



HAUTE AUTORITÉ DE SANTÉ

GUIDE METHODOLOGIQUE

# Test compagnon associé à une thérapie ciblée : Annexe scientifique

Février 2014

Cette annexe scientifique est téléchargeable sur  
[www.has-sante.fr](http://www.has-sante.fr)

**Haute Autorité de santé**

Service documentation – Information des publics  
2, avenue du Stade de France – F 93218 Saint-Denis La Plaine Cedex  
Tél. : +33 (0)1 55 93 70 00 – Fax : +33 (0)1 55 93 74 00

Cette annexe scientifique a été validée par le Collège de la Haute Autorité de santé en février 2014.  
© Haute Autorité de santé - 2014.

# Sommaire

|  |           |
|--|-----------|
| Abréviations et acronymes .....  | 5         |
| Lexique .....  | 6         |
| <b>1. Principaux concepts .....</b>  | <b>8</b>  |
| 1.1 Avant-propos .....   | 8         |
| 1.2 Notion de marqueur prédictif .....   | 9         |
| 1.2.1 Définition, nature, intérêt .....  | 9         |
| 1.2.2 Différence entre facteur de risque et marqueur prédictif .....   | 10        |
| 1.3 Notion d'interaction statistique .....   | 14        |
| 1.4 Notion de test compagnon.....  | 16        |
| 1.5 Notions de thérapies « ciblée » et « stratifiée » .....  | 17        |
| 1.6 Les différentes situations associées en pratique à la restriction d'utilisation d'un traitement à une sous-population de patients .....  | 18        |
| 1.7 Les intérêts potentiels du ciblage d'une thérapeutique .....   | 19        |
| 1.7.1 Ciblage à raison .....   | 20        |
| 1.7.2 Ciblage à tort .....   | 21        |
| 1.7.3 Ciblage sur un marqueur partiellement prédictif (pseudo-ciblage).....  | 22        |
| 1.7.4 Stratification en fonction d'une interaction sur la sécurité .....   | 23        |
| 1.7.5 Synthèse .....   | 25        |
| <b>2. Principes généraux de validation d'un test diagnostique déterminant le statut des patients pour un marqueur présumé prédictif .....</b>  | <b>26</b> |
| 2.1 Les trois niveaux de validation d'un test diagnostique.....  | 26        |
| 2.2 Les trois conditions impérativement requises pour démontrer l'utilité clinique d'un test diagnostique .....  | 26        |
| 2.2.1 Mise en évidence de l'interaction marqueur/traitement.....   | 27        |
| 2.2.2 Mise en évidence de l'effet du traitement chez les patients marqueur (+) .....   | 28        |
| 2.2.3 Mise en évidence de l'absence d'effet chez les patients marqueurs (-).....   | 28        |
| 2.3 Les critères de jugement .....   | 30        |
| 2.4 Conséquences de mauvaises performances diagnostiques du test sur la démonstration de son utilité clinique .....  | 31        |
| <b>3. Principaux schémas d'études cliniques : analyse critique .....</b>   | <b>33</b> |
| 3.1 L'étude d' « interaction marqueur/traitement » : seul schéma permettant de démontrer l'utilité clinique d'un test identifiant un marqueur présumé prédictif et le caractère ciblé d'une thérapie ..... | 33        |
| 3.2 Les principaux autres schémas d'étude : détermination de l'effet du traitement et identification exploratoire de marqueurs présumés prédictifs.....  | 35        |
| 3.2.1 L'étude « ciblée » .....   | 35        |
| 3.2.2 L'étude de « comparaison de stratégies » .....   | 38        |
| 3.2.3 Analyses en sous-groupes <i>a posteriori</i> .....   | 41        |
| 3.3 Synthèse .....   | 48        |
| <b>4. Méthodologie d'évaluation par la HAS de l'utilité clinique d'un test diagnostique associé à une thérapie stratifiée - Exemples illustratifs .....</b>  | <b>49</b> |
| 4.1 Considérations générales.....  | 49        |
| 4.2 Tableaux de synthèse .....   | 51        |
| 4.3 Exemples théoriques d'application des principes énoncés .....  | 53        |
| 4.3.1 Différentes situations de cas d'interaction.....   | 53        |
| 4.3.2 Exemples de cas théoriques d'évaluation de dossiers .....  | 56        |
| 4.3.3 Exemples de cas issus de la littérature.....   | 61        |
| 4.4 Considérations réglementaires.....   | 62        |

|                                   |    |
|-----------------------------------|----|
| Conclusion - Perspectives .....   | 63 |
| Références .....                  | 64 |
| Annexe 1. Index .....             | 66 |
| Annexe 2. Liste des tableaux..... | 67 |
| Annexe 3. Liste des figures ..... | 68 |
| Fiche descriptive .....           | 69 |

## Abréviations et acronymes

**ALK** .....*anaplastic lymphoma kinase*

**AMM** .....autorisation de mise sur le marché

**ARR** .....*absolute risk reduction*

**AVC** .....accident vasculaire cérébral

**cf**.....confer

**EGFR**.....*epidermal growth factor receptor*

**EIG**.....effet indésirable grave

**EMA**.....*European Medicine Agency*

**FISH**.....*fluorescence in situ hybridization*

**HER2** .....*human epidermal growth factor receptor-2*

**PFS**.....*progression free survival*

**IC** .....intervalle de confiance

**IHC**.....immunohistochimie (détection immunohistochimique)

**Mq+**.....marqueur positif

**Mq-** .....marqueur négatif

**NNT** .....*number needed to treat*

**RR**.....risque relatif

**TFN**.....taux de faux négatifs

**TFP** .....taux de faux positifs

**vs**.....*versus*

## Lexique

- **Biomarqueur** : « caractéristique qui est objectivement mesurée et évaluée comme un indicateur de processus biologiques normaux ou pathologiques, ou de réponses pharmacologiques à une intervention thérapeutique » (définition du *National Institute of Health*). cf. terme « marqueur ».

- **Facteur (marqueur) de risque** : facteur (marqueur) associé à une augmentation de fréquence (couramment appelé risque) du critère de jugement considéré (exemples : décès, survenue d'une complication) chez les sujets qui en sont porteurs par rapport aux autres. Dans les essais cliniques, le « risque de base » ou « risque constitutionnel » désigne le risque du critère de jugement d'intérêt dans le groupe contrôle (sans traitement ou avec le traitement de référence).

- **Facteur (marqueur) prédictif** : facteur (marqueur) influençant l'effet du traitement et dont l'aptitude à modifier le devenir clinique des patients est démontrée. La démonstration de la valeur prédictive d'un marqueur est équivalente à celle de l'utilité clinique du test diagnostique associé. Dans l'idéal, un marqueur prédictif permettrait de distinguer parfaitement les patients bénéficiant du traitement de ceux qui n'en bénéficient pas.

- **Facteur (marqueur) pronostique** : la notion de facteur pronostique est proche de celle de facteur de risque. Un facteur pronostique peut être vu comme un facteur de risque de survenue d'une complication chez des sujets malades, le terme facteur de risque s'utilisant plutôt pour des sujets non malades. Ainsi on parle dans un essai de prévention de facteurs de risque et dans un essai curatif de facteurs pronostiques. Ex : la pression artérielle est un facteur de risque dans un essai de prévention cardiovasculaire et un facteur pronostique de la survie dans un essai à la phase aiguë d'infarctus du myocarde.

- **Hétérogénéité** : situation statistique où un effet (par exemple celui du traitement) varie en fonction d'un tiers facteur (synonyme d'interaction).

- **Interaction** : situation statistique où un tiers facteur modifie les effets d'un autre. En l'occurrence, un marqueur prédictif modifie l'effet du traitement.

- **Marqueur** : dans le contexte de ce document, caractéristique des patients pouvant représenter un indicateur de risque (facteur de risque), de pronostique (facteur pronostique), d'effet du traitement (modificateur de l'effet du traitement) ou d'utilité de traiter (marqueur prédictif). Remarque : les termes biomarqueur et marqueur sont utilisés indifféremment dans le langage courant.

- **Médecine « personnalisée » (*personalized medicine*)** : terme consistant en un abus de langage à proscrire et à remplacer par le terme médecine « stratifiée » compte tenu du fait que la validation du couple marqueur/traitement est fondée sur une approche populationnelle classique de validation de thérapie. La particularité de cette approche est limitée au fait que la population des patients est stratifiée en sous-populations en fonction du marqueur pressenti comme prédictif. cf. terme (médecine) « stratifiée ».

- **Médecine « stratifiée » (*stratified medicine*)** : correspond à une approche thérapeutique où l'objectif est de sélectionner les patients auxquels administrer un traitement en fonction d'un marqueur prédictif, afin de ne traiter que la sous-population susceptible de recevoir un bénéfice du traitement.

- **Modificateur de l'effet du traitement (« *moderator* »)** : facteur influençant la taille de l'effet du traitement. Il s'agit donc d'un facteur associé avec une interaction statistique à l'effet du traitement. Exemple : un marqueur prédictif est un modificateur de l'effet du traitement.

- **Test compagnon** : test diagnostique permettant de sélectionner, en fonction de leur statut pour un marqueur prédictif identifié par ce test, uniquement les patients chez lesquels le traitement est suscep-

tible d'apporter un bénéfice parmi ceux diagnostiqués pour une maladie donnée. Le test est considéré comme « compagnon » d'utilisation du traitement.

- **Théranostic (ou théragnostic)** : néologisme qui dérive de la contraction des termes « thérapeutique » et « diagnostic » ; c'est l'utilisation d'un test diagnostique, identifiant un marqueur, pour orienter la thérapeutique du patient en fonction de son statut pour le marqueur (statut positif ou négatif pour un marqueur binaire).

- **Thérapie ciblée** : traitement qui n'a démontré de bénéfices que chez certains patients identifiés par un marqueur prédictif détecté par un test compagnon.

- **Utilité clinique (*clinical utility*)** : aptitude à améliorer le devenir clinique des patients en événements cliniques mesurables, et à apporter une valeur ajoutée en termes d'optimisation de décision de traitement et en corollaire de stratégie thérapeutique. Valider le caractère prédictif d'un marqueur équivaut à démontrer l'utilité clinique de son test diagnostique.

- **Validité analytique (*analytic validity*)** : aptitude d'un test diagnostique à réaliser la mesure d'intérêt avec exactitude et fiabilité. Cette validité inclut la sensibilité et la spécificité analytiques, la reproductibilité, la robustesse et la satisfaction des contrôles de qualité.

- **Validité clinique (*clinical validity*)** : aptitude d'un test diagnostique à prédire avec précision et fiabilité le phénotype clinique d'intérêt (exemple : survie globale ou survie sans progression des patients recevant un traitement donné). Elle inclut la sensibilité et la spécificité cliniques, ainsi que les valeurs prédictives positive et négative du test.

# 1. Principaux concepts

## 1.1 Avant-propos

L'objectif de ce document est d'exposer les fondements théoriques et méthodologiques de l'évaluation d'un test diagnostique, nommé **test compagnon**, permettant de sélectionner par l'identification d'un **marqueur prédictif** les patients seuls susceptibles de recevoir un bénéfice d'une **thérapie dite ciblée**.

Du fait de sa relation étroite avec le traitement, dont il est nommément dit « compagnon », l'évaluation de ce type de test est singulière car elle présente la particularité de ne pas pouvoir être dissociée de celle de l'efficacité du traitement. Il ne peut s'agir que d'une **évaluation synchrone** à celle du traitement. De ce fait résultent nécessairement des conséquences directes en termes de méthodologie d'étude permettant de répondre aux exigences de cette évaluation.

En termes d'**exigences**, la problématique majeure réside dans la nécessité d'apporter la **preuve que le marqueur** identifié par le test évalué est réellement **prédictif** de l'efficacité (ou la toxicité) du traitement revendiqué comme ciblé. Il faut sur ce point souligner que l'appellation de thérapie ciblée constitue une terminologie impropre en l'absence de niveau de preuve satisfaisant du ciblage concerné, donc de la démonstration du caractère prédictif du marqueur de sélection des patients. En l'absence de preuve de ce caractère prédictif, il peut seulement être affirmé que la thérapeutique est **stratifiée**, c'est-à-dire que son efficacité a été étudiée et montrée dans une sous-population (strate) de patients. En outre, démontrer le caractère prédictif du marqueur équivaut à démontrer l'**utilité clinique** du test l'identifiant, c'est-à-dire l'aptitude du test à améliorer le devenir clinique des patients en aidant à la prise de décision dans le cadre de leur prise en charge thérapeutique. C'est ce niveau de validation qui doit être atteint pour que le test puisse être **qualifié de test compagnon et la thérapie de ciblée**.

La réflexion rapportée dans ce document a été élaborée à partir de la littérature disponible sur le sujet, relativement rare, mais également et surtout à partir des principes méthodologiques fondamentaux sur lesquels se fondent plus largement le développement et l'évaluation des thérapeutiques. Cette démarche a permis de proposer les **exigences d'évaluation pour les tests compagnons** en cohérence avec les exigences actuellement imposées pour les médicaments, permettant ainsi de préciser la nature des connaissances apportées par différents schémas d'études, ainsi que les incertitudes qui peuvent leur être associées.

D'une façon générale, les **principales sources d'erreurs** qui peuvent conduire à conclure à tort à l'utilité clinique d'un test diagnostique sont les mêmes que pour les médicaments. Ce sont essentiellement :

- une approche exploratoire *post-hoc* pouvant amener à considérer comme preuve un artefact trouvé à l'occasion de l'exploration d'une abondance de résultats ;
- une erreur « statistique » due au hasard par l'intermédiaire des fluctuations aléatoires d'échantillonnage ;
- un biais dans la démonstration. Les principes méthodologiques de la randomisation, du double aveugle, de l'analyse en intention de traiter permettent de prévenir l'introduction de biais dans les études ;
- un défaut de pertinence clinique des résultats.

Par conséquent, l'évaluation doit s'assurer que les résultats sur lesquels repose la revendication d'utilité clinique d'un test diagnostique respectent les **principes suivants** :

- démarche hypothético-déductive (1);
- contrôle du risque d'erreur de première espèce alpha ;

- contrôle du risque de biais : stricte application des principes méthodologiques des essais contrôlés randomisés (randomisation, double aveugle, analyse en intention de traiter) ;
- pertinence clinique des résultats : une amélioration tangible et probante du devenir des patients doit être apportée, sur des critères de jugement cliniques (parfois clinico-biologiques) adéquats.

## 1.2 Notion de marqueur prédictif

### 1.2.1 Définition, nature, intérêt

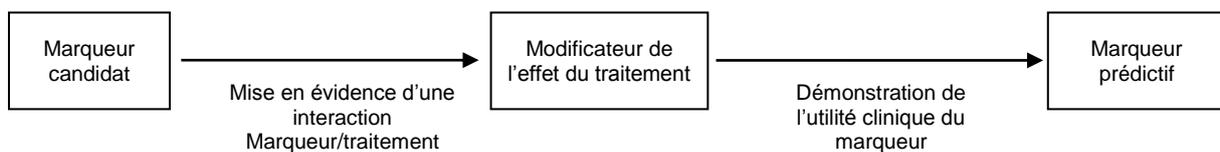
Un marqueur prédictif est une caractéristique permettant de **différencier les patients recevant un bénéfice d'un traitement de ceux qui n'en bénéficient pas**. Il peut être de différentes natures : variant génétique, récepteur particulier, caractéristique métabolique ou biochimique autre, etc. Le terme consacré en français est plutôt celui de marqueur prédictif mais celui de facteur prédictif existe également.

L'idée de l'utilisation des marqueurs prédictifs n'est pas en tant que telle particulièrement récente. En effet, dès 1972, les récepteurs aux œstrogènes avaient été identifiés comme marqueur utile pour sélectionner les femmes potentiellement répondeuses aux traitements hormonaux dans le cancer du sein. Depuis d'autres marqueurs de ce type ont été proposés et du fait des développements récents dans le domaine de la génétique, les marqueurs prédictifs sont très essentiellement des variants génétiques. Le principal domaine de développement de ces marqueurs est l'oncologie, où de nombreux nouveaux marqueurs candidats proposés correspondent à des altérations génétiques au sein de proto-oncogènes à l'origine de modifications d'activités protéiques impliquées dans le processus d'oncogenèse : mutations des gènes EGFR, KIT, KRAS ou BRAF par exemple.

La définition d'un marqueur prédictif réunit **deux conditions** (cf. Figure 1).

**La première** est qu'un marqueur prédictif doit être un « **modificateur** » de l'**effet** d'un traitement, c'est-à-dire qu'il doit prédire en fonction de sa valeur l'effet d'un traitement donné pour un critère de jugement donné (2,3). En termes statistiques, cette condition correspond à l'existence d'une interaction marqueur/traitement, explicitée plus loin dans le document.

**La deuxième** est que la sélection des patients à traiter par l'identification du marqueur prédictif (donc le test diagnostique) doit avoir montré son **utilité clinique** : optimisation de l'efficacité chez les patients portant le marqueur et absence d'intérêt du traitement chez les autres patients, à qui il est donc inutile de l'administrer.



**Figure 1 - Illustration des concepts de modificateur de l'effet du traitement et de marqueur prédictif.**

En réalité, l'optimisation de l'utilisation du traitement recherchée à l'aide d'un marqueur prédictif peut concerner l'optimisation de l'efficacité, de la tolérance ou des deux (cf. Tableau 1).

**Tableau 1 - Principales motivations conduisant à la recherche de marqueurs prédictifs permettant le ciblage de traitements.**

| Type de ciblage                | Motivations, buts recherchés, conséquences  |
|--------------------------------|---|
| Ciblage en termes d'efficacité | <ul style="list-style-type: none"> <li>Augmenter la rentabilité populationnelle du traitement en augmentant le nombre d'événements évités par traitement administré</li> <li>Eviter d'exposer indument aux effets indésirables les patients ne pouvant pas bénéficier des effets bénéfiques du traitement</li> <li>Diminuer le nombre total d'événements indésirables induits (par réduction de la population traitée)</li> </ul>   |
| Ciblage en termes de sécurité  | <ul style="list-style-type: none"> <li>Minimiser le nombre total d'événements indésirables induits (par diminution du risque d'évènements indésirables par sujet et réduction de la population traitée)</li> <li>Eviter de traiter les patients chez lesquels la balance bénéfice/risque est défavorable</li> <li>Ne joue pas forcément sur la rentabilité populationnelle du traitement</li> <li>Prive du traitement certains patients mais chez lesquels la balance bénéfice/risque n'est pas considérée comme favorable<sup>1</sup></li> </ul> |

Les modificateurs de l'effet du traitement peuvent être de nature continue (par exemple un dosage). Dans ce cas, le bénéfice ou le sur-risque d'effet indésirable varie de manière continue avec la valeur du marqueur. Cependant, pour des raisons pratiques, les marqueurs continus sont généralement transformés en marqueurs binaires après fixation d'un seuil de positivité pour leur utilisation comme marqueur prédictif. **Le plus souvent**, les marqueurs prédictifs sont **binaires** et définissent ainsi deux types de « statuts » pour les patients : marqueur positif (noté marqueur (+) ou Mq+ dans la suite du document) et marqueur négatif (noté marqueur (-) ou Mq-). **Marqueur (+)** correspond à la **modalité d'intérêt**, c'est-à-dire celle des patients chez lesquels le traitement est efficace ou à risque de survenue d'une toxicité. Dans le document ci-présent, nous nous essentiellement intéressés aux marqueurs prédictifs d'efficacité d'un traitement. Cependant, les principes y étant décrits sont tout à fait transposables à la validation des marqueurs prédictifs d'effets indésirables, les patients marqueurs (+) étant alors ceux les plus à risque d'effets indésirables avec le traitement.

### 1.2.2 Différence entre facteur de risque et marqueur prédictif

La notion de **facteur de risque** est ancienne et signifie qu'un facteur conditionne le niveau de risque des individus. Facteur de risque est le terme consacré par l'usage mais il est aussi possible de parler de marqueur de risque. Le terme facteur de risque est rencontré dans de nombreux contextes mais il réfère chaque fois substantiellement au même phénomène : **l'augmentation de fréquence du critère d'intérêt**. Cette fréquence du critère d'intérêt est couramment appelée risque (du critère d'intérêt). En fonction de l'objectif de l'étude, ce risque du critère d'intérêt pourra se traduire concrètement par la fréquence de la maladie d'intérêt en épidémiologie étiologique, la fréquence d'une complication (ou du décès) dans les études pronostiques ou la fréquence du critère de jugement dans un essai thérapeutique. Les sujets présentant le facteur de risque sont à plus haut risque que les autres. En cas de

<sup>1</sup> Le fait que certains patients ne seront pas traités avec le nouveau traitement met en exergue l'importance d'une évaluation de la balance bénéfice/risque la plus objective possible.

traitement réduisant le risque, le facteur de risque permet encore de discriminer des sujets à plus haut risque que les autres. En effet, le facteur de risque opère de la même façon chez les patients traités et non traités (cf. Tableau 2). La force d'un facteur de risque se quantifie le plus souvent par un **risque relatif (RR)**, qui est le facteur multiplicatif par lequel la présence du facteur de risque multiplie le risque des individus. Ce risque relatif peut être mesuré par un risk ratio, un odds ratio ou un hazard ratio (cf. Encadré ci-dessous). Le Tableau 2 illustre numériquement cette notion de facteur de risque.

**Encadré - Le risque relatif et ses estimations : risk ratio, odds ratio et hazard ratio**

Le **risque relatif (RR)** est la principale mesure d'association statistique utilisée dans les essais thérapeutiques (et en épidémiologie plus généralement). Le risque relatif est le paramètre d'un modèle multiplicatif liant le risque des patients présentant le facteur d'intérêt (R1) à celui des patients ne le présentant pas (R0). C'est donc, avec le modèle multiplicatif le plus simple, le **facteur multiplicatif par lequel le facteur d'intérêt augmente le risque.**

$$R1 = R0 * RR$$

Dans l'essai thérapeutique, le facteur d'intérêt est le traitement. Le risque relatif est donc le paramètre du modèle multiplicatif simple liant le risque sous traitement (RT) à celui sous traitement contrôle (RC).

$$RT = RC * RR$$

où RT est le risque sous traitement (celui des patients recevant le traitement étudié) et RC est le risque du groupe contrôle (dit aussi risque de base).

Ainsi un risque relatif de 0.5 signifie que le risque sous traitement est la moitié du risque sans ce traitement (sous traitement contrôle). Le traitement divise le risque par 2 (par rapport au traitement contrôle). L'absence d'effet est matérialisé par un risque relatif de 1 (le traitement ne modifie pas le risque).

Dans une étude, ce **risque relatif peut être estimé à l'aide de** différents indices qui sont le **risk ratio, l'odds ratio et le hazard ratio.**

Le **risk ratio** (ratio des risques) est le **rapport du risque sous traitement divisé par celui dans le groupe contrôle.**

$$RR = \frac{R_T}{R_C}$$

Jusqu'à récemment le risk ratio était appelé risque relatif. À présent, on considère que le terme risque relatif désigne le concept qui en pratique peut être estimé à l'aide du risk ratio, de l'odds ratio ou du hazard ratio en fonction de la méthode statistique utilisée (par exemple, le modèle de Cox donne un hazard ratio et la régression logistique un odds ratio). Il est ainsi possible de rapporter un odds ratio en le nommant risque relatif.

|                 | Nb d'évènements | Effectif   | Risque        |
|-----------------|-----------------|------------|---------------|
| Groupe traité   | 234             | 789        | 29.7 %        |
| Groupe contrôle | 348             | 793        | 43.9 %        |
|                 |                 | Risk ratio | 0.68          |
|                 |                 | IC 95 %    | [0.59 ; 0.77] |

**Tableau 2 - Illustration numérique du concept de facteur de risque (chez des sujets traités et non traités).**

|   | Risque de l'évènement critère de jugement                       |                         |  |  |
|---|---|-------------------------|--|--|
|   | Sujets traités (T+)   | Sujets non traités (T-) |  |  |
| Sujets présentant le facteur de risque (FR+)        | 16 %  | 20 %                    | Effet du facteur de risque<br>chez les T- : 0.20/0.10=2.00 | Effet du facteur de risque<br>chez les T+ : 0.16/0.8=2.0 |
|   | Effet du traitement chez les FR+<br>risk ratio : 0.16/0.20=0.80 |                         |  |  |
| Sujets ne présentant pas le facteur de risque (FR-) | 8 %   | 10 %                    |  |  |
|   | Effet du traitement chez les FR-<br>risk ratio : 0.08/0.10=0.80 |                         |  |  |

Un facteur ou **marqueur prédictif** est un facteur influençant l'effet du traitement et dont l'intérêt est de permettre d'identifier les patients susceptibles de recevoir ou non un bénéfice du traitement (cf. Section 1.2.1). Le concept de marqueur prédictif est illustré numériquement dans le Tableau 3 sous la forme d'un « vrai » marqueur prédictif, c'est-à-dire totalement prédictif de l'effet ou non du traitement. On remarque dans ce cas que le marqueur prédictif n'est pas un facteur de risque pour les patients non traités (RR=1), mais qu'il le devient chez ceux recevant le traitement (RR=0.5).

**Tableau 3 - Illustration numérique du concept de marqueur prédictif.**

|  | Risque de l'évènement critère de jugement                         |                         |   |   |
|--|---|-------------------------|---|---|
|  | Sujets traités (T+)   | Sujets non traités (T-) |   |   |
| Sujets présentant le marqueur prédictif (Mq+)        | 10 %  | 20 %                    | Effet du marqueur prédictif<br>chez les T- : 0.20/0.20=1.00 | Effet du marqueur prédictif<br>chez les T+ : 0.10/0.20=0.50 |
|  | Effet du traitement chez les Mq+ :<br>risk ratio = 0.10/0.20=0.50 |                         |   |   |
| Sujets ne présentant pas le marqueur prédictif (Mq-) | 20 %  | 20 %                    |   |   |
|  | Effet du traitement chez les Mq- :<br>risk ratio = 0.20/0.20=1.00 |                         |   |   |

Il faut souligner qu'il existe aussi tout un *continuum* de cas où le marqueur, souvent alors considéré à tort comme prédictif, est effectivement un modificateur de l'effet du traitement (« *moderator* »), mais ne répond pas au second critère définissant un marqueur prédictif (cf. Section 1.2.1). Ainsi, il existe des situations où le traitement est efficace dans les deux strates, celles des patients Mq+ et Mq- (cf. Tableau 4), ou bien efficace dans une strate et délétère dans l'autre (cf. Tableau 5).

