HAUTE AUTORITÉ DE SANTÉ

ÉTAT DES LIEUX

Niveau de preuve et gradation des recommandations de bonne pratique

Avril 2013

Les recommandations et leur synthèse sont téléchargeables sur
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

Ce document a été validé par le Collège de la Haute Autorité de Santé en avril 2013.

Table des matières

Abréviations et acronymes

AAP	*American Academy of Pediatry*
ACP	*American College of Physicians*
AGREE	*Appraisal of Guidelines for Research and Evaluation*
AHRQ	*Agency for Healthcare Research and Quality*
FNCLCC	Fédération nationale des centres de lutte contre le cancer
GIN	*Guidelines International Network*
GRADE	*Grading of Recommendations Assessment, Development and Evaluation*
NHMRC	*National Health and Medical Research Council*
NICE	*National Institute for Health and Clinical Excellence*
NZGG	*New Zealand Guidelines Group*
PICO	*Population Intervention Comparison Outcome*
QUADAS	*Quality Assessment of Diagnostic Accuracy Studies*
RBP	Recommandation de bonne pratique
SIGN	*Scottish Intercollegiate Guidelines Network*
USPSTF	*US Preventive Services Task Force*

Introduction

Parmi ses missions, la Haute Autorité de Santé (HAS) est chargée d'« élaborer les guides de bon usage des soins ou les recommandations de bonne pratique, procéder à leur diffusion et contribuer à l'information des professionnels de santé et du public dans ces domaines, sans préjudice des mesures prises par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) dans le cadre de ses missions de sécurité sanitaire » (loi du 13 août 2004 relative à l'Assurance maladie, titre II, chapitre I^{er} bis, article L. 161-37).

Les « recommandations de bonne pratique » (RBP) sont définies dans le champ de la santé comme « des propositions développées méthodiquement pour aider le praticien et le patient à rechercher les soins les plus appropriés dans des circonstances cliniques données ».

Les RBP sont des synthèses rigoureuses de l'état de l'art et des données de la science à un temps donné. Une démarche rigoureuse et explicite doit être appliquée pour élaborer des recommandations de bonne pratique valides et crédibles. La rigueur méthodologique et la transparence du processus d'élaboration des RBP peuvent être évaluées à partir de critères internationaux. Par exemple dans la grille AGREE II, il y a parmi les éléments retenus pour l'évaluation de la rigueur d'élaboration d'une recommandation, les éléments suivants (1) :

Élément 9 : Les forces et les limites des données scientifiques sont clairement définies.

Élément 11 : Les bénéfices, les effets indésirables et les risques en termes de santé ont été pris en considération dans la formulation des recommandations.

Élément 12. Il y a un lien explicite entre les recommandations et les données scientifiques sur lesquelles elles reposent.

Le système de niveau de preuve et de gradation des recommandations de bonne pratique utilisé par la HAS a été développé par l'Agence nationale d'accréditation et d'évaluation en santé (Anaes) en 1999 (2).

La question est de savoir s'il est nécessaire d'adopter un système existant ou d'élaborer un nouveau système de niveau de preuve et gradation répondant aux attentes des différents partenaires pour les études d'intervention, études diagnostiques (Gradation HAS, SOR, GRADE, SIGN, etc.).

Le projet a consisté dans un premier temps à la rédaction d'un état de lieux sur le sujet.

Les propositions de la HAS seront présentées dans le guide d'analyse de la littérature et gradation des recommandations (en cours).

1. Niveau de preuve et gradation des recommandations : les principaux systèmes actuels

La revue de la littérature a porté sur les différents systèmes actuellement en vigueur, d'élaboration des RBP fondées sur les données scientifiques, publiées en français ou en anglais (annexe 1). Elle a inclus :

- soit des manuels mis à disposition sur le site internet d'organismes nationaux et internationaux qui produisent des RBP (2-8) ;
- soit des articles publiés dans des revues (9-12).

La présentation de ces systèmes suit un plan chronologique.

1.1 Haute Autorité de santé

Ce paragraphe a été rédigé à partir du *Guide d'analyse de la littérature et gradation des recommandations* publié par l'Anaes en 2000 (2).

La rédaction des recommandations aboutit à un texte de synthèse des connaissances et des pratiques à partir des données de la littérature scientifique et de l'avis d'experts. La démarche consiste à identifier les niveaux de preuve scientifique fournis par la littérature et à formaliser des recommandations prenant en compte les informations fournies.

► Niveau de preuve

► Niveau de preuve d'une étude

Le niveau de preuve d'une étude caractérise la capacité de l'étude à répondre à la question posée.

La capacité d'une étude à répondre à la question posée est jugée sur la correspondance de l'étude au cadre du travail (question, population, critères de jugement) et sur les caractéristiques suivantes :

- l'adéquation du protocole d'étude à la question posée (annexe 3) ;
- l'existence ou non de biais importants dans la réalisation ;
- l'adaptation de l'analyse statistique aux objectifs de l'étude ;
- la puissance de l'étude et en particulier la taille de l'échantillon.

Selon le domaine exploré (diagnostic, pronostic, dépistage, traitement, etc.) un fort niveau de preuve peut être donné par des études dont le type de protocole sera différent.

Le tableau 1 présente une classification générale du niveau de preuve d'une étude.

Tableau 1. Classification générale du niveau de preuve d'une étude

Niveau de preuve	Description
Fort	- le protocole est adapté pour répondre au mieux à la question posée ; - la réalisation est effectuée sans biais majeur ; - l'analyse statistique est adaptée aux objectifs ; - la puissance est suffisante.
Intermédiaire	- le protocole est adapté pour répondre au mieux à la question posée ; - puissance nettement insuffisante (effectif insuffisant ou puissance *a posteriori* insuffisante) ;

Niveau de preuve	Description
	- et/ou des anomalies mineures.
Faible	Autres types d'études.

► Évidence scientifique

L'évidence scientifique est appréciée lors de la synthèse des résultats de l'ensemble des études sélectionnées. Elle constitue la conclusion des tableaux de synthèse de la littérature. La **gradation de l'évidence scientifique** s'appuie sur :

- l'existence de données de la littérature pour répondre aux questions posées ;
- le niveau de preuve des études disponibles ;
- la cohérence de leurs résultats.

Pour une question donnée, il est possible de classer les études en fonction de leur niveau de preuve.

Pour chaque niveau, l'attention est portée aux résultats des études en ce qui concerne les critères de jugement définis préalablement pour répondre aux questions posées. Une analyse descriptive donne les résultats et les explications nécessaires pour comprendre les éventuelles divergences.

Si les résultats sont tous cohérents entre eux, des conclusions peuvent facilement être formulées.

En cas de divergence des résultats, il appartient aux « experts » de pondérer les études en fonction de leur niveau de preuve, de leur nombre, et pour des études de même niveau de preuve en fonction de leur puissance.

► Accord d'experts

En 2010, l'accord d'experts a été précisé lors de l'actualisation des méthodes d'élaboration des recommandations de bonne pratique. L'accord d'experts correspond, en l'absence de données scientifiques disponibles, à l'approbation d'au moins 80 % des membres du groupe de travail.

► Grade des recommandations

Le guide rappelle que :

- une classification des recommandations doit s'adresser aux professionnels destinataires de celles-ci ;
- la classification a pour but d'expliciter les bases des recommandations (volonté de transparence) ;
- la gradation proposée est la même que les recommandations soient d'ordre thérapeutique, diagnostique ; **elle peut se fonder sur plusieurs gradations pour le niveau de preuve des études.**

Les recommandations proposées sont classées en grade A, B ou C selon les modalités suivantes (tableau 2) :

- une recommandation de grade A est fondée sur une **preuve scientifique** établie par des **études de fort niveau de preuve :** PAR EXEMPLE, essais comparatifs randomisés de forte puissance et sans biais majeur, méta-analyse d'essais contrôlés randomisés, analyse de décision fondée sur des études bien menées ;
- une recommandation de grade B est fondée sur une **présomption scientifique** fournie par des **études de niveau intermédiaire de preuve :** PAR EXEMPLE, essais comparatifs randomisés de faible puissance, études comparatives non randomisées bien menées, études de cohortes ;

- une recommandation de grade C est fondée sur des études de **moindre niveau de preuve :** PAR EXEMPLE, études cas-témoin, séries de cas.

En l'absence de précision, les recommandations proposées ne correspondent qu'à un accord d'experts.

L'existence d'une évidence scientifique forte entraîne systématiquement une recommandation de grade A quel que soit le degré d'accord d'experts.

En l'absence d'étude de fort niveau de preuve et d'accord d'experts, les alternatives seront exposées sans formulation de recommandations en faveur de l'une ou de l'autre.

Tableau 2. Grade des recommandations

Grade des recommandations	Niveau de preuve scientifique fourni par la littérature
A Preuve scientifique établie	Niveau 1 - essais comparatifs randomisés de forte puissance ; - méta-analyse d'essais comparatifs randomisés ; - analyse de décision fondée sur des études bien menées.
B Présomption scientifique	Niveau 2 - essais comparatifs randomisés de faible puissance ; - études comparatives non randomisées bien menées ; - études de cohortes.
C Faible niveau de preuve scientifique	Niveau 3 - études cas-témoins.
	Niveau 4 - études comparatives comportant des biais importants ; - études rétrospectives ; - séries de cas ; - études épidémiologiques descriptives (transversale, longitudinale).

