Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Livre blanc SFPT

De la nécessité de la méthodologie *dans l'évaluation des médicaments*

Document compagnon

Dossier 25 - L'évaluation de la médecine personnalisée

V1.0, 17 juillet 2023

Comité de rédaction et relecture (par ordre alphabétique)

Jean Luc Cracowski

Michel Cucherat

Dominique Deplanque

Behrouz Kassai

Silvy Laporte

Clara Locher

Florian Naudet

Edouard Ollier

Matthieu Roustit

Table des matières

1 Introduction

Définir la médecine dite personnalisée n'est pas une tâche facile. Notamment, car ce terme est particulièrement à la mode et utilisé à tout propos pour donner une image de modernité, d'innovation, de sophistication scientifique [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. De plus, le terme est quelque peu excessif et il conviendrait plutôt de parler de médecine stratifiée [11] ou de médecine de précision (cf. section 1.3).

Quoi qu'il en soit, l'hypothèse générale qui se cache plus ou moins derrière l'idée de « médecine personnalisée » en thérapeutique est que, peut-être, tous les patients ne tirent pas le même bénéfice des traitements. L'objectif pratique de la médecine dite « personnalisée » serait alors de personnaliser le traitement de chaque patient, en lui proposant le traitement qui lui est le plus adapté, c'est-à-dire celui qui est susceptible de lui apporter le plus de bénéfice. Charge alors à l'évaluation clinique des thérapeutiques de déterminer, pour chaque traitement, les facteurs qui pourraient prédire le bénéfice clinique que celui-ci apporte aux patients en fonction de leurs caractéristiques démographiques, biologiques, génétiques, etc. Ce sont les facteurs (marqueurs) prédictifs de leur efficacité ou de leur sécurité (qu'il serait préférable de dénommer modificateurs d'effet plutôt que facteur prédictif, cf. section 2.1).

> *L'European Council Conclusion on PM[1] a défini la médecine personnalisée comme étant "a medical model using characterisation of individuals' phenotypes and genotypes (eg, molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention".*

Le principe sous-jacent de la médecine personnalisée n'est pas nouveau. De tout temps, la médecine a cherchée à traiter les patients à bon escient et à prescrire le traitement le plus adapté au patient lorsque cela était possible. Le terme médecine personnalisée se généralise dans les années 2000 principalement pour donner de la valeur à de nouvelles molécules en oncologie qui à priori avaient le défaut de ne pouvoir aider qu'une toute petite partie des patients porteur de la maladie visée. Actuellement le besoin de cette personnalisation est mis en avant par exemple dans l'acronyme de la médecine 4P : Prédictive, Personnalisée, Préventive, Participative [12].

1.1 Différentes approches de « personnalisation » des traitements

La « personnalisation » des traitements peut s'envisager de différentes façons.

1. L'utilisation de **médicaments dits « ciblés »** (cf. section 3), qui ont une affinité supérieure pour une cible spécifique, par exemple un variant génétique d'une protéine d'une voie de signalisation surexprimé dans une maladie. Ce variant n'est présent que chez certains patients atteints de la maladie et il est considéré, plus ou moins, comme contribuant à sa physiopathologie (mutation « driver » en oncologie par exemple). Lorsque cette cible particulière n'est pas présente, il n'est pas attendu d'effet pharmacodynamique du traitement, donc de bénéfice clinique. L'utilisation de ces traitements nécessite le typage des patients à

[1] https://www.euapm.eu/council-conclusions.html

l'aide d'un test compagnon, permettant de déterminer s'ils sont porteurs de la variante ciblée par le traitement. La forme ultime des thérapies ciblées est représentée par les vaccins thérapeutiques spécifiquement adaptés à un patient donné [13].

2. La personnalisation basée sur le **risque de base** ou le **pronostic** des patients [14] (cf. section 5). L'idée est d'éviter de traiter des patients qui sont à très faible risque de présenter l'évènement que le traitement cherche à prévenir. Par exemple, dans le cancer du sein précoce, la fréquence de la récidive à distance est assez faible et l'on pressent que certains patients ne récidiveront pas. Dans ce cas, l'administration d'un traitement adjuvant est inutile (bénéfice attendu faible) et sa balance bénéfice risque peut donc être défavorable compte-tenu de la toxicité de ces traitements (chimiothérapie). La personnalisation repose dans ce cas de figure sur l'identification de facteurs pronostiques (« prédictifs » de la survenue des évènements) et non pas de facteurs prédictifs de l'efficacité ou du risque du traitement.
3. La personnalisation basée sur les **facteurs prédictifs** du bénéfice du traitement (cf. section 4). Les facteurs prédictifs validés permettraient d'identifier des patients qui ne tireront pas de bénéfice du traitement soit en raison d'un manque d'efficacité du traitement chez eux ou d'un risque d'effet indésirable augmenté. Par rapport aux traitements ciblés, ces facteurs prédictifs ne sont pas directement liés à la cible du traitement et ne sont donc pas évidents à priori. Ils sont souvent recherchés, mais pas toujours, après la réalisation des essais pivots des traitements.
4. La personnalisation à l'aide d'**outils de prédiction de l'effet du traitement** à partir des caractéristiques du patient, outils élaborés à partir d'une modélisation de l'hétérogénéité de l'effet traitement réalisée à l'aide de techniques classiques de statistique ou par intelligence artificielle (machine learning) (section 4.3). De façon connexe, ces outils sont aussi envisageables pour l'enrichissement d'essais n'incluant que des patients potentiellement répondeurs, sélectionnés à l'aide d'un outil prédictif (cf. section 5).
5. Apparaissent aussi de vrais traitements personnalisés, confectionnés sur mesure pour chaque patient, comme les vaccins à ARMm encodant pour des néo-antigènes spécifiques des tumeurs des patients. Une approche de ce type (mRNA-4157/V940) dans le mélanome à haut risque de récidive a montré une amélioration de la RFS dans l'essai randomisé KEYNOTE-942 (NCT03897881)[2].

Compte tenu des enjeux (cf. section suivante), l'évaluation de l'utilité clinique est un prérequis fondamental à l'utilisation en pratique de ces approches de personnalisation. Pour les thérapies ciblées, cette utilité est démontrée directement dans leurs essais thérapeutiques pivots. Pour les autres approches, elle doit être démontrée à l'aide **d'essais de stratégie** (cf. section 6.5)

[2] https://www.abstractsonline.com/pp8/#!/10828/presentation/10243

Typologie des approches de personnalisation des traitements

1) **Thérapies « ciblées »**
2) **Personnalisation en fonction de la réponse au traitement**
   a) Recherche de marqueurs prédictifs de la réponse
      i) Approche « treatment only » (impropre, correspond à du pronostic sous traitement)
      ii) Recherche par des analyses en sous-groupes d'essais randomisés
         - Analyse en sous-groupe exploratoire
         - Analyse en sous-groupe de confirmation (prévue a priori et intégré au plan de contrôle du risque alpha global)
      iii) Essai randomisé avec un design d'interaction
   b) Modélisation/prédiction de l'effet traitement (cATE, Conditionnal Average Treatment Effect, méthodes statistiques conventionnelles ou IA)
      i) Modélisation du critère de jugement (l'effet traitement est une covariable comme les autres, l'effet du traitement est ensuite déduit de la prédiction du critère de jugement avec et sans traitement)
      ii) Modélisation de l'effet traitement (en intégrant dans le modèle les modificateurs d'effet en termes d'interaction)
3) **Personnalisation en fonction du risque de base** (nécessite un outil prédictif du risque construit soit par des méthodes statistiques conventionnelles soit par IA)

Type de personnalisation	**Principe**	**Méthodologie**	**Exemple**
Thérapie ciblée	Par conception, le médicament vise une cible moléculaire présente chez certains patients uniquement (par exemple une forme mutée d'une protéine d'une voie de signalisation).	Évaluation du bénéfice dans un essai randomisé ciblé d'emblée, n'incluant que des patients porteurs de la cible	Crizotinib pour les cancers du poumon non à petites cellules ALK+
En fonction de la réponse au traitement	Repose sur un marqueur prédictif permettant de déterminer les patients répondeurs ou non répondeurs au traitement (ou à risques d'effets indésirables)	Analyse en sous-groupe de confirmation prévue a priori dans un essai randomisé évaluant le bénéfice clinique du traitement d'intérêt	Restriction d'utilisation du pembrolizumab au patient PD-L1 positif >=50% dans le traitement de première ligne cancer du poumon métastatique[3]
En fonction de la prédiction de l'effet	Prédiction quantitative de l'effet qu'aura le traitement en fonction des caractéristiques du patient	Élaboration d'un outil prédictif validé issu d'une modélisation de l'hétérogénéité de l'effet dans un essai randomisé, validé par une validation externe	Prédiction de la réponse à la canagliflozine dans le diabète de type 2 pour prévenir les évènements cardiovasculaires par le hazard ratio individuel [15]

[3] https://www.ema.europa.eu/en/documents/product-information/keytruda-epar-product-information_en.pdf section 4.1

En fonction du risque de base	Éviter de traiter les patients qui ont un risque d'évolution défavorable très faible	Élaboration d'un outil pronostique de l'évolution défavorable	Signature protéomique pour décider d'un traitement adjuvant dans le cancer du sein précoce

L'approche la plus fréquemment rencontrée dans les publications recherchant des facteurs de réponse ou de non-réponse à un traitement est la recherche d'une valeur pronostique sous traitement (approche « *treatment only* »). Cette approche possède de nombreuses limites conceptuelles et méthodologiques et ne permet pas d'apporter la preuve d'une valeur prédictive d'un marqueur (cf. section 4.1) et, encore moins, celle de leur utilité clinique. Se développent aussi de nombreuses approches par intelligence artificielle et machine learning (cf. section 6.4).

De nombreux designs d'études ont été proposés pour l'évaluation clinique de la médecine personnalisée. Superchi et al. en dénombre 21 [16]. Cependant plusieurs de ces designs n'ont été utilisés que dans des études abandonnées et/ou finalement non publiées. De plus, parmi le lot, quelques études n'avaient pas pour objectif d'évaluer,[4] mais simplement de donner un cadre à l'utilisation de produits n'ayant pas d'indication. En pratique, à partir des travaux actuellement publiés ayant un objectif de personnalisation ou d'identification de marqueurs prédictifs du bénéfice, il apparait que seules les approches listées ci-dessus sont maintenant utilisées pour produire les preuves de l'intérêt cliniques d'approches personnalisées.

1.2 Le risque de la personnalisation

Hormis les traitements ciblés, la majorité des médicaments sont évalués chez des patients dits « tout venant » (*all comers*), c'est-à-dire sans tenir compte de marqueurs prédictifs, pour la simple raison que ces essais sont réalisés chronologiquement avant que soit évoquée l'existence de tels marqueurs.

*L'utilisation d'un marqueur « prédictif » **invalide***
entraine une perte de chance pour les patients

Dans ce contexte, il est indispensable que l'évaluation clinique apporte la démonstration formelle que les candidats marqueurs permettent de personnaliser le traitement sans entrainer une perte de chance pour les patients. En effet, le plus souvent, le traitement concerné est le traitement qui a été montré supérieur aux autres options thérapeutiques (par son essai pivot). Par exemple en oncologie, les nouveaux médicaments majeurs montrent dans leurs essais randomisés pivots qu'ils améliorent la survie globale des patients par rapport au standard de soins précédent. Identifier alors des marqueurs prédictifs de non-réponse conduit à priver de ce nouveau traitement les patients porteurs du marqueur de non-réponse. A la place on leur propose donc un traitement que l'on sait moins efficace tous patients confondus.

Pour justifier cette pratique il est donc indispensable d'avoir la certitude que ces patients n'auraient pas eu de bénéfice supérieur avec le traitement index et, de ce fait, ils n'auront pas de perte de chance à ne pas le recevoir. Comme l'alternative est moins efficace à un niveau tous patient (*all comers*), il

[4] par exemple, randomisation entre plusieurs molécules expérimentales sans groupe de référence, ne permettant aucune comparaison d'intérêt

faut avoir la démonstration que chez ces patients porteurs du marqueur de « non-réponse » ce traitement alternatif est en fait supérieur, car, schématiquement, le nouveau traitement n'apporte chez eux pas plus de bénéfice qu'un placebo.

> Le durvalumab a été évalué dans l'essai PACIFIC versus placebo comme traitement de maintenance après chimiothérapie dans le cancer du poumon non à petites cellules stade III [17]. Une réduction est démontrée avec un hazard ratio à 0.68. Dans l'analyse en sous-groupe en fonction du niveau d'expression du PD-L1, un hazard ratio de 0.92, non nominalement significatif, est observé dans le sous-groupe <=25%. Si ce résultat (exploratoire) est considéré comme montrant la valeur prédictive du niveau d'expression du PD-L1, les patients avec un niveau <=25% ne recevront pas le durvalumab en maintenance malgré le résultat favorable obtenu tout patient confondu, considérant que chez eux ce traitement n'apporte aucun bénéfice par rapport au placebo.

Ainsi, si le nouveau traitement *N* a démontré sa supériorité tous patients confondus sur le traitement précédent *S* en termes de décès avec un risque ratio de 0.7 (*N versus S*). L'utilisation d'un marqueur prédictif conduit à utiliser *S* chez les patients qui présentent le marqueur de « non-réponse » (« non-bénéfice ») à *N*.

- Si le marqueur est réellement prédictif de la réponse au traitement *N*, ces patients auront le bénéfice apporté par *S* (quantifié par exemple par un essai *S* versus *placebo*) alors que s'ils avaient reçu *N* (inefficace chez eux) ils auraient eu un risque de décès équivalent à celui de patients non traité (comme dans le bras placebo de l'essai validant le bénéfice de *S*). Dans ce cas la personnalisation apporte un bénéfice aux patients concernés.
- En revanche, si ce marqueur est **invalide** (non prédictif en réalité du bénéfice de *N*), ces patients recevront un traitement non optimal pour eux, et auront un **risque de décès 1.43 fois supérieur (1/0.70)** à celui qu'ils auraient eu avec *N*. Cette situation représente une personnalisation à tort, délétère pour les patients concernés, qu'il convient d'écarter par une démonstration formelle de l'utilité médicale du candidat marqueur prédictif (cf. section 2.1).

> Par exemple dans l'essai [18] du géfitinib versus carboplatine+paclitaxel dans le cancer du poumon non à petites cellules (cf. page 33), le hazard ratio sur la PFS dans le sous-groupe EGFR non muté est de 2.85, nominalement significatif, suggérant fortement que pour ces patients le carboplatine + paclitaxel est supérieur au gefinitib, même si l'essai, tous patients confondus, montrait initialement la supériorité du géfinitib. In fine, après des essais complémentaires, il a été considéré que la mutation de l'EGFR était un marqueur prédictif démontré du bénéfice du géfitinib et que priver les patients EGFR non mutés de cette molécule n'entrainait pas de perte de change pour eux (dans cette indication et sans arrivée d'un nouveau traitement plus efficace).

L'enjeu de la certitude de la démonstration est moins crucial dans les situations où plusieurs traitements sont considérés comme apportant un bénéfice clinique « équivalent » et où, le choix d'un traitement ou d'un autre, ne fait pas redouter une potentielle perte de chance. Mais il convient de remarquer qu'avec l'essor des méta-analyses en réseau et autres comparaisons indirectes, la notion de bénéfice « équivalent » entre les médicaments d'une même classe tend à être récusée de plus en plus (nonobstant les limites de ces approches).

L'évaluation clinique doit formellement démontrer que les marqueurs prédictifs permettent de personnaliser le traitement sans potentiellement entrainer de perte de chance.

Cet enjeu va faire que la validation complète d'un marqueur prédictif ne se limite pas à la démonstration de sa valeur prédictive, mais aussi à la démonstration de l'absence de perte de chance pour les patients qui seront finalement récusés pour le traitement (cf. section 4).

Lorsque la personnalisation s'effectue sur le risque de base (cf. section 5), la problématique reste identique et il convient de démontrer l'absence de perte de chance pour les patients pour lesquels la décision sera l'abstention thérapeutique.

1.3 L'hétérogénéité des effets traitements

Le postulat de base de la médecine personnalisée est qu'il existe une hétérogénéité des effets traitement individuels. De nombreuses observations empiriques d'une apparente variabilité de « réponse » aux traitements semblent aller dans ce sens.

> *"The* beneficial effects of most treatments vary across individuals. For example, a treatment that reduces mortality from severe COVID-19 saves some patients who would otherwise have died, but others may die despite treatment, others may survive regardless of treatment, and others may die because of *adverse effects of the treatment" [19]*

> *"In clinical practice some patients benefit more than average from treatment, whereas others do not or may even be harmed" [20]*

Ces « réponses » variables d'un patient à l'autre sont, par exemple, des baisses différentes de la pression artérielle avec la même dose d'un antihypertenseur, la survenue d'un évènement cardiovasculaire ischémique chez des patients et pas chez d'autres malgré l'usage d'un antiagrégant plaquettaire ou la survenue d'une hémorragie majeure que chez certains patients.

Cependant ces constats ne permettent pas de conclure *de facto* à une variabilité inter-sujets du bénéfice du traitement, car ces observations peuvent avoir bien d'autres causes [19, 21].

> *Même s'il est effectivement raisonnable de postuler que l'efficacité des traitements peut être modulée par certains facteurs, au cas par cas, ce type d'observations empiriques ne permet pas de conclure à cette variabilité de l'effet traitement, car cette variabilité de « réponse » peut provenir de tout autre chose que d'une véritable variabilité individuelle de l'effet des traitements [22, 23, 24]. Il y a confusion entre deux concepts différents : le devenir du patient (l'outcome) et ce qu'a causé réellement le traitement dans ce devenir (l'effet causal du traitement).*

Ces « réponses » que l'on peut observer au niveau individuel dans les bras traités des essais cliniques, ou dans la pratique courante, découlent de l'évolution d'un critère de jugement au cours du temps chez des patients traités. De nombreuses sources de variabilités affectent le devenir des patients (ou l'évolution de leur état). Par exemple un paramètre physiologique comme la pression artérielle est soumis à une variabilité interindividuelle, intra-individuelle, aux effets traitements. Chacune de ces variabilités conduira à des valeurs différentes de PA après l'instauration du traitement entre les sujets même si ce traitement a eu le même effet chez tous (cf. Figure 1). Rien que la variabilité intra-individuelle autour d'une valeur centrale peut faire qu'un de ces sujets a au moment de la mesure une valeur plus élevée que sa valeur centrale et un autre plus bas.

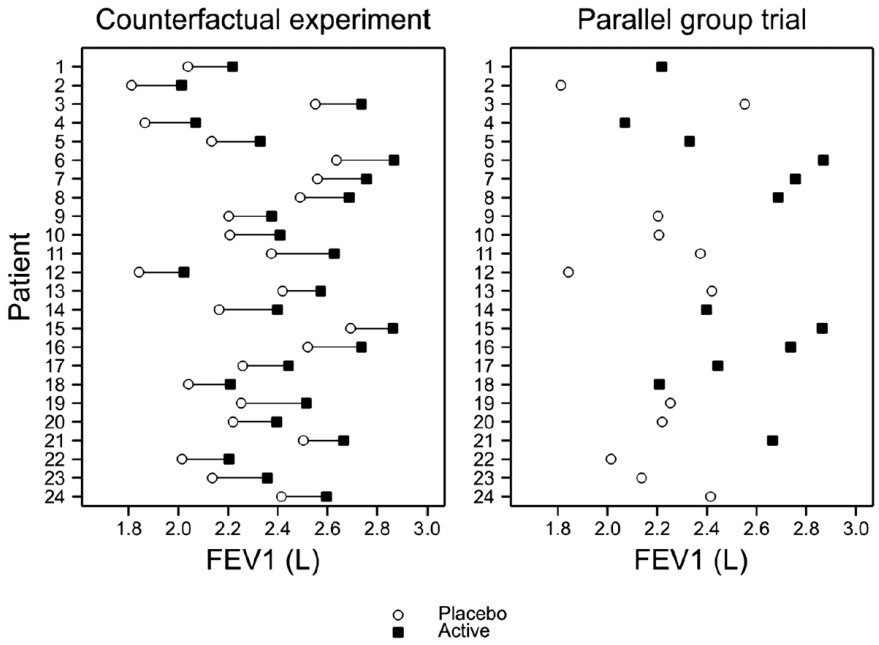


Figure 1 – Les résultats observés dans le groupe traité de l'essai (partie droite de la figure, carrés pleins) pourraient faire penser à une variabilité de la réponse VEMS des patients traités. Mais cette variabilité n'est que le reflet (partie gauche de la figure) de la variabilité intersujet de la valeur de base du VEMS, que la variabilité des VEMS dans le groupe placebo laisse aussi entrevoir. Emprunté à Senn [25].

L'utilisation d'un changement avant-après prendra en compte la variabilité inter-sujets, mais continuera à être affecté par les autres sources de variabilités (régression à la moyenne, évolution naturelle, traitement concomitant, etc.).

> L'identification de différences entre les patients, dans la réponse à un traitement demande au minimum un design randomisé en multiple cross-over afin de prendre en compte la variabilité inter-sujets à l'aide de la randomisation et la variabilité intra-sujets à l'aide de la répétition des périodes de traitement [26] (cf. aussi la figure 213 du supplément 2 du papier décrit ci-dessous).
> Ce design a été utilisé pour montrer une hétérogénéité de réponse à 4 traitements antihypertenseurs [27]. Pour objectiver une variabilité de réponse dans la pression artérielle systolique entre les sujets, chaque patient recevait les 4 traitements dans un ordre aléatoire avec 2 traitements répétés 2 fois (soit 6 périodes de traitement de 7 à 9 semaines par patient). Une hétérogénéité de réponse des patients aux traitements a été mise en évidence avec en moyenne une différence de 4.4mmHg entre le meilleur traitement pour un patient et les autres traitements. Si ces résultats sont confirmés, ils plaideront potentiellement pour une personnalisation du choix du traitement antihypertenseur. En pratique cette personnalisation pourra passer par la réalisation d'essai de taille 1 (N-of-1) (cf. chapitre 22). Cette perspective nécessite cependant de discuter si une optimisation de la baisse de pression artérielle laisse aussi présager d'une optimisation en termes de prévention des évènements cliniques, ce qui débouche sur la question de la valeur de surrogacy de la baisse de pression artérielle et des bénéfices accessoires des différentes classes pharmacologiques d'antihypertenseur.

Pour les événements cliniques, il n'y a pas de notion de variabilité intra-individuelle. Les évènements cardiovasculaires surviennent spontanément de façon erratique chez des patients pourtant très

comparables cliniquement, sans que l'on puisse identifier le ou les déterminants de leur déclenchement. La survenue des évènements indésirables, même très pathognomoniques, pour lesquels le médicament est indéniablement la cause, ne survient pas chez tous les patients, sans que l'on comprenne bien souvent la raison.

Devant cette variabilité irréductible et non inexplicable, les modèles statistiques s'imposent, car ils permettent de modéliser cette variabilité de manière satisfaisante. Car, même si les sujets sont différents entre eux, il y a une assez bonne stabilité, d'un grand groupe de sujets à l'autre, des paramètres statistiques comme la moyenne, la distribution des valeurs, la fréquence, la moyenne des temps de survenue des évènements, etc. Les prédictions faites sur les groupes de sujets sont plus souvent fiables que celle faite au niveau des individus.

Ainsi, à l'exception des essais n-of-1 (qui ne s'appliquent pas si le critère de jugement est un événement), l'observation de l'effet traitement individuel (**ITE**, *Individual Treatment Effect*) est impossible, car cela nécessiterait de connaitre le contrefait, c'est-à-dire ce que serait devenu le patient sans traitement [19, 28]. Le constat d'une variabilité des états et des évolutions des patients sous traitement ne permettent pas de conclure que le traitement a eu des effets différents d'un patient à l'autre.

L'effet du traitement sur un patient particulier est non observable

De ce fait, il est impossible de vraiment personnaliser les effets des traitements et il est donc impossible de faire une véritable médecine personnalisée, au sens strict du terme, basée sur la connaissance de ce qu'apporterait un traitement à un patient donné.

Par contre il est possible de déterminer l'effet moyen d'un traitement (**ATE**, *Average Treatment Effect*), à l'aide d'un essai randomisé (ou une étude observationnelle recevable pour une inférence causale) [29, 30, 31, 32]. L'effet moyen sous-entend l'homogénéité de cet effet à travers la diversité des patients inclus dans l'étude.

Dans ce cadre il est cependant possible de chercher des facteurs qui pourraient modifier l'effet traitement moyen par différentes approches comme l'analyse en sous-groupes (analyse stratifiée) ou la modélisation. Cela introduit le concept de **cATE** (*conditional Average Treatment Effect*) dans le raisonnement : l'effet du traitement ne peut être inféré (estimé) de manière causale que par un effet moyen sur un groupe de sujets (ATE) ; mais cet effet moyen est recherché de manière plus fine en fonction des facteurs qui conditionnent sa valeur (d'où le *conditional*).

L'idée est finalement la même que celle de la médecine personnalisée, identifier les situations où un traitement est plus ou moins adapté aux patients, mais, contrairement à la personnalisation stricto sensu, cette approche est faisable. Il a été proposé de nommer cette approche « médecine stratifiée » compte tenu de son cadre analytique. Dans le langage courant, la distinction entre les termes de médecine stratifiée et médecine personnalisée est rarement faite et le terme médecine personnalisée s'est imposée certainement en vertu de sa valeur marketing.

> Ces notions d'inférence causale, en particulier celle que l'effet traitement est non observable chez un patient donné, déclenchent souvent des discussions impétueuses entre statisticiens et cliniciens.

L'exercice clinique ne peut se faire qu'en raisonnant au niveau des cas isolément. Pour un clinicien, le champ d'observation naturel est un individu, puis un autre, etc. Il doit solutionner le problème d'un patient par les décisions et actions qu'il prendra pour ce patient. Le seul retour de ses actions qu'il aura sera l'observation empirique, chronologique, de l'évolution de l'état du patient. Dans ce cadre, il est évident que, naturellement, du fait de son expérience sensible, on associe l'évolution du patient à la conséquence de ces décisions, actions (en particulier les traitements apportés), même si d'ailleurs on sait parfaitement que cette évolution peut être conditionnée par beaucoup d'autres facteurs non appréhendables.

Il y a donc opposition entre la conceptualisation empirique que fait le clinicien de son expérience sensible et sa conceptualisation des problématiques liées à la variabilité du vivant qui relève du champ de la pensée et de l'abstraction. La pratique médicale est de l'ordre de l'expérience sensible, mais pour appréhender les conséquences de la variabilité du vivant sur cette expérience sensible on doit faire appel aux champs de la pensée et de l'abstraction. Et il est naturel que ces 2 niveaux se heurtent farouchement entre eux, en particulier dans l'instantanéité de l'action.

Ce qui complique encore plus la perception de ces deux niveaux est qu'il existe des effets parfaitement déterministes en médecine, perceptibles par l'expérience sensible, comme les effets toxiques ou des effets pharmacologiques intenses (curares, anesthésiques généraux).

Au niveau individuel, l'incertitude liée à la variabilité du vivant est très difficile à prendre en compte, car dans ce cadre notre cerveau ne gère pas naturellement l'incertitude. Ainsi le facteur de risque devient cause révélée après l'infarctus du myocarde, même si l'élévation du LDL n'augmente que de quelques pourcents le risque d'infarctus. Pour un patient sous statine, la survenue d'un infarctus devient un échec patent de la molécule, même si ce traitement ne fait que réduire, sans l'annuler, le risque d'infarctus. Dans ce cadre, peut-être pour des raisons psychologiques, sociétales, culturelles, il nous semble nécessaire de connaitre la cause, la raison de l'évènement, et même si notre réflexion, en dehors de l'investissement dans un cas particulier, conceptualise très bien l'incertitude qu'il existe dans cette démarche.

2 Les concepts

Le concept fondamental sur lequel repose la médecine stratifiée est celui de modificateur de l'effet (*treatment effect modifier*), aussi appelé marqueur prédictif, et parfois, facteur d'interaction. Dans ce contexte, les termes marqueur, biomarqueur, facteur, variable, covariable sont interchangeables.

2.1 Marqueur prédictif (modificateur d'effet)

Le concept de marqueur/facteur prédictif désigne un facteur associé avec une modification de l'effet du traitement [33, 34, 35]. En d'autres termes, avec un marqueur prédictif binaire (présent/absent, positif/négatif) l'effet du traitement n'est pas le même chez les patients présentant le marqueur (dits patients marqueurs positifs) et chez les patients ne le présentant pas (dits patients marqueurs négatifs).

> *Rien n'interdit que le marqueur prédictif soit de nature continue, modulant progressivement l'effet du traitement. En pratique ces marqueurs sont souvent binarisés pour gagner en simplicité d'utilisation même si cela entraine une perte d'information.*

Le marqueur prédictif idéal serait un marqueur qui permettrait d'identifier les patients chez lesquels le traitement serait sans effet (patients répondeurs/non répondeurs).