**Tableau 4 - Illustration d'une situation où le traitement est efficace dans les deux strates définies par le marqueur modificateur de l'effet du traitement.**

|     | Risque de l'évènement critère de jugement |                         | Effet du traitement (risk ratio) |
|-----|---|-------------------------|----------------------------------|
|     | Sujets traités (T+)                       | Sujets non traités (T-) |                                  |
| Mq+ | 10 %                                      | 20 %                    | 0.50                             |
| Mq- | 16 %                                      | 20 %                    | 0.80                             |

**Tableau 5 - Illustration d'une situation où le traitement est efficace dans la strate des patients présentant le marqueur modificateur de l'effet du traitement et délétère dans l'autre.**

|     | Risque de l'évènement critère de jugement |                         | Effet du traitement (risk ratio) |
|-----|---|-------------------------|----------------------------------|
|     | Sujets traités (T+)                       | Sujets non traités (T-) |                                  |
| Mq+ | 10 %                                      | 20 %                    | 0.50                             |
| Mq- | 25 %                                      | 20 %                    | 1.25                             |

La notion de marqueur prédictif (incluant celle de modificateur de l'effet du traitement) n'implique pas par ailleurs que le risque de base (sans traitement) soit le même en fonction de la présence ou non du marqueur. Rien n'exclut enfin qu'un marqueur prédictif soit aussi un facteur de risque (cf. Tableau 6).

**Tableau 6 - Illustration d'une situation où le marqueur modificateur de l'effet du traitement est aussi facteur de risque.**

|                                 | Risque de l'évènement critère de jugement |                         | Effet du traitement (risk ratio) |
|---------------------------------|---|-------------------------|----------------------------------|
|                                 | Sujets traités (T+)                       | Sujets non traités (T-) |                                  |
| Mq+                             | 10 %                                      | 20 %                    | 0.50                             |
| Mq-                             | 7.5 %                                     | 15 %                    | 0.50                             |
| Effet sur le risque du marqueur | 1.33                                      | 1.33                    |                                  |

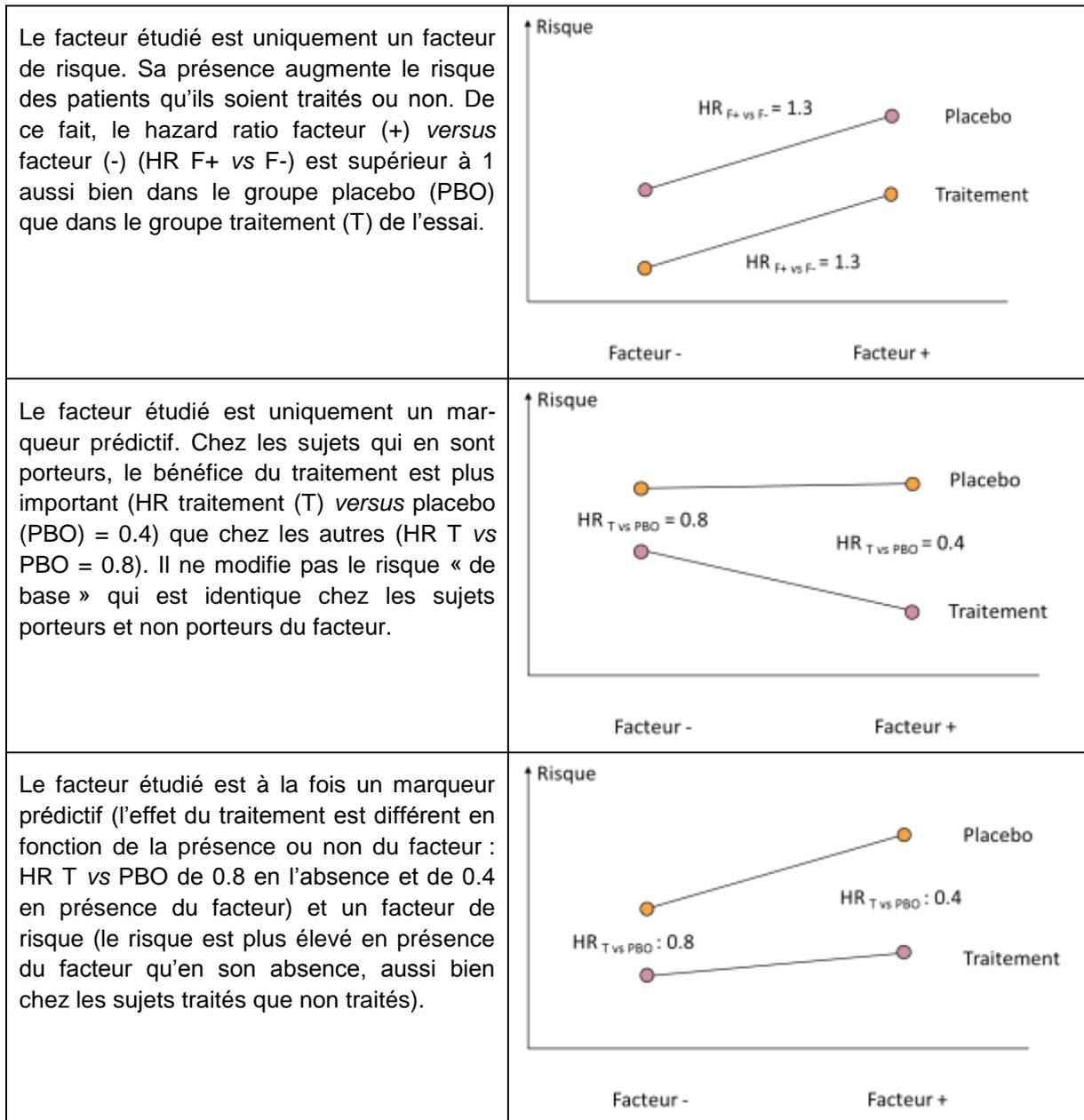


Figure 2 - Illustration graphique des concepts de marqueur prédictif et facteur de risque. Les risques (en ordonnée) sont associés à la présence ou non du facteur étudié (en abscisse) dans les deux groupes d'un essai thérapeutique (contrôles en mauve et traités en orange).

### 1.3 Notion d'interaction statistique

Le terme interaction désigne en statistiques une situation où un effet est influencé quantitativement par un tiers facteur. Il s'agit ici de l'**effet du traitement** qui est **influencé quantitativement par la présence ou non du marqueur prédictif**. Les effets traitements sont quantifiés le plus souvent par un risque relatif calculé comme un rapport des risques représentant le facteur multiplicatif par lequel le traitement augmente le risque des patients le recevant en comparaison de celui des patients du groupe contrôle. Il y a donc interaction lorsque la valeur de ce risque relatif est influencée par le tiers facteur.

Dans le tableau ci-dessous, il existe une interaction entre le facteur 1 et l'effet du traitement, tandis qu'il n'y a pas d'interaction avec le facteur 2. Ce tableau rapporte les vrais effets du traitement, c'est-à-dire l'effet réel du traitement que l'on cherche à estimer dans une étude.

Un essai clinique permet d'obtenir les risques relatifs observés, c'est-à-dire une estimation des vrais risques relatifs. Comme ces estimations peuvent fluctuer autour de la vraie valeur de manière aléatoire, elles sont toujours associées à des intervalles de confiance qui reflètent l'importance de ces fluctuations. En première approximation, l'intervalle de confiance représente l'ensemble des valeurs compatibles avec la vraie valeur.

**Tableau 7 - Exemple de deux facteurs, l'un entraînant une interaction et l'autre non. Les interactions sont visualisées à l'aide des vraies valeurs d'effet du traitement.**

|           |   | Effet du traitement<br>(vrai risque relatif) |
|-----------|---|--|
| Facteur 1 | + | 0.67   |
|           | - | 0.88   |
| Facteur 2 | + | 0.73   |
|           | - | 0.73   |

En pratique, une interaction est recherchée à l'aide des estimations données par un essai thérapeutique. Le design d'étude alors le plus adapté, bien que trop rarement rencontré en pratique, est présenté dans la section 3.1. L'incertitude statistique des estimations de l'effet du traitement obtenues pour chaque modalité du facteur étudié (par exemple un marqueur) doit alors être prise en compte. En effet, ces valeurs d'effet du traitement observées peuvent être différentes sans que cela ne soit une réalité mais simplement un artéfact dues aux erreurs aléatoires. Ainsi, en pratique, afin de prouver que les deux estimations d'effet du traitement obtenues en fonction du facteur étudié (dans le cas où il est binaire) sont réellement différentes l'une de l'autre, il convient de **tester la réalité statistique de l'interaction à l'aide d'un test statistique ad-hoc**. C'est notamment ce qui est couramment fait dans les analyses en sous-groupes. Il ne sera en effet possible de parler d'interaction que si celle-ci est **statistiquement significative**, c'est-à-dire que la différence entre les deux valeurs d'effet du traitement obtenues est statistiquement significative. Dans ce test, l'hypothèse nulle est que le traitement a le même effet quelle que soit la valeur du tiers facteur. Le « p » calculé correspond à la probabilité de rejeter à tort l'hypothèse nulle. On cherche par ce test à savoir si les différences observées entre les valeurs sont raisonnablement imputables ou non au hasard (cf. Figure 3).

Les interactions peuvent être recherchées pour des facteurs binaires, qualitatifs, ou continus. Le **plus souvent** il s'agit de **facteurs binaires**<sup>2</sup> car ce cas est le plus simple et le plus pratique. Quand le marqueur est continu, la recherche de l'interaction utilise des techniques de régression (régression logistique ou modèle de Cox par exemple). Elle peut aussi s'effectuer en transformant la variable continue en variable ordinale.

<sup>2</sup> A ce jour, dans le domaine des tests revendiqués comme « compagnons », les marqueurs identifiés sont très essentiellement de type binaire. Nous nous sommes donc limités à ce type de marqueur dans le cadre des guides méthodologiques portant sur l'évaluation des tests concernés (guide ci-présent et guide intitulé « Test compagnon associé à une thérapie ciblée : définitions et méthode d'évaluation »).

| <p>Compte tenu de la précision de l'estimation de ces deux risk ratios, la différence observée n'est pas suffisante pour obtenir un test d'interaction significatif et donc conclure à l'existence d'une interaction.</p>                    | <table border="1"> <thead> <tr> <th>Strate</th> <th>Risk Ratio</th> <th>RR</th> <th>95%-CI</th> </tr> </thead> <tbody> <tr> <td>marqueur +</td> <td></td> <td>0.92</td> <td>[0.86; 0.99]</td> </tr> <tr> <td>marqueur -</td> <td></td> <td>0.96</td> <td>[0.89; 1.02]</td> </tr> </tbody> </table> <p>Interaction p=0.45</p> | Strate | Risk Ratio   | RR | 95%-CI | marqueur + |  | 0.92 | [0.86; 0.99] | marqueur - |  | 0.96 | [0.89; 1.02] |
|--|--|--------|--------------|----|--------|------------|--|------|--------------|------------|--|------|--------------|
| Strate   | Risk Ratio   | RR     | 95%-CI       |    |        |            |  |      |              |            |  |      |              |
| marqueur +   |  | 0.92   | [0.86; 0.99] |    |        |            |  |      |              |            |  |      |              |
| marqueur -   |  | 0.96   | [0.89; 1.02] |    |        |            |  |      |              |            |  |      |              |
| <p>Le test d'interaction est significatif. Il est raisonnable de considérer qu'il y a interaction et que l'effet du traitement (estimé par le risk ratio) n'est pas le même pour les patients marqueur (+) et les patients marqueur (-).</p> | <table border="1"> <thead> <tr> <th>Strate</th> <th>Risk Ratio</th> <th>RR</th> <th>95%-CI</th> </tr> </thead> <tbody> <tr> <td>marqueur +</td> <td></td> <td>0.78</td> <td>[0.72; 0.84]</td> </tr> <tr> <td>marqueur -</td> <td></td> <td>0.89</td> <td>[0.83; 0.96]</td> </tr> </tbody> </table> <p>Interaction p=0.01</p> | Strate | Risk Ratio   | RR | 95%-CI | marqueur + |  | 0.78 | [0.72; 0.84] | marqueur - |  | 0.89 | [0.83; 0.96] |
| Strate   | Risk Ratio   | RR     | 95%-CI       |    |        |            |  |      |              |            |  |      |              |
| marqueur +   |  | 0.78   | [0.72; 0.84] |    |        |            |  |      |              |            |  |      |              |
| marqueur -   |  | 0.89   | [0.83; 0.96] |    |        |            |  |      |              |            |  |      |              |

**Figure 3 - Prise en compte de l'incertitude des estimations - Test d'interaction.**

À partir de cette notion d'interaction quantitative a été dérivée une notion d' « **interaction qualitative** ». Il y a interaction qualitative lorsqu'une interaction quantitative conduit à un changement de direction dans l'effet du traitement en fonction du facteur étudié. Cette situation est observée lorsque le traitement est bénéfique pour une modalité du facteur et délétère pour l'autre (cf. Tableau 5). La force de cette interaction qualitative est gradée par la signification statistique des effets du traitement. L'interaction qualitative la plus forte est produite par une situation où le traitement est statistiquement significativement bénéfique pour une modalité du facteur et statistiquement significativement délétère pour l'autre.

## 1.4 Notion de test compagnon

Un test compagnon est un test diagnostique permettant de **sélectionner**, en fonction de leur statut pour un marqueur prédictif identifié par ce test, uniquement les patients chez lesquels le **traitement** est **susceptible d'apporter un bénéfice** parmi ceux diagnostiqués pour une maladie donnée. Pour cette raison, le test est dit « compagnon » d'utilisation du traitement.

Comme évoqué en avant-propos dans ce document, l'évaluation de cette approche ne peut reposer logiquement que sur une **évaluation conjointe du test et du traitement** avec un double objectif : montrer que le traitement n'est efficace que chez les patients marqueur (+) et que le test compagnon permet d'améliorer la prise de décision thérapeutique et ainsi l'évolution clinique des patients.

Il est important de souligner que l'intérêt théorique de la sélection des patients présumés meilleurs voire seuls répondeurs au traitement par un test compagnon est aussi d'**optimiser**

la **rentabilité clinique**<sup>3</sup> de ce traitement par rapport à une stratégie d'utilisation du traitement chez tous les patients. En effet, la sélection de patients sur la base de l'identification d'un marqueur peu ou non prédictif par un test diagnostique, qui ne peut en l'occurrence pas être nommé test compagnon, n'optimise pas ou peu cette rentabilité clinique, et peut occasionner une perte de chance pour certains patients si le traitement est en réalité plus ou moins bénéfique chez tous. La suite de ce document montrera que ce problème éthique constitue une préoccupation majeure dans le cadre de l'évaluation de l'utilité de ces tests.

## 1.5 Notions de thérapeutiques « ciblée » et « stratifiée »

L'approche thérapeutique liée à l'utilisation de marqueurs prédictifs est parfois qualifiée de **médecine personnalisée** (« *personalized medicine* ») ou individualisée (« *individualized medicine* »). Ces appellations doivent être cependant considérées comme des abus de langage car l'approche impliquant les marqueurs prédictifs reste populationnelle et statistique, et non individuelle, et correspond en réalité à une approche de type **médecine stratifiée** (« *stratified medicine* ») ou médecine de précision (« *precision medicine* ») (4). Elle affine simplement les estimations en s'intéressant à des sous-populations de patients (d'où le terme de stratification) et produit ainsi une estimation moyenne de l'effet traitement dans ces sous-populations (strates), et non une estimation de l'effet du traitement chez un individu.

On appelle **thérapie ciblée** un traitement n'apportant de bénéfices que chez certains patients identifiés par un **marqueur prédictif** déterminé par un **test compagnon**.

En pratique, il apparaît que la revendication du caractère ciblé d'un traitement peut être observée dans deux situations :

- lorsqu'un ciblage a été recherché par le développeur au moment même de la conception (synthèse) du traitement : on parle alors de ciblage « par construction » ;
- suite à une recherche exploratoire par analyse rétrospective de résultats d'un (ou plusieurs) essai(s) réalisé(s) sur une population de patients non sélectionnés par un marqueur, suivie ou non par un essai prospectif de confirmation (cf. section 3.2.3).

La démarche de **ciblage « par construction »** a été historiquement la première utilisée. Elle l'est toujours aujourd'hui. Le plus souvent, l'efficacité des molécules revendiquées comme relevant de ce concept de ciblage par construction n'est alors étudiée que chez les patients porteurs du marqueur désigné comme cible, l'absence d'efficacité de la molécule chez les patients marqueurs (-) étant considérée comme implicite. Cependant, la valeur prédictive du marqueur ne peut raisonnablement pas en toute rigueur reposer sur des fondements implicites, mais uniquement sur des preuves cliniques. Les données précliniques ne constituent que des données préliminaires, de statut exploratoire, et ne reflétant que plus ou moins bien celles qui seront obtenues chez l'homme. En effet, du fait de la grande complexité et des spécificités de l'organisme humain, aucun type d'études précliniques (lignées cellulaires immortalisées, études animales...) ne peut prévoir précisément et avec fiabilité les effets cliniques d'un médicament. En outre, revendiquer l'absence d'efficacité du traitement chez les marqueurs (-) comme implicite impliquerait de pouvoir apporter la preuve d'une spécificité indiscutable du traitement envers sa cible *in vivo*. Or, il paraît fort difficile de pouvoir apporter une telle démonstration.

En l'occurrence, à ce jour un certain recul sur les premières thérapies qualifiées de ciblées a fait apparaître **plusieurs exemples** où, bien que le traitement ait été initialement considéré comme **a priori spécifique de sa cible** et le marqueur comme central dans le mécanisme d'action du traitement, ce dernier s'est en fait parfois révélé actif chez des patients ne portant pas ce marqueur pour diverses raisons possibles. Cela est notamment le cas avec le trastuzumab dans le cancer du sein, pour lequel une efficacité chez les cancers du sein HER2/neu

---

<sup>3</sup> La rentabilité clinique correspond au nombre d'événements évités pour le traitement d'un certain nombre de patients.

(*human epidermal growth factor receptor-2*) négatifs est envisagée (5–7). Un deuxième exemple est le cetuximab. Cet autre anticorps monoclonal a été synthétisé pour « cibler » le récepteur EGFR (*epidermal growth factor receptor*), récepteur oncogénique détecté dans 70-85 % des cancers colorectaux par technique immunohistochimique (IHC). Or, à ce jour, l'analyse des données de certains essais a montré que jusqu'à 25 % patients atteints d'un cancer colorectal métastatique identifié « EGFR négatif » par IHC pourraient répondre au cetuximab (8). Les explications avancées seraient la mauvaise spécificité de la technique IHC et/ou un phénomène d'ADCC (*Antibody-dependent cell-mediated cytotoxicity*) lié au cetuximab envers les cellules tumorales. Par ailleurs, au sein de la deuxième grande classe de thérapies dites ciblées en oncologie, représentée par les inhibiteurs de récepteurs à activité tyrosine kinase (RTK), il est assez fréquent de ne pas observer, dès les résultats précliniques, une réelle spécificité pour une cible unique. Par exemple, le crizotinib, utilisé en tant qu'inhibiteur du RTK nommé ALK (*anaplastic lymphoma kinase*) et de ses variants oncogéniques dans le cancer du poumon non à petites cellules, inhibe aussi le RTK du facteur de croissance des hépatocytes (c-Met) également impliqué dans les processus oncogéniques. Dans son rapport d'évaluation associé à l'enregistrement de la molécule, l'EMA (*European Medicine Agency* ; EMA/CHMP/497137/2012) évoque clairement les limites de spécificité du crizotinib : « *When tested to its selectivity for different kinases, crizotinib was relatively specific to c-Met/HGFR and ALK fusion proteins* ». Ces exemples montrent bien qu'un « ciblage par construction » n'est pas suffisant pour justifier de l'absence d'étude clinique chez les patients marqueur (-). L'absence d'efficacité du traitement chez ces patients doit être cliniquement prouvée pour **attester de la réalité *in vivo* du ciblage**, le risque d'accepter un ciblage à tort d'un nouveau traitement ayant pour conséquence directe une perte de chance pour ces patients.

L'approche de revendication du caractère ciblé d'un traitement faisant suite à la **recherche exploratoire *post-hoc*** de marqueurs prédictifs est détaillée dans la section 3.2. Il est intéressant de noter que, d'une certaine façon, cette voie de recherche **a élargi le concept de « caractère ciblé »**. En effet, ce concept référait initialement, en lien avec le « ciblage par construction », à la synthèse de traitements orientés vers des cibles moléculaires impliquées dans leur mécanisme d'action, et faisant intervenir une liaison entre le médicament et sa cible. Avec les recherches *post-hoc* de marqueurs prédictifs, le concept de caractère ciblé s'est élargi à la notion plus générale de marqueur prédictif. Cette extension du concept ne pose pas de réel souci de cohérence car les cibles moléculaires sont, en théorie, censées être également des marqueurs prédictifs (même si ce n'est pas toujours le cas ; cf. section 3.2.3 - « Recherche d'un marqueur prédictif après un essai non concluant »). Cependant, il faut tout de même relever qu'avec l'élargissement du concept aux marqueurs prédictifs en général, la compréhension des **mécanismes d'implication** d'un marqueur dans l'efficacité d'un traitement donné, ainsi que du mécanisme d'action global du traitement, est devenue plus complexe.

## 1.6 Les différentes situations associées en pratique à la restriction d'utilisation d'un traitement à une sous-population de patients

Le « ciblage » n'est pas vraiment une démarche nouvelle mais un fil conducteur qui évolue dans la recherche thérapeutique. La restriction d'administration d'un traitement à une sous-population de patients n'est pas limitée en pratique au domaine des thérapies ciblées. Initialement, elle s'est essentiellement effectuée sur la base de critères cliniques et/ou marqueurs biologiques et de mécanismes évidents, sans toujours faire appel à l'utilisation d'un marqueur prédictif (9). Ce qui a fortement changé apparaît être une intensification de la recherche dans ce domaine de thérapeutiques, ceci à un niveau de mécanisme bien plus fin, pouvant faire craindre une explosion de ciblage à tort.

Cinq situations peuvent être principalement distinguées. Ce document ne concerne que les cas 3, 4 et 5.

1. **Restriction d'administration d'un traitement à un sous-groupe à risque** afin d'optimiser la rentabilité (clinique)<sup>4</sup> de ce traitement en termes de nombre d'événements évités par patient traité (mesuré par « l'*absolute risk reduction* » (ARR) ou le *number needed to treat* (NNT)). Ce cas est par exemple celui de la restriction de l'utilisation des statines en prévention primaire aux sujets ayant au moins un autre facteur de risque supplémentaire par rapport à leur élévation de LDL cholestérol. Cette restriction à un sous-groupe particulier peut être décidée *a posteriori* par un raisonnement cherchant à maximiser la rentabilité du traitement en termes d'ARR ou de NNT. Elle est « mathématique » et n'a donc pas besoin de démonstration particulière par les faits. Cet aspect n'est pas abordé dans ce document. Cette approche est par exemple utilisée par Dorresteijn et coll. dans le cadre des statines et des pathologies coronariennes stables (10).

2. **Sélection d'un groupe de patients posant un problème thérapeutique particulier pour l'administration d'un traitement (affinage de l'entité nosologique considérée)**. Par exemple, dans la prévention de la récurrence des accidents vasculaires cérébraux (AVC), il est apparu très tôt que l'existence d'une sténose carotidienne pouvait être une cible thérapeutique prometteuse conduisant à stratifier le post-AVC en différentes entités en fonction de l'existence d'une sténose serrée et de la possibilité de faire une chirurgie d'endartériectomie. De plus, cela a permis d'affiner l'utilisation des antiagrégants plaquettaires en faisant la distinction entre sténose carotidienne et fibrillation atriale. Cette approche a été utilisée sans formalisation dans de nombreux domaines. Elle est même plus naturelle pour un clinicien qu'une approche plus globale mélangeant, dans un même essai, différents types de patients.

3. **Ciblage d'un maillon du mécanisme par lequel agit le traitement, maillon indispensable à ce mécanisme et non ubiquitaire**, par exemple sur une cible moléculaire n'existant pas chez tous les patients en fonction de ses variantes génétiques. Ce ciblage correspond au ciblage dit « par construction », la cible étant la clé de voûte de l'idée thérapeutique. L'idée sous-jacente est qu'il n'y a pas lieu de traiter des patients non porteurs de cette cible car aucun bénéfice n'est espéré chez eux du fait du mécanisme d'action. Ce point très discuté est abordé dans la section 1.5.

4. **Ciblage en fonction d'un marqueur présumé prédictif découvert de façon empirique** (non envisagé *a priori* mais recherché *a posteriori* en particulier devant un résultat non concluant sur la population toute entière). Du fait du caractère *post-hoc* et non hypothético-déductif de l'identification de ce marqueur, un essai de confirmation est indispensable.

5. Proposition de **ciblage *a posteriori* sur un marqueur présumé prédictif pour un traitement ayant déjà fait la preuve de son intérêt** dans un objectif d'optimisation de l'utilisation de ce traitement. Pour exemple, c'est le cas de la mesure de l'activité plaquettaire résiduelle avec le dispositif VerifyNow® pour l'optimisation du traitement par clopidogrel (dans le cadre interventionnel) (11,12).

## 1.7 Les intérêts potentiels du ciblage d'une thérapeutique

L'intérêt du ciblage d'une thérapeutique à l'aide d'un marqueur prédictif est notamment d'optimiser la rentabilité clinique du traitement par rapport à une stratégie d'utilisation du traitement chez tous les patients. Sans ce résultat en termes d'optimisation de la rentabilité,

---

<sup>4</sup> La rentabilité clinique correspond au nombre d'événements évités pour le traitement d'un certain nombre de patients et ne fait intervenir aucune dimension économique. Il s'agit du bénéfice absolu engendré par le traitement d'une population avec une intervention donnée. Elle peut être vue comme une mesure d'efficacité purement clinique.

un pseudo-ciblage peut entraîner des dépenses supplémentaires, coûter du temps et de l'énergie au système de santé, sans retour pour la santé publique, et occasionner une perte de chance pour certains patients si le traitement est en réalité bénéfique chez tous les patients.

Quatre cas de figure de ciblage sont détaillés dans cette section avec leurs conséquences populationnelles :

- un ciblage à raison : le marqueur utilisé pour sélectionner les patients à traiter est un véritable marqueur prédictif ayant une réelle utilité clinique ;
- un ciblage à tort utilisant un marqueur qui n'est même pas un modificateur de l'effet du traitement ;
- un pseudo-ciblage utilisant un marqueur qui est un modificateur de l'effet du traitement mais qui ne différencie pas en totalité les patients recevant un bénéfice ou non du traitement ;
- un ciblage à l'aide d'une interaction sur la sécurité.

Une synthèse des résultats est présentée en section 1.7.5.

### 1.7.1 Ciblage à raison

Le Tableau 8 illustre les conséquences populationnelles du ciblage approprié d'une thérapeutique sur un marqueur réellement prédictif (le traitement est efficace chez les patients marqueur (+) et sans effet chez les patients marqueur (-)).

