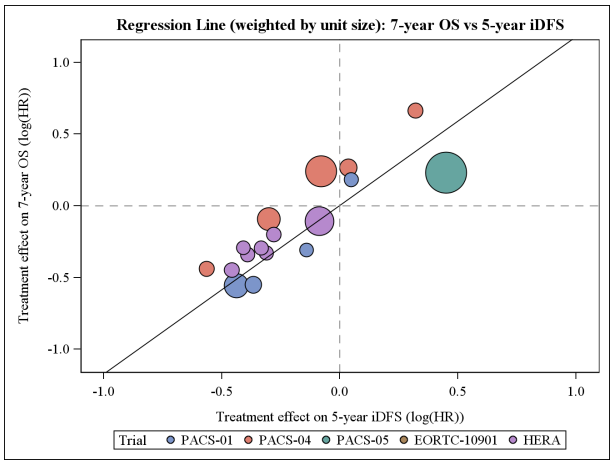
Figure 2. Trial-level association between treatment effects (log(HR)) on 7-year overall survival and 5-year (A) relapse-free-survival, (B) invasive disease-free-survival, (C) locoregional relapse-free survival and (D) distant disease-free survival. Equation of the weighted-linear regression model is provided below. Each circle represents a trial-unit with the surface area proportional to the size of the trial-unit.

A: Relapse-free survival (RFS) – 21 trial-units



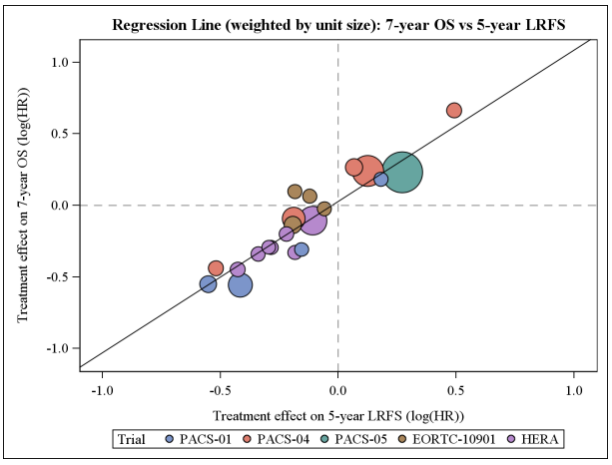
$\log(HR_{OS}) = 0.08\ [-0.02; 0.33] + 1.12\ [1.00; 1.25] * \log(HR_{RFS})$

B: Disease-free survival (DFS) – 17 trial-units



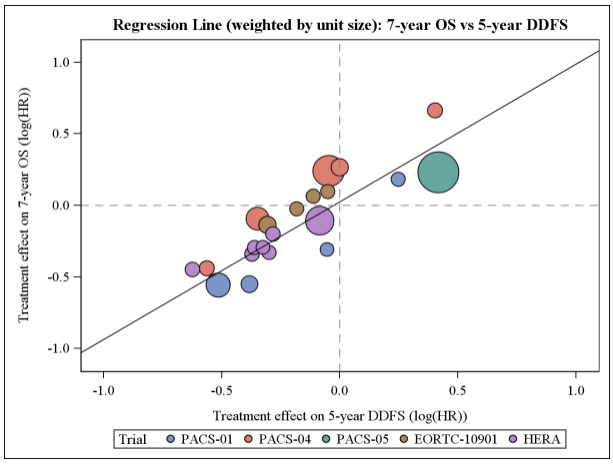
$\log(HR_{OS}) = 0.00\ [-0.38; 0.39] + 1.18\ [0.97; 1.38] * \log(HR_{iDFS})$

C: Locoregional relapse-free survival (LRFS) – 21 trial-units



$\log(HR_{OS}) = 0.03\ [-0.10; 0.15] + 1.06\ [0.9; 1.21] * \log(HR_{LRFS})$

D: Distant disease-free survival (DDFS) – 21 trial-units



$\log(HR_{OS}) = 0.03\ [-0.20; 0.25] + 0.96\ [082; 1] * \log(HR_{DDFS})$

Tables

Table 1. Clinical events and causes of death considered in the definition of relapse-free-survival, invasive disease-free-survival, locoregional relapse-free survival and distant disease-free survival in breast cancer trials in the adjuvant setting. * (8,9).

	Death from…					Clinical events					
Endpoint (*)	**Breast cancer**	**Non-breast cancer**	**Protocol treatment**	**Other cause**	**Unknown cause**	**Locoregional invasive events**	**Invasive contralateral breast cancer**	**Metastatic recurrence**	**2nd primary invasive non-breast cancer**	**Ipsilateral DCIS**	**Contralateral DCIS**
RFS	x	x	x	x	x	x		x		x	
iDFS	x	x	x	x	x	x	x	x	x		
LRFS	x	x	x	x	x	x				x	
DDFS	X	x	x	x	x			x			

* RFS = Relapse-free survival; iDFS = invasive disease-free survival; LRFS = locoregional relapse-free survival; DDFS = distant disease-free survival. Time-to-event endpoints are defined as per international DATECAN guidelines (8,9).
DCIS = Ductal carcinoma in situ.