> *Il est souvent fait référence aux facteurs de « réponse au traitement ». Même si cette appellation pourrait être comprise comme marqueurs prédictifs, elle fait référence à une approche différente et inappropriée qui est celle du pronostic sous traitement (cf. section 4.1). Cette approche ne permet pas d'identifier des marqueurs prédictifs, car elle ne repose pas sur les effets traitements, mais simplement sur le risque (cf. section 2.2).*

La détermination de l'effet du traitement demande une comparaison entre un groupe traité et un groupe contrôle. Le concept de marqueur prédictif recouvre ainsi deux niveaux de comparaison reposant sur 4 groupes de patients.

Classiquement, l'effet du traitement se mesure par un hazard ratio, un risque ratio (risque relatif), un odds ratio, une différence de moyenne, etc.[5]. Avec le risque ratio, par exemple, le concept de marqueur prédictif signifie que le risque ratio du traitement considéré par rapport à son contrôle n'est pas numériquement identique entre les patients marqueurs positifs et ceux marqueurs négatifs. Schématiquement, cela se matérialise numériquement de la façon suivante :

Patients	Décès de toute cause Risque ratio
Marqueurs positifs	0.8
Marqueurs négatifs	1.0

Dans cet exemple, le traitement entraine une réduction relative du risque de -20% chez les patients présentant le marqueur (marqueur positif) tandis qu'il ne semble pas apporter de bénéfice de survie chez les patients marqués négatifs. Bien entendu, l'analyse formelle de ces résultats nécessite de prendre en considération l'incertitude statistique des estimations (cf. infra).

[5] Cf. dossier indices d'efficacités et analyse des courbes de survie

Comme le risque ratio est issu de la comparaison d'un groupe traité et d'un groupe contrôle, le tableau complet comprend 2 colonnes supplémentaires :

Patients	Risque ratio	Décès de toute cause nombre (%)	
		Groupe traité n=1000	Groupe contrôle n=1000
Marqueurs positifs	0.8	64/400 (16%)	80/400 (20%)
Marqueurs négatifs	1.0	120/600 (20%)	120/600 (20%)

Les données sources sont celles des 4 groupes qui apparaissent ainsi (marqueur positif, traité ; marqueurs positifs, contrôle ; marqueur négatif, traité ; marqueurs négatifs, contrôle). Pour mémoire, le risque ratio de 0.8 est obtenu par la division de 16% par 20%.

Il apparait ainsi que le concept de marqueur prédictif s'apparente tout simplement aux analyses en sous-groupes des essais thérapeutiques. Toutes les variables utilisées pour faire des analyses en sous-groupes permettent d'explorer si celles-ci modifient l'effet du traitement et pourraient donc avoir une valeur prédictive. Néanmoins la démonstration de la valeur prédictive d'une variable va aller bien au-delà d'une simple analyse en sous-groupe compte tenu de toutes leurs limites méthodologiques (cf. Dossier 5 – Les analyses en sous-groupes).

En cas de marqueur n'ayant pas de valeur prédictive, un risque ratio similaire serait observé chez les 2 types de patients, comme dans cet exemple :

Patients	Risque ratio	Décès de toute cause nombre (%)	
		Groupe traité n=1000	Groupe contrôle n=1000
Marqueurs positifs	0.7	56/400 (14%)	80/400 (20%)
Marqueurs négatifs	0.7	90/600 (15%)	120/600 (20%)

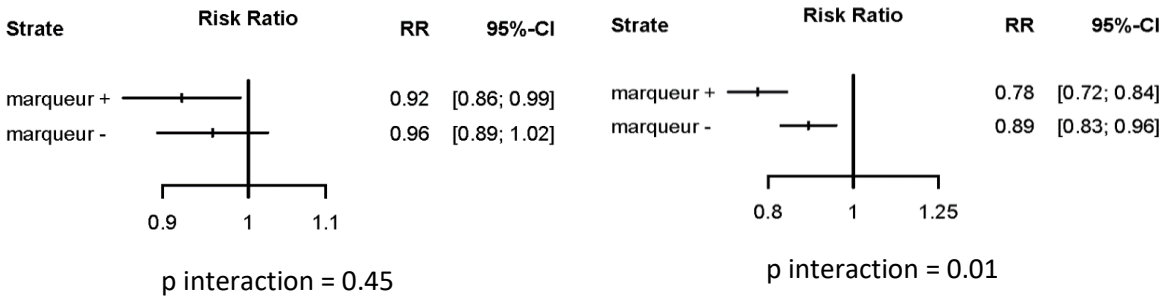
La simple comparaison des effets traitements (estimation ponctuelle) n'est cependant pas suffisante pour conclure à une modification d'effet. Il est nécessaire de prendre en compte l'incertitude statistique de ces deux estimations. En effet, il est nécessaire d'exclure la possibilité que les 2 effets traitements observés ne soient différents que du fait du hasard. Il est donc nécessaire de montrer que la différence d'effet observée entre les marqueurs positifs et les marqueurs négatifs est statistiquement significative, c'est-à-dire qu'elle est bien réelle compte tenu de l'incertitude qui entache chaque estimation.

Intuitivement, il est possible de percevoir que, si les intervalles de confiances des effets traitements se chevauchent largement, il n'est pas possible d'affirmer que les effets sont réellement différents, car, par exemple, la valeur obtenue avec les marqueurs positifs est alors compatible avec les valeurs d'effet possible pour les marqueurs négatifs compte tenu de l'incertitude de l'estimation chez ces derniers patients, et vice versa.

Ce raisonnement graphique, même s'il permet d'intuiter la problématique, n'est que partiellement exact. L'approche formelle de la question « l'effet du traitement est-il statistiquement différent entre les marqueurs positifs et les négatifs » se fait à travers la notion de test d'interaction. Ce test donne une p value de comparaison des effets. Lorsque cette p value est inférieure au seuil de la signification

statistique (5% le plus souvent, mais d'autres valeurs sont parfois rencontrées pour ce test d'interaction) il est alors possible de conclure à une modification de l'effet par le marqueur considéré. Si le test d'interaction est non significatif, il est alors hasardeux de conclure à une différence d'effet comme à une absence de différence (compte tenu de la faible puissance des tests d'interaction le plus souvent [36]).

La figure ci-dessous illustre l'utilisation du test d'interaction. Dans la sous-figure de gauche, le test d'interaction n'est pas statistiquement significatif et ne permet pas de conclure qu'il y a une différence de l'effet entre les marqueurs positifs et les marqueurs négatif (en d'autres termes, compte tenu de leur incertitude respective, le risque ratio de 0.92 ne peut pas être considéré comme différent de 0.96). Dans la sous-figure de droite, le test d'interaction



Ce test d'interaction permet de montrer qu'il existe une interaction entre l'effet du traitement et le marqueur sur la fréquence du critère de jugement (appelé le risque en épidémiologie), ce terme interaction signifiant que l'effet du traitement varie en fonction des modalités du marqueur considéré.

La Figure 2 ci-dessous donne une illustration de cette notion d'interaction. L'ordonnée représente le risque (la fréquence du critère de jugement). Les risques observés dans les quatre groupes (cf. supra) sont positionnés sur cette échelle (les points). Les deux strates de patients définies par le marqueur figurent en abscisse. Dans la strate des patients marqueurs positifs (à droite) le traitement à un effet sur le risque du critère de jugement (le risque est plus faible avec le traitement qu'avec le placebo). Cet effet du traitement peut être quantifié par le risque ratio qui est de 0.4 (rapport des risques, traitement versus placebo). Pour la strate des patients marqueurs négatifs, le traitement a aussi un effet sur le risque puisque celui-ci est, aussi chez ces patients, plus faibles sous traitement que dans le groupe placebo. Cet effet du traitement est caractérisé par un risque ratio de 0.8. Cependant le marqueur n'a pas d'effet sur le risque comme en témoigne la même valeur de risque observé dans le groupe placebo entre les marqueurs positifs et les marqueurs négatifs (il n'est donc pas pronostique). Ainsi la différence de risque chez les patients traités entre les marqueurs positifs et les marqueurs négatifs n'est pas le reflet d'un effet pronostique[6] du marqueur sur le risque lui-même, mais d'une modification par le marqueur de l'effet du traitement, le marqueur interagit sur l'effet du traitement sur le risque.

[6] Cf. section suivante 2.2

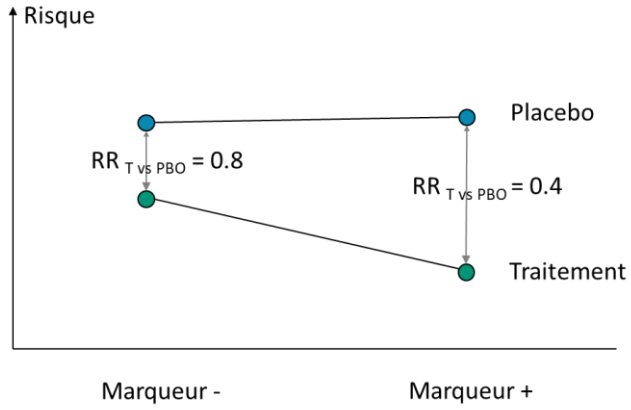


Figure 2 – illustrations du concept d'interaction

L'identification d'un marqueur prédictif nécessite donc la mise en évidence d'une interaction statistiquement significative entre l'effet du traitement et le marqueur. Mais cette condition n'est pas suffisante. Il faut aussi qu'elle témoigne d'une disparition de l'effet du traitement dans une des deux strates

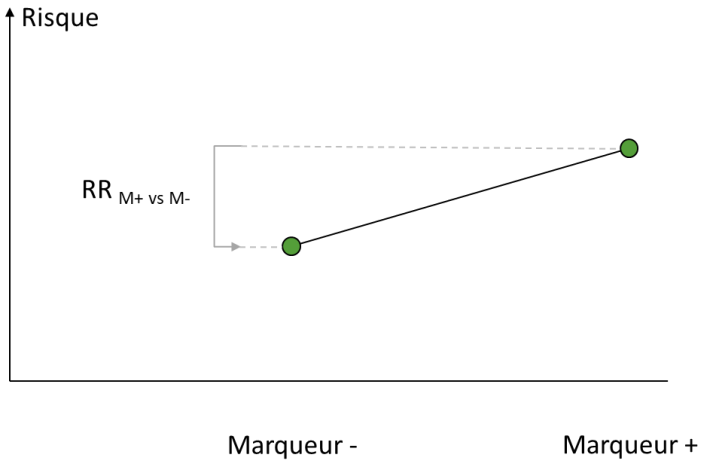
2.2 Marqueur pronostique

Les marqueurs prédictifs ne doivent pas confondu avec les marqueurs pronostiques [33, 37].

Un facteur pronostique est un facteur qui modifie le risque. La notion de traitement n'apparait pas directement dans le concept de marqueur/facteur pronostique. Ainsi pour identifier un facteur pronostique il suffit de comparer le risque entre les patients présentant le marqueur avec celui des patients ne le présentant pas. Les données nécessaires au calcul sont bien moins complexes que celles nécessaires pour un marqueur prédictif :

Décès de toute cause Nombre (%)		Risque ratio
Marqueurs positifs n=202	Marqueurs négatifs n=205	
20(9.9%)	10 (4.9%)	2.03

Dans cet exemple le risque de décès est multiplié par 2.03 chez les patients présentant le marqueur par rapport à ceux ne le présentant pas. Le risque ratio calculé est celui marqueurs positifs versus marqueurs négatifs. Sur une échelle de risque, le concept de facteur pronostique s'illustre de la manière suivante :



La logique voudrait que la valeur pronostique d'un marqueur soit recherchée chez des sujets non traités, car il s'agit d'un concept d'histoire naturelle de la maladie. Cependant cette recherche « pure » est impossible à effectuer dans bien des cas, car elles conduiraient à priver les patients de traitement ayant montré leur bénéfice. De ce fait, la plupart des études pronostiques s'effectuent chez des sujets traités et documentent ipso facto la valeur pronostique **sous traitement** du marqueur/facteur étudié. Cependant cette valeur pronostique sous traitement n'est pas la valeur prédictive du facteur de la réponse au traitement ou de son bénéfice (cf. supra).

La distinction entre facteur pronostique et facteur prédictif est importante dans le champ de la médecine personnalisée, car de nombreuses études parlent de « facteurs de réponse au traitement » qui est, non pas le concept de marqueurs prédictif du bénéfice, mais une recherche de simples facteurs pronostiques de la réponse chez des patients tous traités. Il s'agit donc de facteurs qui discriminent le risque sous traitement et non pas le bénéfice du traitement. Les études basées sur cette approche sont inappropriées de facto pour la recherche des facteurs prédictifs du bénéfice et leurs résultats présentent de nombreuses limites qui ne permettent pas de guider la personnalisation des traitements (cf. section 4.1).

Marqueurs prédictifs et marqueurs pronostiques sont deux concepts différents. Rien n'oblige un marqueur prédictif d'être pronostique et vice-versa. Il arrive cependant qu'un marqueur soit les deux simultanément. Dans la figure issue de la méta-analyse des essais de fibrinolytique sur données individuelles [38] on peut noter :

- Le délai depuis le début des symptômes (hours from onset) est un marqueur prédictif important (test d'interaction très significatif, présenté en dernière colonne) et il n'est absolument pas un marqueur pronostique comme on peut le constater en comparant les risques de décès dans le groupe contrôle qui sont tous de l'ordre de 10%
- La pression artérielle systolique est très pronostique et n'a pas de valeur prédictive du bénéfice sur la mortalité (odds ratio similaire quelle que soit la PAS)
- L'âge est à la fois prédictif du bénéfice et pronostique

E »z'

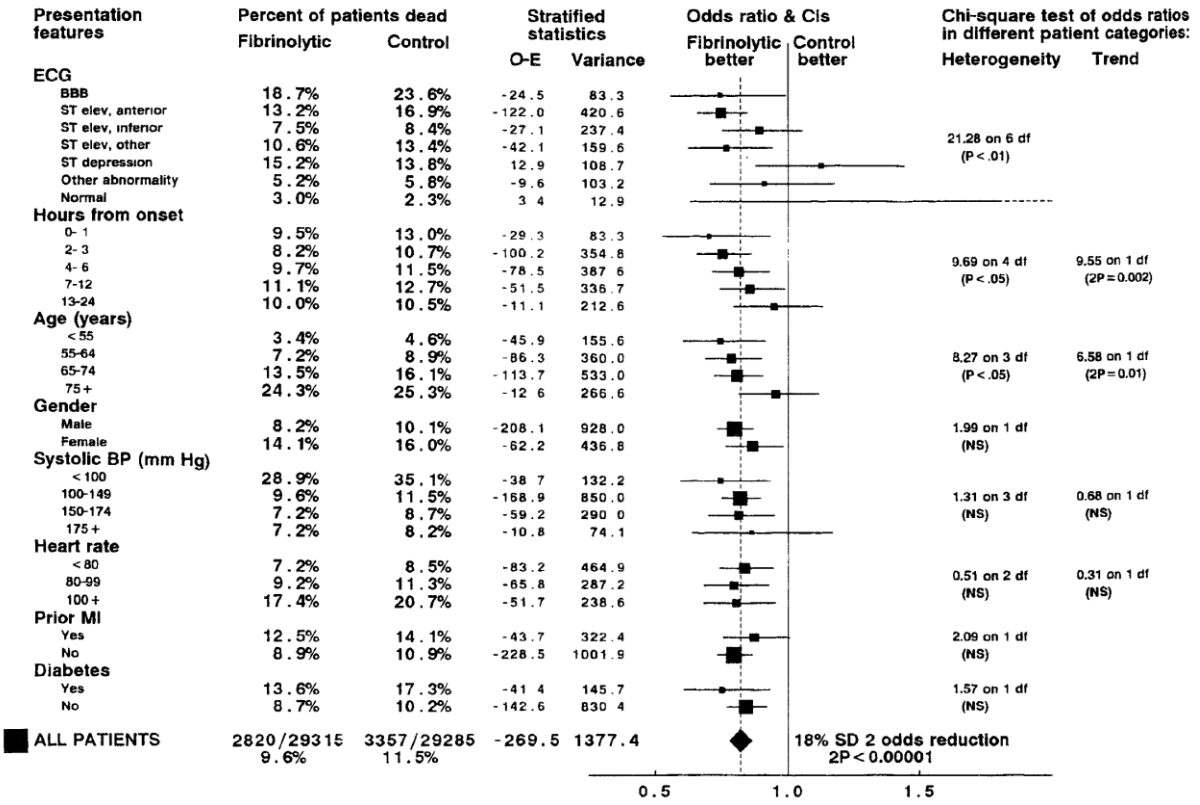


Figure 1: **Proportional effects of fibrinolytic therapy on mortality during days 0–35 subdivided by presentation features**

3 Les thérapeutiques ciblées

3.1 Définition

De nombreuses molécules sont conçues d'emblée pour agir sur un variant moléculaire particulière de leur cible d'action. Cette maximisation de leur affinité pour une forme moléculaire spécifique est recherchée, car cette variante joue un rôle particulier dans la maladie des patients qui en sont porteurs (mutation activatrice d'un dérèglement d'une voie de signalisation par exemple). Ce ciblage est obtenu lors du criblage moléculaire en cherchant parmi les molécules ayant toute une action pharmacologique sur la cible, celle qui présente la plus grande affinité.

> Le sotorasib est un inhibiteur des RAS GTPase, mais qui vise une forme mutée particulière de la protéine K-Ras encodée par le gène KRAS (mutation KRAS p.G12C)

> L'osimertinib est une molécule TKI inhibitrice de l'EGFR qui inhibe sélectivement des formes mutées de cette protéine (mutation EGFR-TKI–sensitizing et EGFR T790M) [39] comme le montre le tableau ci-dessous donnant l'IC50 (nM) de l'activité de phosphorylation de l'EGFR in vitro pour différentes formes d'EGFR mutées et sauvages (WT).

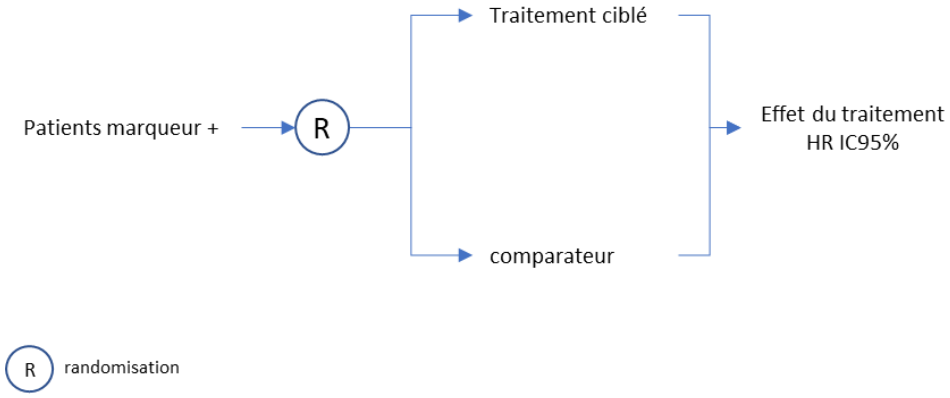
A

	H1975 (L858R/ T790M)	PC-9 VanR (ex19del/ T790M)	PC-9 (ex19del)	H3255 (L858R)	H1650 (ex19del)	LoVo (WT)	A431 (WT)	NCI-H2073 (WT)
AZD9291	15 (10, 20)	6 (3, 13)	17 (13, 22)	60, 49	14, 12	480 (320, 720)	2376, 1193	1865 (872, 3988)

Cette notion de « ciblage » est relative, car certaines molécules ont des affinités assez fortes pour de multiples cibles. Ils peuvent donc avoir une forte affinité pour une cible d'intérêt sans être pour autant « sélectif » pour cette cible. C'est le cas notamment de certains inhibiteurs de protéines kinases, comme le sorafénib par exemple.

3.2 Évaluation clinique des thérapies ciblées

Les essais cliniques nécessaires pour démontrer le bénéfice clinique de ces molécules ciblées sont tout à fait classiques. La présence de la variante est un des critères de sélection de l'essai qui n'inclue que des patients présentant cette variante (patients marqueur positif) :



> L'essai FLAURA [40] a évalué l'osimertinib versus un TKI de référence (gefitinib ou erlotinib) en première ligne du traitement du cancer du poumon non à petite cellule avancé. Pour être inclus dans cet essai, les patients devaient présenter une mutation de l'EGFR (exon 19 deletion ou L858R allele). La méthodologie de l'essai est classique : essais randomisés en 2 bras parallèles, sans double aveugle, avec la PFS et l'OS comme critère de jugement décisionnels. Cet essai a démontré la supériorité de l'osimertinib chez ces patients en termes de survie.

En effet, ici, le but de l'évaluation clinique n'est pas d'établir l'intérêt de la variant moélculaire pour choisir parmi différentes options dans la thérapeutique de la pathologie, mais de démontrer le bénéfice clinique de la molécule chez les patients présentant cette altération moléculaire. Il s'agit d'un parti pris de développement et l'évaluation clinique porte sur le couple molécule et présence de la cible simultanément. Par principe il est considéré que la molécule ne peut apporter aucun bénéfice chez les patients non porteurs de l'altération moléculaire[7].

> *La situation est donc complètement différente de celle où un biomarqueur prédictif (modificateur de l'effet) est recherché pour une molécule qui a déjà démontré un bénéfice chez des sujets tout-venant (« all comers »)*

3.3 Utilisation d'essais non comparatifs

En oncologie particulièrement, les traitements ciblés font souvent l'objet d'études non comparatives (monobras) comme « essai » pivot, principalement car ce type de design moins-disant (cf Dossier 12 - Etudes monobras et essais à contrôle externe) étaient suffisant jusqu'à présent pour obtenir un enregistrement accéléré par la FDA[8].

Le caractère ciblé des molécules ne lève en rien les limites méthodologiques de ce design qui ne permet pas de produire des preuves du bénéfice clinique des traitements. (cf. livre blanc SFPT).

> Par exemple le sotorasib qui est un inhibiteur de tyrosine kinase ciblant la protéine KRAS produite en cas de mutation 12 a fait l'objet d'une première étude dans le cancer du poumon non comparative, codebreak

[7] Ce qui est une hypothèse forte, car l'affinité de ces molécules peut ne pas être nulle pour les autres variants de la cible.

[8] La position de la FDA vis-à-vis de ces études monobras a changé début 2023 avec la publication de recommandations insistant sur la nécessité de recourir à l'essai randomisé pour les accès précoces en oncologie.

100 [41]. Avec une série de 174 patients en 2ème ligne ou plus, tous traités par sotorasib, il a été observé un taux de réponse objective de 40.7% et une médiane de PFS de 6.3 mois.

Dans ces essais monobras (non comparatifs) en oncologie le critère de jugement est la réponse tumorale (*objective response rate*, ORR). Ce critère reflète l'évolution, après l'administration du traitement, de la « taille de la tumeur » par rapport à la taille initiale. Aux stades métastatiques il n'est pas attendu d'évolutions favorables de la taille de la ou des tumeurs [42]. L'observation chez certains patients d'une évolution stable ou d'une régression de la taille tumorale semble ainsi refléter une activité du produit. Mais cette activité « anticancer » ne signifie pas que le produit apporte un bénéfice clinique, par exemple en termes de survie. Il n'a pas été montré que le taux de réponse était un prédicteur (une sorte de surrogate) de la survie en oncologie. De plus, la question n'est pas tant de savoir si un nouveau produit à une activité anticancer, mais s'il est supérieur aux traitements disponibles. L'absence de comparaison ne permet pas de répondre à ce questionnement.

Il existe des exemples de traitements entrainant un taux de réponse important, mais pour lesquels il n'a pas été mis en évidence de bénéfice sur la survie (ibrutinib dans le lymphome à cellules du manteau) et a contrario des produits qui apportent un bénéfice de survie par rapport à l'existant avec un taux de réponse faible (avelumab dans le cancer urothélial métastatique).

	Taux de réponse objective (ORR)	Effet sur la survie (HR)
ibrutinib dans le lymphome à cellules du manteau	69% (NCT01236391)	1.07 [0.81; 1.40] (NCT01776840)
avelumab dans le cancer urothélial métastatique	9.7% (NCT02603432)	0.69 [0.56; 0.86] (NCT02603432)

Un autre argument couramment avancé pour justifier le recours aux essais non-comparatifs est le caractère « innovant » de ces molécules. Même si ces molécules ont un mécanisme d'action nouveau et potentiellement prometteur, le terme innovant est excessif à ce niveau[9]. En effet il est avant tout demandé à un traitement d'apporter un réel bénéfice clinique au patient et non pas d'être basé sur un nouveau mécanisme d'action. Un médicament est innovant que lorsqu'il a démontré qu'il apportait un bénéfice clinique substantiel dans une situation où aucun autre traitement ne l'avait fait auparavant.

Or, un nouveau mécanisme d'action ne préjuge pas de la possibilité d'apporter un bénéfice clinique innovant aux patients (cf. livre blanc SFPT). Il existe maintenant plusieurs cas [44] où les résultats d'essais randomisés de molécules de ce type n'ont pas pu mettre en évidence de bénéfice clinique ou n'ont montré qu'un bénéfice modeste. Ainsi, même avec ces molécules ciblées en oncologie, dépendante d'une altération moléculaire particulière, le caractère prometteur du mécanisme d'action ne donne pas la garantie, à lui seul, que la molécule apportera le bénéfice clinique recherché. Les essais randomisés, comme SHIVA [17] ou MATCH-NCI, entrepris pour montrer une efficience globale des thérapies ciblées ont été négatifs jusqu'à présent (cf. section 3.4). Ainsi la disponibilité d'une molécule à forte affinité pour la forme d'une protéine considérée comme driver dans le processus tumoral ne garantit pas que cette molécule apportera automatiquement un bénéfice clinique.

[9] Le NEJM évite soigneusement que le terme innovant apparaisse dans les publications qu'il accepte concernant ces molécules et fait simplement utiliser le terme prometteur : « Adeno-associated virus (AAV) gene therapy is a **promising treatment approach** for hemophilia B,…. » [43].

CODEBREAK 200, NCT04303780, est un essai randomisé comparant sotorasib versus docetaxel chez des patients porteurs d'un cancer du poumon non à petite cellule au stade métastatique ayant reçu précédemment une immunothérapie et présentant la mutation ciblée par le sotorasib [45].
Cet essai n'a pas mis en évidence de bénéfice en termes de mortalité (hazard ratio 1.01 [0.77; 1.33]), mais seulement en termes de PFS avec un hazard ratio de 0.66 [0.51; 0.86] et une différence de médiane de 1.1 mois.

3.4 Efficience globale des traitements ciblés

L'intérêt général de la médecine personnalisée est débattu [6, 46, 47, 48], en particulier pour les thérapies ciblées qui ne concernent parfois qu'une faible proportion des patients.

> *Le cancer du poumon non à petites cellules est le cancer pour lesquels le plus grand nombre de mutations potentiellement actionnable a été identifié. La proportion des patients porteurs d'au moins une de ces mutations varie entre 25 et 50% entre les régions du monde. Certaines de ces mutations ont une fréquence de quelques pourcents.*

Il reste donc un besoin de traitements au bénéfice ubiquitaire. Or le développement des nouveaux traitements se focalise principalement sur les thérapies ciblées en raison de l'enthousiasme concernant cette voie et peut être de sa rentabilité supérieure. Paradoxalement les patients non porteurs d'une mutation activable deviennent atteints, pour ainsi dire, d'une pathologie orpheline (abandonnée par la recherche thérapeutique).

Le démembrement des pathologies classiques (comme le cancer du poumon non à petites cellules) en de nombreuses « pathologies rares » sur la base de ces variants moléculaires ne couvrira le besoin global que lorsque tous les malades pourront être mis dans une de ces cases spécifiques. En attendant, des traitements à « spectre large » restent encore indispensables (comme les chimiothérapies, les immunothérapies type PD-(L)1 dans une certaine mesure, etc.)

L'impact de la personnalisation sur l'efficience globale de la prise en charge d'une maladie est pour l'instant difficile à cerner. En théorie, cette efficience globale va dépendre de la proportion des patients concernée et de la supériorité de l'efficacité des traitements proposé chez eux par rapport au traitement standard.

L'évaluation de cette efficience globale a fait l'objet de 2 études en oncologie SHIVA et MATCH-MCI (cf. encadré ci-dessous). Dans ces études les patients étaient randomisés entre un groupe utilisation des thérapies ciblées et un groupe traitement conventionnel. Dans le groupe expérimental, les patients recevaient un traitement ciblé lorsqu'un variant moléculaire ciblable avec les traitements retenus par l'étude était présente.

Aucune différence de PFS et de survie n'a été montrée entre ces 2 approches dans ces études ne permettant pas de démontrer l'efficience globale de la personnalisation. Les limites sont représentées par l'ancienneté des thérapeutiques proposées aux patients qui ne comportait pas de nombreuses possibilités actuellement disponibles.

Ces études montrent cependant que la disponibilité d'une molécule ciblée sur une altération moléculaire ne supplante pas, par principe, les thérapeutiques standards et qu'il convient d'évaluer chaque proposition (couple altération et molécule) au cas par cas.