Cette **gradation des recommandations** fondée sur le **niveau de preuve scientifique de la littérature** venant à l'appui de ces recommandations **ne présume pas obligatoirement du degré de force de ces recommandations.** En effet, **il peut exister des recommandations de grade C ou fondées sur un accord d'experts néanmoins fortes malgré l'absence d'un appui scientifique.** Les raisons de cette absence de données scientifiques peuvent être multiples (historique, éthique, technique). Ainsi, ce n'est que récemment que des essais thérapeutiques comparatifs ont apporté la preuve scientifique de l'intérêt des digitaliques dans l'insuffisance cardiaque gauche. Avant ces données scientifiques, les recommandations d'utilisation des digitaliques dans l'insuffisance cardiaque gauche étaient néanmoins des recommandations fortes. Il est donc utile de préciser la relation à laquelle on doit s'attendre entre **gradation et hiérarchisation des recommandations.**

L'appréciation de la **force des recommandations** repose donc sur :

- le niveau d'évidence scientifique ;
- l''interprétation des experts.

L'analyse de la littérature permet rarement de répondre à toutes les questions posées. Les recommandations devront explicitement distinguer les réponses soutenues par une évidence scientifique et celles qui ne le sont pas.

1.2 Institut national du cancer, Unicancer[1] et méthode SOR

La loi du 9 août 2004 prévoit, parmi les missions attribuées à l'INCa, celle de définir au niveau national les référentiels de bonnes pratiques et de prise en charge en cancérologie en France. Son champ peut couvrir tous les aspects de la prise en charge : préthérapeutique (diagnostic initial, bilan d'extension), thérapeutique et de surveillance (récidive : détection et traitement).

Sa méthodologie s'appuie sur celle de l'ancien programme des Standards Options et Recommandations[2]. En effet, l'approche combine une revue systématique de la littérature à l'avis d'experts.

Elle permet la production d'une revue systématique (analyse qualitative des données identifiées), à partir de laquelle les experts formulent un avis pondéré par leur expérience pour la formulation de recommandations.

► Synthèse des données scientifiques

Après recherche de la littérature *a priori* pertinente, les études sélectionnées sont exposées sous forme de tableaux pour présenter l'étude et ses principaux résultats.

Les tableaux d'extraction de données sont accompagnés d'un argumentaire. L'argumentaire propose une synthèse des données avec analyse des biais méthodologiques et cliniques.

Cette évaluation s'appuie sur des critères prédéfinis dans des grilles dédiées à chaque type d'études, qu'elles soient relatives au diagnostic ou au traitement :

- méta-analyse et synthèse méthodique ;
- essai randomisé ou non ;
- étude prospective comparative ou non ;
- étude rétrospective comparative ou non ;
- étude « à visée diagnostique ».

L'analyse de la littérature permet la rédaction de conclusions sur les différents critères de jugement retenus par le groupe de travail comme *a priori* pertinents pour formuler des recommandations (ex. : données de survie, taux de réponse, toxicités sévères). Ces conclusions sont associées à un niveau de preuve (tableau 3), pondérées par la validité de ces études. (Lorsqu'il s'agit d'évaluation de biomarqueurs pronostiques et prédictifs de la réponse aux traitements, d'autres systèmes d'attribution de niveaux de preuve aux données de la littérature existent – annexe 4).

Tableau 3. Les niveaux de preuves

Niveau de preuve	Description
Niveau A	Il existe une (des) méta-analyse(s) de bonne qualité ou plusieurs essais randomisés de bonne qualité dont les résultats sont cohérents. De nouvelles données ne changeront très probablement pas la confiance en l'effet estimé.
Niveau B	Il existe des preuves de qualité correcte (essais randomisés [B1] ou études prospectives ou rétrospectives [B2]) avec des résultats dans l'ensemble cohérents. De nouvelles données peuvent avoir un impact sur la confiance dans l'estimation de l'effet, et peuvent changer l'estimation.
Niveau C	Les études disponibles sont critiquables d'un point de vue méthodologique et/ou les résultats des essais ne sont pas toujours cohérents entre eux. De nouvelles données auront très probablement un impact important sur la confiance dans l'estimation de l'effet et changeront probablement l'estimation.

[1] Anciennement Fédération Nationale des Centres de Lutte Contre le Cancer (FNCLCC).

[2] En 2008, les équipes de ce programme ont été mises à disposition de l'INCa. L'Institut National du Cancer a actuellement en charge le pilotage et la diffusion de ces recommandations, avec le soutien financier d'Unicancer.

Niveau de preuve	Description
Niveau D	Il n'existe pas de données ou seulement des séries de cas. Il existe une forte incertitude sur l'effet estimé.

Complétées par l'expertise clinique du groupe de travail, ces conclusions fournissent la base pour la formulation des recommandations.

► **Formulation des recommandations**

Les recommandations reposent sur les meilleures preuves scientifiques disponibles au moment de leur rédaction, pouvant être selon le sujet des méta-analyses, des essais randomisés ou des études non randomisées.
Dans le programme SOR, les recommandations étaient proposées comme des Standards ou des Options (Standard, Options)[3]. Par symétrie, les recommandations proposent également aujourd'hui deux niveaux de gradation :

- par défaut, la recommandation formulée est l'attitude clinique reconnue à l'unanimité comme l'attitude clinique de référence par les experts ;
- si une attitude clinique a été jugée acceptable sur la base des données de la littérature et de l'avis des experts, mais n'est pas reconnue à l'unanimité comme l'attitude clinique de référence, il est indiqué qu'elle peut être discutée, envisagée ou proposée.

1.3 New Zealand Guidelines Group

Selon le *New Zealand Guidelines Group* (NZGG)[4], les recommandations de bonne pratique sont nécessairement fondées sur les meilleures données scientifiques disponibles (3). Il devrait exister des liens explicites entre la force des données scientifiques disponibles et le grade des recommandations.

La gradation est un processus à deux niveaux fondé sur :

- une évaluation objective du type et de la qualité de chaque étude ;
- un jugement qui peut être plus subjectif, sur la cohérence, la pertinence, et l'applicabilité de tout l'ensemble des données scientifiques relatif aux questions auxquelles la recommandation est censée répondre.

Le système d'évaluation intègre les approches du SIGN (13), de l'*Institute for Clinical Systems Improvement* (ICSI) (14) et de l'USPSTF (15).

Le processus d'évaluation comporte trois étapes.

► **Étape 1 :** évaluation des études pertinentes pour les questions de la RBP

Analyse critique. Chaque étude identifiée par la recherche documentaire est analysée de manière systématique pour identifier les biais potentiels. Si des revues systématiques de bonne qualité sont disponibles, il peut ne pas être nécessaire de faire l'analyse critique des études incluses dans la revue systématique. L'analyse est réalisée avec l'outil GATE (*Generic Appraisal Tool for Epidemiology*). Quel que soit son type, une étude peut être évaluée selon cinq composants qui sont : la population, l'exposition, la comparaison, les résultats, la séquence temporelle (PECOT).

Détermination du type d'étude. Le type d'étude optimal varie selon la question clinique (tableau 4).

[3] Standard : attitude clinique reconnue à l'unanimité comme l'attitude clinique de référence par les experts.
Options : plusieurs attitudes cliniques reconnues comme appropriées par les experts. Une option peut avoir la préférence des experts. Lorsque cela est justifié, une des attitudes cliniques proposées peut être d'inclure le patient dans un essai thérapeutique en cours.

[4] Le NZGG a procédé à une liquidation volontaire en 2012.

Tableau 4. Des questions différentes nécessitent des études de conception différente. D'après le NZGG, 2001 (3)

Question clinique	Type d'étude le plus adapté	Critères de jugement
Diagnostic	Étude transversale. Étude de cohorte.	Sensibilité, spécificité, nombre nécessaire de sujets à tester. Taux d'événement attendu pour un patient.
Risques	Essai contrôlé randomisé ou étude de cohorte ou étude cas-témoins.	Nombre de sujets nécessaire pour un événement indésirable.
Intervention	Revues systématiques ou essais contrôlés randomisés.	Réduction du risque absolu. Nombre de sujets nécessaire à traiter.

Les grilles comportent trois sections :

- validité de l'étude (mesures prises pour diminuer les biais) ;
- résultats de l'étude (taille de l'effet et précision) ;
- pertinence de l'étude (applicabilité, généralisabilité).

► **Étape 2 :** cotation de la qualité de chaque étude et confection des tableaux de synthèse des résultats des études

Un score de qualité peut être attribué pour chaque section de la grille (Plus (+), Moins (-), Neutre (∅)) selon la manière dont la conception de l'étude a satisfait les critères (tableau 5).

Tableau 5. Scores de qualité d'une étude. D'après le NZGG, 2001 (3)

Niveau de preuve	Description
Plus (+)	Étude robuste dans laquelle tous ou la plupart des critères de validité sont remplis.
Moins (-)	Étude faible sur le plan de la méthode dans laquelle très peu de critères de validité sont remplis, et il y a un risque de biais élevé.
Neutre (∅)	Étude pour laquelle tous les critères ne sont pas remplis, mais les résultats de l'étude ne sont probablement pas affectés.

Le NZGG ne donne pas de critères explicites pour attribuer +, -, ou ∅ à une étude.

Le type d'étude et le score de qualité pour les 3 sections sont saisis dans un tableau de synthèse des données scientifiques (*evidence table*).

► **Étape 3 :** développer des recommandations gradées à partir de l'ensemble des données scientifiques

Les recommandations sont développées par le groupe de travail en considérant l'ensemble des données scientifiques, résumées dans des tableaux de synthèse. L'ensemble des données scientifiques est la somme des données de chaque étude, de la cotation de la qualité de chacune des études ainsi que la pertinence de chaque étude par rapport à la question posée.