Table 2. Characteristics of the randomized trials included in the pooled analysis (5 trials, 11 676 subjects)

Trial	Trial-units	N	Inclusion	Control arm	Experimental arm	Median follow-up (*)
PACS-01 (15)	1-4	1999	1997 - 2000	6 FEC	3 FEC + 3 Docetaxel	63.5 months
PACS-04 (16)	5-9	3010	2001 - 2004	FEC	Epirubicin + Docetaxel	53.4 months
PACS-05 (18)	10	1515	2002 - 2007	FEC (6 cycles)	FEC (4 cycles)	73.7 months
EORTC-10901 (14)	11-14	1724	1991 - 1999	No hormonal treatment	Tamoxifen	78.6 months
HERA (17)	15-21	5081	> 1999	Trastuzumab (1 year)	Trastuzumab (2 years)	95.8 months

* Median follow-up is defined as the time for which 50% of patients went out of study, either because of death or lost to follow-up.
FEC = Fluorouracile + Epirubicine + Cyclophosphamide.

Table 3. Frequency of events observed at various time points for overall survival, relapse-free-survival, invasive disease-free-survival, locoregional relapse-free survival and distant disease-free survival *

		Number of events (% of the total number of events) observed after a follow-up of…				
Endpoint [(1)]	**N [(2)]**	**3 years (N = 10 441)**	**5 years (N = 6 938)**	**7 years (N = 3 281)**	**10 years (N = 93)**	**Overall follow-up**
OS	11 676	730 (46.8%)	1238 (79.4%)	1474 (94.5%)	1558 (99.9%)	1560 (100%)
RFS	11 676	1792 (68.9%)	2340 (90.0%)	2550 (98.1%)	2597 (99.9%)	2599 (100%)
iDFS	9 922	1543 (67.2%)	2042 (89.0%)	2249 (98.0%)	2295 (100%)	2295 (100%)
LRFS	11 676	1101 (56.1%)	1639 (83.5%)	1889 (96.3%)	1960 (99.9%)	1962 (100%)
DDFS	11 676	1562 (67.5%)	2074 (89.6%)	2263 (97.8%)	2314 (99.9%)	2315 (100%)

* RFS = Relapse-free survival; iDFS = invasive disease-free survival; LRFS = locoregional relapse-free survival; DDFS = distant disease-free survival. Time-to-event endpoints are defined as per international DATECAN guidelines (8,9).
N: Number of subjects.

Table 4. Individual- and trial-level associations between 7-year overall survival and 5-year relapse-free-survival, invasive disease-free-survival, locoregional relapse-free survival and distant disease-free survival. Results from the pooled analysis (21 trial-units, 11 676 subjects).

	Individual-level association CI [95%]	**Trial-level association CI [95%]**		
Endpoint *	**$\rho_{Spearman}$**	**R^2_{WLR}**	**Unadjusted R^2_{2SM}**	**STE**
RFS	0.99 [0.99; 0.99]	0.96 [0.82 ; 0.99]	0.54 [0.25; 0.83]	0.62
iDFS	0.99 [0.98; 0.99]	0.90 [0.53 ; 0.98]	0.48 [0.14; 0.82]	0.60
LRFS	0.99 [0.99; 0.99]	0.94 [0.73 ; 0.99]	0.89 [0.80; 0.98]	0.79
DDFS	0.99 [0.99; 0.99]	0.93 [0.69 ; 0.98]	0.62 [0.37; 0.88]	0.63

* RFS = Relapse-free survival; iDFS = invasive disease-free survival; LRFS = locoregional relapse-free survival; DDFS = distant disease-free survival. Time-to-event endpoints are defined as per international DATECAN guidelines (8,9).
95%CI: 95% confidence interval. R^2_{WLR}: coefficient of determination as per weighted-linear regression. R^2_{2SM}: coefficient of determination as per the two-stage model. STE: surrogate threshold effect.

Additional files

Additional file 1. Individual- and trial-level associations between overall survival and relapse-free-survival, invasive disease-free-survival, locoregional relapse-free survival and distant disease-free survival, using varying follow-ups. Results from the pooled analysis of 21 trial-units (11 676 subjects).

1a. Individual- and trial-level associations between 7-year overall survival and 3-year relapse-free-survival, invasive disease-free-survival, locoregional relapse-free survival and distant disease-free survival.