Cette simulation a été réalisée avec les paramètres suivants. Ce marqueur est présent chez 60 % des sujets de la population (ligne 1 du tableau). Le risque sans traitement de l'événement critère de jugement est de 10 % (ligne 2) chez les patients marqueur (+) et marqueur (-). Ce marqueur n'est donc pas un facteur de risque. Le traitement n'est efficace que chez les patients marqueur (+) où il entraîne une réduction du risque de 20 % (RR=0.80, ligne 3). Il est sans effet chez les patients marqueur (-) (ligne 4).

Sans traitement, une population de 1 000 patients (ligne 5) va présenter 100 événements (ligne 8) tandis qu'avec le traitement de tous les patients ou seulement des patients marqueur (+) (stratégie ciblée) seulement 88 événements surviendront.

L'**intérêt de la stratégie ciblée** apparaît donc en termes de **nombre d'évènements évités par patient traité** (ligne 17). Par rapport à l'absence de traitement, la stratégie ciblée évite 12 évènements (ligne 15) obtenus à partir du traitement de seulement 600 sujets (60 % de la population) contre 12 évènements évités avec le traitement de la totalité des 1 000 patients. Ainsi le nombre d'évènement évités pour un patient traité est de 0.02 (20 pour 1 000) avec la stratégie ciblée contre 0.012 (12 pour 1 000) avec la stratégie indifférenciée. La stratégie ciblée évite 67 % de plus d'évènements pour le même nombre de sujets traités que la stratégie indifférenciée (ligne 18). De plus, étant donné que le traitement n'a pas d'intérêt chez les sujets marqueur (-), la stratégie ciblée évite la totalité des évènements évitables (ligne 16) avec le traitement de seulement 60 % de la population.

**Tableau 8 - Simulation des conséquences populationnelles de l'utilisation d'un ciblage à l'aide d'un marqueur réellement prédictif.**

 ARR : *absolute risk reduction* ; Nb : nombre ; SO : sans objet

 (\*)  $88 = (600 \times 0.1 \times 0.8) + (400 \times 1 \times 0.1) = (48 \text{ évènements chez les Mq+}) + (40 \text{ évènements chez les Mq-})$ 

| Paramètres                             |   |                  |                                 |                       |
|--|---|------------------|---------------------------------|-----------------------|
| 1                                      | Proportion de Mq+   | 60 %             |                                 |                       |
| 2                                      | Risque de base (Mq+ et Mq-)   | 10 %             |                                 |                       |
| 3                                      | Effet traitement chez les Mq+ (RR)  | 0.8              |                                 |                       |
| 4                                      | Effet traitement chez les Mq- (RR)  | 1                |                                 |                       |
| 5                                      | Taille de la population   | 1 000            |                                 |                       |
| Comparaison des différentes stratégies |   |                  |                                 |                       |
|  |   | Stratégie ciblée | Traitement de tous les patients | Absence de traitement |
| 8                                      | Nb total d'événements sur la population   | 88               | 88                              | 100                   |
| 9                                      | Nb de sujets traités (avec le nouveau traitement)   | 600              | 1 000                           | 0                     |
| 10                                     | Proportion de sujets traités  | 60 %             | 100 %                           | 0 %                   |
| 11                                     | Nb d'événements chez les sujets traités   | 48               | 88                              | 0                     |
| 12                                     | Nb d'événements chez les sujets non traités   | 40               | 0                               | 100                   |
| Risque relatif (RR)                    |   |                  |                                 |                       |
| 13                                     | RR par rapport à l'absence de traitement  | 0.88             | 0.88                            | 1                     |
| 14                                     | RR par rapport au traitement de tous les patients   | 1                | 1                               | SO                    |
| Réduction du risque absolu (ARR)       |   |                  |                                 |                       |
| 15                                     | Nb d'événements évités (par rapport à l'absence de traitement)  | -12              | -12                             | 0                     |
| 16                                     | Proportion des événements évités par rapport à tous les événements évitables (par le traitement de tous les patients) | 100 %            | 100 %                           | SO                    |
| 17                                     | Nb d'événements évités par patient traité   | <b>-0.02</b>     | <b>-0.012</b>                   | SO                    |
| 18                                     | Ratio de rentabilité (d'un traitement) en nb d'événements évités  | 1.67             |                                 |                       |

L'intérêt du ciblage peut aussi être d'**optimiser la balance globale bénéfique/risque** en évitant de soumettre au risque d'effets indésirables les patients ne retirant pas de bénéfice du traitement (les patients marqueur (-) dans ce cas). Par exemple, si le risque d'effet indésirable grave (EIG) est de 1.2 % chez tous les patients (Mq+ et Mq-), le traitement de la population totale entraînera en moyenne la survenue de 12 EIG (pour 1 000 patients traités). La balance bénéfique/risque sera alors de 12 événements évités pour 12 EIG induits, soit une balance bénéfique/risque sans intérêt. Par contre, le traitement ciblé n'engendrera en moyenne que 7.2 EIG (1.2 % \* 600 patients traités) correspondant à une balance bénéfique/risque de 12 événements évités pour 7 EIG induit soit 0.6 EIG induit par événement évité. Afin d'appréhender au mieux cette problématique, il serait souhaitable que les essais utilisent comme critère de jugement principal un critère de bénéfice clinique net.

### 1.7.2 Ciblage à tort

Le Tableau 9 illustre l'utilisation d'une stratégie ciblée à tort sur un marqueur non prédictif. Le traitement a ici la même efficacité chez les patients marqueur (+) et (-) (lignes 3 et 4).

Dans cette situation, le principal inconvénient de la stratégie ciblée est **de ne pas prévenir la totalité des évènements évitables** (par le traitement de tous les patients) mais seulement 60 % d'entre eux (ligne 16).

De plus, en raison de la recherche (inutile) du marqueur chez tous les sujets de la population, le coût de l'événement évité deviendra supérieur à celui obtenu avec la stratégie non ciblée.

**Tableau 9 - Simulation des conséquences populationnelles d'un ciblage à tort à l'aide d'un marqueur non prédictif.**

| Paramètres                              |   |                  |                                 |                       |
|---|---|------------------|---------------------------------|-----------------------|
| 1                                       | Proportion de Mq+   | 60 %             |                                 |                       |
| 2                                       | Risque de base (Mq+ et Mq-)   | 10 %             |                                 |                       |
| 3                                       | Effet traitement chez les Mq+ (RR)  | 0.8              |                                 |                       |
| 4                                       | Effet traitement chez les Mq- (RR)  | 0.8              |                                 |                       |
| 5                                       | Taille de la population   | 1 000            |                                 |                       |
| Comparaison des différentes stratégies  |   |                  |                                 |                       |
|   |   | Stratégie ciblée | Traitement de tous les patients | Absence de traitement |
| 8                                       | Nb total d'événements sur la population   | 88               | 80                              | 100                   |
| 9                                       | Nb de sujets traités (avec le nouveau traitement)   | 600              | 1 000                           | 0                     |
| 10                                      | Proportion de sujets traités  | 60 %             | 100 %                           | 0 %                   |
| 11                                      | Nb d'événements chez les sujets traités   | 48               | 80                              | 0                     |
| 12                                      | Nb d'événements chez les sujets non traités   | 40               | 0                               | 100                   |
| <b>Risque relatif (RR)</b>              |   |                  |                                 |                       |
| 13                                      | RR par rapport à l'absence de traitement  | 0.88             | 0.8                             | 1                     |
| 14                                      | RR par rapport au traitement de tous les patients   | 1.1              | 1                               | SO                    |
| <b>Réduction du risque absolu (ARR)</b> |   |                  |                                 |                       |
| 15                                      | Nb d'événements évités (par rapport à l'absence de traitement)  | -12              | -20                             | 0                     |
| 16                                      | Proportion des événements évités par rapport à tous les événements évitables (par le traitement de tous les patients) | 60 %             | 100 %                           | SO                    |
| 17                                      | Nb d'événements évités par patient traité   | <b>-0.02</b>     | <b>-0.02</b>                    | SO                    |
| 18                                      | Ratio de rentabilité (d'un traitement) en nb d'événements évités  | 1.00             |                                 |                       |

### 1.7.3 Ciblage sur un marqueur partiellement prédictif (pseudo-ciblage)

Cette simulation (cf. Tableau 10) illustre les conséquences populationnelles d'un ciblage effectué avec un modificateur de l'effet du traitement mais qui n'est pas un véritable marqueur prédictif, c'est-à-dire qui ne sépare pas en totalité les patients chez lesquels le traitement est efficace de ceux où il ne l'est pas. Le traitement reste efficace, mais moins, chez les patients marqueur (-).

Dans cette simulation, le traitement entraîne une réduction relative du critère de jugement de 20 % chez les Mq+ et de 10 % chez les Mq- (lignes 3 et 4). Un traitement ciblé sur les Mq+ n'évite que 75 % de tous les événements évitables en traitant tout le monde (ligne 16), mais la rentabilité du traitement (en termes d'événements évités par patients traités) augmente et passe de -0.016 (16 événement évités pour 1 000 traitements) à -0.02 (20 événements évités pour 1 000 traitements). Par contre, le nombre total d'événements observés sur la popu-

lation totale sera plus important (88) que si tous les patients avaient été traités (84 événements).

**Tableau 10 - Simulation des conséquences populationnelles d'un pseudo-ciblage utilisant un marqueur non entièrement prédictif.**

| Paramètres                              |   |                  |                                 |                       |
|---|---|------------------|---------------------------------|-----------------------|
| 1                                       | Proportion de Mq+   | 60 %             |                                 |                       |
| 2                                       | Risque de base (Mq+ et Mq-)   | 10 %             |                                 |                       |
| 3                                       | Effet traitement chez les Mq+ (RR)  | 0.8              |                                 |                       |
| 4                                       | Effet traitement chez les Mq- (RR)  | 0.9              |                                 |                       |
| 5                                       | Taille de la population   | 1 000            |                                 |                       |
| Comparaison des différentes stratégies  |   |                  |                                 |                       |
|   |   | Stratégie ciblée | Traitement de tous les patients | Absence de traitement |
| 8                                       | Nb total d'événements sur la population   | 88               | 84                              | 100                   |
| 9                                       | Nb de sujets traités (avec le nouveau traitement)   | 600              | 1 000                           | 0                     |
| 10                                      | Proportion de sujets traités  | 60 %             | 100 %                           | 0 %                   |
| 11                                      | Nb d'événements chez les sujets traités   | 48               | 84                              | 0                     |
| 12                                      | Nb d'événements chez les sujets non traités   | 40               | 0                               | 100                   |
| <b>Risque relatif (RR)</b>              |   |                  |                                 |                       |
| 13                                      | RR par rapport à l'absence de traitement  | 0.88             | 0.84                            | 1                     |
| 14                                      | RR par rapport au traitement de tous les patients   | 1.05             | 1                               | SO                    |
| <b>Réduction du risque absolu (ARR)</b> |   |                  |                                 |                       |
| 15                                      | Nb d'événements évités (par rapport à l'absence de traitement)  | -12              | -16                             | 0                     |
| 16                                      | Proportion des événements évités par rapport à tous les événements évitables (par le traitement de tous les patients) | 75 %             | 100%                            | SO                    |
| 17                                      | Nb d'événements évités par patient traité   | <b>-0.02</b>     | <b>-0.016</b>                   | SO                    |
| 18                                      | Ratio de rentabilité (d'un traitement) en nb d'événements évités  | 1.25             |                                 |                       |

#### 1.7.4 Stratification en fonction d'une interaction sur la sécurité

Le tableau 11 illustre l'apport d'une stratégie ciblée à l'aide d'un marqueur prédictif des événements indésirables liés à un traitement. Chez les sujets marqueur (+), le traitement est sûr avec une fréquence faible des événements indésirables de 0.5 % (ligne 3), tandis que chez les patients marqueur (-) ce risque est 8 fois plus important (4 %, ligne 4).

Le traitement de tous les patients (stratégie indifférenciée) conduit à une balance bénéfice/risque peu favorable au traitement avec 1.1 événements évités pour 1 effet indésirable induit : 20 événements évités (réduction de 100 à 80) pour 19 événements indésirables induits. Par contre, après ciblage du traitement chez les patients, ce rapport est de 4 événements évités pour 1 effet indésirable induit : 12 événements évités (88 à la place de 100 sans traitement) pour seulement 3 événements indésirables induits. Cependant, la stratégie ciblée est moins efficace en termes d'efficacité en évitant seulement 60 % de tous les événements évitables. Le ciblage réduit de 84 % le nombre d'événements indésirables mais ne prévient que 60 % des événements évitables. En pratique, ce type de ciblage est couramment effectué, par exemple par l'ajout de contre-indications à un traitement.

**Tableau 11 - Simulation des conséquences populationnelles d'un ciblage à l'aide d'un marqueur prédictif de la sécurité.**

| Paramètres                              |   |                  |                                 |                       |
|---|---|------------------|---------------------------------|-----------------------|
| 1                                       | Proportion de Mq+   | 60 %             |                                 |                       |
| 2                                       | Risque de base (Mq+ et Mq-)   | 10 %             |                                 |                       |
|   | Fréquence effet indésirable Mq+   | 0.5 %            |                                 |                       |
|   | Fréquence effet indésirable Mq-   | 4.0 %            |                                 |                       |
| 3                                       | Effet traitement chez les Mq+ (RR)  | 0.8              |                                 |                       |
| 4                                       | Effet traitement chez les Mq- (RR)  | 0.8              |                                 |                       |
| 5                                       | Taille de la population   | 1 000            |                                 |                       |
| Comparaison des différentes stratégies  |   |                  |                                 |                       |
|   |   | Stratégie ciblée | Traitement de tous les patients | Absence de traitement |
| <b>Effet indésirable (EI)</b>           |   |                  |                                 |                       |
|   | Nb total d'EI   | 3                | 19                              | 0                     |
|   | Nb EI induits   | 3                | 19                              | 0                     |
|   | Nb de sujets traités  | 600              | 1 000                           |                       |
|   | Fréquence d'évènements chez les traités   | 0.5 %            | 1.9 %                           | 0.0 %                 |
|   | RR EI par rapport au traitement de tous   | 0.26             |                                 |                       |
|   | ARR EI  | 0.5 %            | 1.9 %                           |                       |
|   | Ratio d'évènements évités par EI induit   | 4                | 1.1                             |                       |
|   | Nb d'évènements induits chez les Mq-  | 0                | 16                              |                       |
|   | % des EI traitement ciblé / traitement tous patients  | 16 %             |                                 |                       |
|   | % des EI épargnés par le ciblage (par rapport au traitement pour tous)  | 84 %             |                                 |                       |
| <b>Critère de jugement d'efficacité</b> |   |                  |                                 |                       |
| 8                                       | Nb total d'évènements sur la population   | 88               | 80                              | 100                   |
| 9                                       | Nb de sujets traités (avec le nouveau traitement)   | 600              | 1 000                           | 0                     |
| 10                                      | Proportion de sujets traités  | 60 %             | 100 %                           | 0 %                   |
| 11                                      | Nb d'évènements chez les traités  | 48               | 80                              | 0                     |
| 12                                      | Nb d'évènements chez les non traités  | 40               | 0                               | 100                   |
| <b>Réduction du risque absolu (ARR)</b> |   |                  |                                 |                       |
| 15                                      | Nb d'évènements évités (par rapport à pas de traitement)  | -12              | -20                             | 0                     |
| 16                                      | Proportion des évènements évités par rapport à tous les évènements évitables (par le traitement de tous les patients) | 60 %             | 100 %                           | SO                    |
| 17                                      | Nb d'évènements évités par patients traités   | -0.02            | -0.02                           | SO                    |

### 1.7.5 Synthèse

|  |  |
|--|--|
| L'utilisation d'un ciblage du traitement à l'aide d'un ... | <b>entraîne par rapport à une stratégie non ciblée ...</b>   |
| véritable marqueur prédictif                               | <ul style="list-style-type: none"> <li>• une diminution du nombre de sujets traités             <ul style="list-style-type: none"> <li>▶ réduction des coûts (à condition que le coût du test ne compense pas l'économie réalisée sur les traitements)</li> <li>▶ réduction du nombre d'effets indésirables</li> </ul> </li> <li>• sans réduction du nombre d'événements évités</li> <li>• au total, amélioration de la balance bénéfice/risque (chez les sujets traités)</li> </ul> |
| marqueur non prédictif                                     | <ul style="list-style-type: none"> <li>• une diminution du nombre d'événements évités (du fait du non traitement à tort des patients marqueurs (-))</li> <li>• un surcoût lié à la recherche du marqueur inutilement chez tous les sujets</li> </ul>   |
| marqueur prédictif partiel (pseudo-ciblage)                | <ul style="list-style-type: none"> <li>• une diminution du nombre d'événements évités (du fait du non traitement à tort des patients marqueurs (-))</li> <li>• un surcoût lié à la recherche du marqueur inutilement chez tous les sujets</li> </ul>   |

## 2. Principes généraux de validation d'un test diagnostique déterminant le statut des patients pour un marqueur présumé prédictif

### 2.1 Les trois niveaux de validation d'un test diagnostique

L'évaluation du test diagnostique déterminant le statut des patients pour un marqueur présumé prédictif comporte trois niveaux : l'évaluation de la validité analytique, de la validité clinique et de l'utilité clinique du test (13–15).

- **La validité analytique (*analytic validity*)** d'un test diagnostique correspond à son aptitude *in vitro* à réaliser la mesure d'intérêt avec exactitude et fiabilité. En d'autres termes : le test mesure-t-il bien ce qu'il est censé mesurer et de façon correcte ? Cette validation inclut l'étude de la sensibilité et la spécificité analytiques, la reproductibilité, la robustesse et la satisfaction des contrôles qualité.
- **La validité clinique (*clinical validity*)** d'un test diagnostique correspond à son aptitude à prédire avec précision et fiabilité le phénotype clinique d'intérêt (ex : survie globale ou survie sans progression des patients recevant un traitement donné) : y-a-t-il une relation, et laquelle, entre les résultats du test et le phénotype d'intérêt ? La validité clinique inclut la sensibilité et la spécificité cliniques, ainsi que les valeurs prédictives positive et négative du test, paramètres que l'on peut réunir sous l'appellation de « performances diagnostiques » du test.
- **L'utilité clinique (*clinical utility*)** d'un test diagnostique correspond à son aptitude à améliorer le devenir clinique des patients en événements cliniques mesurables, et à apporter une valeur ajoutée en termes d'optimisation de décision de traitement et en corollaire de stratégie thérapeutique.

L'étude de la validité analytique et de la validité clinique sont des notions génériques pour les tests diagnostiques, qu'un marqueur soit ou non impliqué. Elles n'entrent donc pas dans le champ de ce document dédié aux tests associés à des marqueurs prédictifs.

La **démonstration de l'utilité clinique** est un **élément indispensable** dans le domaine des tests compagnons. Cette démonstration est indispensable parce qu'elle atteste de la **valeur ajoutée apportée par le test évalué** en comparaison d'un test diagnostique permettant uniquement de sélectionner une sous-population par l'identification d'un marqueur. Le concept même de test compagnon intègre que son utilité clinique soit démontrée puisque il doit permettre de ne pas traiter inutilement des patients chez lesquels le traitement serait inefficace, et à l'inverse de ne pas entraîner de perte de chance par l'identification d'un marqueur non ou peu prédictif en excluant des patients d'un traitement qui leur apporterait en réalité un bénéfice.

### 2.2 Les trois conditions impérativement requises pour démontrer l'utilité clinique d'un test diagnostique

La démonstration de l'utilité clinique d'un test diagnostique implique la **démonstration des trois éléments suivants** (9,16–19) :

- le marqueur identifié par le test doit modifier l'effet du traitement (existence d'une interaction « marqueur/traitement ») ;
- le traitement doit être efficace chez les patients marqueur (+) ;

- le traitement doit être sans intérêt clinique chez les patients marqueur (-).

### 2.2.1 Mise en évidence de l'interaction marqueur/traitement

En termes statistiques, un marqueur prédictif doit impérativement être un « modificateur » de l'effet du traitement (cf. section 1.2.1), c'est-à-dire que la taille de l'effet du traitement doit varier en fonction de la présence ou non du marqueur (dans le cas d'un marqueur binaire). Ainsi, si l'effet du traitement est mesuré par un risque relatif (ou hazard ratio), cela revient à avoir, entre le sous-groupe des patients **marqueur (+)** et celui des patients **marqueur (-)**, **deux risques relatifs statistiquement différents**. Cette différence statistiquement significative entre les deux sous-groupes valide statistiquement la modification de l'effet traitement en fonction de la présence ou non du marqueur, en d'autres termes l'« interaction marqueur/traitement ». Il s'agit donc d'une recherche d'interaction similaire à celle qui est faite dans n'importe quelle analyse en sous-groupes, qu'elle ait ou non comme objectif la mise en évidence d'un marqueur prédictif.

Cette situation est illustrée dans le Tableau 12 avec l'exemple d'un nouveau traitement évalué *versus* placebo en *add-on* de la stratégie habituelle. Le critère de jugement est la mortalité totale.

**Tableau 12 - Illustration d'une situation d'interaction statistique suggérant l'existence d'un marqueur prédictif.**

|            | Traitement N | Placebo   | Risk Ratio (IC 95 %) |                      |
|------------|--------------|-----------|----------------------|----------------------|
| Marqueur + | 325/2 000    | 445/2 000 | 0.73 [0.64 ; 0.83]   |                      |
| Marqueur - | 440/2 000    | 465/2 000 | 0.94 [0.84 ; 1.06]   | Interaction p=0.0033 |

Le test d'interaction ne répond qu'à la question de la mise en évidence ou non de l'existence d'une interaction. Il est aussi possible, dans une certaine mesure, de donner une estimation de l'importance de l'interaction. Une possibilité est de recourir au rapport des risk ratios  $R\alpha RR$ .

$$R\alpha RR = \frac{RR_{mq+}}{RR_{mq-}}$$

Avec les données du Tableau 12,  $R\alpha RR=0.73/0.94=0.78$ . Cela signifie que l'effet du traitement chez les patients marqueur (-) est seulement 0.78 fois celui chez les patients marqueur (+), soit une réduction relative de l'efficacité de 22 %. Il convient cependant en pratique de calculer l'intervalle de confiance de ce rapport pour matérialiser la précision avec laquelle est connu cet indice. Ce calcul fait appel à la même approche que celle des comparaisons indirectes. Cet intervalle de confiance reflète ainsi l'incertitude existant sur l'estimation des deux risques relatifs qui le constituent.

La valeur de ce rapport des risk ratios donne une idée de la **pertinence clinique de l'interaction** détectée. En effet, une interaction de petite ampleur, même significative, ne conduira pas à une optimisation importante de la rentabilité avec la stratégie utilisant le marqueur. Il faut cependant souligner que le lien entre importance quantitative de l'interaction et rentabilité de la stratégie ciblée n'est pas très simple (cf. Tableau 9 ; Tableau 10 ; Tableau 11) et que l'importance quantitative de cette interaction n'est pas un élément en soi de la démonstration de la valeur prédictive d'un marqueur.

## 2.2.2 Mise en évidence de l'effet du traitement chez les patients marqueur (+)

L'essai doit logiquement démontrer l'efficacité du traitement chez les patients marqueur (+), ceci en suivant les **mêmes critères que pour un essai thérapeutique classique** : essai prospectif, à l'abri des biais, statistiquement significatif, cliniquement pertinent et ayant une bonne cohérence externe (plausibilité biologique, cohérence entre les études disponibles).

**Un aspect spécifique** par rapport à un essai classique concernant la démonstration de l'efficacité d'un nouveau traitement est le contrôle du risque alpha. En effet, une multiplicité latente des tests statistiques est inhérente au design dans ce concept d'essai : tests de l'effet du traitement chez les patients marqueur (+), marqueur (-) et/ou tous patients confondus. À chaque fois, un risque de conclure à tort de 5 % peut être pris en l'absence de précaution. Le **contrôle d'inflation du risque alpha** peut être réalisé de plusieurs façons. La première consiste en une analyse séquentielle hiérarchique où l'on teste en premier les patients marqueur (+), puis la totalité des patients. La deuxième est de pré-spécifier explicitement que l'objectif primaire de l'essai est de valider l'interaction et l'efficacité du traitement chez les patients marqueur (+), en concentrant le risque alpha sur l'analyse liée à cet objectif primaire. Cependant, les analyses tous patients confondus et chez les patients marqueur (-) poseront un problème de multiplicité des comparaisons statistiques. Enfin, une troisième possibilité est d'ajuster le risque alpha en tenant compte du nombre des différentes comparaisons possibles à l'aide d'une méthode de type Bonferroni ou autre.

## 2.2.3 Mise en évidence de l'absence d'effet chez les patients marqueurs (-)

### ► Principe

L'existence d'une interaction entre l'effet du traitement et le marqueur ne signifie pas que le traitement n'a pas d'effet chez les patients marqueur (-). En effet, lorsqu'une interaction est détectée, **il est tout à fait possible** (cf. Tableau 13) **que le traitement soit efficace aussi bien chez les patients marqueur (+) que chez les patients marqueur (-)** mais avec une différence de taille d'effet.

**Tableau 13 - Situation où une interaction existe et où le traitement est efficace chez les patients marqueur (+) et marqueur (-).**

|            | Traitement N | Placebo   | Risk Ratio (IC 95 %) |                      |
|------------|--------------|-----------|----------------------|----------------------|
| Marqueur + | 650/5 000    | 890/5 000 | 0.73 [0.66 ; 0.80]   |                      |
| Marqueur - | 800/5 000    | 930/5 000 | 0.78 [0.79 ; 0.94]   | Interaction p=0.0114 |

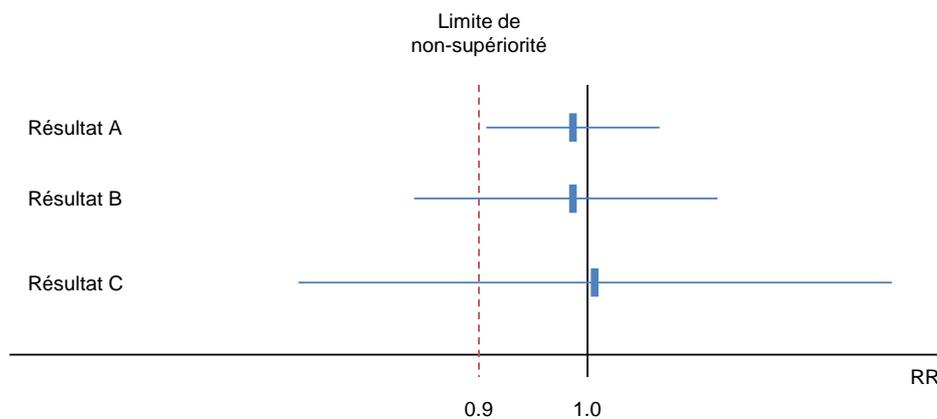
Or, le concept de marqueur prédictif implique impérativement l'absence d'effet du traitement chez les patients marqueur (-). L'obtention d'un rapport des risques (ou risque relatif, RR) proche de 1 dans le sous-groupe des patients marqueurs (-), comme cela est le cas dans l'exemple du Tableau 14, est un argument suggestif mais non suffisant pour exclure avec une certitude raisonnable l'absence d'effet chez ces patients.