L'essai SHIVA [49] est un essai randomisé d'oncologie qui a comparé une stratégie de recours systématique à des traitements ciblés moléculairement par rapport à une stratégie conventionnelle de prise en charge des mêmes cancers. Il ne s'agit donc pas d'un essai d'évaluation d'un traitement ciblé bien

particulier, mais plutôt une évaluation globale des traitements ciblés disponibles à l'époque dans leur globalité et quel que soit le cancer. L'hypothèse sous-jacente était que ces thérapies ciblées étaient, par principe, tellement adaptées à la maladie pour laquelle elles sont prescrites, qu'elles apportaient un bénéfice bien supérieur aux approches traditionnelles et cela, quel que soit le cancer, le stade. L'essai n'a pas montré de différence de PFS entre les stratégies.

NCI-impact [50] est un autre essai du même type. Il n'a pas permis de montrer la supériorité de l'approche ciblée par rapport à l'approche contrôle. IMPACT II (NCT02152254) est aussi un autre essai de ce type en cours de recrutement.

4 L'évaluation des marqueurs prédictifs

Un marqueur prédictif est un biomarqueur conditionnant une modification de l'efficacité ou de la sécurité d'un traitement (cf. section 2.1).

> *Cette problématique est souvent présentée comme la prédiction de la « réponse au traitement », cette réponse étant alors vue comme une évolution favorable chez des patients traités. La section 4.1 démontrera les limites de cette approche et sa non-adéquation avec la définition de marqueur prédictif.*

Le terme prédictif introduit un certain degré d'ambigüité car il peut aussi être utilisé dans d'autre contexte en référence à une valeur pronostique comme dans la médecine prédictive qui cherche à prédire la survenue ou non de la maladie chez des sujets encore sains[10].

Le vocable **modificateur d'effet** (*treatment effect modifier*) est plus approprié mais peine à s'imposer, en dehors du cercle de l'épidémiologie.

On rencontre de plus en plus l'utilisation de l'intelligence artificielle dans cette rechercher des facteurs prédictifs. Le principe reste le même, le machine learning remplaçant les approches de modélisation statistiques conventionnelles pour prédire la « réponse » ou le bénéfice. Les problématiques d'évaluation et de validité restent aussi identiques, ces techniques ne garantissant pas automatiquement la fiabilité des résultats produits. Les aspects particuliers à l'utilisation de l'IA pour la personnalisation des traitements seront abordés dans la section 6.4.

4.1 Etudes « treament only », valeur pronostique sous traitement

4.1.1 Principe des études « treatment only »

Fréquemment un marqueur « prédictif » de réponse à un traitement est cherché en comparant un critère de réponse entre un groupe de patients présentant le marqueur et un autre groupe ne présentant pas le marqueur, tous les patients recevant le traitement concerné. Le critère de réponse peut être de diverse nature. Par exemple en oncologie il peut s'agir soit du taux de réponse objective, soit de la PFS ou même de l'OS.

> Le polymorphisme ERCC1/RRM1 (CT/AC) a été envisagé comme marqueurs de réponse aux sels de platine [51]. L'étude a comparé la PFS entre des patients traités par sels de platine pour un cancer non à petites cellules du poumon présentant ce polymorphisme (CT/AC dans le graphique ci-dessous) et les patients ne le présentant pas (other).

[10] https://fr.wikipedia.org/wiki/M%C3%A9decine_pr%C3%A9dictive

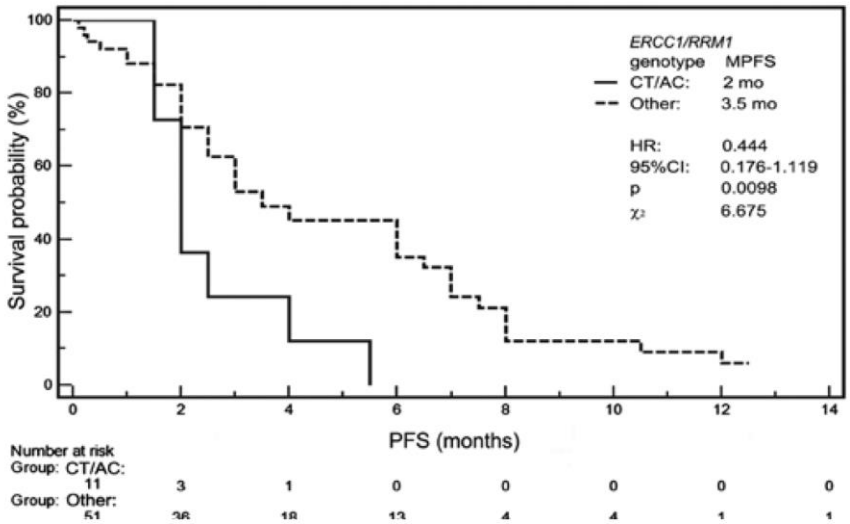


Figure 4. Kaplan-Meier curves of PFS probability according to the *ERCC1* 19007T>C and *RRM1* -37A>C polymorphisms combination.

[11]

Un hazard ration de 0.444 est obtenu en faveur d'un risque instantané de progression ou de décès plus important avec ce polymorphisme que sans. La conclusion pourrait être que ce polymorphisme est un marqueur prédictif de réponse si ce type d'approche n'avez pas des limites intrinsèques.

Cependant ce design d'étude ne permet pas d'appréhender la valeur **prédictive** du marqueur considéré, car il ne permet pas d'objectiver la modification de l'effet du traitement par le biomarqueur.

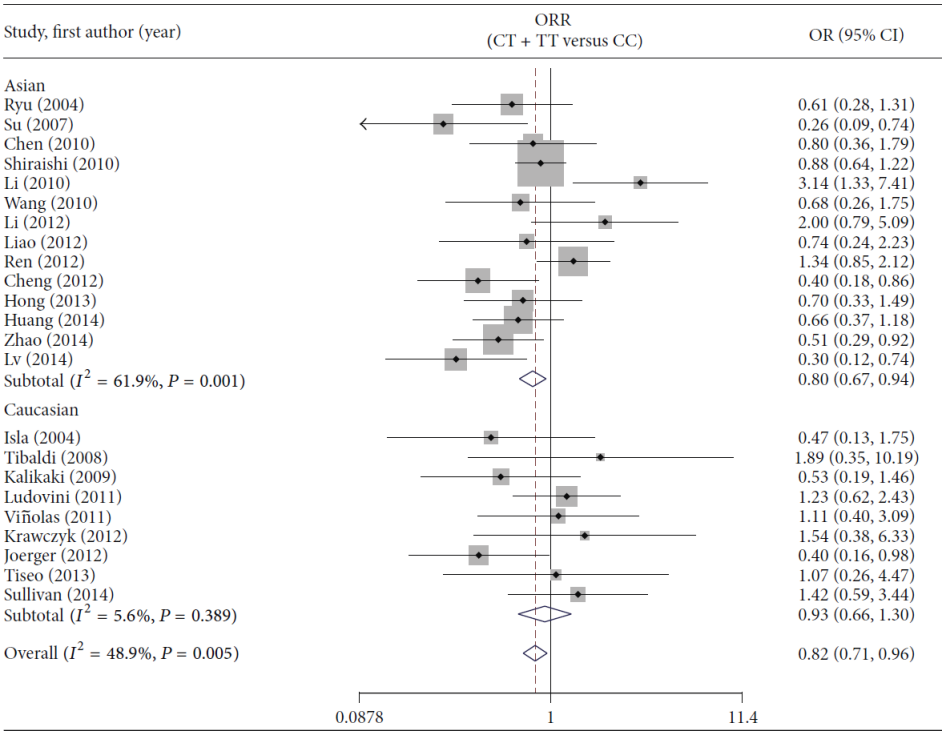
Cette approche montre seulement une valeur **pronostique** du biomarqueur chez des patients traités, ce qui n'implique pas que le biomarqueur module l'effet du traitement. Le marqueur peut par exemple être pronostique de la même façon avec ou sans le traitement considéré. Dans ce cas les patients qui seront étiquetés « répondeurs » au traitement ne seront, en fin de compte, que des patients spontanément à meilleur pronostic que les autres (avec ou sans traitement). La valeur prédictive sera confondue avec la valeur pronostique du marqueur. Son utilisation en pratique reviendra alors à ne traiter que les patients à bon pronostique alors que les autres patients bénéficient aussi du traitement[12]. Le ciblage du traitement sur un tel marqueur sera alors contreproductif et entrainera une perte de chance pour les patients récusés, cela en raison d'une conceptualisation inadaptée. Compte tenu de cette perte de chance possible, il est donc important de ne baser la pratique que sur des résultats forts, mesurant directement la valeur prédictive des biomarqueurs candidats (cf. section 4.2).

Les limites de cette approche liées au fait que la valeur pronostique sous traitement n'est pas la valeur prédictive, sont bien illustrées par le cas du polymorphisme du ERCC1 présenté comme étant prédictif de la réponse aux sels de platines à l'issue de nombreuses études de pronostic sous traitement (cf. exemple

[11] Attention, comme cette étude, ce graphique est mal fait. En abscisse il s'agit du temps (et non pas de la PFS qui est survie donc un %) et l'ordonnée est justement la PFS (% des patients initiaux survivant sans progression). Cet exemple montre bien le faible niveau de rigueur méthodologique que peut avoir ce type d'étude (cf. livre blanc).

[12] Et avec un bénéfice absolu plus important, car leur risque de base est plus élevé (mauvais pronostic)

précédent). En plus de l'étude que nous avons présentée précédemment, de nombreuses autres études de même design ont été publiées. La méta-analyse de toutes ces études conforte leurs résultats. [52].



Cependant, lorsque la valeur prédictive de ce polymorphisme a été recherchée de manière appropriée (cf. section 4.2.3) par un essai randomisé avec un design d'interaction [53], il s'est avéré que ce polymorphisme n'était pas un marqueur prédictif de l'efficacité des sels de platine (cf. section 4.2).

Pour écarter à minima cette problématique, ces études « treatment only » doivent être associées à une étude la valeur pronostique en soi du marqueur, c'est-à-dire chez des patients non traités (ou traités sans le traitement d'intérêt). Comme il s'agit de montrer l'absence de valeur pronostique du marqueur, ces études doivent avoir une puissance statistique importante pour obtenir la précision d'estimation permettant d'écarter une possible valeur pronostique.

4.1.2 Sensibilité au biais de publication

Les études mettant en œuvre cette approche « treatment only » sont réalisées en grand nombre compte tenu de leur relative facilité de réalisation. La nature rétrospective de ces études expose à un risque important de biais de publication. En l'absence d'association entre le marqueur d'intérêt et le pronostic, sur le nombre, plusieurs études montreront malgré tout une association statistiquement significative (résultat faux positif dû au hasard). Ces études seront certainement publiées contrairement aux autres conduisant à un biais de publication classique.

La valeur pronostique du polymorphisme du ERCC1 (cf. supra), qui semblait bien établie à l'issu de la méta-analyse, n'a pas été retrouvée dans l'essai randomisé prospectif mis en place pour chercher sa valeur prédictive d'interaction (cf. section 4.2) : « Unlike many retrospective studies, neither ERCC1 nor XPF were prognostic markers for OS or PFS » [53].

Table 2. ERCC1 or XPF As a Prognostic Marker: Association With OS or PFS

	OS			PFS		
	No. of Patients (events)	HR (95% CI)	*P*	No. of Patients (events)	HR (95% CI)	*P*
Nonsquamous histology						
ERCC1						
Cisplatin and pemetrexed	230 (198)	1.05 (0.80 to 1.40)	.72	230 (215)	1.12 (0.85 to 1.46)	.43
Paclitaxel and pemetrexed	234 (201)	1.14 (0.86 to 1.51)	.36	234 (212)	1.14 (0.86 to 1.49)	.36
Combined*	464 (399)	1.11 (0.91 to 1.35)	.32	464 (427)	1.13 (0.93 to 1.37)	.22

Ces résultats peuvent faire suspecter un biais de publication sur les études rétrospectives d'analyse « treatment only » induisant une méta-analyse positive à tort.

4.1.3 Utilisation en pratique

Malgré ses limites, cette approche est largement employée et même préférentiellement utilisée par rapport aux **approches adaptées** [54].

Une analyse exploratoire de l'étude monobras basket KEYNOTE-158 du pembrolizumab a comparé, entre autres, le taux de réponse objective entre les patients ayant une charge mutationnelle de la tumeur (TMB) importante par rapport aux autres [55]. La conclusion de cette analyse typiquement de pronostique sous traitement, exploratoire au demeurant, précise bien que ses résultats ne sont de l'ordre que de l'association et ne font que suggérer que le TBM puisse être un prédicteur de l'efficacité du pembrolizumab
D'autres publications, de même type méthodologique, prennent moins de précautions scripturales dans leur conclusion.

Une étude fait l'hypothèse qu'une déficience de réparation des mésappariements de l'ADN pourrait conditionner l'efficacité des immunothérapies dans le cancer colorectal [56]. Cette hypothèse est testée prospectivement à l'aide d'une recherche de valeur pronostique sous traitement. La conclusion de l'article est cependant trop forte compte tenu des limites de cette approche " treatment only" et cela malgré la nature hypothético-déductive de l'analyse "This study showed that mismatch-repair status predicted clinical benefit of immune checkpoint blockade with pembrolizumab".

Le terme « valeur prédictive » est fréquemment utilisé à tort dans ces études (comme par exemple dans la ref [57])

4.2 Études évaluant la valeur prédictive d'un marqueur

L'identification correcte des marqueurs prédictifs s'effectue en comparant l'estimation de l'effet du traitement obtenu chez des sujets marqueurs positifs à celle obtenue chez des patients marqueurs négatifs afin de mettre en évidence une interaction entre le marqueur et l'effet traitement (cf. section 2.1). Ces deux estimations de l'effet du traitement peuvent provenir, soit d'une analyse en sous-groupes dans un essai pivot, soit d'un essai dédié à cette recherche utilisant le design d'interaction.

Bien qu'étant la seule adaptée à la mise en évidence de marqueurs prédictifs, l'approche basée sur l'interaction est peu utilisées en pratique, ce qui soulève un certain de considérations éthiques et scientifiques [54].

Les analyses en sous-groupes peuvent être des analyses anticipées et intégrées au plan de contrôle du risque alpha global (dans une démarche d'analyse séquentielle hiérarchique par exemple) lorsque la recherche d'un marqueur prédictif fait partie explicitement des objectifs de l'étude ; ou des analyses en sous-groupes exploratoires « ordinaires », l'essai n'ayant pas dans ses objectifs la validation d'un marqueur prédictif.

> *La recherche des marqueurs prédictifs peut se présenter dans deux situations, très différentes méthodologiquement : soit à priori au moment de l'évaluation d'un nouveau traitement, soit à postériori, après que le traitement a démontré son bénéfice dans un essai incluant des patients « tout venant » («* all comers *») sans sélection sur le/les marqueurs prédictifs.*

4.2.1 Analyses en sous-groupes exploratoires

Souvent l'hypothèse qu'un biomarqueur pourrait être un marqueur prédictif du bénéfice est formulée après la réalisation du ou des essais pivots du traitement concerné. Elle émerge alors que ces essais ont déjà démontré le bénéfice du traitement sur une population de patients non sélectionnés sur ce candidat biomarqueur (« *all comers* »). S'il est possible de déterminer à postériori le statut des patients de l'essai vis-à-vis de ce marqueur candidat, la recherche de sa valeur prédictive est alors envisagée à postériori, de manière rétrospective à partir de ces données.

Par exemple pour les marqueurs génétiques, cette détermination à postériori est possible si du matériel biologique des patients de l'essai a été conservé. Si le candidat marqueur est de nature biologique, sa valeur est peut-être déjà dans les fichiers des données individuelles des patients de l'essai.

Cette recherche à postériori va prendre la forme d'une réanalyse en sous-groupe des données de l'essai en fonction de la présence ou de l'absence[13] du marqueur.

> **Exemple d'analyse en sous-groupe à postériori**
>
> L'essai ATTRACTION-2 démontre que le nivolumab augmente la survie chez des patients ayant un cancer de l'estomac avancé en 3me ligne ou plus. Une analyse post hoc a été réalisée pour évaluer plusieurs candidats marqueurs prédictifs. [58].
>
> *"This study is a subset analysis with patients enrolled in the ATTRACTION-2 study … This study aimed to investigate the predictive values of potential biomarkers such as tumor PD-L1 expression, tumor MSI status, tumor EBV infection, TMB, blood NLR, and serum Na to provide objective guidance in identifying patients with clinical benefits to nivolumab."*

[13] Le plus souvent il a binarisation du marqueur même si celui-ci a une valeur continue à l'origine dans le but d'obtenir un outil simple en pratique. S'il n'y a pas binarisation l'analyse prend alors la forme d'une recherche d'interaction marqueur*traitement dans une modélisation des données.

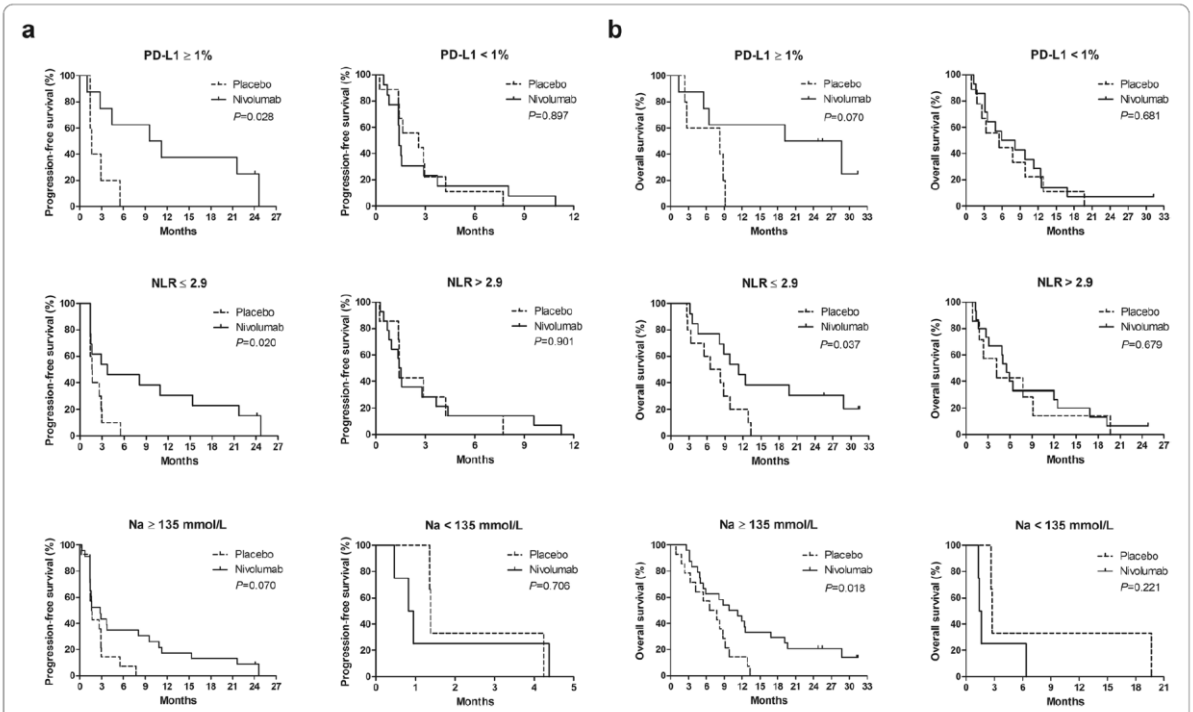
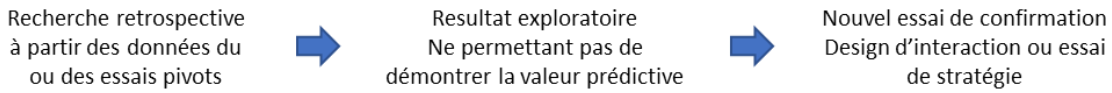


Fig. 2 PFS (**a**) and OS (**b**) in nivolumab vs. placebo group according to each potential predictive biomarker

"Conclusion: Tumor PD-L1 expression, blood NLR, and serum Na could be predictive biomarkers for the efficacy of nivolumab in previously treated cases of AGC."

4.2.1.1 Limites méthodologiques de l'approche exploratoire

Cette recherche de la valeur prédictive à postériori va se heurter à plusieurs problématiques méthodologiques qui limiteront fortement les résultats produits et l'éventuelle conclusion à la valeur prédictive du marqueur. Ces limites, impossibles à lever à travers une approche rétrospective, conduisent à la nécessité de faire une nouvelle étude prospective, spécialement conçue pour mettre en évidence la valeur prédictive du marqueur avec un design d'interaction (cf. section 4.2.3 ou exemple du gefitinib ci-dessous). Cette nouvelle étude permettra de confirmer ou d'infirmer avec solidité l'hypothèse éventuellement soulevée lors de l'analyse exploratoire rétrospective des données de l'essai pivot.



Les limites méthodologiques de la recherche de valeur prédictive d'un marqueur en rétrospectif sont les suivantes.

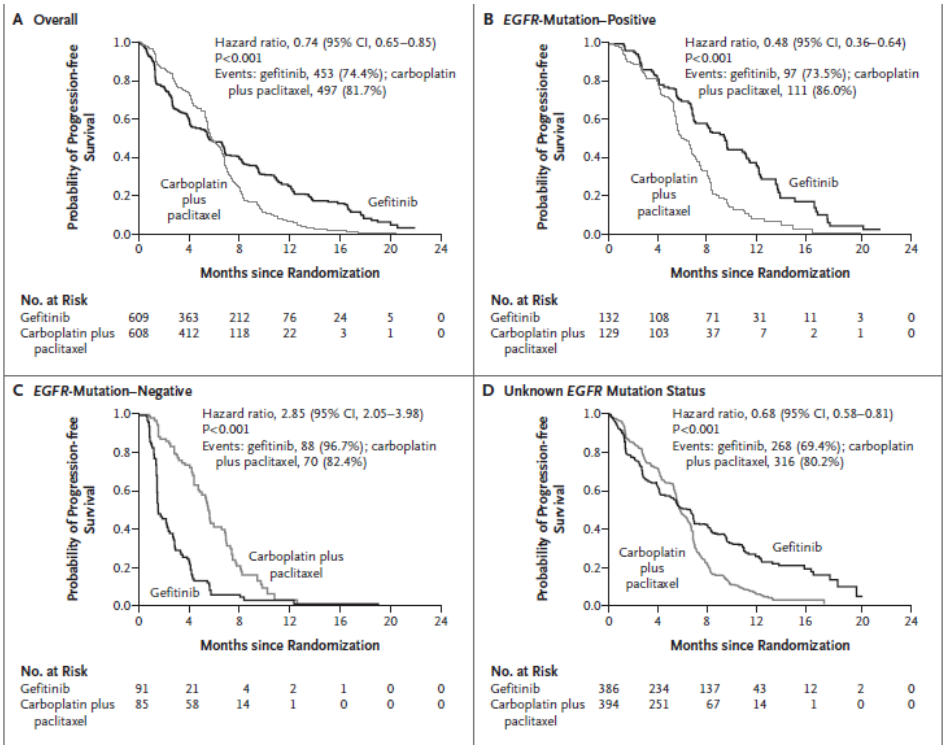
La problématique est identique à celle de toute démarche rétrospective (cf. livre blanc De la nécessité de la méthodologie dans l'évaluation des médicaments) : possibilité de HARking où l'hypothèse a été générée à partir d'une analyse initiale non révélée des mêmes données qui serviront à la tester ; possibilité de p-hacking en choisissant en fonction des résultats obtenus le seuil de dichotomisation des valeurs du marqueur pour en faire un marqueur binaire, choix de la méthode d'analyse, du jeu de données, etc.

La recherche de la valeur prédictive va être effectuée à l'aide d'analyse de sous-groupe qui, par définition, n'étaient pas prévues au protocole et au plan d'analyse statique de l'essai pivot. Ainsi les problématiques de multiplicité des comparaisons, induisant inflation du risque alpha et beta, n'ont pas pu être prises en compte dans le plan de contrôle du risque alpha global de l'essai. La puissance n'a pas été assurée dans les sous-groupes ni au niveau de la recherche de l'interaction.

> *Par moment, il est aussi évoqué une problématique liée à l'absence de stratification de la randomisation sur ce marqueur. Ce point est encore discuté, mais il n'a pas de conséquence en termes de biais de sélection (une randomisation imprévisible garantit l'absence de biais de sélection sur tous les sous-groupes). En revanche des déséquilibres d'effectifs entre groupes de traitement peuvent survenir entre les sous-groupes limitant fortement la précision des estimations et la puissance des comparaisons. Mais même en cas de stratification, les sous-groupes ordinaires posent les mêmes problèmes précédents.*
>
> *La pré-spécification au protocole des sous-groupes est souvent invoquée pour garantir la solidité de leurs résultats. Les pré-spécification n'empêche pas l'inflation des risques alpha et beta et ne solutionne donc en, rien la principale problématique des analyses en sous-groupes .*

Toutes ces limitations font que ces analyses post hoc ne sont, au mieux, qu'exploratoires et exposent à un risque de fausse découverte important (cf. exemple du clopidogrel ci-dessous, section 4.2.1.2). On rejoint ici la question de toutes les analyses en sous-groupes qui semblaient prometteuses et qui n'ont pas pu être confirmées dans des essais prospectifs spécialement mis en place pour confirmer des résultats [59].

> **Exemple du gefitinib**
>
> Le premier essai du gefitinib dans le cancer du poumon a inclus des patients « all comers » [18]. Son résultat a été concluant, mais un croisement précoce des courbes de survies de PFS pouvait faire imaginer la présence de deux sous-populations de patients, bénéficiant de manière inversée du traitement (bénéfice dans une et effet délétère dans l'autre). Simultanément était apparu qu'une mutation sur le récepteur EGFR pourrait conditionner l'efficacité de ce produit, et les analyses en sous-groupes suivant la présence ou non de cette mutation, déterminée de manière rétrospective, était compatibles avec cette hypothèse.



Comme l'ensemble de ces éléments ne reposaient que sur des analyses et hypothèses post-hoc, leur nature de fait exploratoire ne permettait pas de conclure et de restreindre l'utilisation du gefitinib aux patients EGFR muté. Pour confirmer cette hypothèse et démontrer le réel intérêt de ce produit, deux autres essais randomisés ont été entrepris [60, 61], n'incluant que ces patients cibles. Les résultats furent concluants, amenant à cibler le produit sur la mutation du récepteur EGFR.

De nombreux exemples illustrent la nécessité impérative de confirmer prospectivement les résultats des analyses en sous-groupes exploratoires.

Une phase 3 du reldesemtiv dans la SLA a été arrêtée précocement pour futilité lors de sa 2eme analyse intermédiaire[14] . Cette étude avait pourtant été mise en place à la suite d'un résultat très prometteur sur une analyse en sous-groupe d'une précédente étude du produit. La reproductibilité des résultats des sous-groupes dans une étude subséquente est en général faible comme illustre cet exemple comme tant d'autres. [59, 62]. Lorsqu'un essai de confirmation est entrepris, cela ne pose pas de problème (sauf pour le sponsor).

Un premier essai du solanezumab dans la maladie d'Alzheimer s'avère négatif [63]. Cependant un résultat « intéressant » est observé par une analyse en sous-groupe chez les patients ayant une forme légère de la maladie (https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2015.06.1893). Un nouvel essai est entrepris pour confirmer ce résultat et il s'avèrera lui aussi négatif [64].

14 https://www.globenewswire.com/news-release/2023/03/31/2638547/0/en/Cytokinetics-Announces-COURAGE-ALS-Met-Criteria-for-Futility-at-Second-Interim-Analysis.html

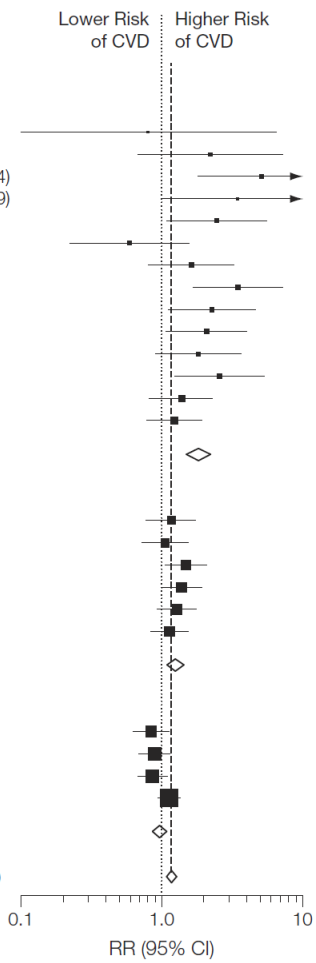
4.2.1.2 Retour sur les limites des analyses « treatment only »

L'exploration du génotype du CYP2C19 comme marqueur prédictif de l'efficacité du clopidogrel sur les évènements cardiovasculaires [65] donne un bon exemple des limites des approches "treatment only" et de leur déconnexion sémantique au concept de marqueur prédictif. Une méta-analyse par Holmes et al. [65] présente simultanément la synthèse d'études « treatment only » et des résultats de recherche d'interaction post hoc dans les essais pivots du clopidogrel.