	Individual-level association [95%CI]	Trial-level association [95%CI]	
Endpoint	**$\rho_{Spearman}$**	**R^2_{WLR}**	**Unadjusted R^2_{2SM}**
RFS	0.99 [0.98; 0.99]	0.71 [0.41; 0.82]	0.61 [0.35; 0.87]
iDFS	0.98 [0.98; 0.98]	0.72 [0.37; 0.83]	0.60 [0.31; 0.89]
LRFS	0.99 [0.98; 0.99]	0.76 [0.51; 0.85]	0.74 [0.55; 0.93]
DDFS	0.99 [0.99; 0.99]	0.70 [0.40; 0.81]	0.69 [0.47; 0.91]

* RFS = Relapse-free survival; iDFS = invasive disease-free survival; LRFS = locoregional relapse-free survival; DDFS = distant disease-free survival. Time-to-event endpoints are defined as per international DATECAN guidelines (8,9).
R^2_{2SM}: coefficient of determination as per the two-stage model. STE: surrogate threshold effect.

1b. Individual- and trial-level associations between 5-year overall survival and 3-year relapse-free-survival, invasive disease-free-survival, locoregional relapse-free survival and distant disease-free survival.

	Individual-level association IC[95%]	Trial-level association IC[95%]	
Endpoint	**$\rho_{Spearman}$**	**R^2_{WLR}**	**Unadjusted R^2_{2SM}**
RFS	0.99 [0.99; 0.99]	0.78 [0.53; 0.86]	0.57 [0.29; 0.85]
iDFS	0.99 [0.98; 0.99]	0.73 [0.39; 0.83]	0.50 [0.17; 0.84]
LRFS	0.99 [0.99; 0.99]	0.76 [0.50; 0.85]	0.70 [0.48; 0.91]
DDFS	0.99 [0.99; 0.99]	0.81 [0.59; 0.88]	0.70 [0.48; 0.91]

* RFS = Relapse-free survival; iDFS = invasive disease-free survival; LRFS = locoregional relapse-free survival; DDFS = distant disease-free survival. Time-to-event endpoints are defined as per international DATECAN guidelines (8,9).
R^2_{2SM}: coefficient of determination as per the two-stage model. STE: surrogate threshold effect.

Additional file 2. Individual- and trial-level associations between overall survival and relapse-free-survival, invasive disease-free-survival, locoregional relapse-free survival and distant disease-free survival, after exclusion of HER+ specific trial. Results from the pooled analysis of 14 trial-units (8 277 subjects).

	Individual-level association IC[95%]	Trial-level association IC[95%]	
Endpoint	**$\rho_{Spearman}$**	**R^2_{WLR}**	**Unadjusted R^2_{2SM}**
RFS	0.99 [0.99; 0.99]	0.98 [0.87; 1.00]	0.48 [0.10; 0.86]
iDFS	0.99 [0.99; 0.99]	0.93 [0.29; 0.99]	0.47 [0.03; 0.92]
LRFS	0.99 [0.99; 0.99]	0.94 [0.61; 0.99]	0.91 [0.82; 1.00]
DDFS	0.99 [0.99; 0.99]	0.95 [0.67; 0.99]	0.56 [0.21; 0.90]

* RFS = Relapse-free survival; iDFS = invasive disease-free survival; LRFS = locoregional relapse-free survival; DDFS = distant disease-free survival. Time-to-event endpoints are defined as per international DATECAN guidelines (8,9).
R^2_{2SM}: coefficient of determination as per the two-stage model. STE: surrogate threshold effect.

6 General discussion

6.1 Conclusion on the thesis work

A surrogate endpoint is one which can be used in lieu of the endpoint of primary interest in the evaluation of experimental treatments or other interventions. Surrogate endpoints are useful when they can be measured earlier, more conveniently, or more frequently than the endpoints of interest. However, before a surrogate endpoint can replace a final endpoint in the evaluation of an experimental treatment, it must be formally validated.

The first part of this work was to draw-up an overview of existing studies assessing the validity of surrogate endpoints for OS in cancer RCTs. We conducted a systematic and critical review of MAs conducted in the field of oncology to evaluate surrogate endpoints for OS. To assess the strength of evidence provided by each MA, we relied on specific grids developed for the validation of surrogate endpoints. Despite the increasing number of studies, only a few MAs provided reliable evidence of surrogacy for OS. In the adjuvant setting, DFS showed good surrogate properties for OS in the context of colon cancer, operable and locally advanced non-small-cell lung cancer, gastric cancer and locally advanced head and neck cancer. In the metastatic setting, PFS showed good surrogate properties for OS in colorectal cancer, locally advanced lung cancer and locally advanced head and neck cancer. No sufficient evidence was available in existing MAs to conclude regarding the validity as surrogates in other disease settings; this was either due to low correlations, missing surrogacy measures, unreliable statistical design or lack of accuracy in the estimations. This review also highlighted the heterogeneity between studies in terms of statistical methodology and reported measures of surrogacy. This issue is closely related to the absence of a validated and objective validation grid for the evaluation of MAs assessing surrogate endpoints. Despite these practical limitations, this work provided key information to researchers involved in the design of RCTs to select appropriate primary endpoints, as well as indications for further research perspectives with regards to surrogate endpoints that require further evaluation, or improvement of existing grids for the assessment of validation studies of surrogate endpoints.