Avant de rédiger des recommandations, il est nécessaire d'évaluer l'ensemble des données scientifiques relatives à la question comme un tout. Il est nécessaire de considérer toutes les données scientifiques en donnant un poids plus grand aux études de qualité élevée. Le processus oblige à faire preuve de jugement fondé sur l'expérience autant que sur la connaissance des données

scientifiques et des méthodes utilisées pour les produire. Il est nécessaire d'associer la qualité de toutes les données scientifiques avec les autres aspects, ce qui requiert un jugement raisonné. Quelques uns des facteurs les plus importants à considérer lors de la rédaction des recommandations sont les suivants (13) :

- quantité (nombre d'études et nombre de participants) et cohérence des résultats entre les études ;
- applicabilité (recommandations fondées sur les meilleures données scientifiques directement applicables au contexte des soins, ou sur des résultats extrapolés d'essais cliniques étrangers) ;
- impact clinique : le bénéfice potentiel d'une intervention est-il suffisamment grand pour justifier une recommandation ? Il dépend de la taille de l'effet et du rapport bénéfice-risque.

Le groupe de travail rédige un énoncé sommaire, fondé sur la synthèse de l'ensemble des données scientifiques relatives à la question posée, et lui attribue un score de qualité des données scientifiques (adapté du SIGN).

La recommandation peut être rédigée à partir de ces énoncés. La rédaction des recommandations implique des considérations qui vont au-delà des données scientifiques et oblige à faire preuve de jugement. Augmenter le rôle du jugement subjectif dans la rédaction des recommandations accroît nécessairement le risque de réintroduire des biais dans le processus. Ce risque est minimisé si les décisions sont fondées sur un consensus du groupe de travail.

La force des données scientifiques considérées dans ce processus de gradation est liée à la question posée.

Le grade de la recommandation est fondé sur les facteurs suivants :

- le type et la qualité des études individuelles identifiées pour répondre à la question posée (converties en un énoncé récapitulatif des données scientifiques reflétant l'ensemble des données scientifiques) ;
- la quantité, la cohérence, l'applicabilité et l'impact clinique de l'ensemble des données scientifiques ;
- le consensus du groupe de travail.

Le grade d'une recommandation est déterminé par la force des données scientifiques venant à l'appui de la recommandation (tableau 6).

Tableau 6. Gradation des recommandations. D'après le NZGG, 2001 (3)

Grade	Description
A	La recommandation (conduite à tenir) est supportée par de bonnes données scientifiques. Les données scientifiques sont composées des résultats d'études de conception robuste pour répondre à la question posée.
B	La recommandation (conduite à tenir) est supportée par des données scientifiques assez bonnes. Les données scientifiques sont composées des résultats d'études de conception robuste pour répondre à la question posée, mais il existe quelques incertitudes sur la conclusion soit à cause d'une hétérogénéité des résultats des études, soit à cause de biais mineurs ; ou les données scientifiques sont composées des résultats d'études de conception moins solide pour la question posée, mais les résultats ont été confirmés dans des études différentes et sont assez cohérents. Il y a des données scientifiques assez bonnes sur le fait que les bénéfices de la conduite à tenir proposée l'emportent sur les risques.
C	La recommandation (conduite à tenir) est supportée uniquement par l'avis

Grade	Description
	d'experts. Pour plusieurs résultats, des essais ou des études ne peuvent pas ou n'ont pas pu être réalisées et la pratique est renseignée uniquement par l'avis d'experts.
I	Aucune recommandation ne peut être faite parce que les données scientifiques sont insuffisantes. Les données scientifiques manquent, ou sont de qualité médiocre, ou elles sont contradictoires, et le rapport bénéfices risques ne peut pas être déterminé.

1.4 *American academy of pediatrics*

La procédure d'élaboration de RBP fondées sur des données scientifiques comporte trois étapes (11) :

- détermination de la qualité des données scientifiques venant à l'appui d'une recommandation ;
- évaluation du rapport bénéfices inconvénients attendu ;
- attribution d'une force à une recommandation.

► Détermination de la qualité des données scientifiques

► Évaluation de la qualité d'une étude

La qualité d'une étude est évaluée sur la conception de l'étude et la rigueur de la méthode de réalisation (tableau 7). Les critères de qualité spécifiques appliqués dépendent du type d'étude et de sa conception.

Tableau 7. Qualité des données scientifiques d'après l'AAP, 2004 (11)

Qualité des données scientifiques	Interventions	Tests diagnostiques
Élevée	Essais contrôlés randomisés bien conçus et bien menés, réalisés sur un groupe issu d'une population similaire à la population cible de la RBP.	Qualité jugée sur : - la représentativité de la population étudiée ; - la description adéquate du test ; - le bien-fondé du test de référence ; - les méthodes utilisées pour éviter les biais d'interprétation.
Intermédiaire	Essais contrôlés randomisés avec des biais non rédhibitoires, ou des limites liées à la méthode (par exemple, réalisé sur un groupe issu d'une population différente de la population cible, et nécessitant une extrapolation des résultats) Études observationnelles : - études de cohortes ; - études cas-témoins.	
Faible	Étude de cas unique, raisonnement issu de principes physiopathologiques, avis	

Qualité des données scientifiques	Interventions	Tests diagnostiques
	d'experts.	

► Évaluation de la qualité des études regroupées portant sur la question

Juger de la force d'un ensemble de données scientifiques nécessite de considérer la cohérence des résultats des études, la taille de l'effet estimé dans les études, et la taille des échantillons de populations individuels et regroupés.

► Évaluation du rapport bénéfices inconvénients attendu

Le second facteur qui influence la force d'une recommandation est constitué par les bénéfices, les inconvénients, les risques et le coût attendus de l'adhésion à une recommandation.

- Quand les données scientifiques indiquent un bénéfice net, non compensé par des inconvénients ou des coûts importants ou des inconvénients nets non atténués par un bénéfice important, des recommandations plus fortes sont possibles.
- Quand le bénéfice est faible ou que les bénéfices sont présents, mais compensés par des effets secondaires importants, l'équilibre entre les bénéfices et les inconvénients empêche une recommandation forte.

Une prépondérance nette du bénéfice ou des inconvénients supporte des recommandations plus forte pour ou contre une conduite à tenir.

Quand le rapport bénéfice-risque est équilibré, peu importe la qualité des études, les médecins devraient offrir des options plutôt que des recommandations.

► Attribution d'une force à une recommandation

Par la force d'une recommandation, le groupe de travail indique l'importance de l'adhésion à une recommandation particulière. Elle est fondée sur la qualité des études venant à l'appui de la recommandation et sur l'ampleur du bénéfice et des inconvénients potentiels.

La classification proposée comporte quatre niveaux : recommandation forte, recommandation, option et pas de recommandation (tableau 8). Elle est fondée sur :

- quatre niveaux de qualité des données scientifiques : A, B, C, D ;
- deux catégories du rapport bénéfice-inconvénients : soit une prépondérance nette du bénéfice ou des inconvénients, soit un équilibre relatif entre les bénéfices et les inconvénients ;
- une catégorie pour des recommandations dans des situations exceptionnelles dans lesquelles des données scientifiques ne peuvent pas être obtenues, mais des bénéfices ou des inconvénients sont nets.

► Interprétation des recommandations

Une « recommandation forte » signifie que les bénéfices de l'approche recommandée excèdent nettement les inconvénients (ou le contraire pour les recommandations fortes négatives), et que la qualité des données scientifiques supportant cette recommandation est soit excellente, soit impossible à obtenir. Les cliniciens devraient suivre une recommandation de ce type, à moins qu'il existe une raison claire et incontestable de faire le contraire.

Une « recommandation » signifie que les bénéfices excèdent les inconvénients (ou le contraire pour les recommandations négatives), mais que la qualité des données scientifiques venant à l'appui de la recommandation n'est pas aussi forte. Les cliniciens devraient généralement suivre une recommandation de ce type, mais ils devraient aussi être attentifs aux nouvelles informations et sensibles aux préférences des patients.

Une « option » signifie soit que la qualité des données scientifiques est discutable, soit que des études bien conçues et bien menées ont montré un petit avantage net d'une approche par rapport à l'autre. Les options offrent une flexibilité aux cliniciens dans leur prise de décision concernant la pratique appropriée, bien qu'elles puissent limiter les alternatives. Les préférences des patients devraient avoir un rôle important en influençant la prise de décision clinique.

La catégorie « pas de recommandation » est attribuée quand il y a à la fois un manque de données scientifiques pertinentes et un rapport bénéfices inconvénients incertain. Les préférences des patients devraient avoir un rôle important en influençant la prise de décision clinique.

Tableau 8. Classification de la force des recommandations d'après l'*American academy of pediatrics*, 2004 (11)

<table>
<tr><th>Qualité des données scientifiques</th><th>Prépondérance des bénéfices ou des inconvénients</th><th>Équilibre des bénéfices et des inconvénients</th></tr>
<tr><td>A. Essais contrôlés randomisés ou études diagnostiques bien conçues sur des populations pertinentes</td><td>Recommandation forte</td><td rowspan="3">Option</td></tr>
<tr><td>B. Essais contrôlés randomisés ou études diagnostiques avec des limitations mineures ; données scientifiques cohérentes à la grande majorité</td><td rowspan="2">Recommandation</td></tr>
<tr><td>C. Études observationnelles (étude cas-témoins, étude de cohorte)</td></tr>
<tr><td>D. Avis d'experts, observations, raisonnement à partir des principes physiopathologiques de base</td><td>Option</td><td>Pas de recommandation</td></tr>
<tr><td>X. Situations exceptionnelles dans lesquelles des études de validation ne peuvent pas être réalisées, et il y a une nette prépondérance du bénéfice ou des inconvénients</td><td>Recommandation forte
Recommandation</td><td></td></tr>
</table>

1.5 *The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group*

Le GRADE *working group* a débuté en 2000 comme une collaboration informelle de personnes intéressées par l'évaluation des défauts des systèmes de gradation actuels dans les soins de santé.