Dans les analyses « treatment only », ce génotype est trouvé associé avec le risque d'évènements cardiovasculaires sous clopidogrel. Dans ce cadre, il est a été conclu que ce génotype permettait de prédire la non-réponse au clopidogrel (en raison d'une modification du métabolisme du clopidogrel conduisant à une faible concentration plasmatique du métabolite actif).

Figure 2. Meta-analysis of *CYP2C19* Genotype and Risk of Composite Cardiovascular Outcome in Individuals Treated With Clopidogrel: "Treatment-Only" Analysis

	Cases, No./Total No.		RR (95% CI)
	*2-*8	*1 or *17	
1-99 CVD Events			
Ono et al,[35] 2011	2/131	0/71	NA
Yamomoto et al,[48] 2011	5/62	0/36	NA
Malek et al,[31] 2008	1/21	5/84	0.80 (0.10-6.49)
Worrall et al,[47] 2009	4/24	6/80	2.22 (0.68-7.23)
Jeong et al,[29] 2011	NR/162	NR/104	5.14 (1.82-14.54)
Yuan et al,[50] 2011	11/137	3/130	3.48 (0.99-12.19)
Campo et al,[24] 2011	10/81	11/219	2.46 (1.09-5.57)
Trenk et al,[45] 2008	5/245	19/552	0.59 (0.22-1.57)
Sawada et al,[37] 2010	13/42	11/58	1.63 (0.81-3.28)
Collet et al,[25] 2009	15/73	11/186	3.47 (1.68-7.21)
Giusti et al,[26] 2009	15/247	14/525	2.28 (1.12-4.64)
Shuldiner et al,[38] 2009	14/67	16/160	2.09 (1.08-4.04)
Malek et al,[32] 2010	10/56	20/205	1.83 (0.91-3.68)
Oh et al,[34] 2011	23/1011	10/1135	2.58 (1.23-5.40)
Harmsze et al,[27] 2011	NR/200	NR/525	1.40 (0.82-2.30)
Komarov et al,[30] 2011	22/108	48/291	1.23 (0.78-1.94)
Subtotal: I^2 = 33% (95% CI, 0%-65%)			1.83 (1.50-2.23)
100-199 CVD Events			
Tello-Montoliu et al,[43] 2011	NR/NR	NR/NR	1.17 (0.78-1.75)
Pare et al (ACTIVE-A),[36] 2010	29/139	83/421	1.06 (0.73-1.54)
Mega et al,[16] 2009	46/395	83/1064	1.49 (1.06-2.10)
Anderson et al,[21] 2009	48/350	89/900	1.39 (1.00-1.93)
Bhatt et al,[22] 2009	54/722	99/1706	1.29 (0.94-1.77)
Sibbing et al,[41] 2009	52/680	121/1805	1.14 (0.83-1.56)
Subtotal: I^2 = 0% (95% CI, 0%-75%)			1.26 (1.09-1.45)
≥200 CVD Events			
Pare et al (CURE),[36] 2010	52/650	178/1880	0.84 (0.63-1.14)
Tiroch et al,[44] 2010	60/248	184/680	0.89 (0.69-1.15)
Simon et al,[42] 2009	76/635	218/1573	0.86 (0.68-1.10)
Wallentin et al,[46] 2010	149/1388	332/3516	1.14 (0.95-1.37)
Subtotal: I^2 = 40% (95% CI, 0%-80%)			0.97 (0.86-1.09)
Overall: I^2 = 60% (95% CI, 38%-75%)			**1.18 (1.09-1.28)**



Comparison of any copy of *CYP2C19* *2 through *8 to wild-type (*1) or *17 (reference) is stratified according to the number of events per study (1-99, 100-199, ≥200). Data-marker sizes indicate the weight applied to each study using fixed-effects meta-analysis. CVD indicates cardiovascular disease; NR, not reported; RR, relative risk.

De manière rétrospective, il a été possible d'établir ce génotype chez les patients de 4 essais pivots comparant clopidogrel au placebo dans différents contextes cliniques : ACTIVE-A, CURE, CHARISMA, CLARITY grâce à du matériel biologique qui avait été conservé pour une partie des patients inclus dans

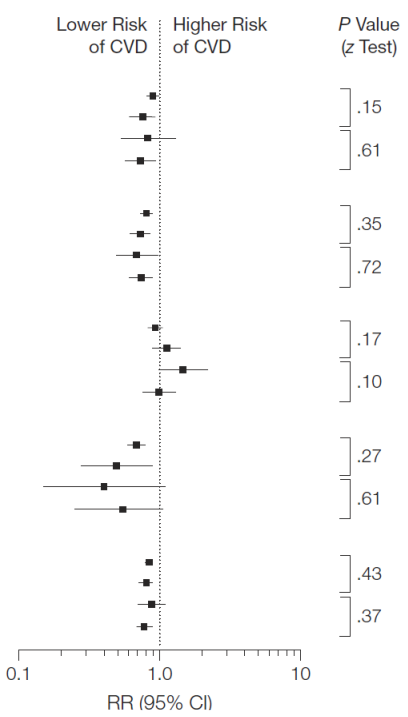
ces essais. Ces données permettent de faire une analyse en sous-groupe en fonction du génotype du CYP2C19 de l'efficacité sur les évènements cardiovasculaires du clopidogrel par rapport au placebo, et de tester l'interaction. Cette analyse correspond donc à la définition d'un marqueur prédictif contrairement aux analyses précédentes « treatment only ».

Les résultats de cette analyse en sous-groupe sont présentés dans la figure 5 de la publication (reproduite ci-dessous) et ne permettent pas de mettre en évidence une interaction entre ce génotype et l'effet traitement du clopidogrel, pour les 4 essais et pour leur regroupement en méta-analyse. Ces résultats ne permettent pas de conclure que ce génotype est un marqueur prédictif.

Figure 5. Analysis of *CYP2C19* Genotype on Composite Cardiovascular End Points and Major Bleeding in Randomized Trials Where Both Clopidogrel and Placebo Groups Were Genotyped: "Effect-Modification" Analysis

A Composite cardiovascular outcome

Trial	Cases, No./Total No. Clopidogrel	Placebo	RR (95% CI)	*P* Value (*z* Test)
ACTIVE-A[36, 54]				
Original RCT	832/3772	924/3782	0.89 (0.81-0.98)	.15
Genetic substudy	112/560	153/574	0.75 (0.61-0.93)	
*2 or *3	29/139	35/139	0.83 (0.54-1.28)	.61
*1 or *17	83/421	118/435	0.73 (0.57-0.93)	
CURE[19, 36]				
Original RCT	582/6259	719/6303	0.80 (0.72-0.90)	.35
Genetic substudy	230/2530	311/2486	0.73 (0.62-0.85)	
*2 or *3	52/650	78/673	0.69 (0.49-0.96)	.72
*1 or *17	178/1880	233/1813	0.74 (0.61-0.89)	
CHARISMA[20, 22]				
Original RCT	534/7802	573/7801	0.93 (0.83-1.05)	.17
Genetic substudy	153/2428	138/2434	1.11 (0.89-1.39)	
*2 or *3	54/722	36/708	1.47 (0.98-2.21)	.10
*1 or *17	99/1706	102/1726	0.98 (0.75-1.28)	
CLARITY-TIMI 28[33, 53]				
Original RCT	262/1752	377/1739	0.69 (0.60-0.80)	.27
Genetic substudy	NR/NR	NR/NR	0.49 (0.28-0.88)	
*2	NR/NR	NR/NR	0.40 (0.15-1.10)	.61
*1	NR/NR	NR/NR	0.55 (0.25-1.05)	
Summary				
Original RCT	2210/19585	2593/19625	0.84 (0.79-0.89)	.43
Genetic substudy	>495/>5518	>602/>5494	0.80 (0.72-0.89)	
*2 or *3	>135/>1511	>149/>1520	0.87 (0.70-1.09)	.37
*1 or *17	>360/>4007	>453/>3974	0.78 (0.69-0.89)	



B Major bleeding event

Meta-analysis of risk of (A) composite cardiovascular outcome and (B) major bleeding event, comparing clopidogrel with placebo, stratified by the following: findings from original randomized clinical trials (RCTs), genetic substudy, and *CYP2C19** allele status into any copy of *2 or *3 and *1 or *17. The *P* value reflects the *z* test for interaction between subgroups, comparing original RCT and genetic substudy, which assesses the representativeness of the genetic substudy to the original cohort, and *2 or *3 compared with *1 or *17, which tests for effect modification of the effect of clopidogrel vs placebo by *CYP2C19* genotype. CVD indicates cardiovascular disease; RR, relative risk.

Cette nette **discordance** entre l'éventuelle conclusion à une valeur prédictive évoquée par l'analyse « treatment only » et l'absence d'interaction dans l'analyse en sous-groupe montre bien les limites des analyses « treatment only » et **la non-congruence des concepts de facteur pronostique sous traitement et de facteur prédictif (modificateur d'effet traitement)**.

Nous retrouverons cet exemple plus tard, car des essais de stratégie ont été entrepris dans ce domaine afin d'évaluer directement l'utilité médicale d'une personnalisation de ces traitements sur le génotype du cytochrome P450 2C19 (cf. section 6.5).

4.2.1.3 Limites de la recherche post hoc de marqueurs prédictifs

La question d'identifier un marqueur prédictif survient souvent dans le contexte d'un nouveau traitement dont l'étude pivot montre un bénéfice[15] dans la population incluse, mais de taille modeste, peu cliniquement pertinente. Émerge alors l'idée que ce faible bénéfice résulte du fait que seulement une sous fraction de la population étudiée bénéfice du traitement, ce qui conduit à la recherche d'un marqueur permettant d'identifier ces patients hypothétiques. Cependant, cette démarche est très hypothétique, relevant plus du vœu pieux que de la démarche scientifique hypothético-déductive. Cette hypothèse, qu'il n'existe qu'une partie de la population qui tire un bénéfice substantiel du traitement, est formulée complètement à postériori, sans aucune prémisse. Lors de la conception de l'essai, les patients à inclure n'ont pas été déterminés « au petit bonheur la chance », mais parce que le rationnel de l'étude laisser prévoir un même bénéfice du traitement pour tous ces patients[16]. Ainsi formuler à postériori l'hypothèse que le traitement n'apporte un bénéfice qu'à une partie des patients est purement du domaine de l'invention créative et en opposition totale avec l'hypothèse initiale de l'essai (qui elle n'a pas été purement inventée, mais déduite des connaissances établies), sauf si de nouveaux éléments physiopathologiques ou pharmacologiques ont été établis entre temps. [24].

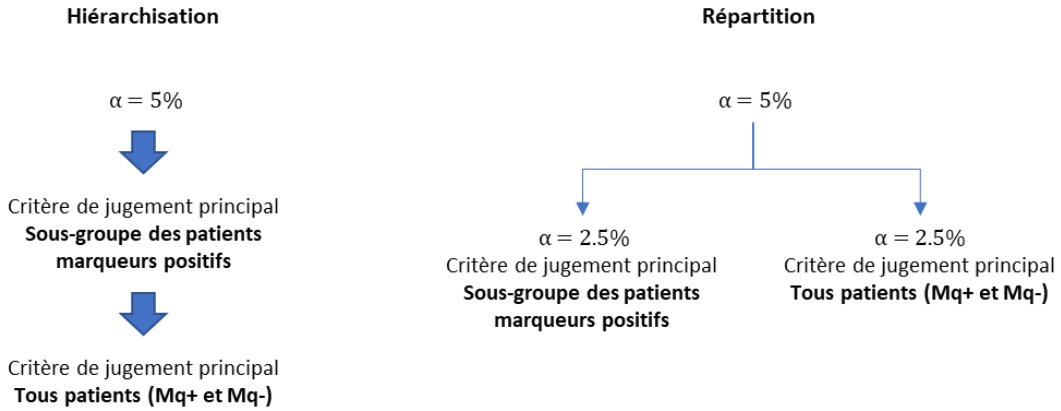
4.2.2 Analyses en sous-groupes de confirmation

La recherche de manière anticipée et prospective de la valeur prédictive d'un marqueur au sein même de l'essai pivot du traitement permet de solutionner toutes les problématiques méthodologiques rencontrées dans la validation à postériori (cf. section 4.2.1). Tous les principes méthodologiques de la recherche de preuves formelles peuvent être ainsi mis en œuvre : respect de la démarche hypothético-déductive avec une formulation à priori de l'hypothèse que le marqueur est prédictif, le sens de l'interaction sera aussi préspécifié par hypothèse ; intégration de la recherche de l'interaction marqueur-traitement dans le plan de contrôle du risque alpha global ; calcul de la puissance nécessaire pour conclure sur cette interaction ; etc.

Ces essais ont ainsi pour objectif de démontrer en premier le bénéfice clinique du traitement, mais aussi d'évaluer un marqueur prédictif candidat. Pour cela, ces essais, incluent des patients aussi bien marqueur positif que marqueur négatif. Très souvent la randomisation est stratifiée sur ce marqueur (et dans ce cas ce type d'étude est proche du design d'interaction, cf. section 4.2.3).

La recherche du bénéfice du traitement va être effectuée chez les patients marqueurs positifs (qui sont en théorie ceux le plus susceptible d'avoir un bénéfice du traitement si celui s'avère efficace) et aussi chez les patients marqueurs négatifs (ou chez tous les patients, quelle que soit la valeur du marqueur). Cette multiplicité de comparaison est alors prise en compte dans le plan de contrôle du risque alpha global de façon tout à fait conventionnelle soit par hiérarchisation soit par répartition du risque alpha.

[15] Le même raisonnement survient aussi quand l'essai a été négatif (cf. dossier sur les analyses en sous-groupes)
[16] Parfois des considérations marketing entrent aussi en ligne de compte.



De plus en plus, un plan complexe de contrôle du risque alpha est employé mêlant hiérarchisation, répartition et recyclage du risque alpha (cf. exemple IMvigor 110 ci-dessous).

Exemple de répartition du risque alpha

L'erlotinib a été évalué versus placebo comme traitement de maintenance après chimiothérapie dans le cancer du poumon par l'essai SATURN [66]. Le critère de jugement principal était la PFS. À priori il était envisagé que le niveau d'expression de l'EGFR pouvait être un marqueur prédictif du bénéfice de l'erlotinib (avec un bénéfice uniquement en cas de surexpression de cette protéine mesurée par immunohistochimie). Cependant il était aussi possible que le traitement apporte un bénéfice indépendamment de ce marqueur et l'essai a inclus les patients positifs et négatifs.

"We designed the phase 3, placebo-controlled Sequential Tarceva in Unresectable NSCLC (SATURN; BO18192) study to investigate the effect of erlotinib as maintenance therapy on PFS in patients with non-progressive disease following first-line platinum-doublet chemotherapy. We assessed PFS in the overall population and in patients with tumours that over-express EGFR."

La recherche de la valeur prédictive de ce marqueur a été intégrée dans le plan de contrôle du risque alpha global en répartissant le risque alpha global de 5% entre le sous-groupe EGFR positif (2%) et l'analyse globale de tous les patients (3%) :

The alpha level of 5% was split between the two coprimary endpoints: 3% for all patients and 2% for patients with EGFR immunohistochemistry-positive tumours.

Un bénéfice de l'erlotinib en termes de PFS est observé aussi bien dans le sous-groupe des patients EGFR positifs (HR 0·69, 0·58–0·82; p<0·0001) que chez tous les patients (HR 0·71, 95% CI 0·62–0·82; p<0·0001). On peut remarquer que le test de l'efficacité de l'erlotinib chez les patients négatifs n'était pas prévu, car celui-ci ne correspond pas à une éventualité de démonstration spécifique de l'intérêt du traitement.

La logique de ces essais, bien illustrée par l'exemple précédent de l'erlotinib, est celle des essais pivots, c'est-à-dire produire des démonstrations de l'intérêt du traitement (d'arguments permettant d'obtenir une AMM ou une place dans la stratégie thérapeutique). Le marqueur prédictif pressenti est présent dans ces études pour gérer l'éventualité que le traitement n'ait une activité qu'en cas de

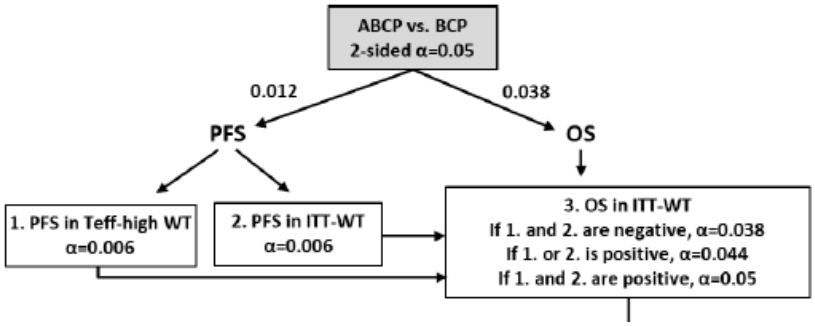
présence de ce marqueur. Dans ce cas, l'étude le démontrera formellement ; contrairement à un essai qui n'aurait pas anticipé cette éventualité et où, les patients marqueurs positifs ne seraient qu'un sous-groupe ordinaire, incapable de fournir des démonstrations. Ainsi si l'essai est concluant sur l'ensemble des patients, la question du marqueur prédictif disparait (en partie, cf. infra). En revanche, si l'essai échoue à montrer un bénéfice sur l'ensemble des patients, mais donne un résultat statistiquement significatif chez les patients marqueurs positifs, il démontrera formellement le bénéfice du traitement chez ces patients. À ce moment la valeur prédictive du marqueur aura aussi été démontrée. Ces 2 résultats permettront de positionner le traitement et son marqueur prédictif dans la stratégie thérapeutique.

Cette approche a cependant une **limite** qui est celle de ne pas pouvoir conclure si le traitement des patients marqueurs négatif est justifié lorsque l'essai conclut à l'effet du traitement chez tous les patients [67]. Même si le traitement n'a aucun effet chez les marqueurs négatifs, le résultat global de l'essai peut être concluant, car il est principalement drivé par le bénéfice obtenu chez les patients marqueurs positifs. Il existe une dilution de l'effet chez les marqueurs positifs par l'absence d'effet chez les marqueurs négatifs, mais qui n'est pas suffisante pour conduire à un résultat global non concluant (par exemple si la proportion de marqueurs négatifs est faible ou si l'effet chez les marqueurs positifs est très important). Comme dans ces essais, le résultat spécifique obtenu chez les marqueurs négatifs n'est pas rapporté, il est difficile d'analyser ce point. Et même s'il est rapporté il va être difficile de conclure à l'exclusion de ces patients de la population cible, car on se retrouve devant un résultat de sous-groupe ordinaire, non nominalement significatif, avec toutes les limites de l'interprétation de ce type de résultat.

De plus en plus fréquemment, ces essais pivots couplant l'évaluation d'un nouveau traitement avec celui d'un ou plusieurs marqueurs prédictifs potentiels de son bénéfice utilisé des plans de contrôle du risque alpha global optimisés donc complexes.

> L'essai IMpower 150 a évalué l'atezolizumab en première ligne du traitement du cancer du poumon non à petite cellule non épidermoïde métastatique [68]. Il était envisagé qu'un marqueur Teff-high pourrait être prédictif sur la base des résultats d'un essai précédent (OAK). La démonstration spécifique du bénéfice chez les marqueurs positifs (Teff-high) a été intégrée à un plan de contrôle du risque alpha global qui repose sur une répartition du risque alpha global entre les 2 critères de jugements PFS et OS. Pour la PFS, ce risque alpha attribué (0.012) est ensuite à nouveau réparti, de manière égale, entre la recherche de l'effet de l'atezolizumab sur la PFS uniquement chez les patients Teff-high WT (alpha 0.006) et la recherche de l'effet de l'atezolizumab chez tous les patients (ITT-WT). L'effet de l'atezolizumab est ensuite recherché chez tous les patients avec une réallocation du risque alpha en provenance de la PFS. Le risque alpha disponible pour l'OS dépend ainsi des résultats obtenus au niveau delà PFS.

Figure S1. Alpha-Spending Algorithm. The study design and α-spending algorithm to control for type I error for PFS and OS in the primary analysis populations for the statistical testing of ABCP vs. BCP followed by ACP vs. BCP is depicted. ABCP denotes atezolizumab + bevacizumab + carboplatin + paclitaxel; BCP, bevacizumab + carboplatin + paclitaxel; ITT, intention-to-treat; OS, overall survival; PFS, progression-free survival; Teff, T-effector gene-signature expression; WT, wild type.



4.2.3 Le design d'interaction biomarqueur-traitement

Le design d'interaction est conçu pour permettre de démontrer qu'un candidat marqueur est bien un marqueur prédictif (modificateur de l'effet traitement). L'objectif spécifique de ces essais est de confirmer que le marqueur est un modificateur de l'effet du traitement d'intérêt en démontrant l'interaction d'où l'appellation.

Le principe de ce design est illustré par la Figure 3. Après inclusion, le marqueur est déterminé chez tous les patients. Ensuite les patients positifs et les patients négatifs sont séparés en deux sous-essais où ils seront randomisés dans les 2 cas entre le traitement d'intérêt A (pour lequel on souhaite confirmer que le marqueur est prédictif du bénéfice) et le traitement de référence S. Comme il n'y a pas d'objectif de déterminer l'effet de A par rapport à S chez tous les patients, il s'agit plus de deux sous essais séparés que d'une randomisation stratifiée, mais en pratique cela revient au même. Seule la finalité de la stratification est différente[17]. En théorie les effectifs des deux strates devraient être aussi déterminés pour garantir une certaine puissance à la recherche de l'interaction biomarqueur-traitement.

L'effet traitement de A est déterminé pour les patients marqueurs positifs (Mq+) et pour ceux qui sont marqueurs négatifs (Mq-). Ces deux estimations permettent ensuite de tester l'interaction statistique recherchée.

[17] Dans un essai où l'objectif est de confirmer le bénéfice du traitement chez tous les patients, la stratification est seulement un moyen d'optimisation de la précision et de la puissance de la comparaison statistique entre le traitement expérimental et son contrôle chez tous les patients.

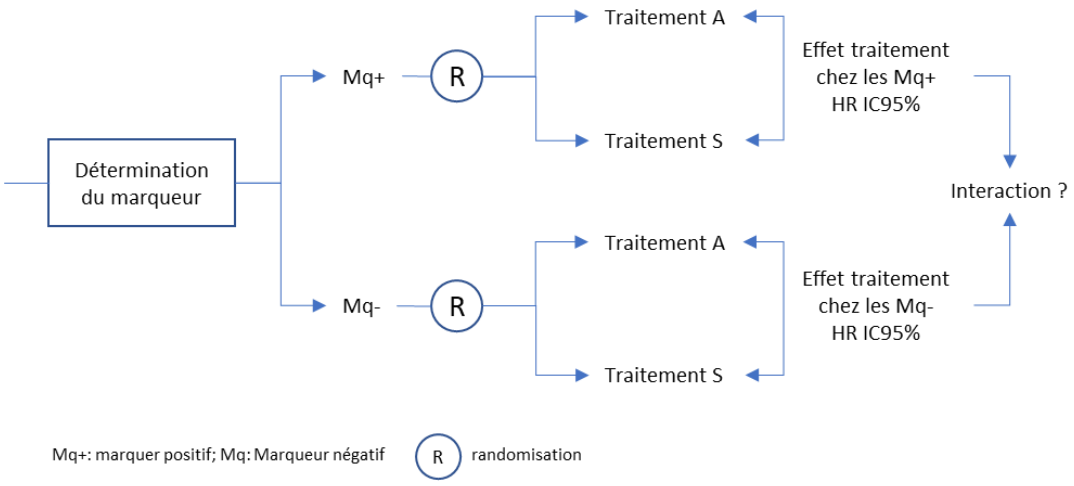


Figure 3 – Principe du design d'interaction biomarqueur-traitement

4.2.3.1 Exemple - Essai PROSE

Une signature protéomique a évalué comme candidat marqueur prédictif de la réponse à l'erlotinib dans le cancer du poumon non à petites cellules à l'aide d'un essai randomisé d'interaction dénommé PROSE [69]. La randomisation entre erlotinib ou chimiothérapie a été stratifiée sur le résultat de la signature protéomine. Le critère de jugement principal était la survie globale (OS, overall survival) et l'hypothèse principale de l'essai était bien la mise en évidence d'une interaction : « and the primary hypothesis was the existence of a significant interaction between the serum protein test classification and treatment.".

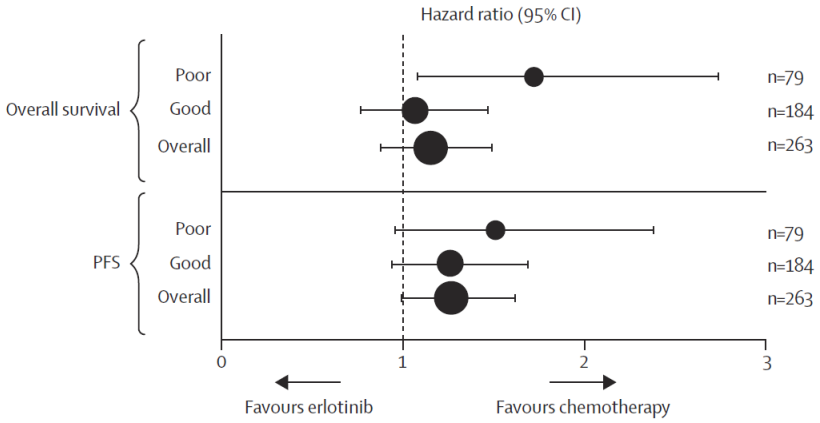
Au total 184 patients pour lesquels le test prédisait une bonne réponse à l'erlotinib (groupe Good, que l'on peut appeler marqueur positif, le but recherché étant d'identifier les patients « répondeurs ») et 79 patients avec un test prédisant une mauvaise réponse (groupe Poor, marqueur négatif) ont été randomisés entre erlotinib et chimiothérapie.

	Chemotherapy group (n=129)	Erlotinib group (n=134)
Proteomic test classification		
Good	88 (68%)	96 (72%)
Poor	41 (32%)	38 (28%)

Les résultats obtenus sont les suivants :

	OS	PFS
Mq+ : Good	HR 1·06 [95% CI 0·77–1·46], p=0·714	HR 1·26 [95% CI 0·94–1·69], p=0·129
Mq- : Poor	HR 1·72 [95% CI 1·08–2·74], p=0·022	HR 1·51 [95% CI 0·96–2·38], p=0·078
Interaction	P interaction=0·017	P interaction=0·268

La figure suivante donne la représentation graphique de ces résultats :



Une interaction statistiquement significative est trouvée pour l'OS (p=0.017). Pour les patients marqueurs négatifs (classification en Poor), l'erlotinib s'avère inférieur à la chimiothérapie. Ce résultat est compatible avec l'hypothèse. En effet, si le marqueur identifie correctement patients répondeurs et patients non répondeurs, les patients négatifs dans le groupe erlotinib se retrouvent comme s'ils n'étaient pas traités. Dans ce cas, l'erlotinib est attendu inférieur à la chimiothérapie (qui apporte un bénéfice indépendamment du statut sur le marqueur).

Cependant pour les patients marqueurs positifs (classification en Good), qui devraient être des patients qui répondent à l'erlotinib, le résultat est paradoxal en montrant une tendance à une infériorité de l'erlotnib par rapport à la chimiothérapie sur la PFS. Globalement, dans une analyse sans tenir compte de la signature protéomique, l'erlotinib ne montre dans cet essai aucun bénéfice en termes de PFS et d'OS.

Comme l'essai ne montre pas la supériorité de l'erlotinib chez les patients positifs et que l'interaction est l'inverse de celle qui était attendue, il est impossible de conclure à la valeur prédictive « pratique » de ce marqueur. Comme le résultat est inverse par rapport à l'hypothèse, il s'agirait d'une conclusion purement exploratoire sans intérêt pratique. Cet exemple illustre bien le fait que l'interaction ne fait pas tout dans la validation d'un marqueur prédictif (cf. section 2.1), encore faut-il qu'un bénéfice soit trouvé chez les patients marqueurs positifs et qu'il soit possible de conclure que les patients marqueurs négatifs ne tirent aucun bénéfice cliniquement pertinent du traitement (pour éviter de les priver à tort d'un traitement potentiellement utile pour eux)

4.2.3.2 Avantages

L'avantage indéniable du design d'interaction est de répondre spécifiquement aux exigences méthodologiques de la démonstration de la valeur prédictive d'un marqueur. Il permet d'évaluer directement cette valeur prédictive grâce à un design qui correspond directement à la définition de la valeur prédictive.