In the second part, we assessed the surrogate properties for OS of three commonly used endpoints – PFS, TTP and TTF – in the context of advanced STS, based on 14 RCTs. The MA did not lead to significant evidence to validate PFS, TTP or TTF as surrogate markers for OS when assessing systemic treatment in advanced STS. OS should therefore remain the primary endpoint in RCTs conducted in this setting. Trial design for advanced STS is particularly challenging due to the rarity and the heterogeneity of the disease and treatments, which may have contributed to weaken the observed correlations between candidate

surrogates and OS. It is however reasonable to assume that an effect on OS can only be achieved if there is an effect also on disease progression. As such, alternative endpoints, e.g. PFS, remain useful in testing new treatments in earlier drug development stages, such as in phase II trials or phase III futility assessment, provided OS data are collected throughout the trial. This analysis did not include recent treatments, in particular immunotherapeutic agents, for which only a small amount of data are available to date. As the mechanism of action of these treatments differs significantly from cytotoxic agents, it would be of interest to conduct specific studies to assess surrogate endpoints for OS.

In the third part, we evaluated TNT as a surrogate endpoint for OS in metastatic STS. Compared to progression-based endpoints, TNT is simple to measure and objectively defined. Indeed, the exact date of new treatment initiation is usually known and it frees itself from the subjectivity related to the evaluation of disease progression. Additionally, TNT includes all causes that could lead to a change in the patient status, and thus to the initiation of a different treatment. One might reasonably assume that deterioration of the patient status, whether directly caused by disease progression or by any kind of toxicity, is closely associated with the patient's survival. Based on a prospective cohort of 1575 patients, we estimated the individual-level association between TNT and OS at different lines of treatment. TNT was highly associated with OS at the patient-level, especially in the first-line setting. Although it was not possible to estimate the trial-level association, this work suggests that TNT would be worth further investigating as surrogate endpoint for OS. Collecting the relevant data in subsequent STS trials would be particularly informative for a proper assessment of this endpoint in future studies.

The last part of this thesis focused on the evaluation of surrogate endpoints in adjuvant BC trials. Relying on a pooled analysis of five RCTs, we assessed the surrogate properties for OS of four commonly used time-to-event endpoints: RFS, iDFS, LRFS and DDFS. The analysis showed strong correlations between the candidate surrogates and OS, both at the individual- and the trial-level. Based on these results, the trial-level association between LRFS and OS was ranked high as per the IQWiG criteria. For the three other endpoints, the trial-level associations were ranked as medium. One should however acknowledge that the use of including centers instead of the trials may have artificially narrowed the reported confidence intervals. Further investigation based on a larger dataset of RCTs would be required to confirm these preliminary findings.

6.2 Critical insight and perspectives

Surrogate endpoints are designed to be easier and quicker to measure than the clinical final endpoint they predict. Their use is driven by the need to reduce the costs and delays related to the approval process by reducing the number of patients and the duration of clinical trials. To date however, despite the increasing number of MA evaluating surrogate endpoints, only a few have been formally validated.

Even though one cannot exclude the possibility that the lack of evidence of surrogacy is due to the absence of prediction capacity of the candidate surrogates, some practical and methodological issues that might have affected the estimated correlations should be considered. First, MAs evaluating surrogate endpoints are based on a fragment of the available evidence, even when an attempt to gain unpublished reports goes along an exhaustive literature search that includes published articles and abstracts (27). As data from unpublished or unreported trials differ from easily located data, it is reasonable to assume that the lack of available data might bias the estimation of the surrogacy measures. Second, the evaluation of surrogate endpoints relies on complex statistical models that are still evolving. Multiple surrogacy measures have been proposed in the past two decades, and it is recognized that the surrogacy question can only be properly addressed through the conduct of a MA. To date however there is no consensus regarding the calculation and interpretation of these surrogacy metrics. As a result, MAs conducted to assess surrogate endpoints are very heterogeneous in terms of statistical methodology and results interpretation, and as such difficult to compare. Most MAs rely on aggregated data rather than IPD. The use of IPD enables estimation of the individual-level association between the candidate surrogate and the clinical endpoint, and harmonization of data in terms of endpoints definition and patient follow-up. Additionally, the most statistically rigorous methods to assess surrogate properties rely on IPD, as the estimation errors associated with the treatment effects cannot be properly taken into account in the model when based on aggregated data. The gathering of IPD is however time-consuming and requires agreements from the sponsors of the trials. As such, IPD-based MAs are usually based on a smaller set of trials, and in particular do not include the most recent trials. Third, commonly used surrogate endpoints, such as PFS and RFS, include clinical events that are difficult to measure precisely. For instance, the evaluation of disease progression relies on radiological imaging, which may be subject to reader's subjectivity. Additionally, it implies that the exact date of progression is known, although we typically use the date of the radiological assessment as a proxy. This uncertainty might thus affect the estimated associations. In such case, endpoints such as TNT might be an interesting alternative to progression-based endpoints.