En 2004, ce groupe a publié une étude rapportant les résultats d'une analyse critique de six systèmes de gradation des données scientifiques et des recommandations (annexe 5) (16). À l'issue des discussions qui ont suivi l'analyse, les conclusions ont été les suivantes :

- il est conseillé de présenter des évaluations séparées pour juger de la qualité des données scientifiques et du rapport bénéfices inconvénients ;

- les données scientifiques sur les inconvénients devraient être évaluées de la même manière que les données scientifiques sur les bénéfices, bien que des données scientifiques différentes puissent être considérées pertinentes pour les inconvénients ;
- les jugements sur la qualité des données scientifiques devraient être fondés sur une revue systématique de la recherche clinique pertinente ;
- les revues systématiques ne devraient pas être incluses dans une hiérarchie du niveau de preuve (comme niveau ou catégorie de preuve). La disponibilité d'une revue systématique bien menée n'est pas équivalente à des données scientifiques de qualité élevée, puisqu'une revue systématique bien menée peut inclure tout, d'aucune étude à des études de qualité médiocre avec des résultats incohérents, à des études de qualité élevée avec des résultats cohérents ;
- le risque de base devrait être pris en considération en déterminant la population à laquelle une recommandation s'applique. Il devrait être utilisé de façon transparente quand on porte des jugements sur le rapport bénéfices-inconvénients. Quand une recommandation varie en fonction du risque de base, les données scientifiques servant à déterminer le risque de base devraient être évaluées de façon explicite. Les recommandations ne devraient pas varier en fonction du risque de base s'il n'y a pas de données scientifiques adéquates pour déterminer le risque de base de façon fiable.

La même année, le GRADE *working group* a publié une étude pilote de son système de gradation de la qualité des données scientifiques et de la force des recommandations (17) suivie de la description *princeps* de ce système (18).

La description du système GRADE présentée ci-dessous a été rédigée à partir des articles à ce sujet publiés en 2004 (18), 2008 (10) et 2010 (19).

Toute question concernant une prise en charge clinique comporte quatre composants : la population de patients, l'intervention d'intérêt, le comparateur, et les résultats d'intérêt (PICO) (10). Une question implique souvent une autre précision : le contexte des soins dans lequel la recommandation sera mise en œuvre (19).

Les auteurs incitent ceux qui élaborent des recommandations à préciser, en début de projet, tous les résultats importants par rapport au patient (bénéfices, inconvénients et coûts) et à distinguer parmi ceux-ci, les résultats décisifs. Ils proposent l'utilisation d'une échelle de 1 à 9 pour juger de l'importance des résultats. (7 – 9 : résultats décisifs ; 4 – 6 : résultats importants mais non décisifs ; 1 – 3 : résultats d'importance limitée).

► Définitions de la qualité des données scientifiques

Dans le contexte de l'élaboration de recommandations, la qualité des données scientifiques reflète notre confiance dans le fait qu'une estimation de l'effet est adéquate pour supporter une recommandation ou une décision.

Pour une revue systématique, la qualité des données scientifiques reflète notre confiance dans le fait qu'une estimation de l'effet est correcte (10,19).

► Qualité des données scientifiques pour chaque résultat important

Le système GRADE est centré sur les résultats.

La qualité des donnés scientifiques pour chaque résultat important peut être déterminée après avoir considéré le type d'études, la qualité des études, l'homogénéité des résultats, le caractère direct des données scientifiques (18).

Pour déterminer la qualité des données scientifiques, le système GRADE **part du type d'étude**. Il classe **initialement** les données en se fondant sur le type d'étude dont elles sont issues. Il distingue deux catégories :

- les essais contrôlés randomisés qui fournissent généralement des données scientifiques de qualité **élevée ;**
- les études observationnelles qui fournissent généralement des données scientifiques de qualité **faible.**

Puis il s'agit de considérer si les études ont des limites sérieuses, s'il y a une hétérogénéité importante des résultats, et si des doutes sur le caractère direct des données sont justifiés.

La définition des niveaux de qualité des données scientifiques pour chaque résultat important est présentée dans le tableau 9. Il s'agit de la qualité des données scientifiques pour chaque résultat important dans toutes les études (*i.e.* : de la qualité d'un ensemble de données scientifiques). Cela ne signifie pas évaluer le niveau de chaque étude individuellement (18,19).

Tableau 9. Niveaux de qualité des données scientifiques pour chaque résultat important d'après Balshem *et al*., 2011 (19)

Niveau de qualité	Définition*
Élevé	**Nous avons une confiance élevée dans l'estimation de l'effet : celle-ci doit être très proche du véritable effet.**
Modéré	Nous avons une confiance modérée dans l'estimation de l'effet : celle-ci est probablement proche du véritable effet, mais il est possible qu'elle soit nettement différente.
Faible	**Nous avons une confiance limitée dans l'estimation de l'effet : celle-ci peut être nettement différente du véritable effet.**
Très faible	Nous avons très peu confiance dans l'estimation de l'effet : il est probable que celle-ci soit nettement différente du véritable effet.

* : ancienne définition des niveaux de qualité d'après Atkins *et al.*, 2004 (18) :

- élevé : il est très improbable que des études futures changent la confiance que nous avons dans l'estimation de l'effet ;

- modéré : il est probable que des études futures aient un impact important sur la confiance que nous avons dans l'estimation de l'effet et qu'elles puissent changer l'estimation de l'effet ;

- faible : il est très probable que des études futures aient un impact important sur la confiance que nous avons dans l'estimation de l'effet et il est probable qu'elles changent l'estimation de l'effet ;

- très faible : toute estimation de l'effet est très incertaine.

► Cinq facteurs peuvent diminuer la qualité des données scientifiques issues d'études observationnelles et d'essais contrôlés randomisés

Un **risque de biais** (anciennement dénommé « limites des études » pouvant biaiser leur estimation de l'effet du traitement) : par exemple, si toutes les études disponibles ont des limites sérieuses, le niveau de qualité des données scientifiques pour le résultat considéré peut être diminué d'un niveau (tableau 10), et si toutes les études ont des limites très sérieuses, le niveau peut être diminué de deux.

La qualité de l'étude se rapporte à un examen détaillé de la méthode de l'étude et de sa réalisation. Ceux qui font l'analyse de la littérature devraient utiliser des critères appropriés afin d'évaluer la qualité de chaque étude pour chaque résultat important. Par exemple, pour les essais contrôlés randomisés : assignation au hasard des patients dans les groupes, insu, suivi, analyse des résultats en intention de traiter et plus récents, l'arrêt précoce de l'essai pour un bénéfice apparent et la publication sélective des résultats (20) ; pour les études observationnelles, mesure adéquate de l'exposition et des résultats et contrôle adapté des facteurs de confusion ; et dans les deux types d'études, prise en compte des perdus de vue. Ils devraient expliquer leurs raisons pour rétrograder une étude.

Une **hétérogénéité** des résultats : de grandes différences de l'estimation de l'effet entre les études (en rapport avec une hétérogénéité ou une variabilité des résultats) suggère des différences dans l'effet du traitement.

Le **caractère indirect** des données scientifiques : soit il s'agit de données scientifiques obtenues par des comparaisons indirectes, soit il y a des différences entre la population, l'intervention, l'intervention de comparaison, les résultats, *d'intérêt* et ceux *des études sélectionnées pour la question donnée*.

Une **imprécision** des données : quand les études incluent relativement peu de patients et peu d'événements et ont des intervalles de confiance larges.

Un **biais de publication**.

- Trois facteurs peuvent augmenter la qualité des données scientifiques issues d'études observationnelles

La force de l'association.

Un gradient dose-réponse.

La présence de facteurs de confusion plausibles qui auraient réduit l'effet observé.

L'évaluation de ces facteurs pouvant intervenir sur le niveau de qualité des données scientifiques a été détaillée dans une série d'articles (20-25).

Tableau 10. Facteurs qui influencent la cotation de la qualité des données scientifiques d'après Atkins *et al.*, 2004 (10,18,19)

Facteurs	Description du facteur	**Niveau de qualité de départ** (nombre de niveaux en moins ou en plus)
Type d'études	essais contrôlés randomisés	**élevé**
	études observationnelles*	**faible**
Facteurs qui peuvent diminuer le niveau de qualité des données scientifiques provenant d'études observationnelles et d'essais contrôlés randomisés	**risque de biais**	
	sérieux	(-1)
	très sérieux	(-2)
	hétérogénéité des résultats	
	importante	(-1)
	très importante	(-2)
	caractère direct des données	
	incertitude	(-1)
	incertitude majeure	(-2)
	imprécision	
	sérieuse	(-1)
	très sérieuse	(-2)
	biais de publication	
	probable	(-1)
	très probable	(-2)
Facteurs qui peuvent augmenter	**force de l'association**	

Facteurs	Description du facteur	**Niveau de qualité de départ** (nombre de niveaux en moins ou en plus)
le niveau de qualité des données scientifiques provenant d'études observationnelles	Données scientifiques solides d'une association – risque relatif significatif > 2 (< 0,5) fondé sur des données cohérentes issues d'au moins deux études observationnelles, sans facteurs de confusion plausible.	(+1)
	Données scientifiques très solides d'une association – risque relatif significatif > 5 (< 0,2) fondé sur des données directes, sans problème majeur de validité.	(+2)
	Données d'un gradient dose-réponse	(+1)
	Présence de facteurs de confusion plausibles	
	qui auraient réduit l'effet observé (ces facteurs n'ayant pas été pris en compte dans l'analyse avec ajustement) ;	(+1)
	qui auraient fait s'attendre à un effet alors que les résultats ne montrent aucun effet.	(+1)

* : étude observationnelle : études de cohortes, études cas-témoins, analyses de séries interrompues, études contrôlées avant-après.