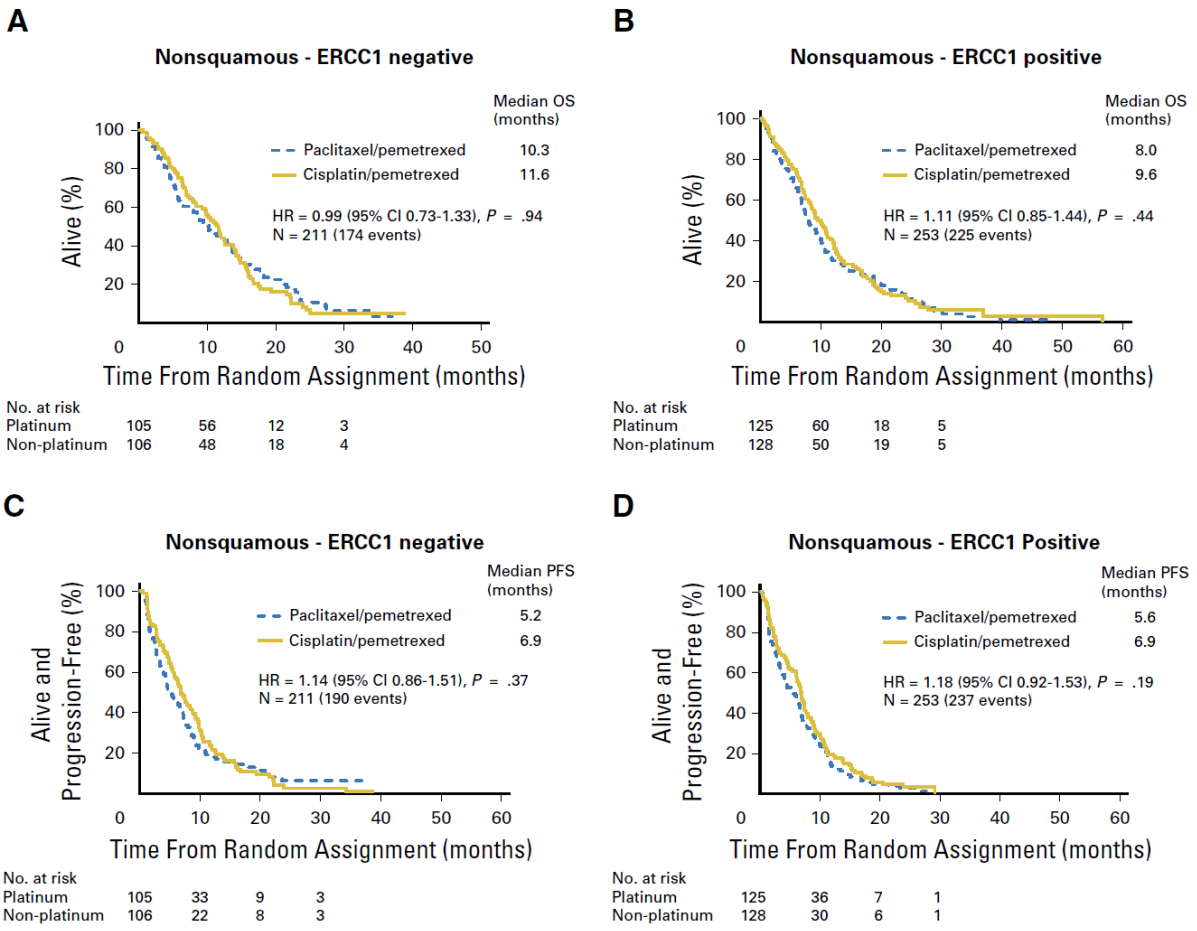
Ce design permet d'effectuer toutes les démonstrations nécessaires pour établir qu'un candidat marqueur est un réel marqueur prédictif utilisable en pratique :

- Existence d'une interaction statistiquement significative
- Le traitement considéré apporte un bénéfice chez les patients marqueurs positifs
- Il est possible de considérer que le traitement n'apporte pas de bénéfice chez les patients marqueurs négatifs (non-supériorité ou infériorité)

En étant prospectif, il garantit le respect de la démarche hypothético déductive vis-à-vis de l'hypothèse que le marqueur est un marqueur prédictif (modificateur d'effet du traitement). Il solutionne la problématique du p-hacking des approches rétrospectives. Ce design lève donc les limites inhérentes aux sous-groupes des essais pivots liées aux caractères post hoc et au non-contrôle des risques d'erreur statistiques (cf. section 4.2.1).

La robustesse méthodologique de ces essais d'interaction permet d'obtenir le plus haut niveau de preuve/crédibilité pour la recherche des marqueurs prédictifs. Grâce à cette robustesse, ces essais ont permis, dans un certain nombre de cas, de montrer la faiblesse des arguments basées sur les autres approches (« treatment only », sous-groupes exploratoires) et d'éviter ainsi la généralisation d'une pratique de personnalisation basées à tort sur des marqueurs sans valeur prédictive.

> L'évaluation de la valeur prédictive de l'expression de la protéine ERCC1 (excision repair cross complementing group 1, ERCC1) sur la réponse au sel de platine fournit un exemple de l'intérêt des designs d'interaction par rapport aux approches rétrospectives.
> À l'issue de ces études rétrospectives « treatment only » faisant suspecter que le polymorphisme de ERCC1 pourrait être un facteur de réponse au sel de platine (cf. exemple section 4.1.1), un essai randomisé d'interaction a été réalisé [53] :
> « This trial had a marker-by-treatment interaction phase III design, with ERCC1 (8F1 antibody) status as a randomization stratification factor. Chemona¨ıve patients with NSCLC (stage IIIB and IV) were eligible. Patients with squamous histology were randomly assigned to cisplatin and gemcitabine or paclitaxel and gemcitabine; nonsquamous patients received cisplatin and pemetrexed or paclitaxel and pemetrexed. Primary end point was overall survival (OS). »
> Aucune différence d'effet du traitement entre les strates ERCC1+ (sous figures) et ERCC1- (sous figures) n'a été montré avec une supériorité du traitement à base de sel de platine dans ces 2 strates.



Ces résultats ne permettent pas de conclure à une valeur prédictive de l'ERCC1. Un autre essai du même type (NCT00801736) non publié a obtenu le même résultat.

Dans un essai d'interaction les risques d'erreurs statistiques sont parfaitement maitrisés 1) par le calcul du nombre de sujets nécessaires qui portera sur chaque « strate » et sur la question de l'interaction et 2) par l'unicité de l'hypothèse testée (le marqueur considéré est-il prédictif du bénéfice du traitement considéré).

> *Même s'il existe plusieurs tests statistiques dans le design d'interaction (test d'interaction et tests de l'effet du traitement chez les Mq+ et Mq-) cela n'entraine pas de multiplicité en termes de risque alpha global. En effet le risque alpha de ce type d'essais est celui de conclure à tort à la valeur prédictive du marqueur du fait du hazard. Ces trois tests ne donnent pas trois occasions différentes de conclure à tort à cette valeur prédictive, car en fait il faut qu'il soit tous les trois simultanément concluant. Si le test d'interaction n'est pas significatif, les 2 autres tests ne permettront pas de « récupérer le coup », l'essai sera définitivement négatif et ne pourra apporter la preuve que le marqueur candidats est un marqueur prédictif par les 2 autres tests. En revanche, si plusieurs candidats marqueurs sont testés simulatément dans un même essai, apparait la problématique de la multiplicité, mais avec peu d'acuité, car la probabilité que sous l'hypothèse nulle généralisée, 3 tests soient significatifs (dont un en non-infériorité) est très peu probable*

4.2.3.3 Inconvénients, limites

L'inconvénient majeur du design d'interaction est que les patients marqueurs négatifs pour lesquels il n'est pas fait d'hypothèse de perte d'efficacité du traitement index (le marqueur est considéré comme un marqueur de non-réponse au traitement index) sont aussi randomisés entre les 2 traitements. Cette

randomisation est indispensable pour établir l'interaction traitement * marqueur mais implique, si l'hypothèse est vérifiée, que des patients reçoivent un traitement moins efficace pour eux que le traitement index. Cependant dans ce nombreuses situations cette situation de moindre efficacité n'est pas établie, comme en l'absence d'essai de comparaison directe des 2 traitements. Ce design ne pose alors pas ce problème.

Une randomisation adaptative sur un biomarqueur surrogate a aussi été proposée pour limiter cette problématique comme dans l'essai PROBIO [70].

4.3 Autres designs

De nombreux autres designs ont été proposés ou utilisés [16, 71, 72, 73] [74] mais ne sont pas couramment utilisés dans des essais pivots (phase 3) jusqu'à présent. Ils se rencontrent plutôt dans des études exploratoires (phase 2) qui n'ont pas comme finalité, en général, d'apporter les preuves nécessaires à la pratique clinique. Pour cette raison ils ne seront pas abordés en détail dans ce document.

4.3.1 Essai basket

L'essai basket[18] en oncologie inclut des patients tous porteurs de la même altération moléculaire, quelles que soient la localisation d'organe et/ou l'histologie de la tumeur pour évaluer une même molécule ciblant cette altération moléculaire. Ce n'est pas en soi un design particulier, mais une hypothèse thérapeutique : « le bénéfice apporté par le traitement est le même quel que soient les organes ou les histologies et seule compte la présence de la cible moléculaire ». Dans ce cas, tous ces patients sont homogènes vis-à-vis du bénéfice du traitement et il est licite de les regrouper sans les distinguer pour évaluer le traitement. Cette hypothèse est très difficile à prouver [75], ce qui limite la portée de cette approche.

Cette hypothèse peut ensuite être évaluée dans un essai monobras qui posera toutes les problématiques des essais monobras, ou dans un essai randomisé comme dans l'essai SHIVA [76] (cf. section 3.4).

4.3.2 Essais plateformes

Les essais « plateformes » [77] sont souvent utilisées dans le développement des thérapies ciblées, principalement en phase 2 comme l'essai SPY-2 [78]. L'aspect plateforme, adaptatif, de l'essai n'est pas un élément spécifique d'évaluation d'une thérapie ciblée ou d'un marqueur prédictif. Pour cette évaluation les éléments spécifiques de méthodologie qui sont présentés dans ce document doivent être mis en œuvre. Ces designs sont détaillés dans le chapitre 16 du livre blanc (https://sfpt-fr.org/livreblancmethodo/part17/file_0.htm).

[18] Cf livre blanc Acceptabilité des « nouvelles méthodologies » pour l'évaluation des médicaments

5 La personnalisation sur le risque de base

Une personnalisation des traitements peut être aussi envisagée en tenant compte du risque de base qu'ont les patients de présenter l'évènement que le traitement cherche à éviter [14]. En effet, chez les patients qui ont spontanément un faible risque, la question de l'intérêt du traitement peut se poser : quel est le réel intérêt de chercher à réduire le risque de survenue d'un évènement qui est déjà spontanément bas. De plus, dans ces conditions, la balance bénéfice risque peut devenir défavorable.

Cette approche conduit à de nombreuses applications, souvent disponible en ligne sur le WEB et utilisable par les médecins et les patients [79, 80, 81, 82, 83, 84]. Se pose alors la question de la validité de ces outils.

> Au stade précoce du cancer du sein, une chimiothérapie adjuvante peut être proposée après le traitement local (chirurgie +/- radiothérapie). À ce stade le pronostic est en général très bon avec un faible taux de récidives. Se pose alors la question de l'intérêt de cette chimiothérapie adjuvante chez les patientes ayant le plus faible risque de récidive. Un traitement adjuvant est probablement non nécessaire. D'où l'idée de stratifier les patientes sur leur risque de récidive pour ne pas exposer au risque d'effets indésirables celles dont le risque de récidive est spontanément proche de zéro.
> Dans ce but de nombreux outils ont été proposés allants de score de risque (implémenter sur le WEB par exemple comme Predict du NHS (https://breast.predict.nhs.uk/) à des signatures génomiques (comme OncotypeDX, MammaPrint, etc.).

5.1 Variation du bénéfice absolu en fonction du risque de base

L'efficacité relative d'un traitement, telle que mesurée par le risque ratio, l'odds ratio ou le hazard ratio, quantifie l'importance de l'efficacité « intrinsèque » du traitement. L'effet traitement relatif prend en compte le risque de « départ » (le risque de base) et s'avère fréquemment constant, quel que soit le risque de base [85, 86]. Cependant cette même réduction relative représentera un changement de risque plus important pour les sujets à haut risque de base que pour ceux à faible risque de base. Une même réduction relative de 50% ramènera les sujets ayant, par exemple, un risque de base de 10% à 5% (différence de -5%), ce qui les fait changer clairement de niveau de risque, tandis que pour ceux ayant, par exemple, un risque de base de 1%, celui-ci restera quasiment inchangé avec le traitement à 0.5% (différence de -0.5%). Apparait ainsi l'importance de la différence de risque (*absolute risk reduction*) pour apprécier le bénéfice qu'apporte un traitement. Cette différence dépend du risque de base et permet de mieux apprécier, pour un niveau de risque de base, l'intérêt du traitement.

Risque de base	**Risque avec le traitement**	**Risque ratio**	**Bénéfice absolu (Différence des risques)**
10%	5%	0.50	-5%
1%	0.5%	0.50	-0.5%

On parle de bénéfices absolus pour les différences de risque et de bénéfices relatifs pour les effets traitements relatifs.

Effet relatif (bénéfice relatif)	Mesuré par un ratio : risque ratio, hors ratio, hazard ratio	Prends en compte le risque de base
Bénéfice absolu	Mesuré par la différence de risque	Dépends du risque de base

Ainsi, un même traitement dont l'efficacité relative ne change pas apportera un plus grand bénéfice absolu aux patients à haut risque qu'aux patients à faible risque.

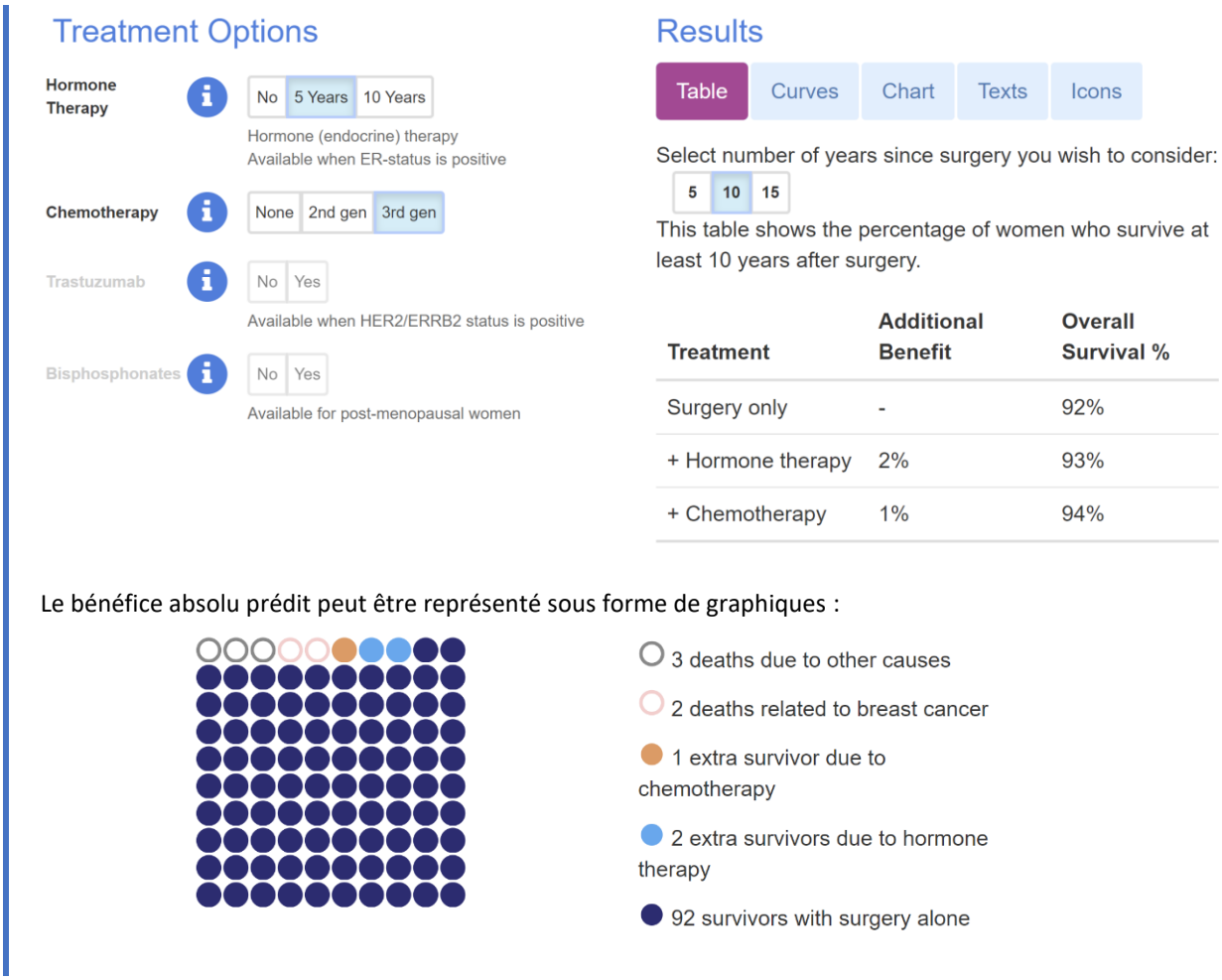
Cette variation du bénéfice absolu survient arithmétiquement, sans que l'effet propre du traitement ne change entre les patients à bas et haut risque. Le niveau de risque de base n'influence pas l'effet relatif du traitement. Ce n'est pas une problématique de modification de la taille de l'effet (comme discuté en section 4.2), mais simplement la conséquence des relations arithmétiques existant entre effet relatif, risque de base et bénéfice absolu.

5.2 Mise en application pour personnaliser les traitements

L'utilisation de ces relations arithmétiques pour personnaliser le traitement des patients va s'effectuer de la manière suivante.

Le risque de base du patient est calculé à partir de ses caractéristiques à l'aide d'un outil prédictif, comme un score de risque ou une équation de risque. Ces outils donnent, par calcul direct ou par l'intermédiaire d'un score, le risque d'événement d'un patient en fonction de ses caractéristiques (facteurs de risque). Ensuite le risque de base obtenu est réduit de l'effet relatif du traitement considéré pour obtenir le risque estimé sous traitement. La différence de risque est ensuite calculée et éventuellement représentée de façon graphique par un diagramme de Cates. Le médecin et/ou le patient pourront ensuite décider de recourir ou non au traitement en fonction de l'importance du bénéfice absolu prédit et des préférences et attentes du patient.

> Le site Web Predict du NHS (https://breast.predict.nhs.uk/tool) permet de calculer le risque de base de décès à 5,10 ou 15 ans en fonction des facteurs pronostiques. Ensuite sont appliqués les effets des traitements choisis, permettant de visualiser en quoi ils changeront de façon importante le taux de survie. Dans le tableau ci-dessous, la survie à 10 ans avec seulement la chirurgie est estimée à 92% compte tenu des caractéristiques de la patiente (tableau suivant). Un traitement hormonal augmentera cette survie de 1% et une chimiothérapie adjuvante de 1% aussi.



Le bénéfice absolu prédit peut être représenté sous forme de graphiques :

5.3 Évaluation clinique d'une personnalisation sur le risque

La mise en pratique de cette approche nécessite un outil prédictif[19] fiable permettant d'estimer correctement le risque de base des patients. En effet, une mauvaise appréciation du risque de base conduit à ne pas traiter à tort des patients en cas de sous-estimation ou conduit à traiter à tort en cas de surestimation. Cette approche deviendrait contreproductive, car elle déboucherait sur une perte de chance pour certains patients (les non traités à tort en raison d'une sous-estimation de leur risque) et elle raterait son objectif chez d'autres (ceux qui continueraient d'être traités à tort en raison d'une surestimation de leur risque). Au total la personnalisation du traitement sur le risque de base avec un outil d'évaluation du risque non performant peut être futile en n'évitant aucun évènement indésirable indu et/ou conduire à une perte de bénéfice avec moins d'évènements évités au total par rapport au traitement de tous les patients.

La performance prédictive n'est pas le seul paramètre conditionnant l'efficience de cette approche. Un élément important est le seuil de risque utilisé pour décider de traiter ou de ne pas traiter. La

[19] Dans ce vocable, le terme prédictif fait appel à la prédiction du risque et non pas à une question de modification de l'effet du traitement. L'ambigüité du terme prédictif a déjà été signalée en section **Erreur ! Source du renvoi introuvable.**.

fixation de ce seuil n'est pas chose aisée. Elle s'appuie en général sur des spéculations intégrant l'efficacité des traitements et la fréquence des effets indésirables. Mais le seuil retenu reste arbitraire.

Il est donc nécessaire de valider le processus entier pour bien s'assurer qu'il produit l'effet escompté. Cette évaluation s'effectue alors par un essai randomisé validant l'utilité médicale de l'approche.

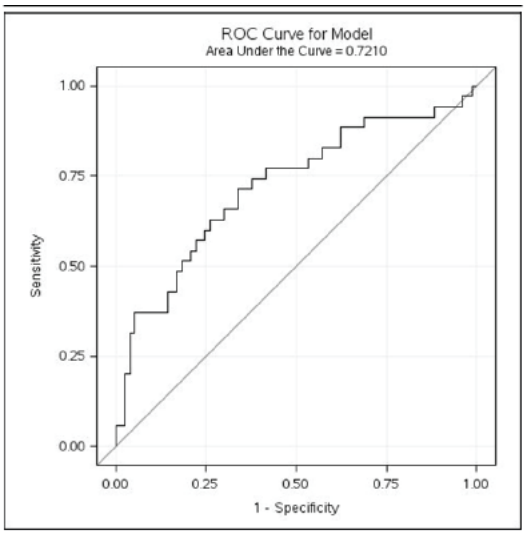
5.3.1 Performance d'un outil de prédiction

Une bonne performance de l'outil prédictif n'est pas une preuve suffisante, mais représente néanmoins un prérequis. La description en détail de la méthodologie d'évaluation de la performance des outils prédictive dépasse le cadre de document. Nous ne présenterons ici que les bases nécessaires à une bonne compréhension de l'évaluation de l'utilité clinique de la personnalisation sur le risque de base. Pour plus de détails, le lecteur pourra consulter les nombreux ouvrages et publications consacrées à cette thématique [87, 88].

La performance prédictive dépend de 2 dimensions : la capacité de discrimination et la calibration [89].

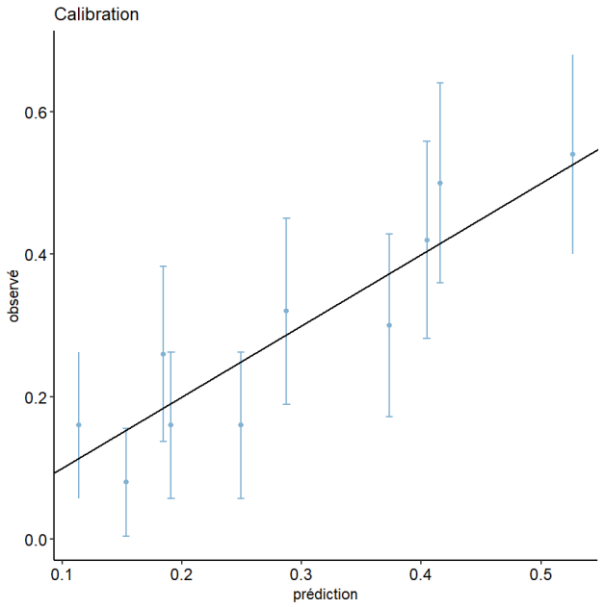
La discrimination est l'aptitude de l'outil à distinguer les patients en fonction de leur risque, c'est-à-dire à prédire une plus forte probabilité pour les patients qui font un évènement que ceux qui ne le font pas. Il existe plusieurs possibilités de mesurer cette capacité comme la courbe ROC avec la surface sous la courbe et la statistique C [90]. Une valeur de C de 0.5 marque l'absence totale de valeur prédictive tandis que la valeur 1 signifie une capacité parfaite à séparer les patients en fonction de la survenue ou non de l'évènement.

Cette courbe ROC est tracée en faisant varier un seuil de risque prédit. Pour chaque valeur de seuil, la sensibilité de la discrimination effectuée par ce seuil est la proportion de patients avec évènement au-dessus de ce seuil et la spécificité est la proportion de patients sans évènement en dessous du seuil.



La capacité de discrimination n'est pas suffisante à elle seule pour évaluer la performance prédictive. Il faut aussi que les risques prédits soient corrects numériquement par rapport au risque observé, c'est-à-dire que les prédictions soient bien calibrées [91]. Le risque observé est calculé dans les strates de risque prédit (par exemple des déciles du risque prédit). Le risque observé (la fréquence de l'évènement) dans ces strates doit être proche de la moyenne des risques prédits pour les patients de la strate. Cette évaluation peut s'effectuer graphiquement avec en abscisse le risque prédit et en ordonnée le risque observé. Les points correspondant à chaque strate doivent être proches de la

bissectrice. Des tests statistiques ont été proposés (comme le test Hosmer-Lemeshow) mais ils présentent de fortes limites.



Pour être utilisé dans une approche de personnalisation des traitements en fonction du risque, la calibration est une propriété très importante. Une mauvaise calibration autour du seuil de décision peut conduire à ne pas traiter des patients pourtant à risque d'évènement ou à traiter à tort des patients à faible risque. En revanche, si l'écart de calibration survient à distance du seuil de décision (de traiter ou de ne pas traiter les patients), l'erreur numérique de la prédiction n'est pas gênante, car la décision prise à partir de cette valeur inexacte ne conduit pas à une décision erronée.

Ces aspects quantitatifs n'assurent pas à eux seuls la validité de l'outil prédictif. Il convient aussi d'écarter la présence de biais dans l'étude d'évaluation. Des outils d'évaluation de ces biais sont disponibles [92].

5.4 L'évaluation de l'utilité clinique

L'évaluation de l'utilité clinique des personnalisations basée sur le risque de base s'effectue principalement à l'aide d'un essai de stratégie en non-infériorité. Le but est, en premier lieu, de montrer qu'il n'y a pas de perte de bénéfice par rapport au traitement de tous les patients, d'où l'approche de non-infériorité. L'essai devra aussi montrer que la personnalisation s'accompagne bien de l'avantage recherché, comme la réduction des effets indésirables, et qu'il y a donc un intérêt patent à ne pas traiter tous les patients. La principale difficulté de cette évaluation réside dans le choix de la limite de non-infériorité.

> L'essai DYNAMIC [93] a évalué l'intérêt clinique de l'ADN circulant pour guider le choix du traitement adjuvant dans le cancer du côlon au stade II.
> À ce stade une chimiothérapie adjuvante (à base d'oxaliplatine ou par fluoropyrimidines) peut être proposée après la chirurgie. Compte tenu du bon pronostic à ce stade, de l'incertitude sur un réel bénéfice sur la mortalité de la chimiothérapie adjuvante et des effets secondaires de celle-ci, la balance bénéfice risque de cette chimiothérapie adjuvante reste débattue. D'où la question d'envisager une désescalade thérapeutique chez les patients ayant le plus faible risque, en ne leur proposant pas de chimiothérapie.

Sur le plan de l'évaluation, le but de cette désescalade est de diminuer la proportion des patients ayant des effets indésirables, sans pour autant perdre en efficacité, c'est-à-dire sans augmentation de la fréquence des récidives par rapport à la prise en charge actuelle de ces patients. L'ADN circulant (ADNc, biopsie liquide) représente une piste pour guider cette désescalade en ne réservant la chimiothérapie qu'en cas de résultat positif 4 ou 7 semaines après la chirurgie.
Cet essai a donc comparé une stratégie basée sur l'ADN circulant (ADNc) par rapport à la stratégie habituelle. Le critère de jugement, classique en situation adjuvante, était la survie sans récidive (RFS) à 2 ans. L'essai était un essai de non-infériorité afin de montrer que le bras désescalade n'entrainait pas une perte de chance trop importante par rapport à la stratégie actuelle.

Au premier abord, cet essai semble donc parfaitement adapté à son objectif et à même d'apporter une réponse de haut niveau de preuve. Cependant, plusieurs limites méthodologiques apparaissent à la lecture approfondie de cet article.
Dans l'essai de non-infériorité les biais marchent à l'envers de ceux de l'essai de supériorité : sera cause de biais tout ce qui contribue à réduire la différence existante entre les 2 groupes et qui conduit à ne pas pouvoir mettre en évidence que le nouveau traitement est en réalité inférieur au traitement standard. Cette situation survient, entre autres, lorsque les deux groupes ne reçoivent pas les traitements de l'étude. Dans ce cas, le critère de jugement dépendra uniquement du risque de base des patients inclus dans l'essai et sera identique entre les 2 groupes conduisant à conclure automatiquement à la non-infériorité. Cette situation est caricaturale, mais permet de bien comprendre le *primum movens* de la problématique.
Dans cet essai, seulement 28% et 15% des patients, respectivement des groupes "stratégie classique" et "stratégie basée" sur l'ADNc, on reçut un traitement adjuvant (cf. table 2 de l'article). Même si la stratégie ADNc prive à tort de traitement adjuvant des patients à risque de récidive, les taux de récidive (RFS) de ces 2 groupes seront très proches, car la majorité des patients de ces 2 groupes ne reçoivent pas de traitement adjuvant de toute façon (décision basée principalement sur la clinique).
En effet, si la stratégie ADNc est inférieure et augmente le nombre de récidives ou de décès, car elle prive de chimiothérapie des patients à haut risque contrairement à l'approche classique, le surcroit de récidives que cause cette stratégie ne concernera qu'au maximum 13% des patients (28% - 15% = 13%). Cette différence sera alors diluée par les patients pour lesquels rien n'est différent entre les 2 groupes et qui sont majoritaires. Ainsi, la conclusion à la non-infériorité est quasiment assurée même en cas de nette infériorité du fait de cette dilution.
Pour éviter cette problématique, il aurait fallu inclure des sujets pour lesquels l'approche actuelle implique le traitement adjuvant et les randomiser entre un groupe où ce traitement est appliqué et un groupe où l'indication du traitement adjuvant est réévaluée à l'aide du ADNc. Cette approche permet de répondre à la question : chez des patients qui actuellement recevraient un traitement adjuvant, l'ADNc permet-il d'identifier ceux chez lesquels il est en réalité inutile ?