**Tableau 14 - Marqueur prédictif associé à une forte suggestion d'absence d'effet chez les patients marqueur (-).**

|            | Traitement N | Placebo   | RR (IC 95 %)       |                      |
|------------|--------------|-----------|--------------------|----------------------|
| Marqueur + | 390/2 000    | 534/2 000 | 0.73 [0.65 ; 0.82] |                      |
| Marqueur - | 565/2 000    | 558/2 000 | 1.02 [0.91 ; 1.12] | Interaction p<0.0001 |

**Il faut une démonstration statistique** de l'absence d'effet dans ce sous-groupe basée sur l'étude de l'intervalle de confiance du risque relatif de ces patients. Un résultat non statistiquement significatif peut simplement provenir d'un manque de puissance. De plus les analyses en sous-groupes engendrent une inflation du risque bêta. Cette démonstration doit faire appel à une étude de non infériorité basée sur l'intervalle de confiance du risque relatif de ces patients, en montrant qu'il est possible d'exclure avec un degré de certitude de 97.5 % que l'effet du traitement puisse être supérieur à une certaine valeur correspondant au bénéfice minimal cliniquement intéressant. Le raisonnement est en fait en pratique un raisonnement en miroir par rapport à celui-ci, consistant en un **raisonnement de non supériorité** (avec une limite de non supériorité clinique). On peut conclure à la non supériorité lorsque la borne inférieure de l'intervalle de confiance de l'effet du traitement dans la strate des patients marqueur (-) est supérieure à la limite de non supériorité. On cherche en effet à éliminer la possibilité que le traitement soit efficace (c'est-à-dire qu'il présente un effet substantiellement supérieur au placebo) chez les patients marqueur (-).

La Figure 4 illustre l'application de ce raisonnement de non supériorité à la mise en évidence d'un marqueur prédictif. Dans le cas du résultat C, il est impossible de conclure que le traitement n'a pas d'effet chez les sujets marqueurs (-) car l'intervalle de confiance associé à ce résultat est très large et compatible avec des effets importants (toute sa partie inférieure à la limite de non supériorité) ne permettant pas d'exclure que le traitement soit sans effet chez ces patients. La situation recherchée est celle du résultat A où il est possible d'exclure que le traitement puisse être suffisamment efficace pour être cliniquement intéressant chez ces patients. Le résultat B ne permet pas de conclure car il reste compatible avec une efficacité cliniquement pertinente du traitement chez ces patients.



**Figure 4 - Application du raisonnement de non supériorité à la mise en évidence des marqueurs prédictifs. Exemples de différents types de résultats possibles dans le sous-groupe des patients marqueurs (-).**

Il convient de remarquer que ce raisonnement n'a pas été régulièrement suivi lors de l'enregistrement des premiers traitements de ce type. Par exemple, le cetuximab a été enregistré dans le traitement du cancer colorectal métastatique limité aux patients sans mutation du gène *KRAS* à partir d'une simple estimation proche de 1 du hazard ratio pour la survie sans progression dans le sous-groupe marqueur (-) (cf. section 4.3.3).

#### ► Discussion sur cette approche

Il est intéressant de noter qu'en corollaire de ce raisonnement de non supériorité et en toute rigueur, la démonstration de l'intérêt du traitement chez les patients marqueur (+) imposerait

que l'intervalle de confiance soit entièrement inférieur à la limite de non supériorité. En effet, le résultat obtenu chez les patients marqueur (+) pourrait sinon être compatible avec des tailles de bénéfice considérées par définition comme cliniquement insuffisantes (la partie de l'IC au-dessus de la limite de non supériorité). Cela revient à obtenir une interaction forte qui sépare complètement les deux intervalles de confiance obtenus chez les patients marqueur (+) et les patients marqueur (-).

Par ailleurs, cette démonstration de non supériorité présente en pratique deux points critiques requérant une attention particulière.

Le **premier point** concerne la **fixation de la limite de non supériorité**, c'est-à-dire de la plus petite différence cliniquement pertinente. La fixation de cette limite est arbitraire et doit être faite *a priori*, avant d'avoir connaissance des résultats de l'étude. De plus elle doit être justifiée cliniquement. La fixation de cette limite est grandement facilitée en raisonnant, non plus en terme d'efficacité, mais de balance bénéfique/risque. En effet, il est très probable d'avoir le même profil de tolérance quelle que soit le statut du marqueur<sup>5</sup>, donc que la balance bénéfique/risque soit peu favorable chez les patients marqueur (-). Ainsi la conclusion de non efficacité chez les patients marqueur (-) pourra être bien plus facile à obtenir en utilisant un critère de bénéfice net agrégeant à la fois l'efficacité et le ou les risques (par exemple un critère regroupant événements cardiovasculaires et hémorragies majeures pour un antiplaquettaire en prévention des événements chez le coronarien). Cette sensibilisation de l'outil de mesure décalera l'intervalle de confiance vers la droite. Le cas le plus simple sera celui où une infériorité statistiquement significative sera mise en évidence (intervalle entièrement au-dessus du 1). Si ce critère de bénéfice net est utilisé à ce niveau, il convient bien sûr d'utiliser le même critère pour la recherche de l'interaction et pour la mise en évidence de l'intérêt du traitement chez les patients marqueur (+).

Le **deuxième point** est lié à l'objectif de cette recherche. En effet, l'objectif de démontrer qu'un nouveau traitement ne serait pas efficace chez les patients marqueur (-) peut être susceptible de soulever une **question éthique**. Ce point est discuté dans la section 3.1 de ce guide, dans le cadre du schéma d'étude d' « interaction marqueur/traitement ».

## 2.3 Les critères de jugement

Les critères de jugement des études mises en œuvre pour la validation de l'utilité clinique du test diagnostique sont **les mêmes critères que ceux utilisés pour l'évaluation de l'efficacité** du traitement, c'est-à-dire des **critères cliniques**.

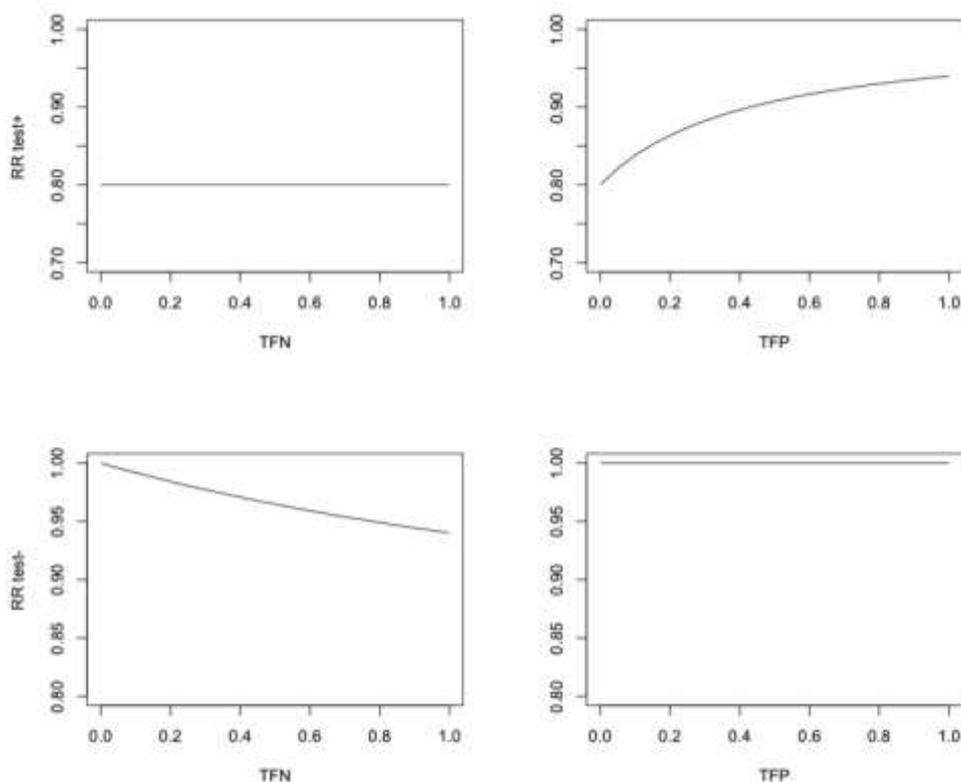
Certains travaux recherchent les marqueurs prédictifs uniquement sur des critères intermédiaires tels que le LDL cholestérol pour un traitement hypocholestérolémiant de prévention des événements cardiovasculaires, ou la réponse tumorale pour un traitement du cancer. Cette approche n'est pas satisfaisante car elle fait l'hypothèse que ce qu'il se passe au niveau du critère intermédiaire est directement prédictif de ce qu'il se passe au niveau du critère clinique (en positif et en négatif). Un effet positif du traitement sur le critère intermédiaire serait prédictif de son bénéfice clinique et l'absence d'effet sur ce critère intermédiaire serait prédictive de l'absence de bénéfice clinique. Cette hypothèse est très forte et n'est pas acceptée pour la démonstration de l'efficacité des traitements. Il n'y a aucune raison qu'elle soit plus acceptable pour la recherche d'un marqueur prédictif.

---

<sup>5</sup> Comme c'est le cas habituellement avec les marqueurs prédictifs d'efficacité.

## 2.4 Conséquences de mauvaises performances diagnostiques du test sur la démonstration de son utilité clinique

Les conséquences de mauvaises performances diagnostiques du test d'identification d'un marqueur sur les estimations de l'efficacité d'un traitement sont illustrées graphiquement sur la Figure 5. Ces graphiques représentent l'évolution de l'effet du traitement (rapport des risques) observé dans les strates test (+) (correspondant aux patients identifiés marqueur (+) par le test) et test (-) (patients identifiés marqueur (-)) en fonction du taux de faux positifs (TFP) et du taux de faux négatifs (TFN) donnés par le test utilisé pour déterminer le statut des patients vis-à-vis du marqueur. Le vrai rapport des risques chez les vrais marqueurs (+) est de 0.8 et celui chez les vrais marqueurs (-) est de 1.00.



**Figure 5 - Évolution des estimations de l'effet du traitement (rapport des risques, RR) dans les strates test (+) et test (-) en fonction des taux de faux positifs (TFP) et de faux négatifs (TFN). Le vrai rapport des risques chez les patients marqueur (+) est de 0.8 et chez les patients marqueur (-) de 1.00.**

Il apparaît qu'un **défaut de sensibilité** (augmentation du TFN) ne perturbe pas l'estimation de l'effet du traitement dans la strate test (+), mais entraîne une surestimation de l'effet dans la strate test (-) pouvant conduire à l'atténuation de l'interaction entre les deux strates et ne pas faire conclure à tort à la valeur prédictive du marqueur (mesuré par ce test).

Un **défaut de spécificité** (augmentation des faux positifs) n'entraîne pas d'erreur de mesure de l'effet du traitement dans la strate test (-) mais entraîne une sous-estimation de l'effet dans la strate test (+) pouvant conduire à une conclusion à tort d'absence d'effet du traitement dans la strate test (+) et donc globalement à une absence d'effet du traitement (rapport des risques proche de 1 dans les deux cas).

On notera que le défaut de spécificité est conservateur et ne conduit pas à accepter à tort la valeur prédictive d'un marqueur et/ou l'efficacité d'un traitement. Par contre, le défaut de sensibilité est conservateur sur la valeur prédictive mais peut en théorie amener à conclure à tort à l'efficacité du traitement chez tous les patients (test (+) et test (-)) alors qu'il ne serait en réalité efficace que dans la strate test (+).

En pratique, il s'avère que ces problèmes de défauts de performances diagnostiques du test utilisé pour déterminer le statut des patients vis-à-vis d'un marqueur constituent **plutôt une problématique pour le développeur** que pour les autorités de régulation. En effet, les défauts de sensibilité et spécificité vont à l'encontre de l'hypothèse de la valeur prédictive du marqueur testée dans l'essai d'interaction. La seule situation intéressant directement les autorités est celle du défaut de sensibilité conduisant à conclure à l'intérêt du traitement chez tous les patients y compris ceux de la strate test (-), où l'efficacité apparaît en raison de la présence dans cette strate de patients en réalité marqueur (+) chez lesquels le traitement est efficace. Il convient donc, devant un essai construit initialement comme un essai d'interaction mais concluant finalement à l'efficacité du traitement tous patients confondus, de vérifier le **taux de faux négatifs** généré par le test et de ne pas accepter la revendication d'efficacité globale si ce taux est important (au-dessus de 30 %).

### 3. Principaux schémas d'études cliniques : analyse critique

#### 3.1 L'étude d' « interaction marqueur/traitement » : seul schéma permettant de démontrer l'utilité clinique d'un test identifiant un marqueur présumé prédictif et le caractère ciblé d'une thérapie

##### ► Caractéristiques de l'approche

Un seul schéma d'étude (ou design), appelé schéma d' « interaction marqueur/traitement », d' « essai de recherche d'interaction » ou « *marker by treatment interaction design* », permet de déterminer de façon concomitante l'effet du traitement chez les patients marqueur (+) et marqueur (-), et de tester en corollaire l'existence d'une interaction marqueur/traitement. Ce schéma est ainsi le seul à pouvoir **démontrer conjointement les trois conditions** permettant la démonstration de l'utilité clinique d'un test diagnostique identifiant un marqueur présumé prédictif (3,16–19). Il est en corollaire également le seul à pouvoir démontrer qu'un traitement peut être réellement qualifié de ciblé, et non simplement de stratifié.

En tant que seul plan d'expérience permettant de démontrer strictement l'utilité clinique d'un test, ce schéma d'étude conduit aux résultats de plus haut niveau de preuve et constitue la **référence** dans son domaine.

Cette approche peut être envisagée dans la majorité des situations :

- avant toute démonstration de l'intérêt du traitement. Il s'agit alors d'un développement conjoint du traitement et du test. Le centrage est celui du traitement et la finalité du dossier est de revendiquer l'intérêt du traitement dans la sous-population sélectionnée par le marqueur ;
- à la suite d'une recherche *post-hoc* en sous-groupes d'un marqueur prédictif suite à un essai non concluant portant sur l'intérêt d'un nouveau traitement. L'étude d'interaction correspond dans ce cas à une étude prospective de confirmation de résultats exploratoires.

Elle ne peut cependant pas être envisagée pour confirmer la valeur prédictive présumée d'un marqueur candidat lors d'une recherche *post-hoc* en sous-groupes à la suite d'une étude aux résultats concluants, ceci pour des raisons éthiques. Il n'est en effet pas raisonnable d'envisager de randomiser à nouveau les patients contre le comparateur du premier essai, le nouveau traitement ayant démontré sa plus grande efficacité pour tous les patients indépendamment de leur statut pour le marqueur.

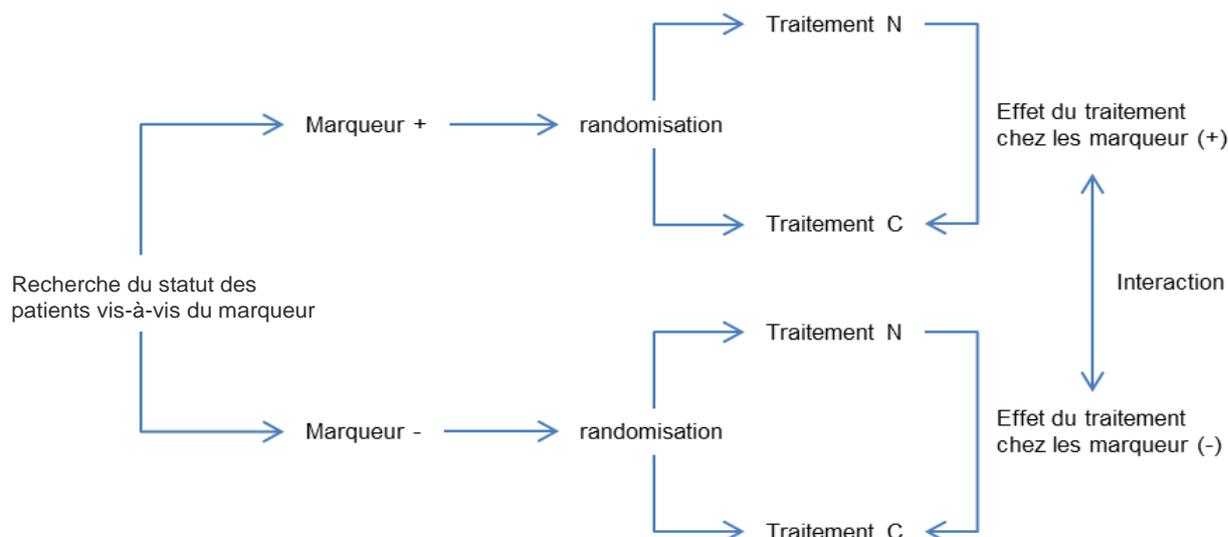


Figure 6 - Illustration du schéma d'étude d' « interaction marqueur traitement ».

### ► Schéma d'étude

Ce schéma comprend une **double randomisation, une dans chaque sous-population identifiée par son statut** positif ou négatif **pour le marqueur**, ce qui permet alors de comparer le nouveau traitement N à un comparateur C aussi bien chez les patients marqueur (+) que chez les patients marqueur (-) (cf. Figure 6). Deux estimations de l'effet de N *versus* C sont produites, chez les patients marqueur (+) et (-). Leur comparaison permet la recherche d'une différence statistiquement significative signifiant le cas échéant l'existence d'une interaction marqueur/traitement.

Différentes stratégies d'analyses peuvent être mises en œuvre. En général, une **approche séquentielle** est utilisée, commençant par le test de l'interaction. Si ce test n'est pas significatif selon un seuil prédéfini, les groupes marqueur (+) et marqueur (-) peuvent être regroupés pour faire une évaluation globale de l'efficacité du traitement. S'il y a significativité, les deux strates définies par le marqueur sont analysées. Freidlin *and* Simon (20) proposent de répartir aussi le risque alpha entre le test global et le test par strate.

### ► Discussion sur cette approche

Il faut noter que dans ce design, seuls sont inclus les patients avec un marqueur parfaitement évaluable. Le pourcentage de patients ayant un test non exploitable est donc à prendre en compte dans la lecture critique des résultats.

Il est également important de noter que, si le traitement est réellement efficace dans une large proportion de la population des malades, ce design peut permettre de conclure même si le marqueur n'est pas prédictif.

Les **enjeux éthiques** soulevés par l'analyse de l'absence d'efficacité du traitement chez les patients marqueur (-) sont similaires à ceux que soulève la randomisation dans les essais cliniques en général. La justification de cette randomisation d'un point de vue éthique repose sur le **principe d'équipoise** : il est justifié de conduire une étude clinique randomisée lorsque la communauté médicale ne peut pas déterminer, selon le principe de la médecine basée sur les preuves, la meilleure option entre deux traitements. La randomisation n'est acceptable que tant que l'incertitude demeure (21–23). Les patients marqueur (-) ne doivent

pas être exposés à des effets indésirables sans espoir escompté de bénéfique. Il paraît alors important de garder à l'esprit, dans le contexte des thérapies stratifiées, que **l'absence d'efficacité** du nouveau traitement chez les patients marqueur (-) n'est pas **un résultat connu a priori**, mais une condition nécessaire à la démonstration du caractère ciblé de la thérapie. **L'hypothèse** testée est celle de la **non supériorité du nouveau traitement au comparateur**, placebo ou traitement de référence en cours, et comme n'importe quelle hypothèse testée par une étude clinique, elle sera susceptible d'être *in fine* rejetée ou non. En l'occurrence, c'est pour cette raison qu'il est recommandé de prévoir l'analyse de l'efficacité du traitement tous patients confondus (Mq+ et Mq-) en plus de l'analyse sous-populationnelle, dans le plan d'analyse statistique du schéma d' « interaction marqueur/traitement ». Cette position est renforcée par le fait qu'à ce jour des **contre-exemples existent** parmi les traitements qui étaient revendiqués comme *a priori* « ciblés par construction » et pour lesquels l'absence d'efficacité chez les patients marqueur (-) a été remise en question. L'absence d'efficacité d'un traitement chez des patients marqueur (-) qui serait « prédite » par un marqueur identifié par des analyses *post-hoc* est aussi très déraisonnable compte tenu de la nature exploratoire de la recherche et du caractère globalement purement théorique du **mécanisme d'implication** du marqueur dans l'efficacité du traitement considéré (cf. section 1.4). Au final, il apparaît que ne pas réaliser cette recherche ferait globalement courir un risque aux patients marqueur (-) en favorisant **l'acceptation à tort de marqueurs non ou insuffisamment prédictifs** susceptibles par la suite de leur causer une perte de chance en les privant d'un traitement efficace pour eux.

Enfin, pour **exemple**, ce schéma est celui utilisé dans l'essai en cours « *Marker Validation of Erlotinib in Lung Cancer (MARVEL)* » (NCT00738881), où les patients sont d'abord stratifiés en fonction de la détection ou non par technique FISH (*fluorescence in situ hybridization*) de l'expression tumorale du récepteur EGFR puis randomisés entre erlotinib et pemetrexed en seconde ligne du traitement du cancer du poumon non à petites cellules.

## 3.2 Les principaux autres schémas d'étude : détermination de l'effet du traitement et identification exploratoire de marqueurs présumés prédictifs

En pratique, le type de méthodologie utilisée dans la majorité des études a été et reste à ce jour centré sur la démonstration de l'efficacité du nouveau traitement mais ne permet pas de démontrer celle de l'utilité clinique du test diagnostique associé.

### 3.2.1 L'étude « ciblée »

#### ► Caractéristiques de l'approche

Le schéma d'étude de type « ciblé » a été à la base des premiers enregistrements de traitements de prescription limitée à une sous-population identifiée (« ciblée ») par la présence d'un marqueur présumé prédictif de l'efficacité de ce traitement. Il est toujours couramment utilisé.

Ce schéma est tout à fait approprié pour démontrer l'efficacité d'un nouveau traitement dans une sous-population de patients définie par la présence d'un marqueur détecté par un test diagnostique. Cependant, en toute rigueur, il ne permet pas de valider l'utilité clinique du test en question, ni par conséquent de revendiquer le caractère ciblé du traitement, car il ne permet pas de démontrer **explicitement** l'absence d'effet chez les patients marqueur (-) (24–26). En effet, pour les raisons explicitées dans la section 1.5, il est peu raisonnable de considérer l'absence d'efficacité chez les patients marqueur (-) comme **implicite** sur la base de données précliniques et/ou d'un rationnel biologique théorique.

Lorsque l'efficacité du nouveau traitement s'avère démontrée par ce schéma dans le sous-groupe sélectionné, donc chez les patients marqueur (+), l'**AMM dans l'indication** étudiée est logiquement **restreinte à ce sous-groupe** de patients chez lesquels l'efficacité du traitement a été étudiée. Une prescription du médicament, pour être conforme à l'AMM, rend donc nécessaire la réalisation du test diagnostique, afin de discriminer les patients marqueur (+) et (-). En effet, prescrire le médicament à un patient marqueur (-) consiste alors en une prescription hors AMM. Il faut souligner sur ce point que la restriction de prescription de l'AMM ne doit pas être considérée comme une preuve d'inefficacité du traitement chez les patients marqueur (-) mais qu'elle résulte seulement de l'absence d'évaluation de cette efficacité. Le **principal risque** de ce design est par conséquent d'entraîner, en l'absence de preuves cliniques, une **perte de chance** pour les patients **marqueur (-)** en les privant d'un traitement susceptible de leur apporter un bénéfice, en plus de réaliser un test inutile.

*In fine*, ce schéma n'est **pas adapté pour démontrer l'utilité clinique** d'un test diagnostique associé à une thérapie stratifiée. Néanmoins, au cas par cas, le niveau de conviction accordé aux éléments précliniques et cliniques précoces lors de l'évaluation peut soutenir dans une certaine mesure cette démonstration. Ce schéma présenterait plutôt essentiellement un intérêt lorsque des données précliniques et/ou cliniques précoces amènent à craindre une toxicité particulière du nouveau traitement chez les patients marqueur (-). Cependant, l'utilité clinique du test ne sera pas démontrée et il peut donc être préférable, lorsque cela est possible, de choisir un design d' « interaction marqueur/traitement » associé à un protocole permettant d'assurer aux patients le niveau de sécurité exigible pour tout essai thérapeutique.

### ► Schéma d'étude

Le schéma d'étude de type ciblé, également nommé « design ciblé », « *targeted design* » ou bien encore « *enrichment design* », consiste à évaluer le nouveau traitement *versus* le traitement comparateur C uniquement chez des sujets marqueur (+) (cf.

Figure 7).

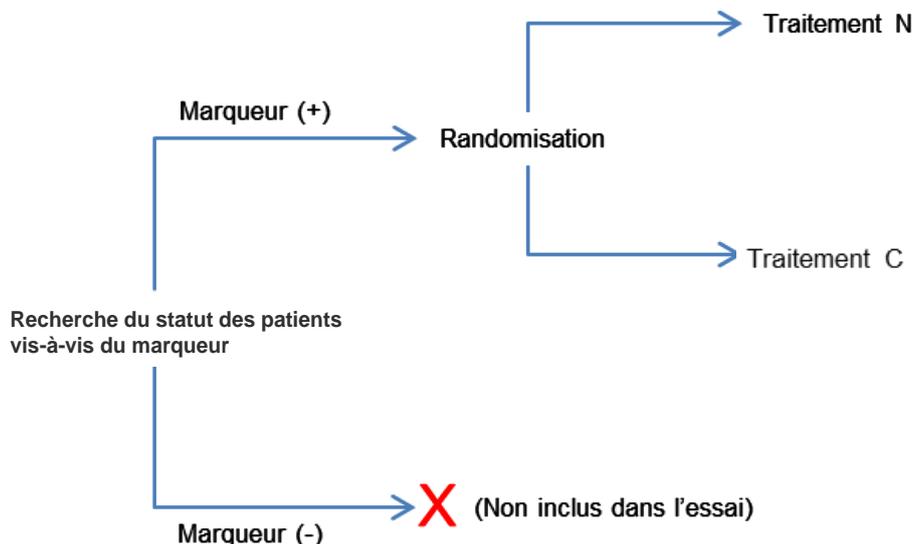


Figure 7 - Schéma d'étude « ciblée ».