The validation of surrogate endpoints has been a controversial issue and is not limited to oncology drugs. As highlighted by Buyse et al., difficulties have arisen on several fronts (Buyse *Biostatistics* 2000). Firstly, some endpoints used as surrogates have been shown to provide misleading predictions of the treatment effect upon the important clinical endpoints, as with the example of bevacizumab in metastatic BC patients, initially approved by the FDA and subsequently withdrawn, or with encainide and flecainide, two harmful drugs also approved by the FDA based on their anti-arrhythmic effects and subsequently withdrawn. Secondly, some endpoints that have not been so catastrophically misleading have still failed to explain the totality of the treatment effect upon the final endpoints: the case of the CD4+ lymphocyte counts in patients with AIDS is an example (63). Many of these problems were already mentioned by Prentice (1989). All these reasons have led some authors to express reservations about attempts to validate surrogate endpoints statistically (35,64). Their reservations rest to a large extent on biological considerations: a good surrogate must be shown to be causally linked to the true endpoint, and even so, it is implausible that the surrogate will ever capture the whole effect of treatment upon the true endpoint. These reservations are well taken, but biologically complex situations lend themselves to statistical evaluations that may shed light on the underlying mechanisms involved. The two-stage modeling approach addresses these issues: a large individual-level coefficient of determination indicates that the endpoints are likely to be causally linked to each other, while a large trial-level coefficient of determination indicates that a large proportion of the treatment effect is captured by the surrogate. Large numbers of observations are however needed for the estimates to be sufficiently precise, while multiple studies are needed to distinguish individual-level from trial-level associations between the endpoints and effects of interest. Finally, even if the results of a surrogate evaluation seem encouraging based on several trials, applying these results to a new trial would still require a certain amount of extrapolation that may or may not be deemed acceptable.

References

1. Kola I. The state of innovation in drug development. Clin Pharmacol Ther. 2008 Feb;83(2):227–30.

2. Hirschfeld S, Pazdur R. Oncology drug development: United States Food and Drug Administration perspective. Crit Rev Oncol Hematol. 2002 May;42(2):137–43.

3. European Medicines Agency - Marketing authorisation - Conditional marketing authorisation [Internet]. [cited 2017 Nov 2]. Available from: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000925.jsp

4. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar;69(3):89–95.

5. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival end point reporting in randomized cancer clinical trials: a review of major journals. J Clin Oncol Off J Am Soc Clin Oncol. 2008 Aug 1;26(22):3721–6.

6. Kim C, Prasad V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. JAMA Intern Med. 2015 Dec;175(12):1992–4.

7. Kim C, Prasad V. Strength of Validation for Surrogate End Points Used in the US Food and Drug Administration's Approval of Oncology Drugs. Mayo Clin Proc. 2016 May 10;

8. Miller K, Wang M, Gralow J, Dickler M, Cobleigh M, Perez EA, et al. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. N Engl J Med. 2007 Dec 27;357(26):2666–76.

9. Miles DW, Chan A, Dirix LY, Cortés J, Pivot X, Tomczak P, et al. Phase III study of bevacizumab plus docetaxel compared with placebo plus docetaxel for the first-line treatment of human epidermal growth factor receptor 2-negative metastatic breast cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2010 Jul 10;28(20):3239–47.

10. Robert NJ, Diéras V, Glaspy J, Brufsky AM, Bondarenko I, Lipatov ON, et al. RIBBON-1: randomized, double-blind, placebo-controlled, phase III trial of chemotherapy with or without bevacizumab for first-line treatment of human epidermal growth factor receptor 2-negative, locally recurrent or metastatic breast cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2011 Apr 1;29(10):1252–60.

11. Carpenter D, Kesselheim AS, Joffe S. Reputation and precedent in the bevacizumab decision. N Engl J Med. 2011 Jul 14;365(2):e3.

12. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989 Apr;8(4):431–40.

13. Burzykowski T, Molenberghs G, Buyse ME. The evaluation of surrogate endpoints. New York; London: Springer; 2011.

14. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med. 1992 Jan 30;11(2):167–78.

15. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics. 1998 Sep;54(3):1014–29.

16. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat. 2006 Sep;5(3):173–86.

17. Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports. Health Technol Assess Winch Engl. 2009 Jan;13(8):iii, ix–xi, 1-50.

18. Institute for Quality and Efficiency in Health Care (IQWiG). Validity of surrogate endpoints in oncology: Executive summary of rapid report A10-05, Version 1.1. In: Institute for Quality and Efficiency in Health Care: Executive Summaries [Internet]. Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2005 [cited 2016 Dec 30]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK198799/

19. Lassere MN, Johnson KR, Schiff M, Rees D. Is blood pressure reduction a valid surrogate endpoint for stroke prevention? An analysis incorporating a systematic review of randomised controlled trials, a by-trial weighted errors-in-variables regression, the surrogate threshold effect (STE) and the Biomarker-Surrogacy (BioSurrogate) Evaluation Schema (BSES). BMC Med Res Methodol. 2012 Mar 12;12:27.