Une autre interrogation survient avec l'analyse de la limite de non-infériorité utilisée. Celle-ci est de 8.5% en termes de différence de RFS à 2 ans. La stratégie ADNc sera considérée comme non-inférieure tant qu'elle n'entrainera pas une augmentation de plus de 8.5% du taux de récidive à 2 ans (ou une diminution de 8.5% de la RFS).
Pour évaluer l'acceptabilité des limites de non-infériorité, il est nécessaire de se référer à l'efficacité du comparateur afin de déterminer la limite maximale acceptable, qui correspond à la perte de 100% de l'apport du traitement standard. Une méta-analyse Cochrane du traitement adjuvant au stade 2 est disponible[20] et donne un risque ratio de 0.83 (IC 95% entre 0.75 et 0.92). Dans le contexte de cet essai, la RFS dans le groupe standard est de 92.4% correspondant à un risque de récidives ou décès de 7.6%. Dans ce contexte, l'effet du traitement adjuvant en termes de différence des risques (ARR) peut être estimé à (1 – 1/0.92) * 7.6%, soit 0.66%, montrant que la limité utilisée est excessivement tolérante.

[20] Figueredo A, Coombes ME, Mukherjee S. Adjuvant Therapy for completely resected Stage II Colon Cancer. Cochrane Db Syst Rev 2008; 2010: CD005390.

Dans ces essais de désescalade, s'assurer que réduire le traitement n'entraine pas une perte de chance rédhibitoire par une logique de non-infériorité est indispensable, mais ce n'est pas la seule démonstration à apporter. Il convient aussi de montrer que cette désescalade s'accompagne bien d'un bénéfice pour le patient, au niveau des effets indésirables. Dans cet essai, cet objectif n'est pas abordé directement. Seule la fréquence de recours à un traitement adjuvant est mesurée et met en évidence la moindre utilisation déjà citée : 28% versus 15% (cf. table 2 de l'article ci-dessous). Cependant, dans le groupe ADNc, les investigateurs choisissent plus fréquemment en cas de positivité du ADNc un traitement adjuvant à base d'oxaliplatine, de moins bonne tolérance qu'une monothérapie avec une fluoropyrimidine. Ce résultat interroge vraiment sur l'obtention du but recherché avec la désescalade.

Table 2. Treatment Delivery and Adherence.*

Treatment Characteristic	Standard Management (N=147)	ctDNA-Guided Management (N=294)	Relative Risk (95% CI)
Adjuvant chemotherapy received — no. (%)			
No	106 (72)	249 (85)	
Yes	41 (28)	45 (15)	1.82 (1.25–2.65)
Chemotherapy regimen received — no./total no. (%)			
Oxaliplatin-based doublet	4/41 (10)	28/45 (62)	
Single-agent fluoropyrimidine	37/41 (90)	17/45 (38)	2.39 (1.62–3.52)

Une tout autre approche aurait pu être envisagée : ADNc ne permettrait-il pas de trouver des sujets qui bénéficieraient du traitement adjuvant alors qu'ils ne sont pas identifiés par l'approche classique ? Pour le montrer, des patients sans indication de traitement adjuvant avec la stratégie actuelle seraient alors randomisés entre un groupe contrôle et un groupe expérimental où un traitement adjuvant serait quand même mis en place en cas d'ADNc positif afin de montrer la supériorité de l'approche basée sur l'ADNc.

Lorsque l'objectif est d'éviter d'exposer tous les patients à la toxicité du traitement, une approche de bénéfice net en supériorité est aussi envisageable à la place d'une approche de non-infériorité. Elle aura pour principal intérêt d'évincer la difficulté de la fixation de la limite de non-infériorité.

Pour l'instant (juillet 2023), il n'existe que quelques autres exemples comme l'essai TAILORx [94] ou l'essai MINDACT [95] pour la décision de traitement adjuvant dans le cancer du sein précoce.

6 Approches prédictives de l'hétérogénéité des effets traitements (*heterogeneity of treatment effects*, HTE)

6.1 Principes

Partant du postulat que les effets des traitements variaient d'un patient à l'autre, différentes approches ont été proposées pour modéliser cette hétérogénéité des effets traitements « individuels » (heterogeneity of treatment effect, HTE) à partir des données des essais thérapeutiques ou des méta-analyses sur données individuelles [96, 97, 98].

Ces approches se proposent de modéliser le cATE (conditionnal average treatment effect, cATE) soit en fonction du risque de base prédit des patients [96, 99] soit directement en fonction d'un modèle de prédiction direct du cATE [100].

Cette modélisation est réalisée à l'aide d'approches biostatistiques classiques (régression logistique, modèle de Cox) ou avec des méthodes d'IA [101]. Les données peuvent être celles d'essais randomisés ou de type observationnel (« real world data »). Pour chaque patient, la prédiction de son cATE est effectuée à partir de ses caractéristiques sur les covariables prises en compte dans le modèle. Ces prédictions ne représentent pas, stricto sensu, le bénéfice individuel qui est impossible à observer [19] en raison de l'impossibilité d'avoir le contrefait individuel[21] [32, 102], mais attribue au patient une estimation moyenne (ATE) **conditionnelle** à leurs caractéristiques sur les facteurs influençant l'effet du traitement déterminé dans une strate de patients similaires.

Cette caractérisation de l'hétérogénéité de l'effet traitement est aussi envisagée par une recherche de sous-groupes de patients drivée par les résultats [103, 104, 105, 106]. Cette recherche expose aux mêmes problématiques de HARking et de multiplicité que les analyses en sous-groupes exploratoires, mais ces problématiques peuvent être partiellement prises en considération par des techniques récentes [105, 106]. Cette approche reste, par essence, purement exploratoire et ne peut pas produire des résultats directement utilisables en pratique. Avant d'être utilisables en pratique, les outils obtenus par ces approches devront démontrer leur utilité médicale dans un essai randomisé de stratégie (cf. section 6.5).

6.2 Modèles utilisés

Une mapping review des approches proposées pour rechercher et/ou modéliser l'hétérogénéité des effets [107] inventorie de nombreuses propositions reposant principalement sur deux grandes catégories de modèles : les modélisation de l'effet (*treatment effect modelling*) et les modèles basés sur le risque de base des sujets (*risk based method*). Une troisième voie est représentée par des techniques de classification des patients suivant qu'ils bénéficieraient ou pas du traitement (*optimal treatment regime methods*).

[21] Même pour les essais individuels (« n-of-1 »), qui sont des crossovers répétés, e contrefait individuel n'existe pas à proprement parler, puisque les conditions ne sont pas strictement identiques entre chaque période.

6.2.1 Modélisation de l'effet (*treatment effect modelling*)

Dans ces modèles, la probabilité de survenu de l'événement clinique servant de critère de jugement *Y* est modélisé en fonction du traitement et des caractéristiques de bases des sujets.

Le modèle le plus simple est appelé modèle à effet homogène car l'effet du traitement sur *Y* est considéré comme constant, homogène pour tous les patients. Dans ce cas l'hétérogénéité des effets traitements individuels $\delta(x_i)$ provient de l'hétérogénéité de pronostic des patients (hétérogénéité du risque de base) du fait que cet effet est défini en termes de différence.

Pour un patient de caractéristiques xi, l'effet traitement individualisé est défini par :

$$\delta(x_i) = Pr\left(Y_i^{a=1} \middle| \boldsymbol{X} = \boldsymbol{x_i}\right) - Pr\left(Y_i^{a=0} \middle| \boldsymbol{X} = \boldsymbol{x_i}\right)$$

Où $Y_i^{a=1}$ et $Y_i^{a=0}$ désigne la valeur potentielle du critère de jugement *Y* (*potential outcome*) du patient avec le traitement (*a=1*) et sans le traitement (*a=0*)[22].

En reprenant les notations de la référence [98], pour un sujet *i*, recevant le traitement a_i (1 pour le traitement étudié et 0 pour le traitement contrôle) et de caractéristiques $\boldsymbol{x_i}$ (qui est un vecteur de toutes les caractéristiques individuelles retenues pour la modélisation), la probabilité de l'évènement est représentée par :

$$logit\left(Pr(Y_i = 1 | A = a_i, \boldsymbol{X} = \boldsymbol{x_i})\right) = \beta_0 + \beta_t a_i + \boldsymbol{\beta}^T \boldsymbol{x_i}$$

Où β_t représente l'effet du traitement et $\boldsymbol{\beta}$ le vecteur des coefficients rattachés aux facteurs pronostiques (et $\boldsymbol{\beta}^T$ sa transposée).

L'effet traitement individuel est estimé à partir des coefficients du modèle par

$$\hat{\delta}(x_i) = \frac{1}{1 + e^{-(\hat{\eta}_i + \hat{\beta}_t)}} - \frac{1}{1 + e^{-\hat{\eta}_i}}$$

Ou $\hat{\eta}_i = \widehat{\beta_0} + \widehat{\boldsymbol{\beta}}^T \boldsymbol{x_i}$ et représente la partie pronostique du modèle.

L'autre catégorie de modèle comporte des termes d'interaction traitement covariable permettant de modéliser des modificateurs d'effet (comme les marqueurs prédictifs) :

$$logit\left(Pr(Y_i = 1 | A = a_i, \boldsymbol{X} = \boldsymbol{x_i})\right) = \beta_0 + \beta_t a_i + \boldsymbol{\beta_m^{\mathrm{T}}} \mathbf{x_i} + \boldsymbol{\beta_z^{\mathrm{T}}} \mathbf{z_i} \mathrm{a_i}$$

Ou z_i représente le vecteur des covariables d'interaction avec le traitement (qui est un sous ensemble de x_i).

L'effet traitement individuel est estimé à partir des coefficients du modèle par

$$\hat{\delta}(x_i) = \frac{1}{1 + e^{-(\hat{\eta}_i + \hat{\beta}_t + \boldsymbol{\beta_z^{\mathrm{T}}} \mathbf{z_i})}} - \frac{1}{1 + e^{-\hat{\eta}_i}}$$

Le modèle intègre les facteurs modificateurs de l'effet (TEM) sous la forme de termes d'interaction. Cette modélisation des interactions présente de nombreuses limites (similaires à celles des analyses

[22] La théorie de l'inférence causale permet de démontrer que sous l'hypothèse d'échangeabilité et de consistency, ces valeurs potentielles peuvent être estimées par les valeurs observées, par exemple, dans un essais randomisés analysés en ITT pour assurer l'hypothèses d'échangeabilité.

en sous-groupes traditionnelles) et augmente la complexité des modèles. Un risque de surdétermination (overfitting) est possible et des techniques de régularisation sont couramment utilisées pour réduire l'optimisme de ces modèles.

L'estimation de ces modèles peut s'effectuer à l'aide de nombreuses méthodes traditionnelles ou de machine learning [108]. Les techniques de machine learning (intelligence artificielle) présentent les mêmes limites que les approches statistiques habituelles et leur usage ne garantit pas d'obtenir automatiquement un modèle fiable.

La fiabilité des estimations proposées, quelle que soit la méthode utilisée, doit être démontrée par des études de validation externe satisfaisantes. Ensuite une démonstration de l'utilité médicale de ces outils, dépendant de leurs performances propres et du bien fondés des décisions thérapeutiques qu'ils engendrent, doit être apportée par des essais randomisés de stratégie (cf. section 6.5).

6.2.2 Modèles basés sur le risque de base (Risk-based methods)

L'approche basée sur l'effet consiste à modéliser le critère de jugement en fonction des caractéristiques des patients et du traitement reçu, en introduisant des facteurs de modification de l'effet du traitement.

$$logit\big(Pr(Y_i = 1|A = a_i, \hat{\eta}_i)\big) = \beta_t a_i + \hat{\boldsymbol{\eta}}_{\boldsymbol{i}} + \boldsymbol{f}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})\boldsymbol{a}_{\boldsymbol{i}}$$

Où $\hat{\eta}_i$ est une estimation du risque de base du patient, c'est-à-dire une estimation de $Pr(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x}_{\boldsymbol{i}})$ et $\boldsymbol{f}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})\boldsymbol{a}_{\boldsymbol{i}}$ un terme permettant de modéliser une interaction risque traitement.

Une autre approche reposant sur le risque de base [99] consiste à construire un modèle prédictif du risque de base des patients à partir des données. Un risque de base prédit est ensuite calculé pour tous les patients (des deux groupes de traitement). Les patients sont ensuite répartis en déciles de ce risque prédits. L'effet du traitement est alors calculé à partir des données observées dans chacune des strates ainsi créées. Dans cette logique l'effet du traitement est mesuré par des différences de risques étant donné que le risque de base conditionne arithmétiquement cette différence de risque. Si un effet traitement relatif est utilisé, cela sous-entend une hypothèse que le risque de base est un facteur modificateur de l'effet (*treatment effect modifiers, TEM*). L'hétérogénéité d'effet du traitement est ensuite recherchée en comparant les différentes strates de risque de base. Une modélisation globale est aussi réalisable sans passer par cette décomposition en déciles.

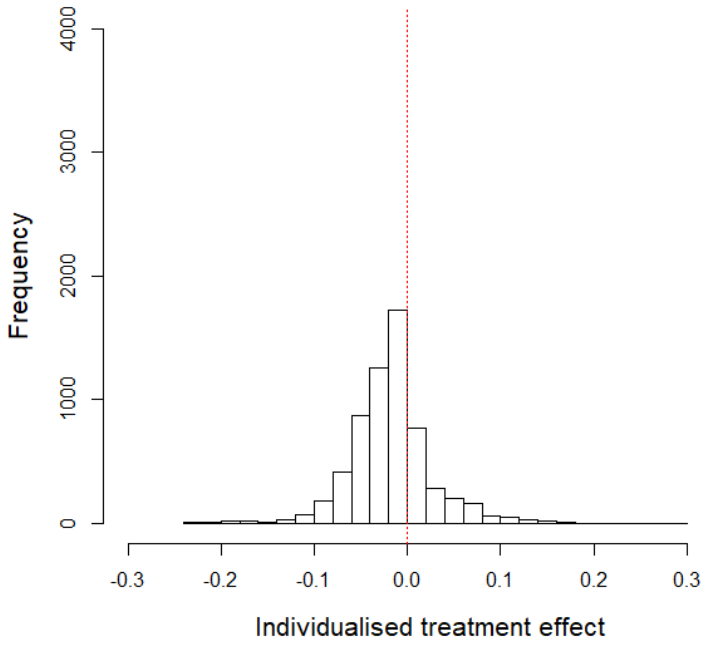
L'effet « personnalisé » est obtenu en comparant le risque prédit avec et sans traitement et peut ainsi être exprimé en relatif ou en différence absolue. La question fondamentale est celle de la validité de ces prédictions qui doit faire l'objet d'une évaluation et d'une démonstration spécifique.

6.3 Exemples d'application

6.3.1 Exemple 1

L'essai randomisé IST a évalué le bénéfice de l'aspirine dans l'AVC. Un bénéfice modeste a été observé sur un critère composite de décès et de dépendance. Dans une analyse secondaire [109] l'hétérogénéité de l'effet a été recherchée à partir de deux modèles (un pour le groupe traité par aspirine, l'autre pour le groupe contrôle) de prédiction du critère de jugement basés sur 23

prédicteurs. Le jeu de données a été divisé en deux parties pour permettre une validation du modèle. Aucune validation externe n'a été effectuée. Sur le jeu de validation, les deux modèles présentent de relative bonne performance avec des C-scores de 0.798 (95% CI: 0.782 to 0.813) et 0.794 (95% CI: 0.778 to 0.809) et une bonne calibration dans les 2 cas. Les effets sont ensuite recherchés en appliquant ces 2 modèles sur les mêmes patients. L'histogramme de différences de risques prédits montre une hétérogénéité importante avec seulement une partie des patients bénéficiant de l'aspirine (différence <0) :



6.3.2 Exemple 2

La recherche de l'hétérogénéité de l'effet traitement (heterogeneity of treatment effect, HTE) a été appliquée aux données des essais évaluant chez les patients hospitalisés pour COVID l'intérêt des doses curatives d'héparine (fortes doses similaires à celles utilisées dans le traitement des TVP) par rapport aux doses prophylactiques (thromboprophylaxie chez des patients alités pour raison médicale).

Cette question a été évaluée dans trois essais similaires, faisant partie du même essai multiplateforme, ATTAC, ACTIV-4a, REMAP-CAP. Les résultats de ces études et d'autres sont contradictoires avec des essais en faveur des doses curatives et d'autres non concluants. Devant ces discordances l'hypothèse d'une forte hétérogénéité d'effet de l'héparine à dose curative est avancées et une recherche exploratoire d'hétérogénéité de l'effet traitement a été réalisée de manière post hoc [110].

Ce papier propose les 2 méthodes décrites précédemment, l'approche basée sur le risque (Risk-Based Approach) et l'approche basée sur les effets (Effect-Based Approach) en plus de l'approche sous-groupe conventionnelle. Un cadre bayésien a été utilisé pour être en cohérence avec l'approche utilisée dans ces essais (cf. dossier essais bayésiens).

Pour l'approche basée sur le risque, un score prédictif a été dérivé d'un modèle de risque construit de manière très classique à l'aide d'une approche basée sur une régression logistique. Les résultats de cette modélisation sont les suivants :

Table 2. Risk Model for Organ Support–Free Days Used in Risk-Based Heterogeneity of Treatment Effect Analysis

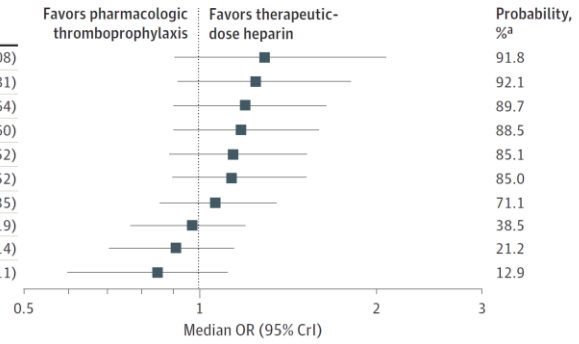
Model variable[a]	Risk model coefficients	
	Odds ratio (95% CI)[b]	Log odds ratio[c]
Age, per 10 y	0.73 (0.68-0.77)	−0.32
Female sex	1.29 (1.10-1.51)	0.25
Body mass index, per 5 units	0.92 (0.88-0.97)	−0.08
Diabetes	0.78 (0.67-0.91)	−0.25
Immunosuppressive disease or therapy[d]	0.58 (0.44-0.77)	−0.54
Baseline organ support		
Vasopressor requirement	0.49 (0.35-0.70)	−0.71
High-flow nasal oxygen	0.07 (0.05-0.08)	−2.70
Noninvasive ventilation	0.05 (0.04-0.06)	−3.10
Invasive mechanical ventilation	0.04 (0.03-0.05)	−3.29
Neutrophil count, per 5 × 10^9/L	0.74 (0.68-0.81)	−0.30
Lymphocyte count, per 5 × 10^9/L	1.35 (1.10-1.66)	0.30
Platelet count, per 25 × 10^9/L	1.07 (1.04-1.09)	0.06

Une analyse en sous-groupe avec lissage des estimations a ensuite été réalisée en se basant sur les déciles de ce score de risque et non pas le risque de base lui-même (colonne « score interval »). Les résultats de cette approche basée sur le risque sont représentés sur la figure ci-dessous

Figure 3. Heterogeneity of Treatment Effect Evaluation by Risk-Based Analysis

A Organ support-free days

Risk group	Score interval	No. of patients: Pharmacologic thromboprophylaxis	No. of patients: Therapeutic-dose heparin	Median OR (95% CrI)
1 (Lowest)	0.00-1.93	164	168	1.29 (0.90-2.08)
2	1.93-2.26	143	189	1.25 (0.92-1.81)
3	2.26-2.49	166	166	1.19 (0.90-1.64)
4	2.49-2.73	169	163	1.17 (0.90-1.60)
5	2.73-2.99	137	195	1.14 (0.89-1.52)
6	2.99-3.46	156	176	1.13 (0.90-1.52)
7	3.46-5.14	151	181	1.06 (0.85-1.35)
8	5.14-5.71	189	143	0.97 (0.76-1.19)
9	5.71-6.25	152	180	0.91 (0.70-1.14)
10 (Highest)	6.25-8.41	186	146	0.85 (0.59-1.11)

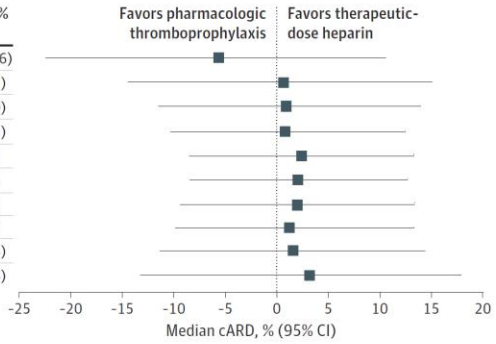


Dans cette approche basée sur le risque, il apparait que l'effet de la dose thérapeutique diminue en fonction du risque de base des patients. Aucun décile ne donne un résultat « statistiquement significatif » avec une probabilité a posteriori >97.5%.

L'approche basée sur les effets repose sur une modélisation des bénéfices absolus sur la survie hospitalière (absolute rate difference, ARD) à l'aide d'une méthode de machine learning (forest method). Les résultats sont présentés par déciles de bénéfice absolu dans la figure 4 du papier.

Figure 4. Heterogeneity of Treatment Effect Evaluation by Effect-Based Analysis

Predicted benefit group	Predicted cARD interval	No. of patients: Pharmacologic thromboprophylaxis	No. of patients: Therapeutic-dose heparin	Observed cARD, % (95% CI)
1 (Lowest)	-1.000 to -0.025	87	79	-5.7 (-22.4 to 10.6)
2	-0.025 to -0.011	81	85	0.7 (-14.4 to 15.1)
3	-0.011 to -0.004	79	87	0.9 (-11.5 to 14.0)
4	-0.004 to 0.001	80	86	0.8 (-10.3 to 12.5)
5	0.001 to 0.006	78	88	2.4 (-8.5 to 13.3)
6	0.006 to 0.010	80	86	2.0 (-8.4 to 12.7)
7	0.010 to 0.015	80	86	2.0 (-9.4 to 13.4)
8	0.015 to 0.020	81	85	1.2 (-9.8 to 13.3)
9	0.020 to 0.030	79	87	1.6 (-11.3 to 14.4)
10 (Highest)	0.030 to 1.000	81	85	3.2 (-13.2 to 17.8)



Effect-based heterogeneity of treatment effect for hospital survival shown by deciles of predicted conditional absolute rate difference (cARD) in hospital survival derived from repeated cross-validation using a causal machine-learning algorithm (n = 100 repetitions).

L'effet pour le premier décile suggère un effet délétère (harm) non « statistiquement significatif ». Pour la mise en pratique éventuellement de ce résultat il est nécessaire d'avoir les caractéristiques des patients constituant ce 1er décile : « *Patients in the lowest cARD decile group (in whom the treatment was associated with possible harm) (Figure 4) tended to have high BMI and were more likely to require ICU admission at baseline* ».

6.4 Utilisation de l'intelligence artificielle

Le domaine de la personnalisation des traitements n'échappe pas à l'enthousiasme, parfois excessif [111], soulevé par l'intelligence artificielle (IA) avec le *machine learning* ou le *deep learning*. Il existe d'ores et déjà de multiples propositions d'outils de prédiction du bénéfice des traitements pour l'aide à la décision thérapeutique [112, 113][114]. D'autres approches dites innovantes sont proposées, comme l'utilisation des jumeaux numériques (*digital twins*). En dehors d'effet d'annonce [115, 116], aucune publication d'application concrète n'est pour l'instant disponible (en juillet 2023).

L'intelligence artificielle, et le machine learning ou le deep learning en particulier, est une voie alternative à la modélisation biostatistique habituelle pour construire des outils prédictifs. Par rapport aux méthodes habituelles (régression logistique par exemple), les méthodes d'IA pourraient être plus flexibles en termes de complexité des modèles sous-jacents et offrent aussi la possibilité d'exploiter de l'imagerie à but prédictif. Il s'agit de techniques différentes de modélisation et d'utilisation des modèles construits pour prédire, mais au-delà de ces aspects techniques l'IA n'apporte rien de plus que les approches classiques au niveau conceptuel dans cette problématique de la prédiction des bénéfices des traitements. Elle présente cependant quelques avantages comme la possibilité d'exploiter l'imagerie, la possibilité de modéliser des espaces de haute dimension (très nombreuses variables) et de régularisation (prévention de la surdétermination, de l'optimisme des modèles par overfitting).

Les outils ainsi produits sont parfois librement disponibles en ligne (comme https://www.cards-lab.org/insight). Se pose alors la question de la fiabilité de ces prédictions, compte tenu du risque de perte de chance induit par une suggestion inopportune d'un traitement non optimal pour le patient

(cf. section 1.2). La validation de leur utilité médicale (cf. section 6.4.3) devrait être un pré requis indispensable à l'utilisation de ces outils compte tenu des risques et enjeux médicaux sous-jacents.

6.4.1 Prédiction du pronostic sous traitement

Les techniques dites d'intelligence artificielle, comme le machine learning, sont utilisées pour construire des outils prédictifs de « l'efficacité des traitements » [101, 117]. La plupart de ces outils sont simplement prédictifs du pronostic sous traitement. Ces techniques permettent de produire un outil informatique (souvent appelé à tort algorithme) qui à partir des caractéristiques du patient (cliniques, biologiques, génétiques et/ou d'imageries) prédisent la survenue ou non d'un évènement assimilé à la non-réponse/réponse au traitement.

> Trebeschi et al. proposent un outil d'IA de prédiction de la réponse aux immunothérapies dans le cancer du poumon non à petites cellules à partir des images du scanner [118].

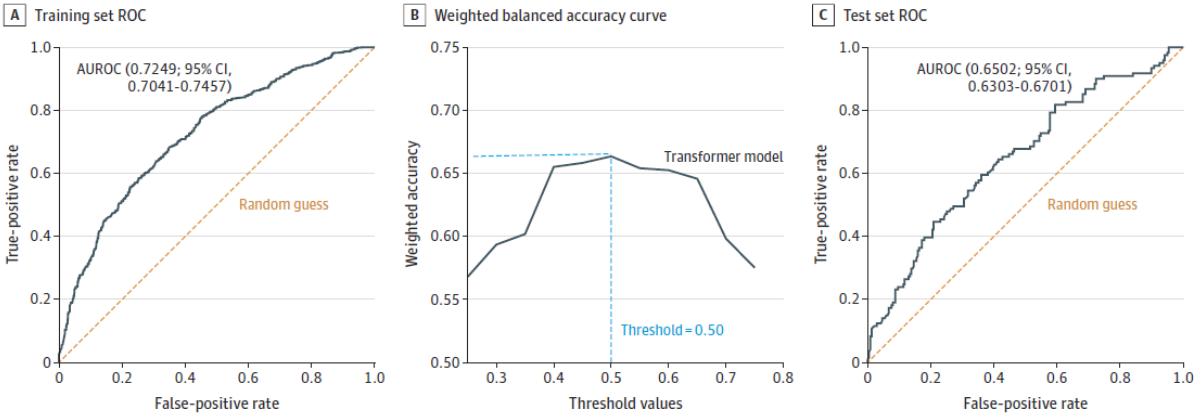
Ces outils ont les limites évoquées précédemment des approches de pronostic sous traitement et ne prédisent donc pas l'efficacité des traitements.

Le fait qu'il s'agisse de techniques d'intelligence artificielle ne lève en rien ces limites. Les approches de machine learning sont en effet très proches de l'approche classique de modélisation statistique et prédiction à partir du modèle [119]. Elles ont les mêmes limites comme l'indisponibilité dans le jeu de données d'apprentissage des réels prédicteurs de l'outcome, imprédictibilité intrinsèque partielle de l'outcome (phénomène aléatoire san réel prédicteur), etc. Ainsi rien ne permet d'espérer que ces approches fassent des miracles dans cette situation et puissent appréhender la réelle valeur prédictive d'un marqueur à travers une simple mesure de sa valeur pronostique sous traitement.