## ► Exemples d'études

### Exemple : étude GRAVITAS

L'essai GRAVITAS (12) avait pour objectif d'évaluer l'intérêt d'un traitement par clopidogrel à haute dose par rapport à un traitement à dose standard chez des sujets présentant une hyperactivité plaquettaire résiduelle après initiation d'un traitement par une dose standard de clopidogrel faisant suite à la mise en place d'un stent actif. L'hyperactivité plaquettaire était évaluée à l'aide du dispositif VerifyNow®. Ce test a été effectué chez 5 429 patients et a détecté 2 214 patients avec une hyperactivité plaquettaire sous traitement. Ces patients ont ensuite été randomisés entre une forte dose et une dose standard de clopidogrel pour 6 mois. Aucune différence statistiquement significative ( $p=0.97$ ) n'a été observée entre les deux groupes sur le critère de jugement principal d'événement cardiovasculaire.

Le design de l'étude est bien un design ciblé, seuls des patients marqueur (+) ont été inclus. L'étude s'est avérée non concluante. Cependant, si elle avait donné un résultat positif, il aurait été impossible d'en déduire l'intérêt clinique du test VerifyNow® car ce design n'aurait pas documenté l'interaction, ni l'effet d'une forte dose de clopidogrel chez les patients sans hyperactivité plaquettaire résiduelle. Rien n'aurait donc permis d'exclure qu'une forte dose de clopidogrel n'aurait pu être également plus efficace chez les sujets sans hyperactivité plaquettaire résiduelle.

### Exemple : historique du cetuximab (anticorps monoclonal anti-EGFR)

Le récepteur au facteur de croissance épidermique (EGFR) est une glycoprotéine transmembranaire de la famille des récepteurs à tyrosine kinase. Il présente des propriétés oncogéniques et joue un rôle connu dans l'apparition et le développement des cancers. Le cetuximab, premier anticorps anti-EGFR, fut initialement « ciblés par construction » afin d'agir comme un inhibiteur du récepteur EGFR, l'expression de ce récepteur par immunohistochimie (IHC) étant alors considérée comme un marqueur prédictif de l'efficacité de ce traitement. En raison de ce « ciblage par construction » et de la spécificité considérée comme acquise compte tenu du caractère monoclonal de l'anticorps, l'absence d'efficacité du traitement chez les patients dits « EGFR-négatifs » (c'est-à-dire obtenant une détection négative d'EGFR par le test IHC) fut en corollaire considérée *a priori* comme implicite et les premiers essais du cetuximab furent réalisés en n'incluant exclusivement des patients exprimant ce récepteur (27). Des études ultérieures ont cependant mis en évidence une réponse au cetuximab chez des patients n'exprimant pas le récepteur EGFR en détection par IHC (8). La recherche de l'expression de ce récepteur n'est même plus requise dans l'AMM des anticorps anti-EGFR plus récents comme le panitumumab. Cet exemple illustre bien la toute relativité de nos connaissances mécanistiques (phénomène d'*antibody-dependent cell-mediated cytotoxicity* ?) ou de la performance de nos outils diagnostiques (problème de faux négatifs ?). Quelle que soit l'explication, la non validation initiale de l'absence d'efficacité du cetuximab chez les patients EGFR-négatifs par un design d'interaction a pu entraîner le non traitement à tort de patients, subissant ainsi une perte de chance.

### 3.2.2 L'étude de « comparaison de stratégies »

#### ► Principe et schéma d'étude

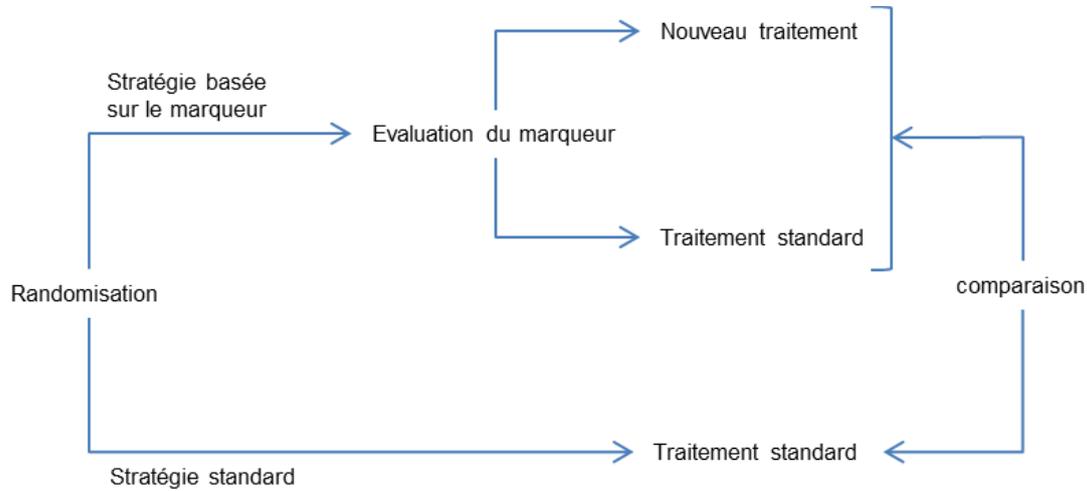


Figure 8 - Schéma d'étude de « comparaison de stratégies »

Le schéma d'étude correspondant à l'évaluation de stratégies est appelé design de « comparaison de stratégies » ou « *marker-based strategy design* » (3). La stratégie standard de traitement de tous les patients par le traitement standard S est comparée à la stratégie consistant à traiter les patients marqueur (+) avec le nouveau traitement N et les patients marqueur (-) avec le traitement standard S (cf. Figure 8). L'objectif (théorique) est de montrer que la nouvelle stratégie utilisant l'évaluation du marqueur permet d'obtenir de meilleurs résultats que la stratégie standard. Il convient donc de comparer les résultats obtenus avec la stratégie basée sur le marqueur (en regroupant les patients traités par N et par S) avec celle de la stratégie standard.

#### ► Discussion sur l'approche

Comme pour le design ciblé, ce schéma ne permet pas de mesurer l'effet du nouveau traitement chez les patients marqueur (-) et **ne permet donc pas de démontrer l'utilité clinique** du test diagnostique, ni par conséquent le caractère ciblé du nouveau traitement.

En réalité, malgré son appellation, ce schéma ne permet pas non plus de démontrer *stricto sensu* la supériorité d'une stratégie stratifiée « marqueur/traitement », où le choix du traitement est associé au résultat du test compagnon, en comparaison d'une stratégie standard non stratifiée. En effet, obtenir un meilleur résultat avec la stratégie « marqueur/traitement » évaluée par rapport à la stratégie standard non stratifiée peut être observé si le **nouveau traitement est plus efficace** que le traitement standard chez les patients marqueur (+), mais également si le nouveau traitement est plus efficace que le traitement standard chez l'ensemble des patients, qu'ils soient marqueur (+) ou (-), donc que le test soit cliniquement utile ou non. Par conséquent, une étude utilisant ce schéma et présentant des résultats concluants **ne permet de conclure que sur la supériorité de la stratégie globale** basée sur l'utilisation du marqueur et du nouveau traitement par rapport à l'utilisation du traitement standard chez tous les patients, mais ne permet pas de savoir si, dans le cadre de la stratégie basée sur le marqueur, la sélection des patients devant recevoir le nouveau traitement en fonction de la présence du marqueur est cliniquement utile.

Il convient donc de **se méfier** particulièrement de ce design lorsque l'objectif primaire est de valider le marqueur et non l'efficacité du nouveau traitement (comme dans l'exemple de l'essai ARTIC, cf. infra). Une étude utilisant ce schéma et présentant des résultats concluants ne permet de conclure que sur la supériorité de la stratégie basée sur le marqueur et le nouveau traitement par rapport à l'utilisation du traitement standard chez tous les patients, mais ne permet pas de savoir si, dans le cadre de la stratégie basée sur le marqueur, la sélection des patients devant recevoir le nouveau traitement en fonction de la présence du marqueur est réellement utile.

Ce design ne permettant pas de s'assurer que le traitement est sans bénéfice chez les patients marqueur (-), son principal **risque** associé est, comme pour le design ciblé, de conclure à tort qu'il ne faut pas traiter ces patients, donc de générer une **perte de chance** pour ces patients conjointement à la mise en œuvre d'un test inutile. Néanmoins, on peut noter qu'en terme populationnel, la stratégie basée sur le marqueur resterait la plus efficace. De plus, il est possible de modifier ce design en randomisant aussi les patients du groupe contrôle (stratégie sans évaluation du marqueur) entre le nouveau traitement et le traitement standard afin de vérifier que le nouveau traitement n'est pas plus efficace que le standard chez tous les patients (28).

Une autre source de risque de priver à tort du traitement des patients en utilisant ce design concerne les « faux négatifs » lorsque le test de détermination du marqueur manque de sensibilité. Un design d'interaction permet dans certains cas de détecter les conséquences de ce type de défaut du test par l'existence d'un bénéfice chez les patients marqueurs (-) (bénéfice induit par la proportion de patients réellement Mq+ et inclus dans cette strate en raison des résultats faussement négatifs). Ce n'est pas le cas avec un design de comparaison de stratégies. Il convient donc avec ce type de schéma d'être particulièrement vigilant sur la **sensibilité du test**.

Enfin, dans ce design, une partie des patients reçoivent le même traitement dans les deux bras. Cela dilue l'effet du traitement entre les deux bras car seule la fraction des patients marqueur (+) sera à même, si le traitement évalué est efficace, de faire baisser le nombre d'événements de la stratégie. Ce design est donc **moins performant dans l'absolu** que les designs d'interaction ou ciblés. Par contre ce type de design évalue directement la démarche qui sera utilisée en pratique si le couple marqueur/traitement est recommandé et documente donc parfaitement l'utilité clinique de la stratégie basée sur le marqueur (mais en aucun cas l'utilité clinique du marqueur).

### ► Illustrations numériques

Le Tableau 15 illustre numériquement ce type de design dans une situation où le marqueur est réellement prédictif. Par hypothèse, l'effet du traitement standard est le même pour tous les patients et caractérisé par un rapport des risques de 0.90. Le nouveau traitement est sans effet chez les patients marqueur (-) (rapport des risques de 1.00) et conduit à un rapport des risques de 0.60 chez les patients marqueur (+). La proportion de sujets marqueur (+) est de 40 %. À l'issue de l'essai, la comparaison de la stratégie d'individualisation basée sur le marqueur *versus* la stratégie standard (utilisation du traitement standard chez tous les patients) donne un rapport des risques de  $7.8 \% / 9.0 \% = 0.87$ , alors que l'utilisation du nouveau traitement chez tous les patients, y compris chez les patients marqueur (-) chez lesquels il est sans effet, aurait conduit à un moins bon résultat (fréquence d'événements de  $8.4 \%$  *versus*  $7.8 \%$ ). La stratégie d'individualisation conduit aussi à un meilleur résultat comparativement à l'utilisation du traitement standard ( $7.8 \%$  *versus*  $9.0 \%$ ) malgré l'absence d'effet du nouveau traitement chez les patients marqueur (-).

**Tableau 15 - Illustration numérique du schéma de comparaison de stratégies (dans une situation où le marqueur est bien prédictif).**

|  | Proportion de patients | Nombre de patients | Risque de base (sans traitement) | Traitement appliqué | Efficacité du traitement (risk ratio) | Risque sous traitement |
|--|------------------------|--------------------|----------------------------------|---------------------|---------------------------------------|------------------------|
| <i>Individualisation (Traitement en fonction du marqueur)</i>  |                        |                    |                                  |                     |                                       |                        |
| marqueur +   | 40 %                   | 400                | 10.0 %                           | N                   | 0.60                                  | 6.0 %                  |
| marqueur -   | 60 %                   | 600                | 10.0 %                           | Std                 | 0.90                                  | 9.0 %                  |
| <b>total</b>   | <b>100 %</b>           | <b>1 000</b>       | <b>10.0 %</b>                    |                     |                                       | <b>7.8 %</b>           |
| <i>Tous les patients traités avec le traitement standard</i>   |                        |                    |                                  |                     |                                       |                        |
| marqueur +   | 40 %                   | 400                | 10.0 %                           | Std                 | 0.90                                  | 9.0 %                  |
| marqueur -   | 60 %                   | 600                | 10.0 %                           | Std                 | 0.90                                  | 9.0 %                  |
| <b>total</b>   | <b>100 %</b>           | <b>1 000</b>       | <b>10.0 %</b>                    |                     |                                       | <b>9.0 %</b>           |
| <i>Tous les patients traités avec le nouveau traitement (N)</i>  |                        |                    |                                  |                     |                                       |                        |
| marqueur +   | 40 %                   | 400                | 10.0 %                           | N                   | 0.60                                  | 6.0 %                  |
| marqueur -   | 60 %                   | 600                | 10.0 %                           | N                   | 1.00                                  | 10.0 %                 |
| <b>total</b>   | <b>100 %</b>           | <b>1 000</b>       | <b>10.0 %</b>                    |                     |                                       | <b>8.4 %</b>           |
| Le nouveau traitement (N) a une efficacité caractérisée par un risk ratio de 0.60 chez les patients marqueur (+) et de 1.00 chez les patients marqueur (-). Le traitement standard (Std) a la même efficacité chez les patients Mq+ et Mq-, correspondant à un risk ratio de 0.90. |                        |                    |                                  |                     |                                       |                        |

Le Tableau 16 illustre une situation où le marqueur n'est pas prédictif de l'effet du nouveau traitement, celui-ci étant efficace de la même manière chez les Mq+ et les Mq- (rapport des risques de 0.6 dans les deux cas). Malgré cela, il apparaît que la stratégie d'individualisation obtient un meilleur résultat (8.1 % *versus* 9.0 %, RR=0.90). Cependant, l'utilisation du nouveau traitement chez tous les patients aurait conduit à un résultat encore meilleur (6.0 % vs 9.0 %, RR=0.67).

**Tableau 16 - Illustration numérique du schéma de comparaison de stratégies (dans une situation où le marqueur n'est pas prédictif).**

|   | Proportion de patients | Nombre de patients | Risque de base (sans traitement) | Traitement appliqué | Efficacité du traitement appliqué (risk ratio) | Risque sous traitement |
|---|------------------------|--------------------|----------------------------------|---------------------|--|------------------------|
| <b>Individualisation (Traitement en fonction du marqueur)</b>   |                        |                    |                                  |                     |  |                        |
| marqueur +  | 30 %                   | 300                | 10.0 %                           | N                   | 0.60   | 6.0 %                  |
| marqueur -  | 70 %                   | 700                | 10.0 %                           | Std                 | 0.90   | 9.0 %                  |
| <b>total</b>  | <b>100 %</b>           | <b>1 000</b>       | <b>10.0 %</b>                    |                     | <b>0.81</b>                                    | <b>8.1 %</b>           |
| <b>Tous les patients traités avec le traitement standard</b>    |                        |                    |                                  |                     |  |                        |
| marqueur +  | 30 %                   | 300                | 10.0 %                           | Std                 | 0.90   | 9.0 %                  |
| marqueur -  | 70 %                   | 700                | 10.0 %                           | Std                 | 0.90   | 9.0 %                  |
| <b>total</b>  | <b>100 %</b>           | <b>1 000</b>       | <b>10.0 %</b>                    |                     | <b>0.90</b>                                    | <b>9.0 %</b>           |
| <b>Tous les patients traités avec le nouveau traitement (N)</b> |                        |                    |                                  |                     |  |                        |
| marqueur +  | 30 %                   | 300                | 10.0 %                           | N                   | 0.60   | 6.0 %                  |
| marqueur -  | 70 %                   | 700                | 10.0 %                           | N                   | 0.60   | 6.0 %                  |
| <b>total</b>  | <b>100 %</b>           | <b>1 000</b>       | <b>10.0 %</b>                    |                     | <b>0.60</b>                                    | <b>6.0 %</b>           |

### ► Exemple extrait de la littérature

L'essai ARTIC (11) a évalué l'intérêt du clopidogrel, traitement antiagrégant plaquettaire, en fonction de la réactivité plaquettaire des patients bénéficiant de la pose d'un stent. Le marqueur était ici directement la réactivité plaquettaire sous traitement par clopidogrel ou aspirine mesurée au « pied du lit » (dispositif VerifyNow®). En cas d'activité plaquettaire résiduelle élevée, une intensification du traitement antiagrégant était effectuée dans le groupe de la stratégie d'ajustement. Aucune évaluation de la réactivité plaquettaire n'était effectuée dans le groupe de la stratégie contrôle et les patients étaient pris en charge à la discrétion des investigateurs, suivant les recommandations de pratiques en cours. Le critère de jugement était les événements cardiovasculaires intégrant les revascularisations et les thromboses de stent à un an. Aucun bénéfice de la stratégie basée sur le marqueur n'a été mis en évidence (avec un hazard ratio de 1.13 ; IC à 95 % compris entre 0.98 et 1.29 ; p=0.10).

Bien que les résultats de cet essai ne soient pas significatifs, il illustre bien cette approche basée sur l'évaluation des stratégies.

### 3.2.3 Analyses en sous-groupes *a posteriori*

#### ► Respect de la démarche hypothético-déductive

De plus en plus, il apparaît assez nettement que la majorité des études concluent au caractère prédictif d'un marqueur à partir de résultats de recherche de marqueurs réalisée *a posteriori* (recherche dite également *post-hoc* ou rétrospective). Cette démarche n'est pas méthodologiquement satisfaisante car, comme pour la démonstration de l'efficacité d'un traitement hors implication de marqueurs, la validation de la valeur prédictive d'un marqueur (ou de l'utilité clinique de son test diagnostique) doit respecter une démarche hypothético-déductive, c'est-à-dire s'effectuer de manière prospective (étude dite *ad-hoc*). La démonstration doit ainsi être obtenue dans le cadre d'un ou de plusieurs essais confirmatoires (« *con-*

*firmatory trial* » dans la terminologie ICH) spécialement mis en place pour tester l'hypothèse conjointe « valeur prédictive du marqueur/efficacité du traitement ».

Sans le respect de ce principe, les deux problèmes suivants surviennent :

- sans hypothèse, il n'est pas possible d'effectuer une double randomisation en fonction du marqueur (cf. section sur les schémas d'études) donc d'utiliser un design d'interaction marqueur/traitement. La démonstration de la valeur prédictive du marqueur et de l'utilité clinique du test associé ne peut donc pas être apportée ;
- seule la formulation *a priori* d'une hypothèse permet de s'assurer de l'absence de multiplicité statistique. Sans cette hypothèse préalable, rien ne garantit que les résultats présentés ne proviennent pas d'une recherche intensive parmi de nombreux marqueurs candidats, même si ces marqueurs avaient été listés *a priori*. Ce type d'approche, qui est celle des analyses en sous-groupes *a posteriori*, est en effet soumis à une inflation non contrôlée du risque alpha de fausse découverte. Elle s'apparente à une démarche purement exploratoire (de type « pêche à la ligne »), qui peut être utile voire nécessaire au repérage de marqueurs candidats d'intérêt, mais doit rester limitée à la génération d'hypothèses devant être impérativement validées par une (des) étude(s) *ad-hoc* de confirmation.

### ► Analyses en sous-groupes - Les limites

Les analyses exploratoires en sous-groupes ne sont pas adaptées à la recherche d'un marqueur prédictif en raison de leur nature exploratoire (non hypothético-déductive) et de l'inflation du risque alpha de 1<sup>ère</sup> espèce qu'elles génèrent.

Les limites des analyses en sous-groupes sont connues principalement en cas de recherche d'arguments pour conclure à l'efficacité du nouveau traitement lorsque l'essai n'est pas concluant par lui-même au niveau de la population incluse toute entière. Cette démarche qui consiste à rechercher un résultat statistiquement significatif en faveur de l'efficacité du nouveau traitement au niveau d'un sous-groupe, entraîne une **multiplicité des comparaisons** statistiques et une **inflation du risque alpha**, rendant ininterprétable le résultat. En multipliant le nombre de sous-groupes explorés, le risque alpha de 1<sup>ère</sup> espèce de conclure à tort à l'efficacité du traitement dans un sous-groupe particulier est augmenté sans maîtrise. La prise en compte de ce phénomène nécessite de **fixer au préalable** dans le protocole **la liste des sous-groupes** qui seront testés ainsi que de **prévoir une méthode de contrôle de l'inflation du risque alpha** telle que la méthode séquentielle hiérarchique ou celle de type Bonferroni. En pratique, ces précautions sont très rarement prises (29–33).

Dans la recherche des marqueurs prédictifs, ce n'est pas l'inflation non contrôlée du risque alpha de 1<sup>ère</sup> espèce mais plutôt celle du **risque bêta** de 2<sup>ème</sup> espèce qui pose problème. En effet, les analyses en sous-groupes augmentent de façon non contrôlée le risque bêta de 2<sup>ème</sup> espèce de ne pas mettre en évidence à tort l'efficacité d'un nouveau traitement. Ainsi, en présence d'un traitement efficace chez tous les patients, l'exploration de plusieurs sous-groupes finira toujours par trouver des résultats non statistiquement significatifs. Si ces derniers sont utilisés pour conclure à la découverte d'un marqueur prédictif, le risque de recommander à tort un marqueur en fait inutile devient bien supérieur au risque habituellement consenti (de 5 %).

Ce problème est particulièrement bien illustré par l'exemple proposé par Rothwell (31) pour mettre l'accent sur le côté hasardeux des conclusions issues des sous-groupes. La Figure 9 montre les résultats obtenus dans les sous-groupes créés en fonction du jour de naissance des patients inclus dans l'essai ECST. Cet essai démontre l'intérêt de l'endartériectomie carotidienne chez des patients porteurs d'une sténose serrée symptomatique. Ce bénéfice disparaît cependant pour les patients nés le dimanche, le mardi, le jeudi et le samedi. L'absence complète de plausibilité biologique montre bien que, par hasard, des résultats de sous-groupes peuvent entraîner des résultats sans réalité et conduire à des conclusions

erronées. Il est à noter que pour les sujets nés le dimanche, l'effet est proche de zéro ce qui serait le résultat recherché par excellence dans une recherche de marqueur prédictif. De plus, ce phénomène est amplifié par la réduction mathématique de puissance qui survient au niveau des sous-groupes qui comportent, par définition, toujours moins de sujets que l'essai global. Il devient ainsi très probable de ne pas avoir de signification statistique. En outre, ce moindre effectif conduit à une moindre précision statistique qui favorise les fluctuations aléatoires et l'obtention d'un hazard ratio très proche de 1 même en cas de traitement toujours efficace dans ce sous-groupe.

| Day of birth | Events/patients |               | ARR (%)     | 95% CI             | p value          |
|--------------|-----------------|---------------|-------------|--------------------|------------------|
|              | Surgical        | Medical       |             |                    |                  |
| Sunday       | 7/56            | 6/41          | 3.1         | -11.3 to 17.5      | 0.34             |
| Monday       | 4/66            | 10/44         | 16.7        | 3.0 to 30.3        | 0.008            |
| Tuesday      | 8/76            | 6/28          | 10.5        | -6.9 to 27.9       | 0.12             |
| Wednesday    | 8/67            | 13/47         | 18.3        | 2.3 to 34.2        | 0.01             |
| Thursday     | 9/75            | 9/36          | 12.8        | -3.8 to 29.4       | 0.07             |
| Friday       | 1/56            | 6/37          | 15.1        | 2.3 to 27.9        | 0.01             |
| Saturday     | 6/51            | 8/41          | 9.5         | -6.6 to 25.6       | 0.12             |
| <b>Total</b> | <b>43/447</b>   | <b>58/274</b> | <b>12.3</b> | <b>6.5 to 18.1</b> | <b>&lt;0.001</b> |

Heterogeneity: p=0.83

**Figure 9 - Exemple de résultats de sous-groupes montrant le risque accru d'erreur bêta de 2<sup>ème</sup> espèce lié aux analyses en sous-groupes : analyse de l'intérêt de l'endartériectomie carotidienne chez les patients inclus dans l'essai ECST présentant une sténose symptomatique  $\geq 70$  % en fonction de leur jour de naissance. (D'après (31))**

Dans ce premier exemple, l'absence d'interaction (hétérogénéité) préviendra les conclusions abusives. Cependant, le test d'interaction est lui aussi soumis à l'inflation du risque alpha en cas de répétition (comme tout test statistique). Ainsi, si de nombreuses explorations de sous-groupes sont conduites, des interactions statistiquement significatives finiront toujours par être observée à tort. La Figure 10 illustre cela à partir du même essai, en montrant des interactions obtenues en croisant la sévérité de la sténose et le mois de naissance. Là aussi, le manque total de plausibilité biologique des résultats observés met en évidence l'unique rôle du hasard dans la genèse de ces interactions.

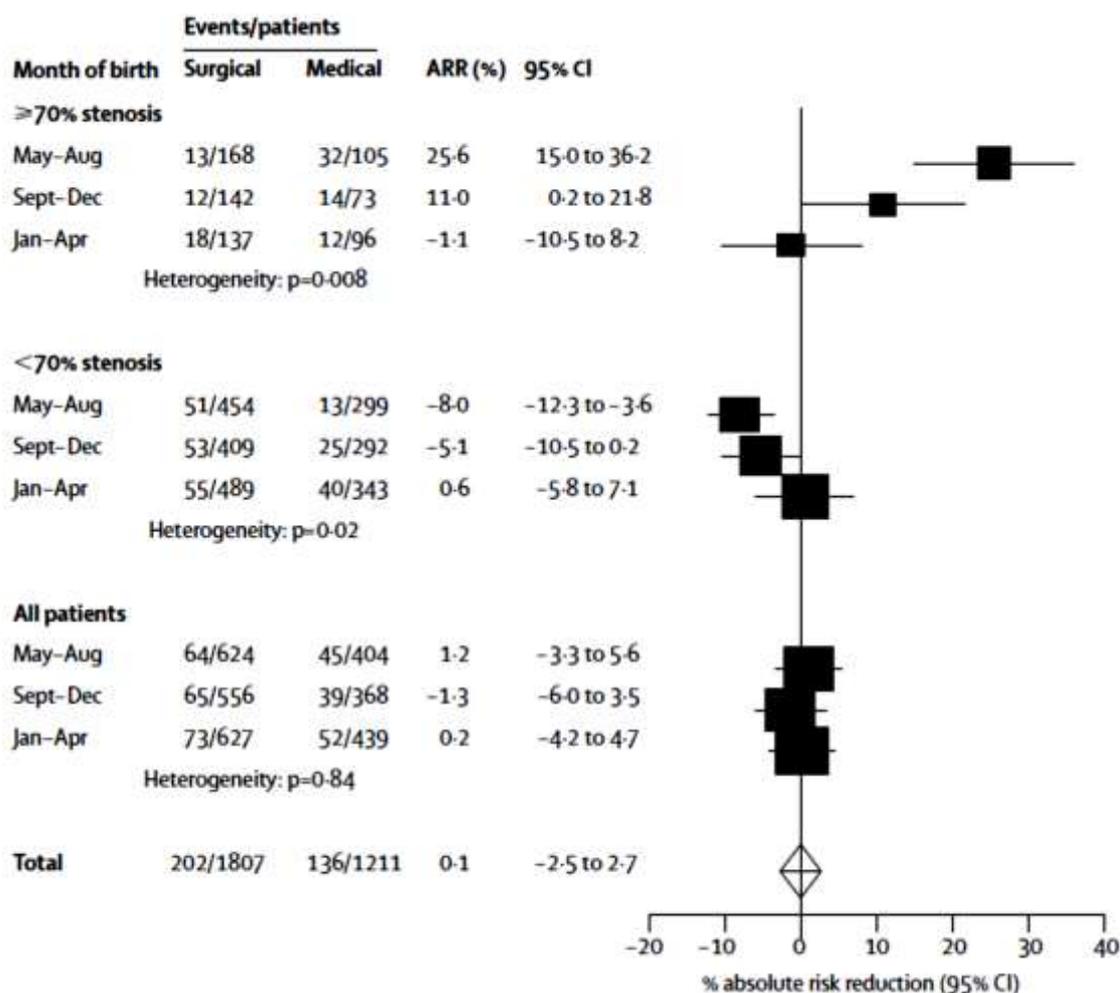


Figure 10 - Influence de la largeur des intervalles de confiance sur la signification statistique de l'interaction (à estimations ponctuelles constantes) (D'après (31))

Ces contre-exemples illustrent la dangerosité des analyses en sous-groupes répétées pour la recherche des marqueurs prédictifs. Pour éviter l'inflation des risques alpha et bêta, les analyses en sous-groupes doivent faire appel à des méthodes statistiques appropriées impliquant une planification *a priori* de ces analyses. Les marqueurs candidats ainsi que le sens de l'interaction doivent ainsi avoir été parfaitement identifiés dans le protocole même de l'essai. La publication antérieure à la date de complétion de l'essai est indispensable pour justifier du caractère prospectif de la recherche (et non pas *post-hoc* à la prise de connaissance des résultats donnés par l'essai).