20. Coindre JM, Terrier P, Guillou L, Le Doussal V, Collin F, Ranchère D, et al. Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: a study of 1240 patients from the French Federation of Cancer Centers Sarcoma Group. Cancer. 2001 May 15;91(10):1914–26.

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

22. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? J Clin Oncol Off J Am Soc Clin Oncol. 2012 Apr 1;30(10):1030–3.

23. Padhani AR, Ollivier L. The RECIST (Response Evaluation Criteria in Solid Tumors) criteria: implications for diagnostic radiologists. Br J Radiol. 2001 Nov;74(887):983–6.

24. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer Oxf Engl 1990. 2009 Jan;45(2):228–47.

25. Gourgou-Bourgade S, Cameron D, Poortmans P, Asselain B, Azria D, Cardoso F, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials)†. Ann Oncol Off J Eur Soc Med Oncol. 2015 May;26(5):873–9.

26. Gourgou-Bourgade S, Cameron D, Poortmans P, Asselain B, Azria D, Cardoso F, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative

(Definition for the Assessment of Time-to-event Endpoints in CANcer trials)†. Ann Oncol. 2015 May;26(5):873–9.

27. Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? BMC Med. 2017 Jul 21;15(1):134.

28. Friedman LM, Furberg C, DeMets DL, Reboussin D, Granger CB. Fundamentals of clinical trials. 2015.

29. Eisenhauer EA, Twelves C, Buyse ME, Eisenhauer EA. Phase I cancer clinical trials: a practical guide. 2015.

30. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004;3(8):711–5.

31. Schatzkin A. Intermediate markers as surrogate endpoints in cancer research. Hematol Oncol Clin North Am. 2000 Aug;14(4):887–905.

32. Johnson JR, Ning Y-M, Farrell A, Justice R, Keegan P, Pazdur R. Accelerated approval of oncology products: the food and drug administration experience. J Natl Cancer Inst. 2011 Apr 20;103(8):636–44.

33. Lu E, Shatzel J, Shin F, Prasad V. What constitutes an “unmet medical need” in oncology? An empirical evaluation of author usage in the biomedical literature. Semin Oncol. 2017 Feb;44(1):8–12.

34. Sridhara R, Johnson JR, Justice R, Keegan P, Chakravarty A, Pazdur R. Review of oncology and hematology drug product approvals at the US Food and Drug Administration between July 2005 and December 2007. J Natl Cancer Inst. 2010 Feb 24;102(4):230–43.

35. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996 Oct 1;125(7):605–13.

36. Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. Stat Med. 2005 Jan 30;24(2):163–82.

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

38. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. J R Stat Soc Ser C Appl Stat. 2001 Jan 1;50(4):405–22.

39. Schweizer B. Thirty Years of Copulas. In: Dall’Aglio G, Kotz S, Salinetti G, editors. Advances in Probability Distributions with Given Marginals [Internet]. Dordrecht: Springer Netherlands; 1991 [cited 2017 Nov 2]. p. 13–50. Available from: http://link.springer.com/10.1007/978-94-011-3466-8_2

40. Nelsen RB. An introduction to copulas. New York, N.Y.: Springer; 1999.

41. Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. Biometrics. 1995 Dec;51(4):1384–99.

42. Shih JH. A goodness-of-fit test for association in a bivariate survival model. Biometrika. 1998 Mar 1;85(1):189–200.

43. Akaike H, Parzen E, Tanabe K, Kitagawa G. Selected papers of Hirotugu Akaike. New York: Springer; 1998.

44. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978 Mar;6(2):461–4.

45. Clayton DG. A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. Biometrika. 1978 Apr;65(1):141.

46. Hougaard P. Frailty models for survival data. Lifetime Data Anal. 1995;1(3):255–73.

47. Plackett RL. A Class of Bivariate Distributions. J Am Stat Assoc. 1965 Jun 1;60(310):516–22.

48. Buyse M, Burzykowski T, Carroll K, Michiels S, Sargent DJ, Miller LL, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2007 Nov 20;25(33):5218–24.

49. Cartier S, Zhang B, Rosen VM, Zarotsky V, Bartlett JB, Mukhopadhyay P, et al. Relationship between treatment effects on progression-free survival and overall survival in multiple myeloma: a systematic review and meta-analysis of published clinical trial data. Oncol Res Treat. 2015;38(3):88–94.

50. Petrelli F, Coinu A, Borgonovo K, Cabiddu M, Barni S. Progression-free survival as surrogate endpoint in advanced pancreatic cancer: meta-analysis of 30 randomized first-line trials. Hepatobiliary Pancreat Dis Int HBPD INT. 2015 Apr;14(2):124–31.

51. Flaherty KT, Hennig M, Lee SJ, Ascierto PA, Dummer R, Eggermont AMM, et al. Surrogate endpoints for overall survival in metastatic melanoma: a meta-analysis of randomised controlled trials. Lancet Oncol. 2014 Mar;15(3):297–304.

52. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostat Oxf Engl. 2000 Mar;1(1):49–67.

53. Green E, Yothers G, Sargent DJ. Surrogate endpoint validation: statistical elegance versus clinical relevance. Stat Methods Med Res. 2008 Oct;17(5):477–86.

54. Renfro LA, Shi Q, Sargent DJ. Surrogate End Points in Soft Tissue Sarcoma: Methodologic Challenges. J Clin Oncol Off J Am Soc Clin Oncol. 2016 Aug 22;

55. Sherrill B, Kaye JA, Sandin R, Cappelleri JC, Chen C. Review of meta-analyses evaluating surrogate endpoints for overall survival in oncology. OncoTargets Ther. 2012;5:287–96.

56. Ciani O, Davis S, Tappenden P, Garside R, Stein K, Cantrell A, et al. Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. Int J Technol Assess Health Care. 2014 Jul;30(3):312–24.

57. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. JAMA Intern Med. 2015 Aug;175(8):1389–98.

58. Zer A, Prince RM, Amir E, Abdul Razak A. Evolution of Randomized Trials in Advanced/Metastatic Soft Tissue Sarcoma: End Point Selection, Surrogacy, and Quality of Reporting. J Clin Oncol. 2016 May 1;34(13):1469–75.

59. Ng R, Pond GR, Tang PA, MacIntosh PW, Siu LL, Chen EX. Correlation of changes between 2-year disease-free survival and 5-year overall survival in adjuvant breast cancer trials from 1966 to 2006. Ann Oncol Off J Eur Soc Med Oncol. 2008 Mar;19(3):481–6.

60. Chudley L, McCann K, Mander A, Tjelle T, Campos-Perez J, Godeseth R, et al. DNA fusion-gene vaccination in patients with prostate cancer induces high-frequency CD8(+) T-cell responses and increases PSA doubling time. Cancer Immunol Immunother CII. 2012 Nov;61(11):2161–70.

61. Liang C, Li L, Fraser CD, Ko A, Corzo D, Enger C, et al. The treatment patterns, efficacy, and safety of nab (®)-paclitaxel for the treatment of metastatic breast cancer in the United States: results from health insurance claims analysis. BMC Cancer. 2015 Dec 29;15:1019.

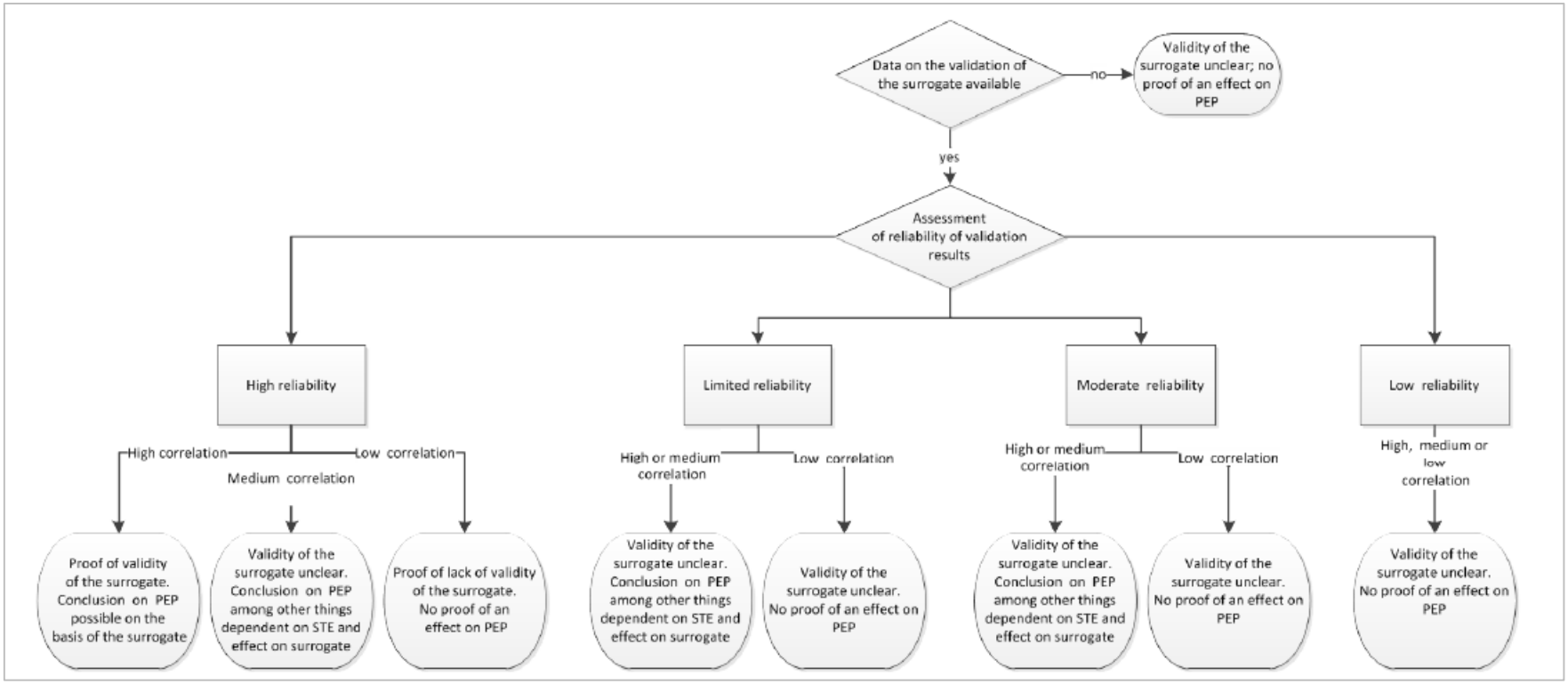
62. Teng C-LJ, Wang C-Y, Chen Y-H, Lin C-H, Hwang W-L. Optimal Sequence of Irinotecan and Oxaliplatin-Based Regimens in Metastatic Colorectal Cancer: A Population-Based Observational Study. PloS One. 2015;10(8):e0135673.