> Hakeem et al. [120] rapporte la construction d'un outil de prédiction de la réponse au traitement dans l'épilepsie nouvellement diagnostiquée. La réponse est définie par l'absence de crise durant l'année suivant l'instauration du premier traitement. Un réseau neural de type « transformer » a été entrainé à partir des données d'une cohorte de patients traités par un anticonvulsivant. Le modèle a ensuite été testé sur 4 autres cohortes pour sa validation externe[23].
> Les performances du modèle s'avèrent assez modestes avec une aire sous la courbe pour la partie validation (test) à 0.65. La conclusion du papier est qu'une amélioration de la prédiction est nécessaire.

[23] Mais le papier laisse suggérer qu'il y a un extra tuning (complément d'apprentissage) lors de cette étape, ce qui transforme cette étape en étape d'apprentissage et non plus en étape de validation externe.

Figure. Receiver Operating Characteristic (ROC) Curves and Weighted Balanced Accuracy Curve for the Transformer Model Developed Using a Pooled Cohort



A, ROC curve on the training set. B, Weighted balanced accuracy curve at different threshold values of probability. The highest weighted balanced accuracy was obtained at a threshold of 0.5. The optimal threshold value is indicated by the intersection of dashed blue lines. C, ROC curve on the test set. AUROC indicates area under the receiver operating characteristic curve.

Tous les patients des cohortes utilisées étant des patients traités, ce travail ne prédit donc pas l'effet des traitements, mais bien l'évolution sous traitement (pronostique sous traitement). Aucun contrôle n'a été utilisé afin de vérifier que l'outil ne prédisait pas simplement le niveau de sévérité de l'épilepsie. Comme discuté en section 4.1 il y a une impossibilité dans cette approche « treatment only » de distinguer les facteurs de réponse des facteurs pronostiques.
On peut aussi remarquer que le modèle ne permet pas de prédire la réponse à chaque molécule, car la réponse modélisée est indépendante du traitement utilisé. Cet outil dans cette forme ne permettrait donc pas de faire de la personnalisation de traitement.

6.4.2 Prédiction du bénéfice, modélisation de l'hétérogénéité des effets traitements

Comme pour la modélisation de l'hétérogénéité des effets traitement avec les modèles standards (cf. section 5), il est possible d'utiliser les techniques d'IA pour prédire l'effet. Au-delà de la simple modélisation, d'autres approches cherchent à identifier les sous-groupes des patients bénéficiant du traitement, principalement en cherchant un seuil optimal pour chaque covariable pertinente. Les techniques les plus abouties [101] veillent à respecter plusieurs principes : contrôle ou évaluation du risque alpha au niveau de toute la stratégie de recherche des sous-groupes, prise en compte de l'incertitude à chaque étape de l'identification des sous-groupes, prévention de la surdétermination et des biais dans la sélection des covariables, s'assurer de la reproductibilité de l'identification des sous-groupes et veiller à la fiabilité des estimations de l'effet traitement dans les sous-groupes.

Pour prédire la réponse individuelle aux traitements de la sclérose en plaque, Falet et al. utilise une approche de machine learning [121]. Un modèle prédictif de l'évolution du score EDSS au cours du temps a été construit à partir des données cliniques et d'imagerie de 6 essais cliniques. L'effet traitement individualisé a été calculé par la différence entre la prédiction du modèle sans et avec traitement (cATE, conditionnal average treatment effect). Cette approche est proposée, non pas pour guider le choix du traitement pour des patients, mais pour sélectionner des patients « répondeurs » pour des essais cliniques de futures molécules (enrichissement des essais cliniques [122]).

L'utilisation de l'IA pour ces modélisations n'est pas indispensable. Par exemple ce travail reproduit celui réalisé par Bovis et al. sur la même question et qui s'appuyait sur une modélisation basée sur la méthode conventionnelle du modèle de Cox [123].
Bien qu'actuellement proposée uniquement pour la construction de phase 2 exploratoire, cette approche ouvre la voie au développement de nouvelles thérapeutiques ciblées pour lesquelles les essais de confirmation (phase 3) sélectionneront les patients, non pas sur la base d'une variant moélculaire, mais sur la prédiction d'un modèle. Une fois validées de cette façon, ces molécules devront être utilisées en pratique chez des sujets identifiés de la même manière, ce qui ouvrira un champ nouveau de problématiques à traiter dans l'évaluation (health technology assessment) de ces produits.

Oikonomou et al. [15] proposent un outil de prédiction du bénéfice cardiovasculaire individuel de la canagliflozine chez le diabétique de type 2 à partir des données de l'essai randomisé CANVAS (4327 patients). Cet outil est accessible en ligne sur le WEB (https://www.cards-lab.org/insight). Contrairement à beaucoup d'autres propositions, l'effet du traitement au niveau individuel n'est pas déduit de la modélisation du critère de jugement avec et sans traitement, mais modélisé directement. Le critère de jugement étant les évènements cardiovasculaires (MACE) l'effet du traitement est appréhendé sous forme de hazard ratio. Le hazard ratio n'a pas d'existence au niveau individuel (contrairement à l'évolution temporelle de l'EDSS de l'exemple précédent). Son calcul nécessite impérativement un groupe de sujets traités et un groupe de sujets non traités. Pour toutefois déterminer un hazard ratio par patient l'approche suivante a été utilisée. Pour chaque patient sont déterminés les patients qui lui sont le plus similaires sur les caractéristiques de base retenues. À partir de ce groupe, il est possible de calculer le hazard ratio à l'aide d'un modèle de Cox classique. L'hypothèse est que parmi des sujets ayant les mêmes caractéristiques le hazard ratio est identique. Le hazard ratio obtenu à partir de groupe de patients est une estimation du hazard ratio individuel (à deux réserves près, ce calcul ne permet pas de faire une inférence causale et il faut que tous les modificateurs de l'effet soient pris en compte dans les caractéristiques envisagées). Les patients similaires à un patient donné sont déterminés parmi la population de l'étude en prenant les plus proches patients dans l'espace multidimensionnel des caractéristiques de bases considérées. Dans ce papier, 75 caractéristiques de base ont été considérées, créant ainsi un espace à 75 dimensions. Pour chaque patient, les 5% de ceux les plus proches de lui sont identifiés et utilisés pour calculer le hazard ratio considéré comme reflétant l'effet traitement individuel du patient. Ensuite ces hazard ratio individuel sont modélisés afin d'obtenir un outil prédictif de l'effet traitement individuel en fonction des caractéristiques des patients. Cette modélisation a été effectuée par un XBG afin d'obtenir un algorithme explicable. Ensuite un outil prédictif a été construit à partir des variables les plus fortement associées avec l'effet de la canagliflozine sur les événements cardiovasculaires.
Une validation externe de cet outil de prédiction a été effectuée avec les données de l'essai CANVAS-R (qui est un essai randomisé de la canagliflozine comparable à CANVAS, les résultats de ces 2 essais ont d'ailleurs été publiés après pooling). Cette validation externe a consisté à classer les patients de CANVAS R en 2 groupes de répondeurs (high et low responders) en fonction du HR prédit, les high responders étant défini par un hazard ratio prédit inférieur à 0.5 écart type (dans la distribution de tous les hazard ratio prédits). Les hazard ratio observés sont ensuite calculés pour chacun de ces 2 groupes de patients avec leur p-value d'interaction. Il est effectivement trouvé que l'effet de la cangliflozine versus placébo sur les MACE a été significativement plus important en moyenne chez les patients identifiés comme hautement répondeurs (high responders), avec un HR de 0.60, que chez les autres (low responders) où le HR est de 0.99, p d'interaction = 0.04. Cette validation en 2 catégories est assez rudimentaire, mais il faut noter que le hazard ratio individuel n'ayant pas de réalité il ne peut directement d'observer dans l'étude de validation externe. Il est donc impossible de valider les prédictions par rapport à l'observé patient par patient.

Malgré son élégance cette approche présente plusieurs limites. Elles reposent sur au moins 2 hypothèses fondamentales invérifiables : 1) l'homogénéité des hazard ratio individuels des plus proches voisins, qui implique que tous les modificateurs d'effet sont couverts par les caractéristiques de base prises en considération et 2) le modèle de prédiction est bien spécifié. Une autre limite est l'impossibilité de faire une validation externe directe des prédictions du hazard ratio individuel.
La dernière limite est celle de la validité du seuil de binarisation. Ce seuil a été défini d'après la distribution des hazard ratio individuel obtenu avec la population de CANVAS-R. Ce seuil est-il transposable à de futurs patients ne provenant pas de cette étude ? Pour une utilisation en pratique médicale quel est le seuil à utiliser ? Une option pourrait être de considérer simplement la valeur du hazard ratio et d'écarter le recours à la canagliflozine quand le hazard ratio prédit est égal ou supérieur à 1. Se pose alors la question du traitement de remplacement qui pourrait être une autre gliflozine (dapagliflozine ou empagliflozine), mais 1) il est possible de penser que toutes les molécules de la classe partagent les mêmes modificateurs d'effet et 2) dans ce cas le choix serait effectué sans se baser sur la prédiction de leur bénéfice (pas manque d'outil). Ce dernier point montre bien la faible utilité pratique pour l'instant de ces approches qui reste du domaine de la recherche et ne peuvent pas être considérés comme des aides à la décision opérationnelle. Les outils nécessaires à la décision en pratique nécessitent une prédiction du bénéfice de tous les traitements envisageables afin d'homogénéiser le principe du choix entre toutes ces molécules. Un outil prédisant le bénéfice que d'un traitement n'est acceptable que lorsqu'il n'existe pas d'alternative à ce traitement et que la question est finalement y-a-t-il un bénéfice à ajouter cette molécule au traitement de ce patient.
Finalement, il convient aussi de remarquer que la modélisation effectuée par cette approche pouvait être effectuée par un modèle de Cox avec intégration des modificateurs d'effet potentiels par des termes d'interaction traitement covariable. Comme le nombre final de variables retenues pour l'outil de prédiction a été de 15, l'instanciation d'un tel modèle avec autant de patients aurait été certainement réalisable sans grande difficulté. Dans ce cas de figure, l'avantage de recourir à l'IA et au machine learning n'est pas évident. Cependant lorsque la prédiction se base sur de l'imagerie [124], l'IA avec le deep learning s'impose.
Le même groupe a utilisé la même méthodologie pour la prédiction du bénéfice de l'intensification du traitement antihypertenseur à partir des données de l'essai randomisé SPRINT [125].

6.4.3 Validation

L'IA n'apporte pas de solution magique à la question de la personnalisation des traitements ou de l'identification des marqueurs prédictifs, contrairement à ce qui pourrait laisser penser le buzz qui entoure cette approche actuellement [10].

L'IA, avec le machine learning, offre simplement des techniques de classification et de prédiction un peu différentes des méthodes statistiques habituelles [120]. Mais rien dans l'IA ne fait disparaitre les problématiques de base (effet traitement individuel non observable, performance des prédictions dépendant entièrement de l'existence de réels déterminants du bénéfice et de leur capture dans les variables du jeu de données d'entrainement, modélisation du bénéfice absolu ou de l'effet relatif). Ces techniques sont donc utilisées pour construire des modèles de prédiction du cATE soit par la modélisation du risque ou directement de l'effet (cf. section 6.1).

Au niveau de la fiabilité des prédictions, là aussi rien de magique avec l'IA, et les performances prédictives des outils construits par ces approches doivent être validées de la même façon que ceux issus des techniques statistiques habituelles [126]. Après leur construction, ces outils doivent faire l'objet d'études de validation externe correctement conçues et réalisées. Ils ne pourront être déployés en pratique médicale courante que si ces études de validation externe confirment la généralisabilité de leur performance. Comme la validation externe de la performance de leur prédiction est limitée par

l'impossibilité d'observer la vraie valeur de l'effet traitement individuel, la démonstration de leur utilité médicale produite par des essais de stratégie est indispensable compte tenu des enjeux sous-jacents (perte de chance possible).

Il convient d'être très vigilant sur ces aspects, car la méta-épidémiologie met en évidence une grande faiblesse de l'évaluation clinique des outils d'IA produits à l'heure actuelle. La validation externe est loin d'être systématique et la majorité des études de validation sont à haut risque de biais [127, 128, 129, 130]. Au total, il est souvent impossible de conclure sur le réel niveau de performances des outils développés, ce qui bloque leur utilisation en pratique.

6.5 Limites de l'approche

Par essence ces analyses ne sont qu'exploratoires et présentent les mêmes limites que les analyses en sous-groupes conventionnelles : multiplicité avec inflation du risque de faux positif et de faux négatifs. De plus, lorsque de nombreuses covariables sont prises en compte simultanément survient un risque de surajustement (overfitting). Il semble difficile d'inscrire cette approche de modélisation dans une démarche hypothético-déductive de confirmation. Quand des marqueurs candidats sont identifiés à priori, il semble bien plus logique d'utiliser une des autres approches qui sont conçues pour apporter des démonstrations (design d'interaction, sous-groupes de confirmation, essai de stratégie).

La modélisation en fonction du risque présente l'avantage d'utiliser le risque de base, ou plus exactement un score de risque, pour diminuer le nombre de covariables et le risque de surajustement. Mais cette approche est cependant basée sur l'acceptation implicite que le score de risque est un marqueur prédictif de l'effet relatif (odds ratio dans l'exemple précédent). Or rien ne garantit cela étant donné que pronostic et modificateurs d'effet du traitement (interaction) sont deux concepts différents (cf. section 2). De plus, l'effet relatif est ajusté sur le risque de base étant donné que le devenir des patients sous un traitement est exprimé relativement à celui sans traitement (par exemple avec un risque ratio). Ainsi ce type de modélisation de l'effet relatif basée sur le risque n'aboutit que dans quelques cas particuliers où le risque est modificateur de l'effet relatif soit directement soit indirectement, car un ou plusieurs marqueurs pronostiques sont aussi marqueurs prédictifs. De plus tous les réels marqueurs prédictifs qui ne sont pas marqueurs pronostiques ne seront pas considérés, conduisant à une explication parcellaire de l'hétérogénéité de l'effet traitement. En revanche, le bénéfice absolu dépend directement du risque de base et ce principe est le fondement de la personnalisation sur le risque de base (cf. section 5).

La modélisation des interactions de l'approche basée sur les effets expose à des problématiques bien connues de surparamétrisation, de taille d'échantillon et de pénalisation des modèles [98].

À cela s'ajoutent des difficultés techniques [131], avec le constat que les différentes méthodes produisent des résultats différents [132]. En pratiques ces méthodes sont encore peut utilisées avec une hétérogénéité des méthodes employées [108].

La difficulté principale liée à cette approche de modélisation est, qu'avant de mettre en pratique le modèle prédictif du bénéfice et ainsi choisir le traitement d'un patient en fonction du résultat prédit, il est impératif d'avoir une **validation externe du modèle** [98]. La validation directe de ces modélisations est impossible étant donné que le vrai bénéfice personnalisé est non observable. La validation externe ne peut s'effectuer qu'avec une granularité de sous-groupes, par exemple des quintiles ou des déciles de risque prédits [98].

Pour effectuer cette validation il faut un jeu de données complètement différent de celui ayant permis de construire le modèle (une séparation en deux parties d'un même jeu de donnée avec une partie servant à identifier le modèle et l'autre servant pour la validation ne constitue pas une vraie validation externe et n'est pas suffisante). Une réelle difficulté survient alors, car le jeu de données de validation devrait être celui-ci d'un autre essai randomisé, comparant les mêmes traitements dans le même contexte clinique, avec les mêmes critères de jugements et un effectif suffisant. En pratique ce jeu de données n'existe pas (sauf peut-être dans les rares cas où deux phases 3 identiques sont réalisées). Des réflexions sont en cours pour explorer la possibilité d'obtenir ces jeux de validation avec les données de vraie vie [133]. La validation externe des modèles issus de cette approche de modélisation semble très difficile, enlevant beaucoup d'intérêt à cette approche. En effet l'absence de réelle validation fait courir le risque de priver à tort des patients d'un traitement efficace (cf. section 1.2).

7 L'évaluation de l'utilité clinique par les essais de stratégie

7.1 Principe des essais de stratégie

Les essais de stratégie ont pour objectif de démontrer l'utilité médicale des marqueurs prédictifs ou de tout autre type d'outils proposés pour prédire le bénéfice des traitements. Ils permettent de savoir si l'utilisation du marqueur prédictif dans un but décisionnel (c'est-à-dire de personnalisation des traitements) permet effectivement, en pratique, d'améliorer le devenir des patients (davantage d'évènements prévenus, moins d'effets indésirables sans perte d'efficacité, etc.) par rapport à une stratégie n'intégrant pas le marqueur (standard de soin).

Le design de ces études est représenté Figure 4. Les patients éligibles sont randomisés entre un bras « stratégie basée sur le marqueur prédictif » et un bras « stratégie habituelle ».

> *We conducted a randomized, open-label, assessor-blinded trial in which patients undergoing primary PCI with stent implantation were assigned in a 1:1 ratio to receive either a P2Y12 inhibitor on the basis of early CYP2C19 genetic testing (**genotype-guided group**) or standard treatment with either ticagrelor or prasugrel (**standard-treatment group**) for 12 months [134]*

Dans le bras expérimental, le marqueur est évalué chez tous les patients et le traitement est choisi en fonction du résultat. Plusieurs variantes sont possibles à ce niveau, le choix peut se faire entre deux traitements conventionnels déjà disponibles ou entre un nouveau traitement ou une nouvelle modalité de ce traitement et le traitement conventionnel (comme le doublement de la dose de clopidogrel dans l'exemple de l'essai TAILOR-PCI développé ci-dessous).

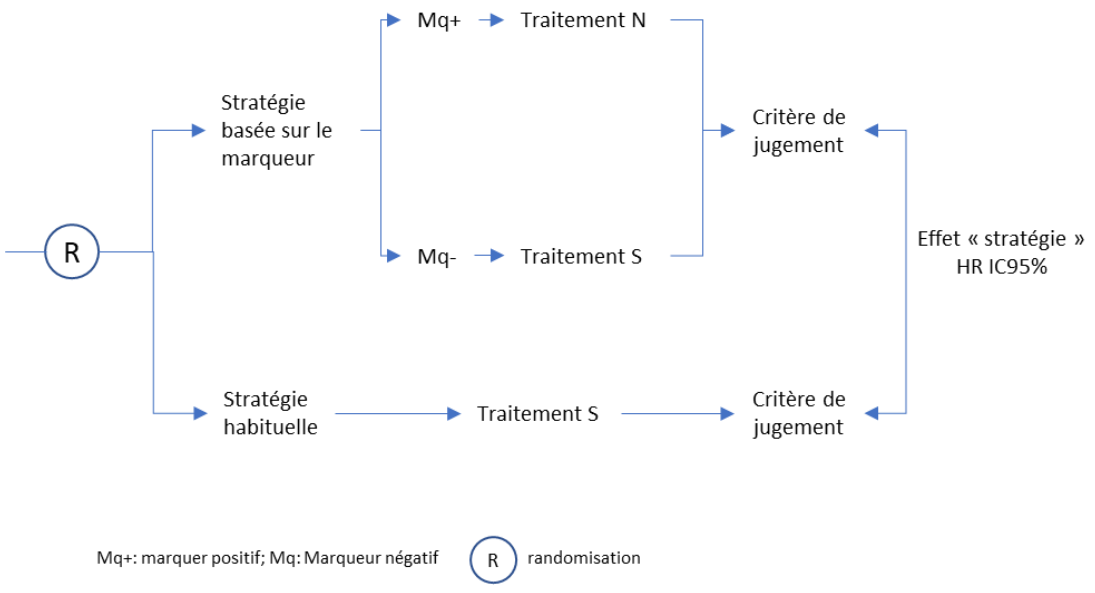


Figure 4 – Principe des essais de stratégie

De nombreuses variantes de ce design sont utilisées [16]. En particulier, se pose la question du dénominateur. La logique du design voudrait que la fréquence des évènements, calculée avec le nombre de patients randomisés dans chaque bras comme dénominateur, soit comparée globalement entre le bras expérimental et le bras contrôle. Cette analyse répond à la question « en quoi l'adoption de la stratégie de personnalisation éviterait globalement la survenue d'évènements », sur l'ensemble de la population des malades correspondants. Cependant parmi tous ces malades le bénéfice attendu ne concerne structurellement qu'une partie d'entre eux (plus ou moins importante en fonction de la prévalence de la positivité du marqueur). Une autre option d'analyse est de ne considérer dans les deux groupes que les patients marqueurs positifs et de comparer ainsi la fréquence des évènements chez les patients marqueurs positifs dans le bras expérimental versus celle dans le bras contrôle aussi uniquement chez les mêmes patients marqueurs positifs. Cette approche nécessite que le marqueur soit là aussi recherché dans le groupe contrôle mais non exploité (avec un risque de contamination intergroupe si le résultat est disponible pour les investigateurs).

Exemple – TAILOR-PCI

L'essai TAILOR-PCI [135] a comparé, après pose d'un stent coronarien, une stratégie d'adaptation de la dose du clopidogrel basée sur le génotype par rapport à la pratique habituelle d'une dose fixe. Le critère de jugement était les évènements cardiovasculaires. Le génotype recherché était les variants du gène CYP2C19*2 or *3 considérés comme des marqueurs de perte de fonction (cf. supra).
L'essai a inclus 5302 patients. L'analyse primaire préspécifiée ne portait que sur les patients porteurs des variants de perte de fonction, ce qui conduisait à exclure de l'analyse 3453 patients. L'analyse primaire porte finalement sur 1849 patients. Une analyse sur tous les patients inclus était prévue comme analyse secondaire.
L'essai échoue à montrer une réduction des évènements cardiovasculaires avec un résultat de l'analyse préspécifiée non statistiquement significatif alors que l'essai avait la puissance voulue pour mettre en évidence l'effet traitement recherché. Cet essai n'a donc pas apportée de preuve formelle de l'utilité médicale de la mise en pratique d'une stratégie basée sur le génotype du CYP2C19 dans ce contexte clinique.
D'autres essais similaires [134, 136] ont donné des résultats comparables.

7.2 Exemple 1 – ARTIC

L'essai ARTIC [136] a évalué une stratégie d'intensification de la dose de clopidogrel sur la base d'un test d'agrégation plaquettaire lors de la pose d'un stent coronarien dans la prévention des évènements cardiovasculaires chez des patients bénéficiant de la pose d'un stent. Le rationnel de l'étude est de détecter avec le test de fonction plaquettaire les patients qui, malgré les doses standards d'antiagrégant plaquettaire (aspirine + un inhibiteur d'agrégation), n'ont pas d'inhibition suffisante de l'agrégation de leurs plaquettes (résistance) et d'intensifier la dose chez eux pour obtenir un niveau d'inhibition suffisant.

Les patients ont été randomisés entre un bras où le traitement était adapté en fonction du test d'agrégation et un bras contrôle où tous les patients recevaient une dose standard.

L'essai n'a pas pu mettre en évidence de bénéfice de stratégie d'adaptation basée sur le test de fonction plaquettaire avec un hazard ratio de 1.06; 95% CI, 0.74 to 1.52; P=0.77.

7.3 Exemple 2 – L'essai COAG, génotypage pour l'ajustement des doses de la warfarine

L'essai COAG [137] est un essai de stratégie qui a évalué l'apport du génotypage dans le choix des doses initiales lors de l'instauration d'un traitement anticoagulant par warfarine. Le génotypage portait sur deux gènes, CYP2C9 et VKORC1, qui avaient été trouvés associés avec la dose de maintenance dans des études observationnelles de type « treatment only » (cf. section 4.1). Les patients, chez lesquels devait être initié un traitement par warfarine étaient randomisés entre un bras où la dose initiale était déterminée en fonction de leur génotypage et un bras où la détermination de la dose initiale reposait sur un algorithme clinique. Le critère de jugement était le temps passé dans la zone d'INR entre 2 et 3 durant les 4 premières semaines de traitement. L'essai n'a pas pu mettre en évidence que l'ajustement de dose basé sur le génotypage améliorait le contrôle de la coagulation durant les 4 premières semaines. Deux autres essais de stratégie analogues publiés simultanément ont donné des résultats similaires. [138, 139].

7.4 Exemple 3 – pharmacogénétique pour la prévention des effets indésirables

L'essai européen PREPARE [140] a évalué l'intérêt médical d'un panel pharmacogénétique de 12 gènes dans la prévention des effets indésirables médicamenteux. La stratégie évaluée reposait sur le génotypage des patients qui devaient recevoir une première prescription d'un médicament pour lesquels étaient connus un ou des variants génétiques de susceptibilité aux effets indésirables. En fonction du résultat du génotypage (présence ou non d'une interaction gène-médicament) un changement de traitement était envisagé. Le critère de jugement était la survenue d'un effet indésirable cliniquement pertinent au cours du suivi de 12 semaines. Cette stratégie était comparée à la pratique habituelle. Une réduction de la fréquence des effets indésirables a été notée dans la sous-population des patients chez lesquels une interaction gène-médicament était présente 152/725 (21·0%) versus 231/833 (27·7%) (odds ratio [OR] 0·70 [95% CI 0·54–0·91]; p=0·0075). Une réduction a aussi été obtenue sur la totalité des patients de l'étude (porteur ou non d'une interaction gène-médicament) 628/2923 (21·5%) versus 934/3270 (28·6%) (OR 0·70 [95% CI 0·61–0·79]; p <0·0001).

Cependant cet essai ne répond qu'à la moitié de la question. La possibilité d'une perte de bénéfice en switchant pour un autre traitement moins efficace n'est pas documentée. Or il s'agit d'un aspect essentiel de la problématique. Le but n'est pas d'avoir moins d'EI mais bien d'optimiser la balance bénéfice risque. L'aspect bénéfice est donc tout aussi primordial.

Cet aspect n'est pas documenté dans PREPARE et les résultats ne sont vraiment probants, en terme décisionnel, qu'en faisant l'hypothèse que pour les couples considérés toutes les alternatives thérapeutiques sont aussi efficaces les unes que les autres (tous les anticonvulsivants sont équivalents en efficacité, tous les antipsychotiques, tous les hypocholestérolémiants, etc.).

La prise en compte de cette double question impose donc soit la réalisation d'un essai de non-infériorité (comme dans l'essai du clopidogrel par Claassens [134]) soit la démonstration d'un bénéfice clinique net en supériorité (et non pas en non-infériorité).

L'enjeu d'une évaluation rigoureuse de ces stratégies provient du fait que ces approches peuvent induire une perte de chance pour les patients en privant certains d'entre eux du traitement optimal de leur pathologie alors que le risque d'EI n'invalide pas le bénéfice clinique net chez eux. En effet chaque fois qu'un traitement est placé en tête dans la stratégie thérapeutique, cela a été fait sur la base d'une balance bénéfice risque favorable. Ainsi la possibilité d'EI ne remet pas en cause a priori le

bénéfice net du traitement chez un patient particulier lorsque l'on instaure le traitement sauf si l'on est capable d'identifier avec une forte certitude ceux qui ont un surrisque suffisamment important pour contrebalancer le bénéfice. En effet ces tests ne discriminent jamais, sauf exception, en tout ou rien et sont le plus souvent que des marqueurs pronostiques et dans quelques cas des marqueurs prédictifs. Récuser certains patients uniquement sur l'aspect EI conduit à lui prescrire, dans beaucoup de domaines, le traitement précédent que l'on sait inférieur en balance bénéfice risque au traitement récusé, jusqu'à preuve du contraire pour ce patient. Il y a donc a priori une perte de chance induite par cette pratique sauf si un essai de haut de niveau de preuve démontre formellement le bénéfice clinique de cette approche. D'où l'importance de cette évaluation.

7.5 Intérêt des essais de stratégie

L'avantage de ces essais de stratégie est d'évaluer simultanément toute la chaine sous-jacente à l'utilisation d'un marqueur prédictif : performance prédictive du marqueur et bénéfice/risque des possibilités thérapeutiques proposées en fonction du marqueur. Cette évaluation intègre aussi les éventuels effets délétères propres au marqueur prédictif (non-traitement à tort des patients, effets indésirables propres à leur évaluation pour les marqueurs invasifs ainsi que ceux découlant des traitements). Cette utilité médicale sera d'autant mieux appréhendée que le critère de jugement intègrera les bénéfices et les effets indésirables (bénéfice clinique net).

Ces essais évaluent directement le changement de devenir des patients que pourrait produire la stratégie intégrant le marqueur prédictif afin de déterminer s'il y a amélioration de la prise en charge.

Ces essais raisonnent au niveau de la population totale des patients qui seront « dépistés » (qu'ils soient ou non marqueurs positifs et concernés par le changement de thérapeutiques) ce qui a l'avantage d'intégrer dans l'évaluation la prévalence du marqueur positif dans la population générale des patients concernés. Cette approche permet ainsi d'évaluer la rentabilité de soumettre à un test la totalité des patients alors que seulement une partie, plus ou moins importante de ceux-ci, sont susceptibles d'être concernés.