#### ► Recherche *a posteriori* après un essai concluant

Cette première situation peut se produire lorsqu'après un essai montrant l'efficacité d'un nouveau traitement sur une population de patients non sélectionnés, une **recherche *post-hoc* de déterminants de la réponse** peut faire suspecter que le traitement ne serait en réalité efficace que chez certains patients porteurs d'un marqueur défini.

Cette situation présente les problèmes méthodologiques des démarches *post-hoc*, en particulier de multiplicité, avec une inflation du risque bêta de ne pas mettre en évidence à tort l'efficacité du traitement chez les patients marqueur (-), et donc de récuser à tort le traitement de ces patients initialement inclus dans l'essai ayant démontré l'efficacité du traitement

sans sélection des patients en fonction de leur statut pour le marqueur. Un essai *ad-hoc* est donc indispensable pour confirmer ces résultats.

Une **problématique éthique** apparaît néanmoins dans ce cas. En effet, par l'essai initial réalisé selon un schéma standard d'étude de thérapeutique non stratifiée et concluant, le nouveau traitement est devenu le traitement de référence dans le domaine concerné. Randomiser à nouveau des patients entre le nouveau traitement et un placebo (ou l'ancien traitement de référence) pose alors un problème éthique. Un schéma de type « interaction marqueur/traitement » ne peut donc pas être recommandé, en particulier à l'égard des patients marqueur (+) qu'il n'est pas raisonnable d'envisager de randomiser à nouveau *versus* le comparateur initial alors que l'étude de confirmation aurait pour objectif de démontrer que le traitement n'est en fait efficace que chez eux. Au final, la recherche d'un marqueur prédictif à la suite d'un essai concluant en population générale aboutit le plus souvent à des résultats exploratoires non confirmés prospectivement donc de niveau de preuve non acceptable pour démontrer l'utilité clinique du test diagnostique permettant d'identifier le marqueur.

Néanmoins, des situations peuvent se présenter où un **rational fort** peut apparaître dans le sens d'une probable absence d'efficacité chez les patients non porteurs d'un marqueur, justifiant ainsi éthiquement la conduite d'une étude prospective ayant pour objectif de confirmer (ou non) l'absence présumée d'efficacité du traitement chez ces patients. Dans ce cas, la démarche à suivre est celle d'une étude prospective de non supériorité, de type ciblée incluant uniquement des patients marqueur (-), ou bien de type « comparaison de stratégies » avec un bras « stratégie standard » correspondant à l'utilisation du nouveau traitement sans identification du marqueur. Une excellente qualité métrologique d'essai est alors impérative, le risque majeur étant de conclure à tort à l'exclusion du traitement de la sous-population des patients marqueur (-). Si l'absence d'intérêt clinique du traitement chez les patients marqueur (-) est confirmée, il est possible de conclure à l'utilité clinique du test, ce dernier apportant, conformément à la définition de l'utilité clinique, une optimisation de la prise en charge des patients, en permettant en particulier d'éviter l'exposition inutile de patients à des effets indésirables potentiels sans bénéfices pour leur santé et de recourir, s'il en existe, à d'autres thérapeutiques potentiellement plus efficaces.

### ► Recherche *a posteriori* après un essai non concluant

La recherche d'un marqueur prédictif après l'obtention de résultats non concluants dans un essai évaluant un nouveau traitement pose d'importants problèmes méthodologiques. Elle expose en particulier à un **fort risque de découverte artéfactuelle de marqueurs non prédictifs**, pouvant amener à priver à tort du traitement des patients susceptibles d'en recevoir des bénéfices et d'alourdir la prise en charge de tous les patients avec la réalisation d'un test inutile.

En effet, dans le contexte d'un essai non concluant, la recherche d'un marqueur prédictif résulte généralement directement de l'obtention de ce résultat non satisfaisant, qui amène à proposer l'hypothèse d'une efficacité non ubiquitaire mais limitée à un certain type de patients. Cette recherche est alors entreprise par une analyse rétrospective des données de l'essai non initialement prévue. Cependant, ces mêmes données seront ensuite également utilisées comme démonstration de la valeur prédictive de ce marqueur, créant ainsi une situation tautologique. En outre, faisant suite à l'obtention de résultats non concluants, une démarche *post-hoc* de type « pêche à la ligne » est utilisée et génère une multiplicité de comparaisons à l'origine d'une inflation du risque alpha sur l'hypothèse d'efficacité du traitement, revenant à s'octroyer une deuxième chance de conclure, à tort, à cette efficacité.

Par conséquent, comme pour toute analyse en sous-groupes, les résultats obtenus dans cette situation doivent être considérés comme exploratoires et impérativement validés par une étude de **confirmation prospective**. En l'occurrence, la démonstration du caractère

prédicatif d'un marqueur présumé prédictif par les analyses exploratoires requiert l'utilisation d'un **schéma de type « interaction marqueur/traitement »**.

Il faut souligner que la démonstration isolée de l'efficacité du nouveau traitement chez les patients marqueur (+) avec un autre schéma d'étude ne permettrait pas d'exclure qu'un premier essai conduit avec un nouveau traitement en réalité efficace chez tous les patients ait pu s'avérer « faussement » non concluant en raison d'un manque de puissance, puis qu'un second essai alors réalisé uniquement chez les patients porteurs d'un marqueur identifié dans les analyses en sous-groupes du premier essai ait pu ensuite être concluant uniquement en raison d'une meilleure puissance, que le marqueur soit prédictif ou non.

#### ► Valeur pronostique dans le groupe traité (« *treatment-only analysis* »)

Il n'est pas rare de rencontrer des travaux où la recherche de marqueurs prédictifs s'apparente à une exploration des facteurs de risque de l'événement correspondant au critère de jugement dans le groupe recevant le nouveau traitement au sein d'un essai thérapeutique. Cette démarche n'incluant pas les patients du groupe contrôle est appelée « *treatment-only analysis* » et revient en réalité à chercher un facteur de risque d'événement et non à mettre en évidence une modification de l'effet du traitement (24). Elle ne permet donc en aucun cas la mise en évidence de marqueurs prédictifs et expose au risque de fausse identification d'un effet thérapeutique spécifique dans une sous-population de patients marqueur (+).

#### Exemple : variantes génétiques des cytochromes P-450 et effet du clopidogrel

Pour rechercher les facteurs susceptibles de modifier les effets du clopidogrel parmi les variantes génétiques des cytochromes P450 ayant un retentissement fonctionnel, Mega et coll. ont utilisé les patients du groupe traité par le clopidogrel de l'essai TRITON TIMI 38 (34). À l'origine, l'objectif de cet essai était de comparer le ticagrelor au clopidogrel dans la prise en charge des syndromes coronariens aigus. Le critère de jugement principal était le critère composite « événements coronariens ». Pour cette recherche de facteurs, les patients du groupe clopidogrel ont été génotypés afin d'identifier les porteurs de variants génétiques de cytochromes P450 connus chez le volontaire sain comme pouvant altérer la transformation du clopidogrel en son composé actif.

L'analyse a consisté à comparer la fréquence des événements cardiovasculaires entre les patients porteurs ou non de variants génétiques de cytochromes P450 entraînant une altération de la biotransformation du clopidogrel en son composé actif (cf. Figure 11).

La conclusion des auteurs a été que les formes génétiques variantes déficientes du CYP2C19 seraient susceptibles d'être un marqueur prédictif de la réponse au clopidogrel.

| Gene                                | Carriers of Reduced-Function Allele | Noncarriers of Reduced-Function Allele | Hazard Ratio (95% CI) | P Value |
|-------------------------------------|-------------------------------------|--|-----------------------|---------|
|                                     | no./total no. (%)                   |  |                       |         |
| Composite primary efficacy outcome† |                                     |  |                       |         |
| CYP2C19                             | 46/395 (12.1)                       | 83/1064 (8.0)                          | 1.53 (1.07–2.19)      | 0.01    |
| CYP2C9                              | 22/230 (10.0)                       | 107/1226 (9.0)                         | 1.09 (0.69–1.73)      | 0.41    |
| CYP2B6                              | 36/370 (10.0)                       | 68/777 (9.0)                           | 1.11 (0.74–1.67)      | 0.78    |
| CYP3A5                              | 95/1130 (8.7)                       | 14/151 (9.5)                           | 0.89 (0.51–1.57)      | 0.69    |
| CYP1A2                              | 5/59 (8.5)                          | 95/1099 (8.9)                          | 0.97 (0.40–2.39)      | 0.96    |

**Figure 11 - Fréquence de survenue d'évènements cardiovasculaires chez des porteurs *versus* non porteurs de variants génétiques de cytochromes P450 chez des patients recevant du clopidogrel. (D'après (34))**

Cette approche est impropre à la mise en évidence de marqueurs prédictifs pour plusieurs raisons :

- l'utilisation limitée à des patients traités ne permet pas d'appréhender l'effet du traitement en raison de l'existence de **facteurs de confusion**. Seule la comparaison à un groupe contrôle subissant les mêmes effets des facteurs de confusion est en mesure d'identifier l'effet du traitement ;
- le facteur identifié peut très bien être un facteur pronostique réel, qui sous et sans traitement est associé à une moindre fréquence des évènements considérés. Dans ce cas, retrouver cette relation ne signifie pas que le facteur en question modifie l'effet du traitement<sup>6</sup> ;
- au-delà des limitations du point précédent, interpréter un facteur pronostic de la survenue de l'évènement « critère de jugement » sous traitement comme facteur modifiant l'effet du traitement fait l'hypothèse forte, qu'après la prise en compte de tous ses facteurs modificateurs, le traitement évite en totalité la survenue des évènements.

Ainsi, il serait possible de trouver par cette approche des marqueurs prédictifs à un traitement sans effet (avec lequel aucune différence significative n'a été mise en évidence entre le groupe traité et le groupe contrôle). Tous les marqueurs pronostics ressortiront car étant prédictifs de la survenue des évènements.

Pour toutes ces raisons, cette approche est inappropriée pour rechercher des marqueurs prédictifs et expose à de fausses découvertes comme l'illustre bien l'histoire du clopidogrel (cf. Encadré ci-dessous).

<sup>6</sup> Ce point n'est qu'une variante du point précédent et conduit à un biais de confusion.

### Encadré - Histoire du clopidogrel

De nombreuses études similaires à celle de Mega et coll. (cf. exemple précédent) ont été réalisées. Une méta-analyse de ces études (35) a conclu qu'étant donnée l'association entre ce marqueur et les évènements, celui-ci devrait être utilisé pour l'individualisation du traitement : « *In conclusion, the findings of this collaborative meta-analysis demonstrate that common genetic variants in the CYP2C19 gene are associated with almost 1 in 3 patients not receiving ideal protection from ischemic events when treated with standard doses of clopidogrel for PCI. Given how widely clopidogrel is used to treat patients with cardiovascular disease, determination of the optimal antiplatelet treatment doses or regimens for individual patients is needed to tailor therapy appropriately* ».

Cependant, dans une autre méta-analyse parue également dans le JAMA, Holmes *et al.* (36) abordent cette question du point de vue de l'interaction marqueur/traitement. Pour les mêmes variants génétiques précédemment identifiés dans les travaux de valeur pronostique (qu'ils dénomment « *treatment-only analysis* »), ils ont recherché des données d'interaction à partir des essais randomisés *versus* placebo (« *effect-modification analysis* »). Aucune interaction n'a alors été confirmée, montrant ainsi les limites de l'analyse « *treatment-only* ».

### 3.3 Synthèse

Le Tableau 17 présente une synthèse des types d'essais appropriés en fonction du contexte de l'évaluation d'un marqueur prédictif candidat.

**Tableau 17 - Essais appropriés en fonction du contexte de l'évaluation d'un marqueur prédictif candidat.**

| Évaluation de la valeur prédictive d'un marqueur candidat ...   | Schéma d'étude approprié  |
|---|---|
| ... simultanément à l'évaluation de l'efficacité du traitement  | Essai prospectif d'interaction  |
| ... après un premier essai non concluant évaluant l'efficacité du traitement  | Essai prospectif d'interaction  |
| ... après un premier essai concluant évaluant l'efficacité du traitement et un rationnel fort en faveur d'une absence d'efficacité chez les patients marqueur (-) | Essai prospectif de non supériorité de type ciblé ou de comparaison de stratégies |

## 4. Méthodologie d'évaluation par la HAS de l'utilité clinique d'un test diagnostique associé à une thérapie stratifiée - Exemples illustratifs

### 4.1 Considérations générales

Ce chapitre propose les principes de conduite par la Haute Autorité de santé de l'évaluation d'un test diagnostique permettant d'identifier un marqueur présumé prédictif associé à une thérapie stratifiée. Une synthèse pragmatique de ces principes est présentée dans le guide méthodologique intitulé « Test compagnon associé à une thérapie ciblée : définitions et méthode d'évaluation ».

Le principe de base de l'évaluation de ce type de test par la HAS réside dans le fait que la reconnaissance des appellations « test compagnon » et « thérapie ciblée » repose sur la démonstration, apportée ou non, de l'utilité clinique du test. Par conséquent, sont présentées dans ce chapitre les **preuves requises pour la démonstration d'utilité clinique d'un test diagnostique associé à une thérapeutique stratifiée**. Cette démonstration prouve conjointement la valeur prédictive du marqueur vis-à-vis de l'efficacité du traitement, le ciblage de la thérapie et l'appellation de test compagnon.

L'utilité clinique d'un test diagnostique est parfois revendiquée à partir d'arguments ou de données issues d'études qui ne permettent pas de la démontrer mais qui sont plus facilement obtenues que des **éléments de preuve** rigoureux. En l'occurrence, la reconnaissance ou non de cette revendication dépend directement du **schéma utilisé pour l'étude supportant cette demande**.

Il est de plus important de souligner qu'un **cadre méthodologique rigoureux** répondant aux standards actuels de l'évaluation des thérapeutiques est nécessaire compte tenu de l'implication directe des tests compagnons dans la décision de traiter ou non des patients (cf. Encadré ci-dessous).

Pouvoir garantir l'utilité clinique d'un test compagnon et concomitamment le ciblage de la thérapie constitue un **enjeu de santé** important car les risques en l'absence d'évaluation sont nombreux :

- le risque de recommander à tort un traitement en réalité sans efficacité (résultat biaisé ou erreur de première espèce/risque alpha) ;
- le risque de recommander à tort un traitement en réalité sans intérêt clinique (absence de pertinence clinique).

Ces deux premiers risques sont identiques à ceux que l'on cherche à exclure dans l'analyse d'un dossier classique de thérapeutique.

- Le risque de recommander à tort une stratégie stratifiée impliquant la réalisation chez tous les patients d'un test diagnostique supplémentaire inutile.
- Le risque de ne pas utiliser à tort chez les patients marqueur (-) un traitement qui pourrait avoir en réalité un certain intérêt clinique chez eux, résultant en une perte de chance pour ces patients.

*In fine*, ne pas s'assurer de l'utilité clinique d'un test identifiant les sous-populations concernées par un traitement stratifié peut conduire à recommander une stratégie discriminante d'utilisation d'un nouveau traitement et non optimale en termes de bénéfice populationnel.

**Encadré - Conditions requises pour la démonstration rigoureuse de l'efficacité d'un traitement**

**1) Validité interne**

- a) Absence de biais
  - i) Essai à l'abri d'un biais de sélection : randomisation imprévisible
  - ii) Essai à l'abri d'un biais de réalisation
  - iii) Essai à l'abri d'un biais de mesure
  - iv) Essai à l'abri d'un biais d'attrition
  
- b) Réalité statistique de la démonstration d'efficacité : résultat statistiquement significatif avec maîtrise du risque alpha
  - i) Gestion correcte de la multiplicité des comparaisons
  - ii) Critère de jugement principal et, si multicritères, méthode de contrôle du risque alpha définis *a priori*
  - iii) Pas de sous-groupes
  - iv) Pas d'inflation introduite par les analyses intermédiaires
  
- c) Validité hypothético-déductive

**2) Pertinence clinique**

- a) Pertinence clinique du comparateur
- b) Pertinence clinique du critère de jugement
- c) Balance bénéfice/risque acceptable

**3) Cohérence externe**

- a) Plausibilité biologique
- b) Cohérence entre les résultats des différentes études disponibles

## 4.2 Tableaux de synthèse

Le Tableau 18 liste les principales réserves méthodologiques susceptibles d'invalider la démonstration, et donc la revendication, de l'utilité clinique d'un test diagnostique.

**Tableau 18 - Réserves méthodologiques pouvant invalider la revendication d'utilité clinique d'un test diagnostique.**

| Critère   | Situation observée  | Réserve méthodologique  | Justification de la réserve méthodologique   |
|---|---|---|--|
| Démarche hypothético-déductive pour la démonstration d'interaction  | Utilisation de résultats <i>post-hoc</i> pour suggérer l'interaction  | Les résultats utilisés pour démontrer l'interaction doivent provenir d'un essai spécifiquement conçu pour tester cette hypothèse.   | Si l'essai a été initialement prévu pour un autre objectif, l'analyse rétrospective ( <i>post-hoc</i> ) des résultats qu'il produit ne permet pas de démontrer l'interaction, ni de revendiquer l'utilité clinique du test diagnostique, mais seulement de générer des données à caractère exploratoire.   |
| Démarche hypothético-déductive pour la démonstration d'efficacité du traitement chez les patients marqueur (+)      | Utilisation de résultats <i>post-hoc</i> pour suggérer l'efficacité du traitement chez les patients marqueur (+)      | Les résultats utilisés pour démontrer l'efficacité du traitement chez les patients marqueur (+) doivent provenir d'un essai spécifiquement conçu pour tester cette hypothèse. | Si l'essai a été initialement prévu pour un autre objectif, l'analyse rétrospective ( <i>post-hoc</i> ) des résultats qu'il produit ne permet pas de démontrer l'efficacité du traitement chez les patients marqueur (+), ni de revendiquer l'utilité clinique du test diagnostique, mais seulement de générer des données à caractère exploratoire. |
| Démarche hypothético-déductive pour la démonstration d'absence d'effet du traitement chez les patients marqueur (-) | Utilisation de résultats <i>post-hoc</i> pour suggérer l'absence d'effet du traitement chez les patients marqueur (-) | Les résultats utilisés pour démontrer l'absence d'effet chez les patients marqueur (-) doivent provenir d'un essai spécifiquement conçu pour tester cette hypothèse.          | Si l'essai a été initialement prévu pour un autre objectif, l'analyse rétrospective ( <i>post-hoc</i> ) des résultats qu'il produit ne permet pas de démontrer l'absence d'effet chez les patients marqueur (-), ni de revendiquer l'utilité clinique du test diagnostique, mais seulement de générer des données à caractère exploratoire.          |
| Absence de multiplicité au niveau de la recherche du marqueur   | Existence d'une multiplicité au niveau de la recherche du marqueur  | La vérification statistique des trois hypothèses précitées (lignes ci-dessus) doit être issue d'un seul et unique test statistique prévu <i>a priori</i> .                    | La recherche du marqueur a été conduite dans un contexte exploratoire où plusieurs / de nombreux marqueurs candidats ont été envisagés successivement (démarche de type « pêche à la ligne » dans des sous-groupes par exemple) générant une multiplicité des tests statistiques associée à une inflation du risque alpha.                           |
| Absence de multiplicité à d'autres niveaux  | Existence d'une multiplicité à d'autres niveaux   | Le marqueur, le critère de jugement, la population d'analyse, etc. doivent avoir été définis <i>a priori</i> .  | Il existe une multiplicité au niveau des critères de jugement, des populations d'analyse, des analyses intermédiaires, au cours du temps.  |

| Critère          | Situation observée                       | Réserve méthodologique   | Justification de la réserve méthodologique   |
|------------------|--|--|--|
| Absence de biais | Protection insuffisante contre les biais | Le cadre méthodologique de l'essai doit respecter les standards rigoureux de tout essai thérapeutique pivotal. | La recherche des effets du traitement chez les patients marqueur (-) ou (+) a été réalisée au sein d'un essai insuffisamment protégé contre les biais : essai non randomisé ; essai non en double aveugle ; essai non analysé en intention de traiter avec remplacement des données manquantes, etc. |

**Tableau 19 - Critères impérativement requis pour la revendication d'utilité clinique d'un test diagnostique.**

| Critère   | Situation observée  | Justification de la réserve méthodologique  |
|---|---|---|
| Démarche hypothético-déductive pour la démonstration d'interaction  | Les résultats utilisés pour démontrer l'interaction doivent provenir d'un essai spécialement conçu pour tester cette hypothèse.   | Si l'essai a été initialement prévu pour un autre objectif, l'analyse rétrospective ( <i>post-hoc</i> ) des résultats qu'il produit ne permet pas de démontrer l'interaction, mais seulement de générer des données à caractère exploratoire insuffisantes pour soutenir la revendication d'utilité clinique du test. |
| Démarche hypothético-déductive pour la démonstration d'efficacité du traitement chez les patients marqueur (+)      | Les résultats utilisés pour démontrer l'efficacité du traitement chez les patients marqueur (+) doivent provenir d'un essai spécialement conçu pour tester cette hypothèse. |   |
| Démarche hypothético-déductive pour la démonstration d'absence d'effet du traitement chez les patients marqueur (-) | Les résultats utilisés pour démontrer l'absence d'effet chez les patients marqueur (-) doivent provenir d'un essai spécialement conçu pour tester cette hypothèse.          |   |
| Absence de multiplicité au niveau de la recherche du marqueur   | La vérification statistique des trois hypothèses précitées (lignes ci-dessus) doit être issue d'un seul et unique test statistique prévu <i>a priori</i> .                  |   |
| Absence de multiplicité à d'autres niveaux  | Le marqueur, le critère de jugement, la population d'analyse, etc. doivent avoir été définis <i>a priori</i> .  |   |
| Absence de biais  | Le cadre méthodologique de l'essai doit respecter les standards rigoureux de tout essai thérapeutique pivotal.  | Essai non randomisé ; essai non en double aveugle ; essai non analysé en intention de traiter avec remplacement des données manquantes, etc.  |

Le Tableau 20 résume très succinctement les démonstrations à apporter pour revendiquer l'utilité clinique du test dans les principaux contextes de développement de couples traitement/test diagnostique. La première situation correspond à un développement conjoint vertueux, tandis que les

deux autres peuvent n'avoir comme seul objectif que de valoriser un développement de traitement (pour le 2<sup>e</sup>), ou de marqueur et test diagnostique associé (pour le 3<sup>e</sup>).

**Tableau 20 - Démonstrations à apporter pour prouver l'utilité clinique du test dans les différents contextes de développement d'un couple traitement/test diagnostique.**

| Contexte de développement du couple traitement/test diagnostique  | Approche   |
|---|--|
| Développement conjoint d'emblée d'un couple nouveau traitement/test diagnostique.   | Ce type de développement doit faire appel à un design d'interaction pour apporter conjointement la démonstration de l'efficacité du traitement et de l'utilité clinique du test diagnostique.  |
| Recherche <i>a posteriori</i> de l'intérêt d'un nouveau traitement dans une sous-population particulière en essayant successivement plusieurs marqueurs après un premier essai non concluant sur une population de patients non sélectionnée par un marqueur. | Le risque de développement opportuniste d'un nouveau traitement est réel. Cette situation pose plusieurs problèmes : absence de démarche hypothético-déductive, multiplicité.<br><br>La démonstration de l'interaction doit donc être formellement apportée.   |
| Recherche <i>a posteriori</i> d'un marqueur avec pour justification d'optimiser l'utilisation d'un traitement initialement démontré efficace sur une population de patients non sélectionnée par un marqueur.   | Le risque de développement opportuniste d'un marqueur est réel et associé à un risque de perte de chance pour les patients non porteurs du marqueur qui pourraient se voir à tort exclus du traitement initialement montré efficace chez eux par le premier essai.<br><br>La démonstration d'absence d'effet du traitement chez les patients marqueur (-) doit donc être formellement apportée si l'optimisation d'utilisation du traitement consiste à ne pas les traiter sans bénéfice escompté. |

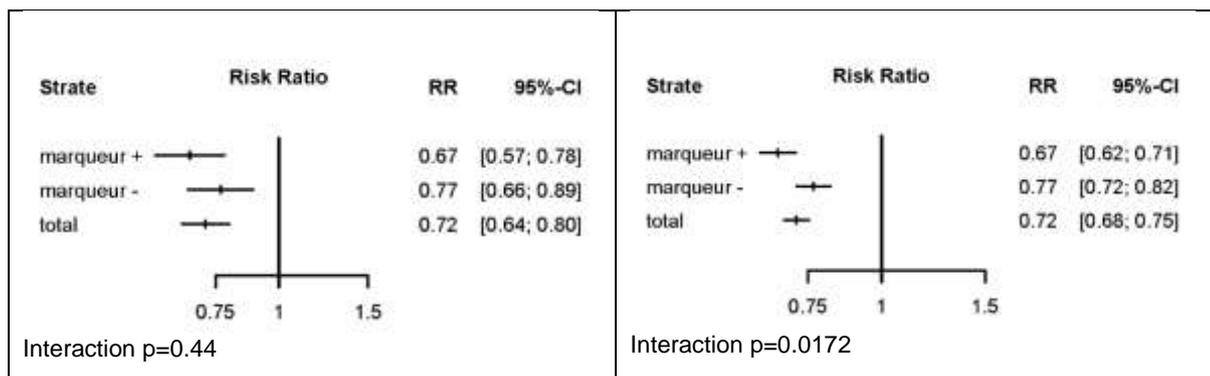
## 4.3 Exemples théoriques d'application des principes énoncés

### 4.3.1 Différentes situations de cas d'interaction

Cette section illustre les différentes situations auxquelles conduisent l'interprétation de l'interaction statistique entre les deux estimations de l'effet du traitement chez les patients marqueur (+) et (-).