63. Jacobson MA, De Gruttola V, Reddy M, Arduino JM, Strickland S, Reichman RC, et al. The predictive value of changes in serologic and cell markers of HIV activity for subsequent clinical outcome in patients with asymptomatic HIV disease treated with zidovudine. AIDS Lond Engl. 1995 Jul;9(7):727–34.

64. De Gruttola V, Fleming T, Lin DY, Coombs R. Perspective: validating surrogate markers--are we being naive? J Infect Dis. 1997 Feb;175(2):237–46.

Appendices

Appendix A: The German Institute of Quality and Efficiency in Health Care (IQWiG) decision tree for the overall conclusion on the surrogate validity (18)



STE = Surrogate threshold effect

Appendix B: Biomarker-Surrogacy (BioSurrogate) Evaluation Schema (19)

BIOMARKER-SURROGATE DOMAINS	
Study Design Domain #	
0	Biological plausibility & lower quality clinical studies e.g. cross-sectional observational studies
1	Rank 0 and at least 2 good quality prospective observational cohort studies measuring S and T
2	Rank 1 and at least 2 high quality adequately powered RCTs measuring S and T
3	Rank 1 and all, and at least 5 high quality adequately powered, RCTs measuring S and T
Target Outcome Domain	
0	Target is reversible disease-centered biomarker of harm
1	Target is irreversible disease-centered biomarker of harm
2	Target is patient-centered endpoint of reversible organ morbidity or clinical burden of disease or clinical harm
3	Target is patient-centered endpoint of irreversible organ morbidity or clinical burden of disease or severe irreversible clinical harm or death
Statistical Evaluation of BioSurrogate – Target (B-T) Domain	
0	Poor: Does not meet the criteria for Rank 1
1	Fair: RCT $R^2_{trial} \geq 0.2$ AND STEP* ≥ 0.1 AND $R^2_{ind} \geq 0.2$ OR cohort data $R^2_{ind} \geq 0.4$
2	Good: RCT $R^2_{trial} \geq 0.4$ AND STEP* ≥ 0.2 AND $R^2_{ind} \geq 0.4$
3	Excellent: RCT $R^2_{trial} \geq 0.6$ AND STEP* ≥ 0.3 AND $R^2_{ind} \geq 0.6$ (without data subdivision)**
Generalisability of BioSurrogate-Target Domain: **Clinical evidence across different risk populations & pharmacologic evidence across different drug-class mechanisms**	
0	No clinical or pharmacologic evidence
1	Clinical OR pharmacologic evidence
2	Clinical AND pharmacologic evidence
3	Consistent Clinical RCT AND pharmacologic RCT evidence

Where S is the surrogate / biomarker/ biosurrogate and T is the target / true outcome

* STEP is defined as that proportion of the total range of the surrogate that is equal or larger than the STE

** Some analyses with few trials subdivide into centres to increase the number of data points

LEVEL OF EVIDENCE OF SURROGATE ENDPOINT MULTIDIMENSIONAL VALIDITY

A high rank on any one or more domain should not be allowed to prevail over a low rank on one or more domain when determining the overall level of evidence because at least good evidence of surrogacy across all domains is needed for surrogate validity. An A, B+, B, B- level surrogate endpoint ranks at least 2 on all domains.

Steps to determine the level of evidence:

1. The one and the same 'evidence-base' is applied across all four domains when determining the level of evidence.
2. Sum of the highest rankings achieved across the four domains.
3. If any one domain is less than Rank 2, the level of evidence drops by one alphabetic category irrespective of the initial level. For example, B becomes a C, B- becomes a C- , C- becomes a D and so forth.

12	**level A**
11- 9	**level B+, B, B-**
8 - 6	**level C+, C, C-, D+, D, D-**
5 - 3	**level D+, D, D-, E+, E, E-**
2 - 0	**level E+, E, E- F+, F, F**