Ces considérations sont **cumulatives** (18). Par exemple, si des essais contrôlés randomisés ont à la fois des limites sérieuses et s'il existe une incertitude sur le caractère direct des données scientifiques, le niveau des données scientifiques sera réduit de élevé à faible.

Dans le système GRADE, la catégorie « accord d'experts » n'existe pas. L'élaboration de recommandations nécessite toujours l'avis d'experts aussi bien que les résultats d'essais contrôlés randomisés ou d'études observationnelles. Ceux qui élaborent des recommandations devraient exprimer clairement les données scientifiques qui sous-tendent l'avis des experts, et coter la qualité de ces données scientifiques.

► Qualité des données scientifiques dans leur ensemble

► Les résultats décisifs déterminent la cotation de la qualité des données scientifiques de l'ensemble des résultats

Les recommandations dépendent des données scientifiques pour les résultats importants (incluant des bénéfices et des inconvénients) et de la qualité des données scientifiques pour chacun de ces résultats.

Comment coter la qualité des données scientifiques de l'ensemble des résultats si cette qualité diffère ?

L'approche du GRADE *working group* suggère que la qualité des données scientifiques de l'ensemble des résultats pour une question est celle du résultat décisif ayant les données scientifiques de la qualité la plus faible (10).

Le niveau de qualité des données scientifiques pour chaque recommandation est le plus faible niveau de cotation des résultats décisifs (à distinguer des résultats importants mais non décisifs) (18,26).

Si toutes les données scientifiques pour tous les résultats décisifs sont en faveur de la même alternative, et s'il y a des données scientifiques de qualité élevée pour quelques uns mais pas pour tous ces résultats, la qualité globale des données scientifiques pourrait encore être considérée élevée (18).

► Profil des données scientifiques et résumé des résultats

Un **profil des données scientifiques** comprend une évaluation détaillée de la qualité des données et un **résumé des résultats** (tableau 11). Un profil des données scientifiques inclut un jugement explicite de chaque facteur qui détermine la qualité des données scientifiques pour chaque résultat important et la taille de l'effet pour chaque résultat important.

Tableau 11. Profil des données scientifiques et résumé des résultats

Évaluation de la qualité						Résumé des résultats					
						Nombre de patients			Risque absolu		
Nombre d'études (Type)	Limites	Hétérogénéité	Caractère direct des données	Imprécision	Biais de publication	Témoins	Intervention	Risque relatif IC95 %	Groupe contrôle	Différence de risque	Qualité
Résultat											
Résultat											

► Recommandations

Selon les auteurs, une qualité particulière des données scientifiques n'implique pas forcément une catégorie particulière de recommandation.

La force d'une recommandation reflète la confiance que l'on peut avoir dans le fait que les effets souhaitables d'une intervention l'emportent sur les effets indésirables.

Les effets souhaitables d'une recommandation incluent : une réduction de la morbidité et de la mortalité, une amélioration de la qualité de vie, une réduction de la lourdeur du traitement (par exemple : prendre des médicaments, ou inconvénients des prises de sang), et une utilisation diminuée des ressources.

Les effets indésirables incluent : un impact délétère sur la morbidité, la mortalité, ou la qualité de vie, ou une utilisation augmentée des ressources.

Le système GRADE comporte deux catégories de recommandations : forte et faible (tableau 12).

Tableau 12. Catégories de recommandations d'après le GRADE *working group*, 2008 (10)

Recommandation	Description
Forte	Quand le groupe de travail est confiant dans le fait que les effets souhaitables de l'adhésion à une recommandation l'emportent sur les effets indésirables.
Faible	Indique que les effets souhaitables de l'adhésion à une recommandation l'emportent probablement sur les effets indésirables, mais le groupe de travail est moins confiant.

Cette classification binaire fournit des indications claires aux patients, aux cliniciens et aux décideurs (tableau 13).

Tableau 13. Implication de la force d'une recommandation, selon le public considéré d'après le GRADE *working group*, 2008 (10)

Public	Recommandation forte	Recommandation faible
Patient	La plupart des personnes dans votre situation choisirait la conduite à tenir recommandée et seulement une petite proportion ne la choisirait pas*.	La plupart des personnes dans votre situation choisirait la conduite à tenir recommandée, mais de nombreuses personnes ne la choisiraient pas †.
Clinicien	La plupart des patients devrait être traités selon la conduite à tenir recommandée.	Vous devriez reconnaître que différents choix seront appropriés pour des patients différents et que vous devez aider chaque patient à parvenir à une décision de prise en charge cohérente avec ses valeurs et ses préférences.
Décideur	La recommandation peut être adoptée comme une mesure dans la plupart des situations.	Les décisions nécessiteront des débats et une implication de nombreuses parties prenantes.

* : un outil d'aide à la décision n'est pas nécessaire, presque tous les patients informés feraient le même choix.

† : un outil d'aide à la décision pourrait être utile.

Les options de prise en charge associées à des recommandations fortes peuvent servir de base pour des critères d'évaluation de la qualité.

Quatre facteurs clés déterminent la force d'une recommandation (tableau 14).

Les recommandations impliquent un arbitrage entre des bénéfices et des inconvénients. Faire cet arbitrage implique de placer, de façon implicite ou explicite, une valeur relative sur chaque résultat. Il est souvent difficile de juger quel poids donner aux différents résultats, et différentes personnes auront souvent des valeurs différentes. Est-ce que l'intervention fait plus de bien que de mal ?

Tableau 14. Facteurs déterminant la force d'une recommandation d'après le GRADE *working group*, 2004 et 2008 (10,18)

Facteur	Description
Rapport bénéfices inconvénients	Arbitrage prenant en compte la taille estimée de l'effet pour les principaux résultats, l'intervalle de confiance de ces estimations, et la valeur relative placée sur chaque résultat. Quand les avantages l'emportent largement sur les inconvénients, il est vraisemblable qu'une recommandation forte est justifiée. Quand les avantages et les inconvénients s'équilibrent, une recommandation faible paraît justifiée.
Qualité des données scientifiques	Si nous sommes incertains de l'ampleur des bénéfices et des inconvénients d'une intervention, faire une recommandation pour ou contre une conduite à tenir pose un problème. Plus la qualité des données est élevée, plus il est vraisemblable qu'une recommandation forte est justifiée.
Incertitude sur / ou variabilité des valeurs et des préférences	Étant donné que les stratégies alternatives de prise en charge ont toujours des avantages et des inconvénients, et qu'un arbitrage existe, les valeurs d'un groupe de travail sur les bénéfices, les risques et les inconvénients sont décisives sur la force

Facteur	Description
	d'une recommandation.
Coût	Les coûts élevés réduisent la vraisemblance d'une recommandation forte en faveur d'une intervention, cependant le contexte de l'intervention est décisif. Le groupe de travail doit spécifier le contexte clinique auquel la recommandation s'applique.

► Utilisation du système GRADE pour les tests ou les stratégies diagnostiques

Un groupe de travail considérant un test ou une stratégie diagnostique devrait commencer par identifier les patients, le test ou la stratégie diagnostique, le comparateur, et les résultats d'intérêt. **En termes de résultats importants pour le patient, la validité diagnostique d'un test est un résultat intermédiaire ou de substitution.**

La meilleure manière d'évaluer une stratégie diagnostique est un essai contrôlé randomisé, dans lequel les investigateurs randomisent les patients en deux groupes (approche diagnostique expérimentale ou contrôle) et mesurent la mortalité, la morbidité, les symptômes et la qualité de vie.

Quand des études comportant un diagnostic suivi d'une intervention (voir § 1.10 NICE, études « *Teste et traite* ») sont disponibles (idéalement des essais contrôlés randomisés ou des études observationnelles) comparant l'impact des stratégies diagnostiques alternatives sur des résultats importants pour le patient, il est possible d'utiliser l'approche GRADE décrite précédemment.

Quand ces études ne sont pas disponibles, le groupe de travail doit se centrer sur des études portant sur la validité diagnostique des tests, et faire des inférences relatives à son impact probable sur les résultats importants pour le patient (27). Les questions clés sont : y aura-t-il une réduction des faux négatifs ou des faux positifs et une augmentation correspondante des vrais positifs et des vrais négatifs ? Des patients semblables ou différents sont-ils classés de manière fiable par les stratégies diagnostiques alternatives ? Quels sont les résultats chez les patients classés comme cas et ceux classés comme n'ayant pas la maladie ?

► Jugements sur la qualité des données scientifiques à l'appui des recommandations

Les quatre niveaux de qualité des données scientifiques représentent un gradient de confiance dans l'estimation de l'effet d'un test ou d'une stratégie diagnostique sur des résultats importants pour le patient.

Des études portant sur la validité diagnostique des tests sont classées initialement dans la catégorie qualité élevée. Cependant, beaucoup de ces études apportent au final des données scientifiques de qualité faible pour étayer les recommandations, car il s'agit de données scientifiques indirectes de l'impact du test sur les résultats importants pour le patient (tableau 15).