#### ► Influence de la largeur des intervalles de confiance sur la signification statistique

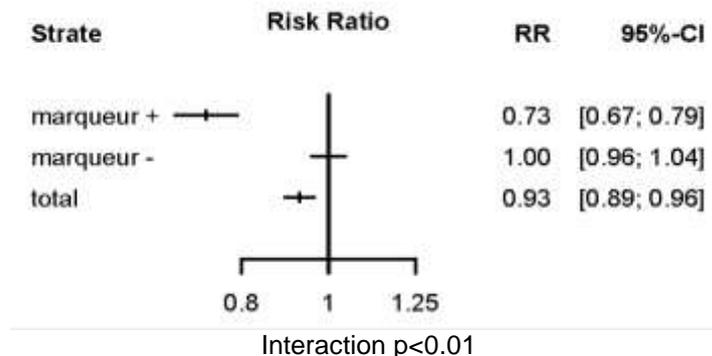
La Figure 12 montre l'influence de la largeur des intervalles de confiance sur la signification statistique de l'interaction à estimations ponctuelles constantes. Il est visible que plus les intervalles sont étroits, plus l'interaction est statistiquement significative. La raison en est que plus on connaît avec précision les effets respectifs du traitement dans les deux strates, plus il est simple de les discriminer, les intervalles de confiance ayant de moins en moins de valeurs en commun. En fait, il y a interaction statistiquement significative à partir du moment où l'intervalle de confiance de l'un n'inclut plus l'estimation ponctuelle de l'autre.



**Figure 12 - Influence de la largeur des intervalles de confiance sur la signification statistique de l'interaction (à estimations ponctuelles constantes)**

### ► Cas idéal en faveur de la valeur prédictive du marqueur

La situation représentée sur la Figure 13 est la situation idéale. Il existe une interaction importante (compte tenu de la séparation des deux estimations ponctuelles des risk ratios) et hautement significative. Le traitement est efficace chez les patients marqueur (+). Chez les patients marqueur (-), l'estimation ponctuelle du risk ratio est de 1 et la borne inférieure de l'intervalle de confiance est assez proche de 1 (0.96) permettant, compte tenu de la situation clinique de ce traitement (non décrite) de conclure à la non supériorité. Cette situation est donc la situation idéale statistiquement pour conclure favorablement à la démonstration de valeur prédictive du marqueur.



**Figure 13 - Existence d'une interaction très en faveur d'un marqueur prédictif**

Il est à noter que globalement sur l'ensemble des patients de l'essai (« total »), un effet significatif est observé, la forte efficacité obtenue chez les patients marqueur (+) n'étant pas totalement diluée par l'absence d'effet chez les autres patients. Cela conduit à s'interroger sur la valeur décisionnelle de ce « total ».

### ► Interaction qualitative

La Figure 14 illustre une situation d'interaction qualitative étant donné que le traitement est significativement bénéfique dans une strate et significativement délétère dans l'autre. Au total, il n'y a pas d'effet car les effets dans les deux strates s'annulent en moyenne.

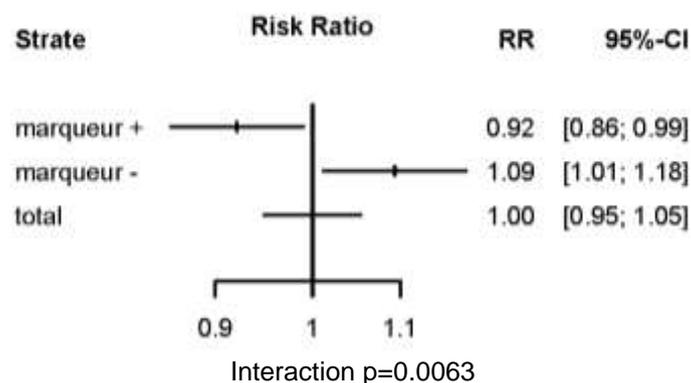


Figure 14 - Interaction qualitative

### ► Existence d'une interaction sans que le marqueur soit prédictif

Dans la situation illustrée par la Figure 15, il existe une interaction significative mais le traitement est efficace dans les deux strates. Il est simplement moins efficace chez les patients marqueur (-) que chez les patients marqueur (+). Cette situation ne permet pas de conclure que le marqueur en question est prédictif.

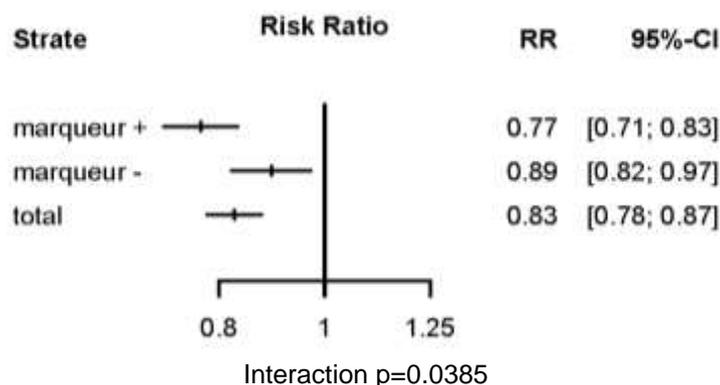
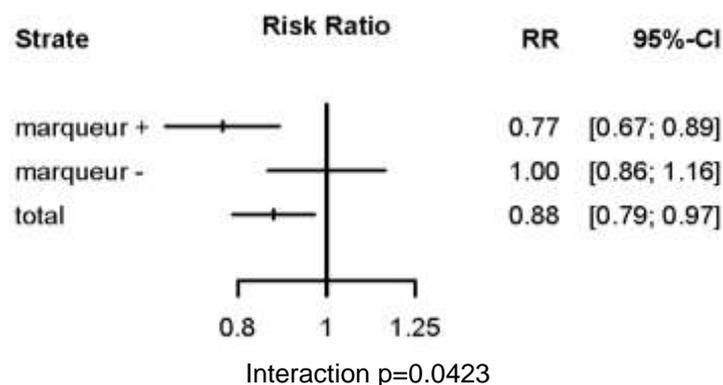


Figure 15 - Existence d'une interaction statistique avec un traitement efficace dans les deux strates marqueurs (+) et (-).

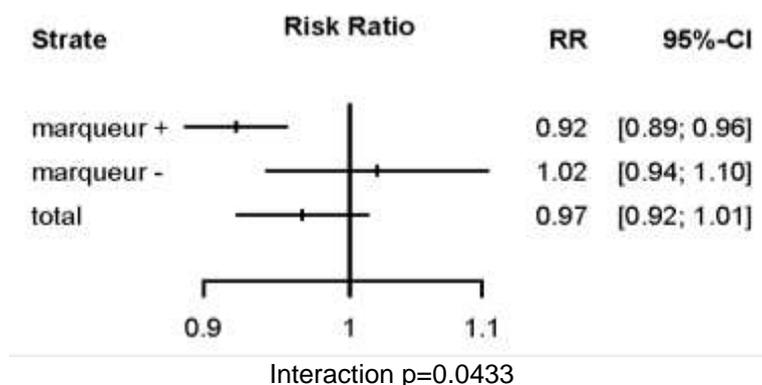
### ► Situation où il est difficile de conclure à la valeur prédictive du marqueur

La Figure 16 illustre la situation d'interprétation la plus difficile. Le traitement est démontré efficace chez les patients marqueur (+). Chez les patients marqueur (-), le risk ratio est de 1, en faveur de l'absence d'effet. Mais compte tenu de la précision de cette estimation, il va être difficile de conclure à la non supériorité et de considérer qu'il y a démonstration de l'absence d'effet du traitement chez les patients marqueur (-). En effet, la borne inférieure de l'intervalle de confiance est de 0.86. Peut-elle être considérée en soi comme cliniquement négligeable et conclure ainsi à la non supériorité ? Si le réel effet du traitement chez les patients marqueur (-) est de 0.86, il pourrait représenter un bénéfice non négligeable. Priver ces patients de ce traitement pourrait conduire à une perte de chance pour eux. De plus, au total, patients marqueur (+) et (-) confondus, le traitement est statistiquement efficace. Est-il alors possible, en présence d'une démonstration de l'efficacité du traitement tous types de patients confondus, d'exclure les patients marqueur (-) du traitement ?



**Figure 16 - Situation où il est difficile de conclure à la valeur prédictive du marqueur.**

La situation illustrée par la Figure 17 pose le même type de difficulté. Il existe une interaction significative et la borne inférieure de l'intervalle de confiance des patients marqueur (-) est compatible avec un bénéfice cliniquement intéressant. Il n'est donc pas possible de conclure que le traitement n'est efficace que chez les patients marqueur (+), ni qu'il y a démonstration de la valeur prédictive, du fait de l'impossibilité de conclure à la non supériorité chez les patients marqueur (-). De plus, il n'est pas non plus possible de conclure que le traitement est efficace chez tous les patients. Cet essai n'est donc pas concluant car il ne répond positivement qu'à la moitié de son objectif (montrer que le traitement est efficace chez les patients marqueur (+) et non efficace chez les patients marqueur (-)).



**Figure 17 - Autre situation posant problème pour conclure à la valeur prédictive du marqueur.**

#### 4.3.2 Exemples de cas théoriques d'évaluation de dossiers

Ces études de cas ont été composées avec pour objectif de clarifier les principes exposés. Le cas échéant, elles renvoient vers des situations réelles s'il a été possible d'en identifier.

##### ► Cas 1 - Design d'interaction montrant une interaction qualitative

Un essai de recherche d'interaction montre l'efficacité du traitement dans la strate des patients marqueur (+) avec un hazard ratio de 0.67 [0.47 ; 0.87]. Dans la strate des patients marqueur (-), les résultats montrent un effet délétère statistiquement significatif avec HR=1.32 [1.01 ; 1.61]. L'interaction est très significative avec  $p<0.01$ .

Cette interaction qualitative témoignant d'un effet délétère chez les patients Mq- **démontre l'utilité clinique du marqueur**, le bénéfice du traitement étant par ailleurs démontré chez les patients Mq+.

Cependant, la potentialité du traitement d'avoir un effet délétère, comme le montre le résultat obtenu dans la strate Mq-, doit générer une réelle **suspicion** vis-à-vis du traitement. Il convient de bien comprendre les mécanismes entraînant cet effet délétère afin d'être sûr que le traitement présente une balance bénéfique/risque favorable chez les patients marqueur (+).

### ► Cas 2 - Design d'interaction satisfaisant

Un essai de recherche d'interaction aboutit aux résultats suivants :

| Strate     | Hazard ratio (IC 95 %) |                    |
|------------|------------------------|--------------------|
| Marqueur + | 0.68 [0.38 ; 0.88]     |                    |
| Marqueur - | 1.01 [0.91 ; 1.11]     | Interaction p=0.02 |

L'utilité clinique de ce marqueur est directement démontrée par ces résultats grâce à la mise en évidence d'une interaction statistiquement significative, associée à une démonstration de l'intérêt du traitement (aucune réserve méthodologique n'étant à apporter à l'essai par ailleurs). Le résultat chez les patients marqueur (-) permet de conclure à la non supériorité si dans le domaine de l'essai une réduction de moins de 10 % de ce critère de jugement peut être considérée comme étant un bénéfice insuffisant cliniquement.

### ► Cas 3 - Design d'interaction mais démonstration de non supériorité non acceptable

Un essai de recherche d'interaction aboutit aux résultats suivants :

| Strate     | Hazard ratio (IC 95 %) |                     |
|------------|------------------------|---------------------|
| Marqueur + | 0.68 [0.38 ; 0.88]     |                     |
| Marqueur - | 0.98 [0.72 ; 1.38]     | Interaction p=0.036 |

Il est connu par ailleurs que le comparateur de cet essai réduit le même critère de jugement de 15 % (HR=0.85, IC 95 % entre 0.70 et 0.98).

Ces résultats mettent en évidence l'interaction et l'intérêt du traitement chez les patients marqueur (+) mais ne permettent pas de conclure à l'absence d'effet chez les sujets marqueur (-). En effet, la borne optimiste de l'intervalle de confiance est de 0.72, correspondant à une réduction potentielle de ce critère de jugement de 28 % qui **ne peut pas être considérée comme cliniquement négligeable** comparée au bénéfice qu'apporte le traitement comparateur ou le traitement dans l'autre strate.

### ► Cas 4 - Identification *post-hoc* par une « pêche à la ligne »

L'étude pivot d'un nouveau traitement en cancérologie ne montre **pas**, par rapport au traitement standard, **de différence statistiquement significative** sur la survie sans progression qui était son critère de jugement principal (HR=1.05, p=0.68). Par contre, une **analyse génétique ancillaire**, réalisée sur 67 % des sujets inclus dans l'essai, montre que, pour les porteurs d'un variant génétique donné, le nouveau traitement **réduit de manière significative** la mortalité totale par rapport au traitement standard (HR=0.40, p=0.014) avec une **interaction** entre ce marqueur génétique et l'effet du traitement statistiquement significative (p=0.023). Aucune interaction n'a été détectée pour les autres génotypes étudiés sur aucun des critères de jugement de l'étude.

Ce résultat est utilisé pour soutenir la revendication de l'intérêt majeur de ce traitement (réduisant la mortalité) chez les sujets marqueur (+).

Cette revendication ne peut pas être acceptée car elle se base sur un résultat d'analyse en sous-groupes réalisée *a posteriori* et vraisemblablement entreprise dans le but de démontrer l'intérêt du

traitement après une analyse primaire négative. Il est non hypothético-déductif car l'hypothèse de l'interaction observée et de son sens n'a pas été faite *a priori*, et se base pour confirmer son hypothèse (le traitement est efficace chez les patients marqueur (+)) sur les mêmes données que celles qui ont servi à générer cette hypothèse. En outre, ce résultat pose également un problème de multiplicité des comparaisons statistiques avec une multiplication des tests statistiques en fonction des différents génotypes et critères de jugement testés. Le fait que ce résultat soit obtenu sur la mortalité totale n'augmente pas son niveau de preuve (la mortalité totale est soumise aux fluctuations aléatoires d'échantillonnages comme tous les autres critères), pas plus que le fait que l'effet soit important (pour être significatif sur un sous-groupe, où la puissance est forcément moindre que celle de l'essai car il y a moins de sujets, un effet doit être bien plus important que l'effet attendu).

### ► Cas 5 - Résultats de sous-groupe confirmés prospectivement par un essai ciblé

Un premier essai du nouveau traitement *versus* placebo dans une population non ciblée de 756 patients n'a pas été concluant sur son critère de jugement principal, la mortalité totale, avec un HR=0.87, p=0.15, IC95 % entre 0.71 et 1.09. Une analyse ancillaire génétique, réalisée chez 70 % des patients, suggère que le traitement pourrait être efficace pour le variant génétique X avec un HR=0.47, p=0.038.

Pour confirmer cette hypothèse, un nouvel essai satisfaisant méthodologiquement est entrepris uniquement chez des patients porteurs du variant génétique X avec le même critère de jugement principal. Neuf cent quatre-vingt-cinq patients ont été inclus dans ce nouvel essai qui s'avère concluant avec HR=0.88, p=0.012. Le résultat de cet essai est utilisé pour revendiquer l'intérêt de ce traitement et la valeur prédictive du variant génétique (marqueur) X.

L'efficacité du traitement dans la sous-population de patients porteurs du variant génétique considéré est démontrée de manière satisfaisante (qualité de l'essai, essai de confirmation de l'hypothèse).

Cependant, cette démarche ne permet pas de démontrer l'utilité clinique du test diagnostique identifiant le variant génétique X. En effet, il se peut simplement que le second essai soit concluant en raison d'une meilleure puissance (plus de patients et risque de base plus élevé) et non parce que le traitement serait réellement plus efficace chez ces sujets. Les éléments plutôt en faveur de cette hypothèse sont l'obtention de la même taille d'effet dans les deux essais (HR de 0.87 et 0.88) *versus* la valeur extrême de HR obtenu dans le sous-groupe du 1<sup>er</sup> essai (HR=0.47).

Cette situation expose à un fort risque d'entraîner la réalisation d'un test inutile, de recommander le marqueur à tort, et de priver des patients d'un traitement potentiellement intéressant pour eux. Un essai en design d'interaction est fortement souhaitable. L'inclusion de patients dans la strate marqueur (-) ne devant en outre pas poser de problème car l'essai est entrepris pour montrer que le traitement est efficace chez eux.

### ► Cas 6 - Identification du marqueur par une analyse rétrospective d'un essai concluant

L'intérêt d'un traitement a été parfaitement montré par un essai académique (HR=0.81, p=0.034) et il a été enregistré à l'aide de cet essai. Lors de la généralisation de son utilisation, une analyse en sous-groupes *post-hoc* de cet essai est entreprise à la recherche de marqueurs génétiques prédictifs de la réponse (ou de la non-réponse). Cette analyse est réalisée à partir de matériel génétique recueilli à l'entrée de l'étude chez tous les patients. Bien que prévue dès le début de l'étude, les hypothèses de cette analyse n'avaient pas été clairement définies. Le but donné à cette étude ancillaire était globalement de rechercher les déterminants génétiques de la réponse. A l'issue de l'analyse, un seul marqueur parmi quatorze testés aboutit à une interaction statistiquement significative (p=0.02) avec chez les patients marqueur (+) HR=0.56, p=0.02 et chez les patients marqueur (-) HR=0.98, p=0.87. La conclusion donnée à cette analyse est de restreindre l'utilisation de ce traitement coûteux aux patients marqueur (+) seulement.

Cette conclusion pose cependant plusieurs problèmes. Premièrement, il s'agit d'une analyse rétrospective, non hypothético-déductive. Deuxièmement, elle pose un problème de multiplicité avec une démarche de « pêche à ligne » parmi quatorze possibilités de trouver à tort une interaction avec un risque de 5 %. Et compte tenu du fait que l'essai démontre un bénéfice sur la population toute entière, toute restriction de la population cible est associée à un risque de priver du traitement des patients appartenant à la population évaluée lors du 1<sup>er</sup> essai. Afin de pouvoir garantir que cette restriction n'entraîne pas une perte de chance notoire pour ces sujets, il convient de montrer que chez ces patients marqueur (-), le traitement est inefficace (ou très peu efficace). Cela nécessite d'analyser l'intervalle de confiance des patients marqueur (-). Chez ces sujets, l'effet du traitement est HR=0.98 ; IC 95 % [0.61 ; 1.42] (tandis que chez les patients marqueur (+) il est HR=0.56 [0.16 ; 0.96]). Au vu de cet intervalle, il n'est pas raisonnable de conclure que le traitement est sans effet chez les patients marqueur (-). En effet, il est encore compatible, au mieux, avec un HR de 0.61 qui représente un bénéfice substantiel. Il n'est donc pas possible d'exclure avec un risque d'erreur de 5 % que le traitement pourrait être très efficace chez ces sujets. Il convient de plus de noter que dans cette situation une valeur de p très proche de 1 ne signifie pas que le traitement n'a, très vraisemblablement, pas d'effet mais simplement que l'intervalle de confiance est large.

### ► Cas 7 - Design ciblé

Le marqueur utilisé pour l'essai pivot du dossier est considéré comme un élément clé dans la physiopathologie de la maladie et dans le mécanisme d'action pressenti pour le traitement. La molécule a été « construite » lors du design moléculaire pour agir sur cette cible (cf. « ciblage par construction », section 1.5) considérée à ce moment comme *a priori* non présente chez tous les malades mais seulement chez certains d'entre eux.

L'essai est un essai ciblé incluant seulement des patients marqueurs positifs (+) comparant le nouveau traitement *versus* placebo « *on top* » la stratégie habituelle en double aveugle. L'essai a comme critère de jugement principal la mortalité totale. Il montre une réduction relative de 20 % de la mortalité avec le nouveau traitement sans effet délétère par ailleurs. L'essai réalisé répond par ailleurs à toutes les exigences de la démonstration de l'efficacité d'un traitement (protection contre les biais, contrôle du risque alpha).

Au total, cet essai est tout à fait acceptable pour valider la démonstration de l'intérêt du traitement chez les sujets marqueur (+). Cependant, il ne permet pas de revendiquer l'utilité clinique du test diagnostique identifiant le marqueur concerné car les données cliniques fournies ne permettent pas d'exclure que le traitement puisse en réalité apporter un bénéfice substantiel chez tous les patients.

### ► Tableau de synthèse des situations génériques possibles

Le tableau suivant synthétise les différents cas de figure génériques vraisemblablement susceptibles de constituer les supports de dossiers de demande d'évaluation de tests diagnostiques associés à des thérapeutiques stratifiées.

**Tableau 21 - Démonstrations et niveau de preuve associés au schéma de développement choisi pour un test diagnostique associé à une thérapie stratifiée au regard des appellations de test compagnon et thérapie ciblée.**

| Schéma de développement choisi pour le test diagnostique   | Démonstration(s) potentiellement apportée(s)   | Niveau de preuve considéré pour les appellations de test compagnon et thérapie ciblée   |
|--|--|---|
| Étude standard avec recherche <i>post-hoc</i> de « marqueurs prédictifs » n'incluant que les patients du groupe traité                           | Recherche de facteurs de risque du critère de jugement et non de marqueurs prédictifs de la réponse thérapeutique  | <b>Pas de conclusion possible</b>   |
| Recherche <i>post-hoc</i> en sous-groupes d'un marqueur prédictif dans une étude sur l'efficacité d'un nouveau traitement en population générale | <ul style="list-style-type: none"> <li>- Efficacité ou non efficacité du traitement dans une population de patients non sélectionnés par un marqueur</li> <li>- Recherche exploratoire d'un marqueur prédictif dont l'utilité clinique du test diagnostique doit être confirmée dans le cadre d'un essai prospectif</li> </ul>   | <b>Pas de conclusion possible</b> (En l'absence d'essai prospectif de confirmation)   |
| Étude de « comparaison de stratégies »   | <ul style="list-style-type: none"> <li>- Efficacité du nouveau traitement chez les patients marqueur (+), ou chez tous les patients quel que soit leur statut pour le marqueur, supérieure à celle du traitement standard chez tous les patients quel que soit leur statut pour le marqueur</li> <li>- Supériorité de la stratégie basée sur le marqueur identifié par le test diagnostique évalué et l'administration du nouveau traitement aux patients Mq+, par rapport à la stratégie d'utilisation du traitement standard sans détermination du marqueur</li> </ul> | <b>Pas de conclusion possible</b>   |
| Étude « ciblée » chez les patients marqueur (+)  | Efficacité du nouveau traitement dans la sous-population de patients identifiés marqueur (+) par le test diagnostique évalué   | <b>Niveau de preuve faible à modéré</b> en fonction du poids accordé aux éléments précliniques et cliniques précoces lors de l'évaluation |
| Étude d' « interaction marqueur/traitement »   | <ul style="list-style-type: none"> <li>- Validation de l'utilité clinique du test diagnostique évalué, incluant la démonstration de l'efficacité du nouveau traitement dans la sous-population de patients identifiés marqueur (+) par le test</li> <li>- Appellations de test compagnon et thérapie ciblée</li> </ul>   | <b>Niveau de preuve élevé</b>   |

### 4.3.3 Exemples de cas issus de la littérature

#### ► Exemple d'essais ciblés - Trastuzumab et cancer du sein

Le trastuzumab (Herceptin®) est un anticorps monoclonal dont la synthèse a été orientée contre le récepteur HER2 (*human epidermal growth factor receptor-2*). Compte tenu de la spécificité attribuée aux anticorps monoclonaux, il a été implicitement considéré lors de son développement clinique que cet anticorps ne pouvait être efficace que chez les patientes porteuses d'un cancer du sein HER2-positif. Ce traitement n'a par conséquent été évalué que dans des essais ciblés n'incluant que des femmes ayant un cancer du sein HER2-positif (5,37). Ce type d'essai a été choisi sur la base du rationnel mécanistique, la démonstration de l'interaction étant alors considérée comme implicite.

Cependant, des résultats d'analyses rétrospectives réalisées sur plusieurs études prospectives ont laissé plus récemment supposer que le trastuzumab pourrait aussi être efficace en l'absence d'expression détectable de HER2. Ces résultats ont suggéré la notion de récepteurs HER2 cachés (6,7).

Ces observations ont conduit à la mise en place de l'essai "TREAT CTC" (*TRastuzumab in HER2-negative Early Breast Cancer as Adjuvant Treatment for Circulating Tumor Cells*) actuellement en cours (NCT01548677).

#### ► Exemple d'essai d'interaction marqueur/traitement - Essai MARVEL

L'essai MARVEL est un essai (NCT00738881) comparant l'erlotinib au pemetrexed dans le cancer du poumon. Il présente un véritable schéma d'étude d'interaction, stratifié sur le statut des patients pour l'amplification du gène EGFR évalué par technique FISH.

#### ► Exemple d'essai de comparaison de stratégies - Essai SPANISH

L'essai SPANISH (38) est un essai de comparaison de stratégies comparant une stratégie basée sur le niveau d'ARNm d'ERCC1 mRNA à une stratégie standard.

#### ► Exemple d'analyse *post-hoc* après une étude concluante - Cetuximab dans le cancer colorectal métastatique

Le cetuximab a été enregistré par la FDA en juillet 2012 pour l'indication « *for use in combination with FOLFIRI (irinotecan, 5-fluorouracil, and folinic acid) for first-line treatment of patients with KRAS mutation-negative (wild-type), EGFR-expressing metastatic colorectal cancer (mCRC) as determined by FDA-approved tests* ».