Références

1 Maughan T. The Promise and the Hype of 'Personalised Medicine'. *New Bioeth* 2017;23:13–20 doi:10.1080/20502877.2017.1314886; PMID:28517988;

2 Feiler T, Gaitskell K, Maughan T, et al. Personalised Medicine: The Promise, the Hype and the Pitfalls. *New Bioeth* 2017;23:1–12 doi:10.1080/20502877.2017.1314895; PMID:28517985;

3 Pelter MN, Druz RS. Precision medicine: Hype or hope? *Trends Cardiovasc Med* 2022 doi:10.1016/j.tcm.2022.11.001; PMID:36375778;

4 Prasad V, Gale RP. Precision medicine in acute myeloid leukemia: Hope, hype or both? *Leuk Res* 2016;48:73–77 doi:10.1016/j.leukres.2016.07.011; PMID:27497757;

5 Polasek TM, Shakib S, Rostami-Hodjegan A. Precision medicine technology hype or reality? The example of computer-guided dosing. *F1000Res* 2019;8 doi:10.12688/f1000research.20489.1; PMID:31754426;

6 Joyner MJ, Paneth N. Promises, promises, and precision medicine. *J Clin Invest* 2019;129:946–48 doi:10.1172/JCI126119; PMID:30688663;

7 Joyner MJ, Paneth N. Seven Questions for Personalized Medicine. *JAMA* 2015;314:999–1000 doi:10.1001/jama.2015.7725; PMID:26098474;

8 Montello M. Announcement from Perspectives in Biology and Medicine. *Perspect Biol Med* 2018;61:465–66 doi:10.1353/pbm.2018.0056;

9 Caulfield T. Spinning the Genome: Why Science Hype Matters. *Perspect Biol Med* 2018;61:560–71 doi:10.1353/pbm.2018.0065; PMID:30613038;

10 Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med* 2018;16:150 doi:10.1186/s12916-018-1122-7; PMID:30145981;

11 Brown MJ. Personalised medicine for hypertension. *BMJ* 2011;343:d4697 doi:10.1136/bmj.d4697; PMID:21798973;

12 Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnology* 2012;29:613–24 doi:10.1016/j.nbt.2012.03.004; PMID:22450380;

13 Rojas LA, Sethna Z, Soares KC, et al. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* 2023:1–7 doi:10.1038/s41586-023-06063-y; PMID:37165196;

14 Kerr DJ, Yang L. Personalising cancer medicine with prognostic markers. *eBioMedicine* 2021;72:103577 doi:10.1016/j.ebiom.2021.103577; PMID:34563926;

15 Oikonomou EK, Suchard MA, McGuire DK, et al. Phenomapping-Derived Tool to Individualize the Effect of Canagliflozin on Cardiovascular Risk in Type 2 Diabetes. *Diabetes Care* 2022;45:965–74 doi:10.2337/dc21-1765; PMID:35120199;

16 Superchi C, Brion Bouvier F, Gerardi C, et al. Study designs for clinical trials applied to personalised medicine: a scoping review. *BMJ open* 2022;12:e052926 doi:10.1136/bmjopen-2021-052926; PMID:35523482;

17 Antonia SJ, Villegas A, Daniel D, et al. Overall Survival with Durvalumab after Chemoradiotherapy in Stage III NSCLC. *N Engl J Med* 2018;379:2342–50 doi:10.1056/NEJMoa1809697; PMID:30280658;

18 Mok TS, Wu Y-L, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 2009;361:947–57 doi:10.1056/NEJMoa0810699; PMID:19692680;

19 Angus DC, Chang C-CH. Heterogeneity of Treatment Effect: Estimating How the Effects of Interventions Vary Across Individuals. *JAMA* 2021;326:2312–13

doi:10.1001/jama.2021.20552; PMID:34905043;

20 Dorresteijn JAN, Visseren FLJ, Ridker PM, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ* 2011;343 doi:10.1136/bmj.d5888; PMID:21968126;

21 Gewandter JS, McDermott MP, He H, et al. Demonstrating Heterogeneity of Treatment Effects Among Patients: An Overlooked but Important Step Toward Precision Medicine. *Clin Pharmacol Ther* 2019 doi:10.1002/cpt.1372; PMID:30661240;

22 Dahly DL. Response Heterogeneity | 2019. Available at: https://darrendahly.github.io/post/2017-08-11-response-heterogeneity/ Accessed March 20, 2023.

23 Harrell F. Statistical Thinking - Viewpoints on Heterogeneity of Treatment Effect and Precision Medicine 2023. Available at: https://www.fharrell.com/post/hteview/ Accessed March 20, 2023.

24 Senn S. Statistical pitfalls of personalized medicine. *Nature* 2018;563:619–21 doi:10.1038/d41586-018-07535-2; PMID:30482931;

25 Senn S. Mastering variation: variance components and personalised medicine. *Stat Med* 2016;35:966–77 doi:10.1002/sim.6739; PMID:26415869;

26 Senn S, Rolfe K, Julious SA. Investigating variability in patient response to treatment--a case study from a replicate cross-over study. *Stat Methods Med Res* 2011;20:657–66 doi:10.1177/0962280210379174; PMID:20739334;

27 Sundström J, Lind L, Nowrouzi S, et al. Heterogeneity in Blood Pressure Response to 4 Antihypertensive Drugs: A Randomized Clinical Trial. *JAMA* 2023;329:1160–69 doi:10.1001/jama.2023.3322; PMID:37039792;

28 Mueller S, Pearl J. Personalized decision making – A conceptual introduction. *Journal of Causal Inference* 2023;11 doi:10.1515/jci-2022-0050;

29 Pearl J. An introduction to causal inference. *The International Journal of Biostatistics* 2010;6:Article 7 doi:10.2202/1557-4679.1203; PMID:20305706;

30 Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health* 2005;95 Suppl 1:S144-50 doi:10.2105/AJPH.2004.059204; PMID:16030331;

31 Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004;58:265–71 doi:10.1136/jech.2002.006361; PMID:15026432;

32 Hernán M, Robins JM. Causal inference. Boca Raton: Chapman & Hall/CRC 2021 ISBN:1420076167;

33 Ballman KV. Biomarker: Predictive or Prognostic? *JCO* 2015;33:3968–71 doi:10.1200/JCO.2015.63.3651; PMID:26392104;

34 Janes H, Pepe MS, McShane LM, et al. The Fundamental Difficulty With Evaluating the Accuracy of Biomarkers for Guiding Treatment. *JNCI Journal of the National Cancer Institute* 2015;107 doi:10.1093/jnci/djv157; PMID:26109106;

35 Oldenhuis CNAM, Oosting SF, Gietema JA, et al. Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer* 2008;44:946–53 doi:10.1016/j.ejca.2008.03.006; PMID:18396036;

36 Dobbin KK, McShane LM. Sample size methods for evaluation of predictive biomarkers. *Stat Med* 2022;41:3199–210 doi:10.1002/sim.9412; PMID:35491401;

37 Clark GM. Prognostic factors versus predictive factors: Examples from a clinical trial of erlotinib. *Mol Oncol* 2008;1:406–12 doi:10.1016/j.molonc.2007.12.001; PMID:19383314;

38 Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. *The Lancet* 1994;343:311–22 ; PMID:7905143;

39 Cross DAE, Ashton SE, Ghiorghiu S, et al. AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer Discov* 2014;4:1046–61 doi:10.1158/2159-8290.CD-14-0337; PMID:24893891;

40 Ramalingam SS, Vansteenkiste J, Planchard D, et al. Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC. *N Engl J Med* 2020;382:41–50 doi:10.1056/NEJMoa1913662; PMID:31751012;

41 Skoulidis F, Li BT, Dy GK, et al. Sotorasib for Lung Cancers with KRAS p.G12C Mutation. *N Engl J Med* 2021;384:2371–81 doi:10.1056/NEJMoa2103695; PMID:34096690;

42 Sachdev A, Sharpe I, Bowman M, et al. Objective response rate of placebo in randomized controlled trials of anticancer medicines. *EClinicalMedicine* 2023;55:101753 doi:10.1016/j.eclinm.2022.101753;

43 Chowdary P, Shapiro S, Makris M, et al. Phase 1-2 Trial of AAVS3 Gene Therapy in Patients with Hemophilia B. *N Engl J Med* 2022;387:237–47 doi:10.1056/NEJMoa2119913; PMID:35857660;

44 Gyawali B, Rome BN, Kesselheim AS. Regulatory and clinical consequences of negative confirmatory trials of accelerated approval cancer drugs: retrospective observational study. *BMJ* 2021;374:n1959 doi:10.1136/bmj.n1959; PMID:34497044;

45 Langen AJ de, Johnson ML, Mazieres J, et al. Sotorasib versus docetaxel for previously treated non-small-cell lung cancer with KRASG12C mutation: a randomised, open-label, phase 3 trial. *The Lancet* 2023;401:733–46 doi:10.1016/S0140-6736(23)00221-0; PMID:36764316;

46 Joyner MJ, Paneth N (2019). Precision medicine's rosy predictions haven't come true. We need fewer promises and more debate. *STAT,* 7 February 2019. Available at: https://www.statnews.com/2019/02/07/precision-medicine-needs-open-debate/ Accessed March 19, 2023.

47 (20180921T123000Z). Precision medicine fails for up to 93% of patients. Are its proponents selling 'false hope'? *Advisory Board,* 20180921T123000Z. Available at: https://www.advisory.com/daily-briefing/2018/09/21/precision-medicine Accessed March 19, 2023.

48 Szabo L (2018). Opinion | Are We Being Misled About Precision Medicine? *The New York Times,* 11 September 2018. Available at: https://www.nytimes.com/2018/09/11/opinion/cancer-genetic-testing-precision-medicine.html Accessed March 19, 2023.

49 Le Tourneau C, Delord J-P, Gonçalves A, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The Lancet Oncology* 2015;16:1324–34 doi:10.1016/S1470-2045(15)00188-6;

50 Chen AP, Kummar S, Moore N, et al. Molecular Profiling-Based Assignment of Cancer Therapy (NCI-MPACT): A Randomized Multicenter Phase II Trial. *JCO Precision Oncology* 2021;5 doi:10.1200/PO.20.00372; PMID:33928209;

51 MLAK R, KRAWCZYK P, RAMLAU R, et al. Predictive value of ERCC1 and RRM1 gene single-nucleotide polymorphisms for first-line platinum- and gemcitabine-based chemotherapy in non-small cell lung cancer patients. *Oncology Reports* 2013;30:2385–98 doi:10.3892/or.2013.2696; PMID:23982437;

52 Han Y, Liu J, Sun M, et al. A Significant Statistical Advancement on the Predictive Values of ERCC1 Polymorphisms for Clinical Outcomes of Platinum-Based Chemotherapy in Non-Small Cell Lung Cancer: An Updated Meta-Analysis. *Disease Markers* 2016;2016:7643981 doi:10.1155/2016/7643981; PMID:27057082;

53 Lee SM, Falzon M, Blackhall F, et al. Randomized Prospective Biomarker Trial of ERCC1 for Comparing Platinum and Nonplatinum Therapy in Advanced Non-Small-Cell Lung Cancer: ERCC1 Trial (ET). *JCO* 2017;35:402–11 doi:10.1200/JCO.2016.68.1841; PMID:27893326;

54 Vivot A, Boutron I, Béraud-Chaulet G, et al. Evidence for Treatment-by-Biomarker interaction for FDA-approved Oncology Drugs with Required Pharmacogenomic Biomarker Testing. *Sci Rep* 2017;7:6882 doi:10.1038/s41598-017-07358-7; PMID:28761069;

55 Marabelle A, Fakih MG, Lopez J, et al. Association of tumour mutational burden with outcomes in patients with select advanced solid tumours treated with pembrolizumab in KEYNOTE-158. *Annals of Oncology* 2019;30:v477-v478 doi:10.1093/annonc/mdz253.018;

56 Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* 2015;372:2509–20 doi:10.1056/NEJMoa1500596; PMID:26028255;

57 Nikola N, Aleksa G, Ana R, et al. Brief Report: Predictive value of PD-L1 Expression in non-Small-Cell Lung Cancer - Should we Set the Bar Higher for Monotherapy? *Clinical Lung Cancer* 2023;0 doi:10.1016/j.cllc.2023.04.010; PMID:37236852;

58 Kim JH, Ryu M-H, Park YS, et al. Predictive biomarkers for the efficacy of nivolumab as ≥ 3rd-line therapy in patients with advanced gastric cancer: a subset analysis of ATTRACTION-2 phase III trial. *BMC Cancer* 2022;22 doi:10.1186/s12885-022-09488-2; PMID:35397540;

59 Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet* 2005;365:176–86 doi:10.1016/S0140-6736(05)17709-5;

60 Maemondo M, Inoue A, Kobayashi K, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* 2010;362:2380–88 doi:10.1056/NEJMoa0909530; PMID:20573926;

61 Mitsudomi T, Morita S, Yatabe Y, et al. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *The Lancet. Oncology* 2010;11:121–28 doi:10.1016/S1470-2045(09)70364-X; PMID:20022809;

62 Kent DM, Nelson J, Dahabreh IJ, et al. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *Int J Epidemiol* 2016;45:dyw118 doi:10.1093/ije/dyw118; PMID:27375287;

63 Doody RS, Thomas RG, Farlow M, et al. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N Engl J Med* 2014;370:311–21 doi:10.1056/NEJMoa1312889; PMID:24450890;

64 Honig LS, Vellas B, Woodward M, et al. Trial of Solanezumab for Mild Dementia Due to Alzheimer's Disease. *N Engl J Med* 2018;378:321–30 doi:10.1056/NEJMoa1705971; PMID:29365294;

65 Holmes MV, Perel P, Shah T, et al. CYP2C19 genotype, clopidogrel metabolism, platelet function, and cardiovascular events: a systematic review and meta-analysis. *JAMA* 2011;306:2704–14 doi:10.1001/jama.2011.1880; PMID:22203539;

66 Federico Cappuzzo, Tudor Ciuleanu, Lilia Stelmakh, et al. Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomised, placebo-controlled phase 3 study ;

67 McShane LM, Rothmann MD, Fleming TR. Finding the (biomarker-defined) subgroup of patients who benefit from a novel therapy: No time for a game of hide and seek. *Clin Trials* 2023:17407745231169692 doi:10.1177/17407745231169692; PMID:37095696;

68 Socinski MA, Jotte RM, Cappuzzo F, et al. Atezolizumab for First-Line Treatment of Metastatic Nonsquamous NSCLC. *N Engl J Med* 2018;378:2288–301 doi:10.1056/NEJMoa1716948; PMID:29863955;

69 Gregorc V, Novello S, Lazzari C, et al. Predictive value of a proteomic signature in patients with non-small-cell lung cancer treated with second-line erlotinib or chemotherapy (PROSE): a biomarker-stratified, randomised phase 3 trial. *The Lancet Oncology* 2014;15:713–21 doi:10.1016/S1470-2045(14)70162-7;

70 Crippa A, Laere B de, Discacciati A, et al. The ProBio trial: molecular biomarkers for advancing personalized treatment decision in patients with metastatic castration-resistant prostate cancer. *Trials* 2020;21:579 doi:10.1186/s13063-020-04515-8; PMID:32586393;

71 Janiaud P, Serghiou S, Ioannidis JPA. New clinical trial designs in the era of precision medicine: An overview of definitions, strengths, weaknesses, and current use in oncology. *Cancer Treat Rev* 2019;73:20–30 doi:10.1016/j.ctrv.2018.12.003; PMID:30572165;

72 Garralda E, Dienstmann R, Piris-Giménez A, et al. New clinical trial designs in the era of precision medicine. *Mol Oncol* 2019;13:549–57 doi:10.1002/1878-0261.12465; PMID:30698321;

73 Tajik P, Zwinderman AH, Mol BW, et al. Trial Designs for Personalizing Cancer Care: A Systematic Review and Classification. *Clin Cancer Res* 2013;19:4578–88 doi:10.1158/1078-0432.CCR-12-3722; PMID:23788580;

74 Mandrekar SJ, Sargent DJ. Design of clinical trials for biomarker research in oncology. *Clin Investig (Lond)* 2011;1:1629–36 doi:10.4155/CLI.11.152; PMID:22389760;

75 Zhou T, Ji Y. RoBoT: a robust Bayesian hypothesis testing method for basket trials. *Biostatistics* 2021;22:897–912 doi:10.1093/biostatistics/kxaa005; PMID:32061093;

76 Le Tourneau C, Delord J-P, Gonçalves A, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The Lancet Oncology* 2015;16:1324–34 doi:10.1016/S1470-2045(15)00188-6; PMID:26342236;

77 Roustit M, Demarcq O, Laporte S, et al. Platform trials. *Therapie* 2023;78:29–38 doi:10.1016/j.therap.2022.12.003; PMID:36529559;

78 Barker A, Sigman C, Kelloff G, et al. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clin Pharmacol Ther* 2009;86:97–100 doi:10.1038/clpt.2009.68;

79 Zhao A, Larbi M, Miller K, et al. A scoping review of interactive and personalized web-based clinical tools to support treatment decision making in breast cancer. *Breast* 2022;61:43–57 doi:10.1016/j.breast.2021.12.003; PMID:34896693;

80 Adsul P, Wray R, Spradling K, et al. Systematic Review of Decision Aids for Newly Diagnosed Patients with Prostate Cancer Making Treatment Decisions. *J Urol* 2015;194:1247–52 doi:10.1016/j.juro.2015.05.093; PMID:26055824;

81 Yu L, Li P, Yang S, et al. Web-based decision aids to support breast cancer screening decisions: systematic review and meta-analysis. *J Comp Eff Res* 2020;9:985–1002 doi:10.2217/cer-2020-0052; PMID:33025800;

82 Shojaie D, Hoffman AS, Amaku R, et al. Decision Making When Cancer Becomes Chronic: Needs Assessment for a Web-Based Medullary Thyroid Carcinoma Patient Decision Aid. *JMIR Form Res* 2021;5:e27484 doi:10.2196/27484; PMID:34269691;

83 Yung A, Kay J, Beale P, et al. Computer-Based Decision Tools for Shared Therapeutic Decision-making in Oncology: Systematic Review. *JMIR Cancer* 2021;7:e31616 doi:10.2196/31616; PMID:34544680;

84 Austin CA, Mohottige D, Sudore RL, et al. Tools to Promote Shared Decision Making in Serious Illness: A Systematic Review. *JAMA Intern Med* 2015;175:1213–21 doi:10.1001/jamainternmed.2015.1679; PMID:25985438;

85 Blette BS, Granholm A, Li F, et al. Causal Bayesian machine learning to assess treatment effect heterogeneity by dexamethasone dose for patients with COVID-19 and severe hypoxemia. *Sci Rep* 2023;13:6570 doi:10.1038/s41598-023-33425-3; PMID:37085591;

86 Zhao Y, Slate EH, Xu C, et al. Empirical comparisons of heterogeneity magnitudes of the risk difference, relative risk, and odds ratio. *Syst Rev* 2022;11:26

doi:10.1186/s13643-022-01895-7; PMID:35151340;

87 Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd edn. Cham: Springer International Publishing; Imprint: Springer 2019 ISBN:9783030164010;

88 Riley RD, van der Windt D, Croft P, et al., eds. Prognosis research in healthcare: Concepts, methods, and impact. Oxford, United Kingdom: Oxford University Press 2019.

89 Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA* 2017;318:1377–84 doi:10.1001/jama.2017.12126; PMID:29049590;

90 Pencina MJ, D'Agostino RB. Evaluating Discrimination of Risk Prediction Models: The C Statistic. *JAMA* 2015;314:1063–64 doi:10.1001/jama.2015.11082; PMID:26348755;

91 van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230 doi:10.1186/s12916-019-1466-7; PMID:31842878;

92 Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* 2019;170:51–58 doi:10.7326/m18-1376; PMID:30596875;

93 Tie J, Cohen JD, Lahouel K, et al. Circulating Tumor DNA Analysis Guiding Adjuvant Therapy in Stage II Colon Cancer. *N Engl J Med* 2022;386:2261–72 doi:10.1056/nejmoa2200075; PMID:35657320;

94 Sparano JA, Gray RJ, Della Makower F, et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N Engl J Med* 2018;379:111–21 doi:10.1056/NEJMoa1804710; PMID:29860917;

95 Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N Engl J Med* 2016;375:717–29 doi:10.1056/NEJMoa1602253; PMID:27557300;

96 Kent DM, Rothwell PM, Ioannidis JPA, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010;11:85 doi:10.1186/1745-6215-11-85; PMID:20704705;

97 Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ* 2018;363:k4245 doi:10.1136/bmj.k4245; PMID:30530757;

98 Hoogland J, IntHout J, Belias M, et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Stat Med* 2021;40:5961–81 doi:10.1002/sim.9154; PMID:34402094;

99 Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 2007;298:1209–12 doi:10.1001/jama.298.10.1209; PMID:17848656;

100 Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Ann. Intern. Med.* 2020;172:35–45 doi:10.7326/M18-3667; PMID:31711134;

101 Lipkovich I, Svensson D, Ratitch B, et al. Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. *Clin Trials* 2023:17407745231174544 doi:10.1177/17407745231174544; PMID:37203150;

102 Kent DM, van Klaveren D, Paulus JK, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement: Explanation and Elaboration. *Ann. Intern. Med.* 2020;172:W1-W25 doi:10.7326/M18-3668; PMID:31711094;

103 Lipkovich I, Dmitrienko A, B R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med* 2017;36:136–96 doi:10.1002/sim.7064; PMID:27488683;

104 Lipkovich I, Dmitrienko A, Denne J, et al. Subgroup identification based on differential effect search--a recursive partitioning method

for establishing response to treatment in patient subpopulations. *Stat Med* 2011;30:2601–21 doi:10.1002/sim.4289; PMID:21786278;

105 Dmitrienko A, Millen B, Lipkovich I. Multiplicity considerations in subgroup analysis. *Stat Med* 2017;36:4446–54 doi:10.1002/sim.7416; PMID:28762525;

106 Lipkovich I, Dmitrienko A, Muysers C, et al. Multiplicity issues in exploratory subgroup analysis. *J Biopharm Stat* 2018;28:63–81 doi:10.1080/10543406.2017.1397009; PMID:29173045;

107 Rekkas A, Paulus JK, Raman G, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Med Res Methodol* 2020;20:264 doi:10.1186/s12874-020-01145-1; PMID:33096986;

108 Rekkas A, Paulus JK, Raman G, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Med Res Methodol* 2020;20:264 doi:10.1186/s12874-020-01145-1; PMID:33096986;

109 Nguyen T-L, Collins GS, Landais P, et al. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials-An illustration with the International Stroke Trial. *Journal of Clinical Epidemiology* 2020;125:47–56 doi:10.1016/j.jclinepi.2020.05.022; PMID:32464321;

110 Goligher EC, Lawler PR, Jensen TP, et al. Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in Patients Hospitalized for COVID-19. *JAMA* 2023 doi:10.1001/jama.2023.3651;

111 Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med* 2018;16:150 doi:10.1186/s12916-018-1122-7; PMID:30145981;

112 Eckardt J-N, Rollig C, Kramer M, et al. Prediction of Complete Remission and Survival in Acute Myeloid Leukemia Using Supervised Machine Learning. *Blood* 2021;138:108 doi:10.1182/blood-2021-149582;

113 Zawadzki P, Woźna A, Sztromwasser P, et al. 1134P Personalized medicine in advanced breast cancer: AI-driven genomic test for CDK4/6 treatment response prediction. *Annals of Oncology* 2021;32:S925 doi:10.1016/j.annonc.2021.08.776;

114 Sechidis K, Papangelou K, Metcalfe PD, et al. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics* 2018;34:3365–76 doi:10.1093/bioinformatics/bty357; PMID:29726967;

115 Kesari G (2021). Meet Your Digital Twin: The Coming Revolution In Drug Development. *Forbes,* 29 September 2021. Available at: https://www.forbes.com/sites/ganeskesari/2021/09/29/meet-your-digital-twin-the-coming-revolution-in-drug-development/?sh=98cc45c745fb Accessed March 20, 2023.

116 Björnsson B, Borrebaeck C, Elander N, et al. Digital twins to personalize medicine. *Genome Med* 2020;12:1–4 doi:10.1186/s13073-019-0701-3;

117 Al-Tashi Q, Saad MB, Muneer A, et al. Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review. *Int J Mol Sci* 2023;24 doi:10.3390/ijms24097781; PMID:37175487;

118 Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Annals of Oncology* 2019;30:998–1004 doi:10.1093/annonc/mdz108; PMID:30895304;

119 Finlayson SG, Beam AL, van Smeden M. Machine Learning and Statistics in Clinical Research Articles-Moving Past the False Dichotomy. *JAMA Pediatr* 2023 doi:10.1001/jamapediatrics.2023.0034; PMID:36939696;

120 Hakeem H, Feng W, Chen Z, et al. Development and Validation of a Deep Learning Model for Predicting Treatment Response in Patients With Newly Diagnosed Epilepsy. *JAMA Neurol* 2022;79:986–96 doi:10.1001/jamaneurol.2022.2514; PMID:36036923;

121 Falet J-PR, Durso-Finley J, Nichyporuk B, et al. Estimating individual treatment effect on disability progression in multiple sclerosis

using deep learning. *Nat Commun* 2022;13:5645 doi:10.1038/s41467-022-33269-x; PMID:36163349;

122 Temple R. Enrichment of clinical study populations. *Clin Pharmacol Ther* 2010;88:774–78 doi:10.1038/clpt.2010.233; PMID:20944560;

123 Bovis F, Carmisciano L, Signori A, et al. Defining responders to therapies by a statistical modeling approach applied to randomized clinical trial data. *BMC Med* 2019;17:113 doi:10.1186/s12916-019-1345-2; PMID:31208412;

124 Mu W, Jiang L, Zhang J, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat Commun* 2020;11:5228 doi:10.1038/s41467-020-19116-x; PMID:33067442;

125 Oikonomou EK, Spatz ES, Suchard MA, et al. Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials. *Lancet Digit Health* 2022;4:e796-e805 doi:10.1016/S2589-7500(22)00170-4; PMID:36307193;

126 van Calster B, Steyerberg EW, Wynants L, et al. There is no such thing as a validated prediction model. *BMC Med* 2023;21:70 doi:10.1186/s12916-023-02779-w; PMID:36829188;

127 Dhiman P, Ma J, Andaur Navarro CL, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. *Diagn Progn Res* 2022;6:13 doi:10.1186/s41512-022-00126-w; PMID:35794668;

128 Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *Journal of Clinical Epidemiology* 2021;138:60–72 doi:10.1016/j.jclinepi.2021.06.024; PMID:34214626;

129 Efthimiou O, Hoogland J, Debray TPA, et al. Measuring the performance of prediction models to personalize treatment choice. *Stat Med* 2023;42:1188–206 doi:10.1002/sim.9665; PMID:36700492;

130 Wynants L, van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328 doi:10.1136/bmj.m1328; PMID:32265220;

131 van Klaveren D, Balan TA, Steyerberg EW, et al. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology* 2019;114:72–83 doi:10.1016/j.jclinepi.2019.05.029; PMID:31195109;

132 Olsen MK, Stechuchak KM, Oddone EZ, et al. Which patients benefit most from completing health risk assessments: comparing methods to identify heterogeneity of treatment effects. *Health Serv Outcomes Res Method* 2021;21:527–46 doi:10.1007/s10742-021-00243-x;

133 Segal JB, Varadhan R, Groenwold RH, et al. Assessing Heterogeneity of Treatment Effect in Real-World Data. *Ann. Intern. Med.* 2023 doi:10.7326/M22-1510; PMID:36940440;

134 Claassens DMF, Vos GJA, Bergmeijer TO, et al. A Genotype-Guided Strategy for Oral P2Y12 Inhibitors in Primary PCI. *N Engl J Med* 2019;381:1621–31 doi:10.1056/NEJMoa1907096; PMID:31479209;

135 Pereira NL, Farkouh ME, So D, et al. Effect of Genotype-Guided Oral P2Y12 Inhibitor Selection vs Conventional Clopidogrel Therapy on Ischemic Outcomes After Percutaneous Coronary Intervention: The TAILOR-PCI Randomized Clinical Trial. *JAMA* 2020;324:761–71 doi:10.1001/jama.2020.12443; PMID:32840598;

136 Collet J-P, Cuisset T, Rangé G, et al. Bedside monitoring to adjust antiplatelet therapy for coronary stenting. *N Engl J Med* 2012;367:2100–09 doi:10.1056/NEJMoa1209979; PMID:23121439;

137 Kimmel SE, French B, Kasner SE, et al. A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N Engl J Med* 2013;369:2283–93 doi:10.1056/NEJMoa1310669; PMID:24251361;

138 Verhoef TI, Ragia G, Boer A de, et al. A randomized trial of genotype-guided dosing of acenocoumarol and phenprocoumon. *N Engl J Med* 2013;369:2304–12 doi:10.1056/NEJMoa1311388; PMID:24251360;

139 Pirmohamed M, Burnside G, Eriksson N, et al. A randomized trial of genotype-guided dosing of warfarin. *N Engl J Med* 2013;369:2294–303 doi:10.1056/NEJMoa1311386; PMID:24251363;

140 Swen JJ, van der Wouden CH, Manson LE, et al. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. *The Lancet* 2023;401:347–56 doi:10.1016/S0140-6736(22)01841-4; PMID:36739136;