Tableau 15. Facteurs qui influencent la qualité des données scientifiques des études sur la validité diagnostique des tests d'après le GRADE *working group*, 2008 (28)

Facteur	Explication
Type d'études	Les données scientifiques issues d'études transversales ou les études de cohortes, incluant des patients chez lesquels il existe des incertitudes diagnostiques (*i.e.* : patients auxquels les cliniciens appliqueraient le test en pratique courante), impliquant une comparaison directe entre le test considéré et un test de référence adapté (*gold standard*) sont considérées de qualité élevée.
Risques de biais (limites)	Des patients consécutifs devraient être recrutés comme une seule cohorte et non classés par stade de la maladie, et la manière dont les patients ont été

<table>
<tr><th>Facteur</th><th>Explication</th></tr>
<tr><td></td><td>adressés ainsi que leur sélection doivent être clairement décrits.
Les tests devraient être réalisés chez tous les patients d’une même population de patients pour le nouveau test et le test de référence, par des investigateurs en insu des résultats de l’autre test.</td></tr>
<tr><td>Caractère indirect des données scientifiques</td><td></td></tr>
<tr><td>- résultats</td><td>Il existe souvent une absence de données scientifiques directes de l’impact du test sur des résultats importants pour le patient.
Le groupe de travail doit faire des déductions à partir des études portant sur la validité des tests pour ce qui est du rapport entre :
- d’une part les influences présumées des différences dans les VP, FP, VN, FN sur résultats importants pour le patient ;
- et d’autre part les complications et les coûts du test.

<table>
<tr><td colspan="3">Exemple de résultats d’un test et son impact attendu sur la prise en charge thérapeutique.</td></tr>
<tr><td></td><td>malades</td><td>non malades</td></tr>
<tr><td>test positif</td><td>vrais positifs
administration d’un traitement d’efficacité connue.</td><td>faux positifs
probable anxiété et possible morbidité liée aux examens supplémentaire et au traitement.</td></tr>
<tr><td>test négatif</td><td>faux négatifs
conséquences possible du retard de diagnostic (ne reçoit pas les traitements disponibles efficaces).</td><td>vrais négatifs
épargne au patient des effets secondaires potentiels du test de référence,
patient rassuré.</td></tr>
</table>

C’est pourquoi le plus souvent, les études portant sur la validité diagnostique des tests n’apportent que des données scientifiques de qualité faible pour faire des recommandations en raison du caractère indirect des résultats (sur des critères de jugement intermédiaire ou de substitution).</td></tr>
<tr><td>- populations, test diagnostique, test de comparaison et comparaisons indirectes</td><td>La qualité des données scientifiques peut être diminuée :
- s’il existe des différences importantes :
. entre la population étudiée et la population cible de la recommandation ;
. dans les tests étudiés et dans l’expertise diagnostique de ceux qui réalisent les tests dans les études comparé à ceux qui les réalisent dans l’environnement clinique cible de la recommandation ;
- si chacun des tests examinés est comparé au même test de référence dans des études différentes, et ne sont pas comparés directement dans les mêmes études.</td></tr>
<tr><td>Hétérogénéité importante des résultats</td><td>Une hétérogénéité inexpliquée de la sensibilité et de la spécificité peut diminuer la qualité des données scientifiques.</td></tr>
<tr><td>Probabilité élevée de biais de publication</td><td>Données scientifiques issues de petites études pour des tests nouveaux ou graphique en funnel plot.</td></tr>
</table>

Facteur	Explication

VP : vrais positifs ; FP : faux positifs ; VN : vrais négatifs ; FN : faux négatifs.

► Recommandations

La balance entre les résultats présumés importants pour le patient, comme le résultat des vrais et des faux positifs et négatifs et les complications du test détermine si le groupe de travail fait une recommandation en faveur ou en défaveur de l'utilisation du test évalué. Les autres facteurs qui influencent la force d'une recommandation incluent : la qualité des données scientifiques, l'incertitude sur les valeurs et les préférences associées aux tests et aux résultats présumés importants pour le patient, et les coûts.

1.6 *Scottish Intercollegiate Guidelines Network*

Les RBP du SIGN sont fondées sur une revue systématique de la littérature (4). Une revue systématique est définie comme « une technique scientifique efficace pour identifier et résumer les données scientifiques concernant l'efficacité réelle des interventions et qui permet d'évaluer la généralisabilité des résultats des études et leur cohérence, et d'explorer des données discordantes ».

L'approche du SIGN est caractérisée par la forte implication des membres du groupe de travail dans la synthèse des données scientifiques, ce qui leur permet d'exercer leur « jugement raisonné » quand ils dégagent/ en tirent les recommandations.

► Évaluation des données scientifiques

Une fois que les articles ont été sélectionnés comme source potentielle de données scientifiques, la méthode de chaque étude est évaluée pour s'assurer de sa validité. Cette évaluation de la méthode est fondée sur un nombre de questions clés, centrées sur les aspects du type de l'étude dont on connaît l'influence sur la validité des résultats et des conclusions. Ces questions clés diffèrent selon le type d'étude. Le SIGN a fondé son évaluation des études sur la méthode MERGE (*Method for Evaluating Research and Guideline Evidence*).

L'évaluation globale d'une étude est fondée sur la qualité de la méthode que l'on code (tableau 16).

Le code attribué à la qualité de la méthode, couplé au type de l'étude, fait décider du niveau de preuve que fournit cette étude.

Tableau 16. Qualité de la méthode d'une étude. D'après le SIGN, 2008 (4)

Code de qualité globale	Description
++	Tous ou la plupart des critères sont remplis. Il est **très peu probable** que les conclusions de l'étude ou de la revue soient affectées par les critères non remplis.
+	Plusieurs critères sont remplis. Il est peu probable que les conclusions de l'étude soient affectées par les critères non remplis ou non décrits de manière adéquate.
-	Peu ou aucun des critères ne sont remplis. Il est **probable** ou **très probable** que les conclusions de l'étude en soient affectées.

Le processus d'évaluation implique une part inévitable de jugement subjectif. La façon dont une étude remplit un critère particulier et l'impact probable de ceci sur les résultats rapportés dépen-

dent du contexte clinique. Pour minimiser tout biais potentiel, chaque étude doit être évaluée de façon indépendante par au moins deux individus. Toute différence dans l'évaluation doit être discutée par l'ensemble du groupe de travail.

► Synthèse des données scientifiques

Les recommandations sont gradées pour différencier celles fondées sur des données scientifiques robustes de celles fondées sur des données scientifiques peu solides. Ce jugement est porté sur la base d'une évaluation (objective) du type et de la qualité de chaque étude, et d'un jugement (peut-être plus subjectif) sur l'homogénéité, la pertinence clinique et la généralisabilité de l'ensemble des données scientifiques. Le but est de produire une recommandation fondée sur les données scientifiques, mais qui est pertinente avec la manière dont les soins sont délivrés en Écosse, et qui va pouvoir être mise en œuvre.

Le grade d'une recommandation ne se rapporte pas à l'importance clinique de la recommandation, mais à la force des données scientifiques sur lesquelles la recommandation est fondée, en particulier, à la valeur prédictive des types d'études dont les données sont issues. **Le grade attribué à une recommandation indique aux utilisateurs la probabilité que le résultat prévu soit atteint si la recommandation est mise en œuvre.**

Les tableaux de synthèse des données scientifiques (*evidence table*) sont remplis par le SIGN à partir des évaluations de la qualité des études individuelles fournies par les membres du groupe de travail (tableau 17). Les tableaux résument toutes les études validées, identifiées à partir de la revue systématique de la littérature pour chaque question. Ils présentent de manière séparée les données scientifiques pour chaque critère de jugement utilisé dans les études publiées.

Tableau 17. Tableau de synthèse des données scientifiques pour des études d'intervention d'après le SIGN, 2008 (4)

Question :										
Référence	Type d'étude	Niveau de preuve	Nombre de patients	Caractéristiques des patients	Intervention	Comparaison	Durée du suivi	Critères de jugement	Taille de l'effet	Source de financement

► Jugement raisonné

Il est rare que les données scientifiques montrent clairement et sans ambiguïté la procédure qui devrait être recommandée pour n'importe quelle question donnée. Il n'est pas toujours clair pour ceux qui n'ont pas été impliqués dans le processus d'élaboration, comment le groupe de travail est arrivé à sa recommandation, étant donné les données scientifiques sur lesquelles il a dû s'appuyer.

Sous la rubrique jugement raisonné, le groupe de travail résume son avis sur tout l'ensemble des données scientifiques couvert par chacun des tableaux de synthèse des données scientifiques (*evidence table*). Cet avis est divisé en trois parties (tableau 18).

► Juger le niveau de preuve

À cette première étape, le groupe de travail donne ses commentaires sur :

- la quantité, la qualité et la cohérence des données scientifiques ;
- la généralisabilité des résultats ;
- l'applicabilité directe à la population cible de la recommandation de bonne pratique.

Et il note les niveaux de preuve globaux concernant une question spécifique.

► Juger l'impact des données scientifiques

À l'étape suivante, on demande au groupe de travail de considérer d'autres facteurs qui pourraient influencer le grade des recommandations. Ces facteurs sont :

- toutes données scientifiques sur des inconvénients potentiels associés à la mise en œuvre d'une recommandation ;
- l''impact clinique (l'impact sur la population cible et les ressources nécessaires au système de soins pour traiter les patients en suivant la recommandation) ;
- si et dans quelle mesure des groupes pourraient être particulièrement avantagés ou désavantagés par les recommandations faites ;
- la mise en œuvre (au sein du système de santé écossais).