Cette indication a été obtenue suite à une analyse rétrospective des résultats de l'essai nommé « Crystal ». L'essai Crystal (NCT00154102) (39) avait comparé en ouvert le cetuximab à l'absence de cetuximab en *add-on* de l'association nommée « FOLFIRI » dans le cancer colorectal métastatique chez des patients exprimant le récepteur EGFR et n'ayant pas reçu de chimiothérapie préalable. Les patients avaient été inclus quel que soit leur statut pour les mutations activatrices de KRAS et aucune analyse en sous-groupes en fonction de ce statut mutationnel n'avait été prévue *a priori*. La survie sans progression (*progression-free survival*, PFS) était le critère de jugement principal.

Sur la totalité des patients inclus (Mq+ et Mq-), une amélioration statistiquement significative de la PFS par le cetuximab avait été observée avec un HR=0.85 (IC95 % compris entre 0.74 et 0.99), p=0.036.

Lors d'une analyse en sous-groupes en fonction du statut mutationnel pour KRAS réalisée chez seulement 45 % des patients inclus, une interaction statistiquement significative a été relevée pour un critère secondaire, la réponse tumorale (p=0.03).

Aucune interaction statistiquement significative n'a été observée sur le critère principal, la PFS, p=0.07, ni pour la survie totale, p=0.44.

Ces données ont abouti *in fine* à la restriction d'indication du cetuximab aux patients KRAS non mutés, bien que les réserves méthodologiques pouvant être émises soient relativement nombreuses :

- démarche non hypothético-déductive/analyse rétrospective ;
- absence de significativité statistique de l'interaction sur le critère de jugement principal ;
- très faible effectif du sous-groupe marqueur (-) (mutation du KRAS) représentant 6 % des patients inclus ;
- intervalle de confiance large chez les patients marqueur (-) englobant la totalité de l'intervalle de confiance du résultat sur le critère de jugement principal et la population globale.

Une publication dans le NEJM en 2008 (40) a présenté une analyse *post-hoc* de l'essai co.17 (41) testant le marqueur constitué par la mutation du gène KRAS sur l'efficacité du cetuximab (*versus* soins palliatifs). Initialement l'essai co.17 avait montré l'efficacité du cetuximab par réduction de la mortalité totale (HR 0.77 ; IC95 %, 0.64-0.92 ;  $p=0.005$ ) chez 572 patients exprimant le récepteur EGFR. L'analyse *post-hoc* a porté sur un échantillon de 394 patients et a montré un HR de 0.55 (IC95 %, 0.41-0.74 ;  $p<0.001$ ) chez les patients sans mutation de KRAS, *versus* 0.98 (IC95 %, 0.70-1.37 ;  $p=0.89$ ) chez ceux ayant une mutation de ce gène. L'interaction était statistiquement significative avec un  $p=0.01$ .

En 2009, sur la base d'une revue systématique de 5 essais portant sur le cetuximab et le panitumumab, l'ASCO (42) a recommandé que les patients ayant une mutation (codon 12 ou 13) de leur gène KRAS ne reçoivent plus d'anticorps anti-EGFR...

#### 4.4 Considérations réglementaires

La problématique des évaluations technologiques dans le domaine des thérapies « ciblées » est une problématique récente mais ayant rapidement pris de l'importance du fait du nombre croissant de ce type de thérapeutiques mises sur le marché ces dernières années et du nombre considérable de molécules (et tests associés) actuellement en cours de développement dans ce domaine. Depuis très récemment, des règles d'évaluation sont en cours d'élaboration dans différents pays.

Cette problématique a, à ce jour, été essentiellement abordée sous l'angle de la pharmacogénomique. Sur cette thématique, une liste de recommandations émises par la FDA essentiellement, mais également par l'EMA (43) ainsi que quelques guidelines ICH sont disponibles sur le site internet de la FDA à l'adresse :

<http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083374.htm>

Par ailleurs, la FDA a publié fin 2012 un document de recommandations dédié à la réalisation des études cliniques de type « ciblées » nommé « *Guidance for industry - Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products* », téléchargeable à l'adresse :

<http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf> (44). Les approches décrites dans ce guide ne sont cependant pas totalement spécifiques de la problématique des marqueurs mais apportent plus généralement des recommandations concernant la réalisation des essais.

## Conclusion - Perspectives

Le concept de **médecine stratifiée** consiste en une approche thérapeutique encore récente s'appuyant sur des couples « **médicament/test diagnostique** ». Au sein de cette approche, le test diagnostique a pour objectif théorique de permettre de **sélectionner**, en fonction de leur statut pour un marqueur présumé prédictif, uniquement les patients chez lesquels le traitement est susceptible d'apporter un bénéfice et de limiter ainsi l'utilisation de ce traitement à ces seuls patients.

Les concepts de **thérapie ciblée** et **test compagnon** s'inscrivent dans le cadre de la médecine stratifiée mais sont beaucoup plus étroits et donc plus exigeants en termes de démonstrations requises. Ainsi, pour permettre de revendiquer qu'une thérapie stratifiée peut être qualifiée de ciblée ou qu'un test diagnostique peut être nommé test compagnon, la démonstration doit être apportée qu'en plus d'être efficace chez les patients porteurs du marqueur, le traitement ne l'est pas chez les patients ne le portant pas. Ces démonstrations sont par ailleurs les fondements de la validation de l'**utilité clinique du test**, c'est-à-dire de son aptitude à améliorer le devenir clinique des patients et apporter une valeur ajoutée en termes d'optimisation de la prise en charge. En l'occurrence, l'utilisation du médicament et la réalisation du test étant dans ce contexte indissociables, l'utilité clinique du test ne peut être évaluée que conjointement à celle de l'efficacité du traitement au sein d'un même essai. Seul le **design de type « interaction marqueur/traitement »** permet alors d'apporter conjointement ces démonstrations. Les autres schémas d'étude, où la détermination d'effet du nouveau traitement est limitée à la sous-population de patients marqueur (+), ne permettent de conclure qu'à l'efficacité (ou non) du traitement dans cette sous-population de patients, mais en aucun cas à l'utilité clinique du test.

Au regard de la HAS, lorsqu'un test diagnostique permet la sélection des patients afin de répondre aux **exigences formulées au sein de l'AMM** d'un traitement, l'évaluation pourra, en fonction des démonstrations apportées, aboutir à **deux types de conclusions** générales :

- si la **démonstration** de l'utilité clinique du test n'a **pas** été **apportée** :
  - le test sera considéré comme un test diagnostique « classique » permettant d'identifier la sous-population de patients dont la sélection est notifiée au sein de l'AMM. Bien que sans utilité clinique démontrée, la réalisation de ce test n'en sera pas moins indispensable d'un point de vue réglementaire afin de permettre l'utilisation du traitement dans le cadre de son AMM,
  - l'appellation « test compagnon » ne pourra pas être revendiquée. La thérapie ne pourra pas être considérée comme ciblée. En effet, la preuve de la valeur prédictive du marqueur n'étant pas apportée, la possibilité d'une efficacité chez tous les patients ne pourra pas être exclue,
- si la **démonstration** d'utilité clinique du test est **apportée** :
  - le test sera considéré comme un test diagnostique permettant d'identifier les patients dont la sélection est notifiée au sein de l'AMM,
  - le test pourra de plus être qualifié de « test compagnon » et la thérapie considérée comme ciblée.

Dans tous les cas, les performances diagnostiques du test évalué doivent assurer la fiabilité des résultats des études menées. Sa mise en œuvre doit prévoir des contrôles de qualité et respecter l'ensemble des normes de qualité en vigueur dans le domaine concerné.

## Références

1. Sargent D, Allegra C. Issues in clinical trial design for tumor marker studies. *Semin Oncol.* juin 2002;29(3):222- 230.
2. Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points--the challenge of statistical validation. *Nat Rev Clin Oncol.* juin 2010;7(6):309- 317.
3. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol Off J Am Soc Clin Oncol.* 20 mars 2005;23(9):2020- 2027.
4. Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nat Rev Drug Discov.* avr 2007;6(4):287- 293.
5. Yin W, Jiang Y, Shen Z, Shao Z, Lu J. Trastuzumab in the adjuvant treatment of HER2-positive early breast cancer patients: a meta-analysis of published randomized controlled trials. *PLoS One.* 2011;6(6):e21030.
6. Ardavanis A, Kountourakis P, Kyriakou F, Malliou S, Mantzaris I, Garoufali A, et al. Trastuzumab plus Paclitaxel or Docetaxel in HER-2–Negative/HER-2 ECD–Positive Anthracycline- and Taxane-Refractory Advanced Breast Cancer. *The Oncologist.* 4 janv 2008;13(4):361- 369.
7. Tuma RS. Cancer stem cell hypothesis and trastuzumab in HER2-negative tumors. *J Natl Cancer Inst.* 3 juill 2012;104(13):968- 969.
8. Chung KY, Shia J, Kemeny NE, Shah M, Schwartz GK, Tse A, et al. Cetuximab shows activity in colorectal cancer patients with tumors that do not express the epidermal growth factor receptor by immunohistochemistry. *J Clin Oncol Off J Am Soc Clin Oncol.* 20 mars 2005;23(9):1803- 1810.
9. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol Off J Am Soc Clin Oncol.* 20 août 2009;27(24):4027- 4034.
10. Dorresteijn JAN, Boekholdt SM, van der Graaf Y, Kastelein JJP, LaRosa JC, Pedersen TR, et al. High-dose statin therapy in patients with stable coronary artery disease: treating the right patients based on individualized prediction of treatment effect. *Circulation.* 25 juin 2013;127(25):2485- 2493.
11. Collet J-P, Cuisset T, Rangé G, Cayla G, Elhadad S, Pouillot C, et al. Bedside monitoring to adjust antiplatelet therapy for coronary stenting. *N Engl J Med.* 29 nov 2012;367(22):2100- 2109.
12. Price MJ, Berger PB, Teirstein PS, Tanguay J-F, Angiolillo DJ, Spriggs D, et al. Standard- vs high-dose clopidogrel based on platelet function testing after percutaneous coronary intervention: the GRAVITAS randomized trial. *JAMA J Am Med Assoc.* 16 mars 2011;305(11):1097- 1105.
13. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group. Recommendations from the EGAPP Working Group: can testing of tumor tissue for mutations in EGFR pathway downstream effector genes in patients with metastatic colorectal cancer improve health outcomes by guiding decisions regarding anti-EGFR therapy? *Genet Med Off J Am Coll Med Genet.* juill 2013;15(7):517- 527.
14. Teutsch SM, Bradley LA, Palomaki GE, Haddow JE, Piper M, Calonge N, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med Off J Am Coll Med Genet.* janv 2009;11(1):3- 14.
15. Febbo PG, Ladanyi M, Aldape KD, De Marzo AM, Hammond ME, Hayes DF, et al. NCCN Task Force report: Evaluating the clinical utility of tumor markers in oncology. *J Natl Compr Cancer Netw JNCCN.* nov 2011;9 Suppl 5:S1- 32; quiz S33.
16. Mandrekar SJ, Grothey A, Goetz MP, Sargent DJ. Clinical trial designs for prospective validation of biomarkers. *Am J Pharmacogenomics Genomics-Relat Res Drug Dev Clin Pract.* 2005;5(5):317- 325.
17. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. *Clin Trials Lond Engl.* oct 2010;7(5):567- 573.
18. Mandrekar SJ, Sargent DJ. Design of clinical trials for biomarker research in oncology. *Clin Investig.* déc 2011;1(12):1629- 1636.
19. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: one size does not fit all. *J Biopharm Stat.* 2009;19(3):530- 542.
20. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res Off J Am Assoc Cancer Res.* 1 nov 2005;11(21):7872- 7878.
21. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med.* 16 juill 1987;317(3):141- 145.
22. Féré C, Hollebécque A, Salleron J, Penel N. [Clinical equipoise and randomised clinical trials in oncology]. *Bull Cancer (Paris).* juin 2009;96(6):727- 731.
23. Gifford F. Community-equipoise and the ethics of randomized clinical trials. *Bioethics.* avr 1995;9(2):127- 148.
24. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Ann Intern Med.* 15 févr 2011;154(4):253- 259.
25. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res Off J Am Assoc Cancer Res.* 15 oct 2004;10(20):6759- 6763.

26. Ma BBY, Britten CD, Siu LL. Clinical trial designs for targeted agents. *Hematol Oncol Clin North Am.* oct 2002;16(5):1287- 1305.
27. Cunningham D, Humblet Y, Siena S, Khayat D, Bleiberg H, Santoro A, et al. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *N Engl J Med.* 22 juill 2004;351(4):337- 345.
28. Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, de Gramont A. Integrating biomarkers in clinical trials. *Expert Rev Mol Diagn.* mars 2011;11(2):171- 182.
29. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA J Am Med Assoc.* 3 juill 1991;266(1):93- 98.
30. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med.* 1 janv 1992;116(1):78- 84.
31. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet.* 8 janv 2005;365(9454):176- 186.
32. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet.* 15 janv 2005;365(9455):256- 265.
33. Cuzick J. Forest plots and the interpretation of subgroups. *Lancet.* 9 avr 2005;365(9467):1308.
34. Mega JL, Close SL, Wiviott SD, Shen L, Hockett RD, Brandt JT, et al. Cytochrome p-450 polymorphisms and response to clopidogrel. *N Engl J Med.* 22 janv 2009;360(4):354- 362.
35. Mega JL, Simon T, Collet J-P, Anderson JL, Antman EM, Bliden K, et al. Reduced-function CYP2C19 genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. *JAMA J Am Med Assoc.* 27 oct 2010;304(16):1821- 1830.
36. Holmes MV, Perel P, Shah T, Hingorani AD, Casas JP. CYP2C19 genotype, clopidogrel metabolism, platelet function, and cardiovascular events: a systematic review and meta-analysis. *JAMA J Am Med Assoc.* 28 déc 2011;306(24):2704- 2714.
37. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med.* 15 mars 2001;344(11):783- 792.
38. Cobo M, Isla D, Massuti B, Montes A, Sanchez JM, Provencio M, et al. Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: a phase III trial in non-small-cell lung cancer. *J Clin Oncol Off J Am Soc Clin Oncol.* 1 juill 2007;25(19):2747- 2754.
39. Van Cutsem E, Köhne C-H, Hitre E, Zaluski J, Chang Chien C-R, Makhson A, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N Engl J Med.* 2 avr 2009;360(14):1408- 1417.
40. Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med.* 23 oct 2008;359(17):1757- 1765.
41. Jonker DJ, O'Callaghan CJ, Karapetis CS, Zalcborg JR, Tu D, Au H-J, et al. Cetuximab for the treatment of colorectal cancer. *N Engl J Med.* 15 nov 2007;357(20):2040- 2048.
42. Allegra CJ, Jessup JM, Somerfield MR, Hamilton SR, Hammond EH, Hayes DF, et al. American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *J Clin Oncol Off J Am Soc Clin Oncol.* 20 avr 2009;27(12):2091- 2096.
43. European Medicines Agency. Reflection paper on methodological issues associated with pharmacogenomic biomarkers in relation to clinical development and patient selection. EMA/446337/2011, june 2011 [cité 8 août 2013]. Disponible sur: [http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2011/07/WC500108672.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/07/WC500108672.pdf)
44. Food and Drug Administration. Guidance for Industry - Enrichment Strategies for Clinical Trials to Support Approval of Human Drug and Biological Products, dec 2012 [cité 8 août 2013]. Disponible sur: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf>

## Annexe 1. Index

|                                 |            |                                |    |
|---------------------------------|------------|--------------------------------|----|
| <b>A</b>                        |            |                                |    |
| absolute risk reduction         | 19, 21, 22 | marqueur de risque             | 10 |
| analyse en sous-groupes         | 42         | marqueur prédictif             | 12 |
| analyse rétrospective           | 45         | MARVEL                         | 61 |
| ARR                             | 19         | MARVEL study                   | 35 |
| <b>B</b>                        |            | médecine individualisée        | 17 |
| balance bénéfice risque         | 30         | médecine stratifiée            | 17 |
| bénéfice net                    | 30         | multiplicité                   | 50 |
| biais                           | 50         | mutation K-ras                 | 61 |
| biais d'attrition               | 50         | <b>N</b>                       |    |
| biais de mesure                 | 50         | NNT                            | 19 |
| biais de réalisation            | 50         | number needed to treat         | 19 |
| biais de sélection              | 50         | <b>O</b>                       |    |
| <b>C</b>                        |            | odds ratio                     | 11 |
| cetuximab                       | 61         | <b>P</b>                       |    |
| clopidogrel                     | 41, 46, 48 | performance diagnostique       | 31 |
| cohérence externe               | 50         | pertinence clinique            | 50 |
| conséquence populationnelle     | 20         | <b>R</b>                       |    |
| critère de bénéfice net         | 30         | rapports de risques ratio      | 27 |
| critère de jugement             | 30         | réalité statistique            | 50 |
| critère de jugement principal   | 50         | récepteur EGFR                 | 37 |
| critère intermédiaire           | 30         | récepteurs œstrogènes          | 9  |
| Crystal (essai)                 | 61         | rétrospective (analyse)        | 45 |
| <b>E</b>                        |            | risk ratio                     | 11 |
| effect-modification analysis    | 48         | <b>S</b>                       |    |
| EGFR                            | 61         | SPANISH                        | 61 |
| erlotinib                       | 35, 61     | stratified medicine            | 17 |
| <b>F</b>                        |            | <b>T</b>                       |    |
| facteur de risque               | 10         | targeted design                | 36 |
| <b>G</b>                        |            | taux de faux négatif           | 31 |
| GRAVITAS                        | 37         | taux de faux positif           | 31 |
| <b>H</b>                        |            | test statistique d'interaction | 15 |
| hazard ratio                    | 11         | TFN                            | 31 |
| <b>I</b>                        |            | TFP                            | 31 |
| individualized medicine         | 17         | thérapeutique ciblée           | 19 |
| inflation du risque alpha       | 45         | ticagrelor                     | 46 |
| interaction                     | 14         | trastuzumab                    | 17 |
| interaction marqueur traitement | 27         | <i>treatment only analysis</i> | 46 |
| interaction qualitative         | 16         | TRITON TIMI 38                 | 46 |
| interaction quantitative        | 16         | <b>V</b>                       |    |
| <b>M</b>                        |            | validité interne               | 50 |
| marker-based strategy design    | 38         | VerifyNow                      | 41 |

## Annexe 2. Liste des tableaux

|  |    |
|--|----|
| Tableau 1 - Principales motivations conduisant à la recherche de marqueurs prédictifs permettant le ciblage de traitements.....  | 10 |
| Tableau 2 - Illustration numérique du concept de facteur de risque (chez des sujets traités et non traités).....   | 12 |
| Tableau 3 - Illustration numérique du concept de marqueur prédictif.....   | 12 |
| Tableau 4 - Illustration d'une situation où le traitement est efficace dans les deux strates définies par le marqueur modificateur de l'effet du traitement.....   | 13 |
| Tableau 5 - Illustration d'une situation où le traitement est efficace dans la strate des patients présentant le marqueur modificateur de l'effet du traitement et délétère dans l'autre.....                                    | 13 |
| Tableau 6 - Illustration d'une situation où le marqueur modificateur de l'effet du traitement est aussi facteur de risque.....   | 13 |
| Tableau 7 - Exemple de deux facteurs, l'un entraînant une interaction et l'autre non. Les interactions sont visualisées à l'aide des vraies valeurs d'effet du traitement.....   | 15 |
| Tableau 8 - Simulation des conséquences populationnelles de l'utilisation d'un ciblage à l'aide d'un marqueur réellement prédictif.....  | 21 |
| Tableau 9 - Simulation des conséquences populationnelles d'un ciblage à tort à l'aide d'un marqueur non prédictif.....   | 22 |
| Tableau 10 - Simulation des conséquences populationnelles d'un pseudo-ciblage utilisant un marqueur non entièrement prédictif.....   | 23 |
| Tableau 11 - Simulation des conséquences populationnelles d'un ciblage à l'aide d'un marqueur prédictif de la sécurité.....  | 24 |
| Tableau 12 - Illustration d'une situation d'interaction statistique suggérant l'existence d'un marqueur prédictif.....   | 27 |
| Tableau 13 - Situation où une interaction existe et où le traitement est efficace chez les patients marqueur (+) et marqueur (-).....  | 28 |
| Tableau 14 - Marqueur prédictif associé à une forte suggestion d'absence d'effet chez les patients marqueur (-).....   | 28 |
| Tableau 15 - Illustration numérique du schéma de comparaison de stratégies (dans une situation où le marqueur est bien prédictif).....   | 40 |
| Tableau 16 - Illustration numérique du schéma de comparaison de stratégies (dans une situation où le marqueur n'est pas prédictif).....  | 41 |
| Tableau 17 - Essais appropriés en fonction du contexte de l'évaluation d'un marqueur prédictif candidat.....   | 48 |
| Tableau 18 - Réserves méthodologiques pouvant invalider la revendication d'utilité clinique d'un test diagnostique.....  | 51 |
| Tableau 19 - Critères impérativement requis pour la revendication d'utilité clinique d'un test diagnostique.....   | 52 |
| Tableau 20 - Démonstrations à apporter pour prouver l'utilité clinique du test dans les différents contextes de développement d'un couple traitement/test diagnostique.....  | 53 |
| Tableau 21 - Démonstrations et niveau de preuve associés au schéma de développement choisi pour un test diagnostique associé à une thérapeutique stratifiée au regard des appellations de test compagnon et thérapie ciblée..... | 60 |

### Annexe 3. Liste des figures

|   |    |
|---|----|
| Figure 1 - Illustration des concepts de modificateur de l'effet du traitement et de marqueur prédictif.....   | 9  |
| Figure 2 - Illustration graphique des concepts de marqueur prédictif et facteur de risque. Les risques (en ordonnée) sont associés à la présence ou non du facteur étudié (en abscisse) dans les deux groupes d'un essai thérapeutique (contrôles en mauve et traités en orange). .....   | 14 |
| Figure 3 - Prise en compte de l'incertitude des estimations - Test d'interaction. ....  | 16 |
| Figure 4 - Application du raisonnement de non supériorité à la mise en évidence des marqueurs prédictifs. Exemples de différents types de résultats possibles dans le sous-groupe des patients marqueurs (-). ....  | 29 |
| Figure 5 - Évolution des estimations de l'effet du traitement (rapport des risques, RR) dans les strates test (+) et test (-) en fonction des taux de faux positifs (TFP) et de faux négatifs (TFN). Le vrai rapport des risques chez les patients marqueur (+) est de 0.8 et chez les patients marqueur (-) de 1.00. ....                                      | 31 |
| Figure 6 - Illustration du schéma d'étude d' « interaction marqueur traitement » .....  | 34 |
| Figure 7 - Schéma d'étude « ciblée » .....  | 36 |
| Figure 8 - Schéma d'étude de « comparaison de stratégies » .....  | 38 |
| Figure 9 - Exemple de résultats de sous-groupes montrant le risque accru d'erreur bêta de 2 <sup>ème</sup> espèce lié aux analyses en sous-groupes : analyse de l'intérêt de l'endartériectomie carotidienne chez les patients inclus dans l'essai ECST présentant une sténose symptomatique ≥ 70 % en fonction de leur jour de naissance. (D'après (31)) ..... | 43 |
| Figure 10 - Influence de la largeur des intervalles de confiance sur la signification statistique de l'interaction (à estimations ponctuelles constantes) (D'après (31)).....   | 44 |
| Figure 11 - Fréquence de survenue d'évènements cardiovasculaires chez des porteurs <i>versus</i> non porteurs de variants génétiques de cytochromes P450 chez des patients recevant du clopidogrel. (D'après (34)).....   | 47 |
| Figure 12 - Influence de la largeur des intervalles de confiance sur la signification statistique de l'interaction (à estimations ponctuelles constantes) .....   | 54 |
| Figure 13 - Existence d'une interaction très en faveur d'un marqueur prédictif .....  | 54 |
| Figure 14 - Interaction qualitative .....   | 55 |
| Figure 15 - Existence d'une interaction statistique avec un traitement efficace dans les deux strates marqueurs (+) et (-). ....  | 55 |
| Figure 16 - Situation où il est difficile de conclure à la valeur prédictive du marqueur. ....  | 56 |
| Figure 17 - Autre situation posant problème pour conclure à la valeur prédictive du marqueur.....   | 56 |

## Fiche descriptive

| Intitulé                   | Descriptif  |
|----------------------------|---|
| Méthode de travail         | Guide méthodologique  |
| Date de mise en ligne      | Avril 2014  |
| Date d'édition             | Uniquement disponible sous format électronique sur <a href="http://www.has-sante.fr">www.has-sante.fr</a>   |
| Objectif(s)                | <p>Exposer les fondements théoriques et méthodologiques de l'évaluation d'un test diagnostique permettant d'identifier un marqueur présumé prédictif associé à une thérapie stratifiée.</p> <p>Définir et exposer les preuves requises par la HAS pour la reconnaissance ou non des appellations « test compagnon » et « thérapie ciblée ».</p> <p>Apporter des outils méthodologiques pour la démonstration de l'utilité clinique d'un test diagnostique associé à une thérapie stratifiée.</p> <p>Illustrer les aspects théoriques par des exemples extraits de la littérature ou à unique but didactique.</p> <p>Informar des niveaux de preuve accordés par la HAS aux schémas d'études des tests diagnostiques concernés par l'évaluation.</p> |
| Demandeur                  | Autosaisine HAS   |
| Promoteur                  | Haute Autorité de santé (HAS), service évaluation des actes professionnels (SEAP)   |
| Pilotage du projet         | <p>Coordination SEAP : Carole GIRAUD, chef de projet, SEAP (chef de service : Michèle MORIN-SURROCA, adjoint au chef de service : Marc GUERRIER)</p> <p>Secrétariat : Christine MAYOL, assistante, SEAP</p> <p>Coordination externe : Michel CUCHERAT, consultant externe</p>   |
| Recherche documentaire     | <p>Carole GIRAUD, chef de projet, SEAP</p> <p>Michel CUCHERAT, consultant externe</p>   |
| Auteurs de l'argumentaire  | <p>Carole GIRAUD, chef de projet, SEAP, sous la responsabilité de Marc GUERRIER, adjoint au chef de service, SEAP</p> <p>Clémence THEBAUT, chef de projet, service évaluation économique et santé publique (SEESP) et Salah GHABRI, chef de projet, SEESP, sous la responsabilité de Catherine RUMEAU-PICHON, chef de service, SEESP</p> <p>Michel CUCHERAT, Consultant</p>   |
| Validation                 | Validation par le Collège de la HAS : février 2014  |
| Autres formats             | Pas d'autre format que le format électronique disponible sur <a href="http://www.has-sante.fr">www.has-sante.fr</a>   |
| Documents d'accompagnement | Guide méthodologique intitulé « Test compagnon associé à une thérapie ciblée : définition et méthode d'évaluation », disponible sur <a href="http://www.has-sante.fr">www.has-sante.fr</a>  |





Toutes les publications de la HAS sont téléchargeables sur  
[www.has-sante.fr](http://www.has-sante.fr)