Finalement, le groupe résume son avis à la fois sur la qualité des données scientifiques et sur leur impact potentiel, avant de rédiger une recommandation qu'il grade. Le résumé doit être succinct (**énoncé des données scientifiques**) et, accompagné de l'avis du groupe de travail sur le niveau de preuve, il représente la première version du texte qui apparaîtra dans la RBP immédiatement avant une recommandation gradée.

Tableau 18. Formulaire de jugement raisonné d'après le SIGN 2008 (4)

Question :	Tableau des données scientifiques numéro :
1. Volume des données scientifiques *Commentez ici les résultats concernant la quantité des données scientifiques disponibles sur le sujet et leur qualité.*	
2. Applicabilité *Commentez ici dans quelle mesure les données scientifiques sont directement applicables au système de soins écossais*	
3. Généralisabilité *Indiquez ici à quel point il est raisonnable de généraliser les résultats des études utilisés comme données scientifiques à la population cible de cette RBP.*	
4. Cohérence *Commentez ici le degré de cohérence des données scientifiques disponibles. Quand il y a des résultats hétérogènes, indiquez comment le groupe fait son jugement et la direction globale des résultats.*	
4. Impact clinique *Commentez ici l'impact clinique potentiel que l'intervention en question pourrait avoir – par exemple, taille de la population des patients, ampleur de l'effet, bénéfice relatif par rapport à d'autres options de soins, implication des ressources, rapport bénéfice-risque.*	
5. Autres facteurs *Indiquez ici les autres facteurs que vous prenez en compte quand vous évaluez l'ensemble des données scientifiques.*	
6. Énoncé des données scientifiques *Résumez la synthèse élaborée par le groupe des données scientifiques relatives à cette question, en prenant en compte tous les facteurs ci-dessus, et indiquez le niveau de preuve qui s'applique.*	**Niveau de preuve**

Question :	Tableau des données scientifiques numéro :
7. Recommandation *Quelle recommandation le groupe de travail peut-il déduire de ces données scientifiques ? Indiquez le grade de la (des) recommandation(s) et de toute opinion différente dans le groupe.*	**Grade de la recommandation**

► Identifier les recommandations clés

On demande au groupe de considérer l'importance de la recommandation qu'il vient de rédiger. **L'importance de la recommandation** n'est pas nécessairement en rapport direct avec la force des données scientifiques, mais **reflète dans quelle mesure le groupe croit que la recommandation aura un impact sur l'état de santé ou la qualité de vie des patients concernés.**

Il est demandé au groupe de justifier en quoi cette recommandation jugée clé doit être mise en valeur dans le texte final des recommandations. Toutes les recommandations clés doivent être identifiées comme telles dans le texte publié et doivent apparaître dans le *Quick Reference Guide*.

► Niveaux de preuve et grades des recommandations

La liste des niveaux de preuve pour l'énoncé des données scientifiques sont présentés dans le tableau 19.

Tableau 19. Niveaux de preuve d'après le SIGN 2008 (4)

Niveaux	Description
1++	Méta-analyses de qualité élevée, revues systématiques d'essais contrôlés randomisés, ou essais contrôlés randomisés avec un risque de biais très faible.
1+	Méta-analyses bien menées, revues systématiques, ou essais contrôlés randomisés avec un risque de biais faible.
1-	Méta-analyses, revues systématiques, ou essais contrôlés randomisés avec un risque de biais élevé.
2++	Revues systématiques de qualité élevée d'études cas-témoins ou d'études de cohortes. Études cas-témoins ou études de cohortes avec un faible risque d'effet de facteurs de confusion ou de biais et une probabilité élevée que la relation est causale.
2+	Études cas-témoins ou études de cohortes bien menées avec un faible risque d'effet de facteurs de confusion ou de biais et une probabilité modérée que la relation est causale.
2-	Études cas-témoins ou études de cohortes avec un risque élevé d'effet de facteurs de confusion ou de biais et un risque significatif que la relation ne soit pas causale.
3	Études non analytiques, par exemple séries de cas.
4	Opinion d'experts.

Occasionnellement, le groupe de travail estime qu'il y a un point pratique important sur lequel il souhaite insister, mais pour lequel il n'y a pas et il n'y aura probablement pas de données scientifiques. C'est typiquement un aspect du traitement considéré comme une bonne pratique clinique

que probablement personne ne remettra en question (tableau 20). Il peut être considéré comme du bon sens clinique.

Tableau 20. Points de bonne pratique d'après le SIGN 2008 (4)

☑	Meilleure pratique recommandée. Fondé sur l'expérience clinique du groupe de travail.

Les grades des recommandations sont présentés dans le tableau 21.

Tableau 21. Grades des recommandations d'après le SIGN 2008 (4)

Grade	Description
A	Au moins une méta-analyse, une revue systématique, ou un essai contrôlé randomisé coté 1++, et directement applicable à la population cible ; ou un ensemble de données scientifiques composé principalement d'études cotées 1+, directement applicable à la population cible, démontrant une homogénéité globale des résultats.
B	Un ensemble de données scientifiques incluant des études cotées 2++, directement applicable à la population cible, et démontrant une homogénéité globale des résultats ; ou données scientifiques extrapolées d'études cotées 1++ ou 1+.
C	Un ensemble de données scientifiques incluant des études cotées 2+, directement applicable à la population cible et démontrant une homogénéité globale des résultats ; ou données scientifiques extrapolées d'études cotées 2++.
D	Niveau de preuve 3 ou 4 ; ou données scientifiques extrapolées d'études cotées 2+.

Ce système de gradation a pour objectif :

- de mettre en exergue la qualité des données scientifiques supportant chaque recommandation, et d'insister sur le fait que l'ensemble des données scientifiques devrait être considéré comme un tout, et ne devrait pas reposer sur une seule étude supportant chacune des recommandations ;
- de permettre d'attribuer plus de poids aux recommandations fondées sur des études observationnelles de bonne qualité quand les essais contrôlés randomisés ne sont pas disponibles pour des raisons pratiques ou éthiques.

Le processus de jugement raisonné permet aussi au groupe de travail de rétrograder une recommandation dont il pense que les données scientifiques sur laquelle elle est fondée ne sont pas cohérentes, ou non généralisables, non directement applicables à la population cible ou, pour d'autres raisons, elle est perçue plus faible que ce que seule une évaluation de la méthode aurait suggéré.

En 2009, le SIGN a décidé de mettre en œuvre l'approche GRADE dans sa méthode d'élaboration des RBP. Les principes que le SIGN appliquera quand il mettra en œuvre le système GRADE (voir § 1.5 GRADE) sont les suivants :

- Toutes les RBP seront fondées sur une revue systématique de la littérature et une évaluation de la qualité des données scientifiques. La qualité des données scientifiques est définie par la confiance que l'on a dans le fait qu'une estimation de l'effet est adéquate pour supporter des recommandations.
- L'évaluation de la qualité des données sera réalisée dans le contexte de leur pertinence par rapport au système de santé écossais. Les critères pour établir la qualité de l'ensemble des

données scientifiques incluront tous les facteurs pour augmenter ou diminuer le niveau de qualité des données scientifiques identifiées par le GRADE *working group*.

- Les données scientifiques identifiées dans une revue systématique seront résumées dans un tableau des données scientifiques listant les caractéristiques clés des études individuelles. Chaque tableau sera à son tour résumé en relation avec l'ensemble des données scientifiques pour chaque résultat important ou décisif identifié par le groupe de travail. Ces résumés formeront la base de toutes les décisions concernant la qualité des données scientifiques ou la force des recommandations. Ces résumés seront produits soit en utilisant le logiciel GRADEpro, soit en notant les décisions du groupe de travail concernant chaque facteur de qualité sur un formulaire de jugement raisonné spécifique à cette étape du processus.
- La qualité des données scientifiques sera cotée selon les quatre catégories définies par le GRADE *working group*.
- La force des recommandations sera fondée sur la prise en considération explicite de chacun des critères retenus par le GRADE *working group,* et notée dans un formulaire de jugement raisonné spécifique de cette étape du processus.
- Les recommandations seront soit inconditionnelles (données scientifiques robustes, pas de biais importants), soit conditionnelles (données scientifiques moins solides, biais potentiels sérieux).

1.7 Organisation Mondiale de la Santé

L'OMS a adopté le système GRADE (5).

1.8 *US Preventive Services Task Force*

L'USPSTF utilise un schéma analytique présentant de manière graphique les questions spécifiques auxquelles la revue de la littérature doit répondre. Les questions sont représentées par des liens entre les interventions et les résultats (6,15).

Le processus de revue de la littérature implique d'évaluer la validité et la fiabilité des données scientifiques recevables à trois niveaux : celui d'une étude, d'une question, d'une intervention de prévention dans sa globalité. Les six questions de l'analyse critique de la littérature sont les suivantes :

1. Les études ont-elles une méthode de recherche adaptée pour répondre à la(les) question(s) ?
2. Les études sont-elles de qualité élevée ? (quelle est leur validité interne ?)
3. Dans quelle mesure les résultats des études sont-ils généralisables à la population générale américaine des soins de premier recours ?
4. Combien d'études évaluent la question(s) ? Quelle est la taille de l'échantillon ? (quelle est la précision des données scientifiques ?)
5. Les résultats des études sont-ils cohérents ?
6. Y a-t-il des facteurs supplémentaires pour nous aider à tirer les conclusions ? (présence ou absence d'effets dose-réponse ; cela correspond-il à un modèle biologique ?)

► Évaluation d'une étude

On utilise les questions 1-3 et 6. Une étude est catégorisée en fonction de sa méthode, sa validité interne et sa généralisabilité.

► Validité interne

La validité interne des études (qualité) est évaluée de façon plus précise avec une liste de critères minimum adaptée à chaque type d'étude (tableau 22).

Tableau 22. Hiérarchie des types d'études d'après l'USPSTF 2008 (6)

Type	Description
I	Essai contrôlé randomisé de conception et de puissance correctes. Revue systématique ou méta-analyse d'essais contrôlés randomisés bien conduites.
II-1	Essai contrôlé bien conduit sans randomisation.
II-2	Étude de cohorte bien conduite ou étude cas-témoin.
II-3	Séries temporelles multiples avec ou sans intervention. Résultats suffisants issus d'études non contrôlées.
III	Opinions de spécialistes fondées sur l'expérience clinique ; études descriptives ou observations ; rapport de comités d'experts.

Les études sont classées en trois catégories :

- « bonne » : l'étude remplit tous les critères ;
- « moyenne » : l'étude ne remplit pas (ou pas clairement) au moins un critère, mais n'a pas de biais majeur ;
- « médiocre » : l'étude a au moins un biais majeur.

La *Task force* envisage le niveau de preuve d'une étude, d'une question, ou d'une intervention de prévention dans sa globalité, indépendamment de l'ampleur de l'effet. Une étude (ou un nombre d'études) peut être classée « bonne », alors qu'elle ne montre pas d'effet de l'intervention de prévention.

► Généralisabilité

La généralisabilité est évaluée en considérant :

- la population de l'étude (population spéciale ou non représentative de la population des soins de premier recours qui pourrait avoir l'intervention de prévention) ;
- la situation : dans quelle mesure l'expérience clinique dans le cadre de l'étude pourra être reproduite dans l'environnement des soins de premier recours ;
- les professionnels : dans quelle mesure des professionnels ayant les capacités et l'expérience de ceux de l'étude seront accessibles pour des soins de premier recours.

La généralisabilité est cotée « bonne », « moyenne », « médiocre ».

► Applicabilité de l'ensemble des données scientifiques à la population/environnement des soins/professionnels

Le but de l'évaluation est de juger s'il est probable que des différences cliniquement importantes soient observées entre les résultats de l'ensemble des études et les résultats attendus quand l'intervention sera mise en œuvre en soins de premier recours (compte tenu de la population, de l'environnement des soins et des professionnels accessibles). Les questions suivantes sont abordées :

- Est-il possible de conclure à partir des données scientifiques à l'efficacité réelle de l'intervention dans la population des soins de premier recours avec l'environnement des soins et les professionnels accessibles ?
- Est-il probable que la taille du bénéfice observée dans les études individuelles qui composent l'ensemble des données scientifiques soit la même en soins de premier recours ?
- Est-il probable que les effets indésirables observés dans les études individuelles soient les mêmes en soins de premiers recours ?

- Quelle sera la probable relation entre les bénéfices et les risques en soins de premier recours à partir des données scientifiques ?
- Est-ce que le temps et les efforts nécessaires pour délivrer les interventions dans les études individuelles sont envisageables en soins de premier recours ?

► Évaluation du niveau de preuve des données scientifiques pour une question

On évalue le niveau de preuve pour une question en utilisant les six questions de l'analyse critique. L'ensemble des données scientifiques est souvent catégorisé en se fondant sur les données scientifiques de niveau de preuve le plus fort.

Le niveau de preuve pour chaque question correspond à l'une des trois catégories : « convaincant », « adéquat », « inadéquat ».

La *Task force* considère la validité interne de l'ensemble des études pour une question. Le jugement est souvent le reflet des meilleures études cliniques concernant un résultat. De la même façon, l'applicabilité reflète dans quelle mesure les résultats des meilleures études peuvent être généralisés aux soins de premiers recours populations /environnement/professionnels.

La cohérence est utilisée pour indiquer qu'un ensemble de données scientifiques a du sens. Elle inclut la cohérence entre les populations, les interventions et les résultats dans les études analysées.

► Évaluation de la certitude de l'ensemble des données scientifiques pour l'intervention de prévention dans sa globalité : synthèse des données scientifiques

À cette étape, le problème est de savoir comment les données scientifiques correspondant aux différentes questions s'accordent pour fournir une estimation valide de l'ampleur des bénéfices, des inconvénients et du bénéfice net attendus lors de la mise en œuvre de l'intervention de prévention.

L'USPSTF définit la **certitude** comme la probabilité que son évaluation du bénéfice net d'une intervention de prévention est correcte.

Le **bénéfice net** est défini comme les bénéfices moins les inconvénients de l'intervention mise en œuvre dans une population des soins de premiers recours.

L'USPSTF attribue un niveau de certitude fondé sur la nature de l'ensemble des données scientifiques disponibles pour évaluer le bénéfice net d'une intervention de prévention. Les niveaux de certitude sont présentés dans le tableau 23.

La synthèse de l'information issue de l'ensemble des données scientifiques correspondant à tout le schéma analytique découle d'un jugement fondé sur :

- les six questions relatives à l'analyse critique des études identifiées ;
- la façon dont les données scientifiques s'accordent pour compléter la chaîne reliant une intervention de prévention à des résultats cliniques.

Tableau 23. Niveaux de certitude par rapport au bénéfice net d'après l'USPSTF 2008 (6)

Niveau de certitude	Description
Élevé	Les données scientifiques disponibles incluent en général des résultats cohérents, issus d'études bien conçues et bien menées dans des populations des soins de premier recours représentatives. Ces études évaluent les effets de l'intervention de prévention sur des résultats cliniques. Il est peu probable que cette conclusion soit fortement affectée par les résultats d'études futures.
Modéré	Les données scientifiques disponibles sont suffisantes pour déterminer les effets

Niveau de certitude	Description
	de l'intervention de prévention sur des résultats cliniques, mais la confiance dans l'estimation est limitée par des facteurs tels que : - le nombre, la taille, ou la qualité des études individuelles ; - une hétérogénéité des résultats des études individuelles ; - une généralisabilité limitée des résultats à la pratique courante des soins de premier recours ; - un manque de cohérence dans la chaîne des données scientifiques. Si des informations supplémentaires deviennent disponibles, l'ampleur ou la direction de l'effet observé pourrait changer, et ce changement pourrait être assez grand pour altérer les conclusions.
Faible	Les données scientifiques disponibles sont insuffisantes pour évaluer des effets sur les résultats cliniques. Les données scientifiques sont insuffisantes à cause : - du nombre limité ou de la taille des études ; - des biais importants dans la conception de l'étude ou des méthodes ; - d'une hétérogénéité des résultats des études individuelles ; - des lacunes dans la chaîne des données scientifiques ; - des résultats non généralisables à la pratique courante des soins de premier recours ; - d'un manque d'information sur des résultats cliniques importants. Davantage d'informations peuvent permettre une estimation des effets sur des résultats cliniques.

La *Task Force* sépare les concepts de certitude des données scientifiques et d'ampleur des bénéfices, des inconvénients et du bénéfice net.

► Évaluation du bénéfice net

Le bénéfice net est évalué avec une gradation à quatre niveaux allant de :

- substantiel : quand les bénéfices l'emportent nettement sur les inconvénients ;
- à zéro/négatif : quand les bénéfices et les inconvénients s'équilibrent ou quand les inconvénients l'emportent sur les bénéfices.

► Grades des recommandations

Les recommandations sont codées pour refléter à la fois la certitude des données scientifiques et l'ampleur du bénéfice net. La *Task Force* a adopté une formulation standardisée des recommandations.

Les grades des recommandations sont présentés dans le tableau 24.

Tableau 24. Grille de recommandation de l'USPSTF : grade de la recommandation ou énoncé de l'insuffisance des données scientifiques pour évaluer la certitude et l'ampleur du bénéfice net d'après l'USPSTF 2008 (6)

Certitude du bénéfice net	Ampleur du bénéfice net			
	Substantielle	Modérée	Petite	Nulle / négative
Élevée	A	B	C	D
Modérée	B	B	C	D
Faible	Insuffisant			

Grade A : l'USPSTF recommande l'intervention. Il y a une certitude élevée d'un bénéfice net substantiel.

	Ampleur du bénéfice net

Grade B : l'USPSTF recommande l'intervention. Il y a une certitude élevée d'un bénéfice net modéré ou il y a une certitude modérée d'un bénéfice net modéré à substantiel.

Grade C : les médecins peuvent offrir cette intervention à des patients sélectionnés selon les circonstances. Il y a une certitude élevée ou modérée d'un petit bénéfice net.

Grade D : l'USPSTF ne recommande pas l'intervention. Il y a une certitude élevée ou modérée de l'absence de bénéfice net ou que les inconvénients l'emportent sur les bénéfices.

Grade I : les données scientifiques sont insuffisantes pour évaluer le rapport bénéfices-inconvénients de l'intervention. Les données scientifiques manquent, sont de qualité faible ou contradictoires, et le rapport bénéfices-inconvénients ne peut pas être déterminé (*i.e :* le bénéfice net). La proposition est considérée comme un énoncé et non comme une recommandation.

1.9 *National Health and Medical Research Council*

► Hiérarchie des niveaux de preuve

Les recommandations peuvent avoir différents objectifs, répondant à des questions cliniques d'intervention, de diagnostic ou autres (7). Pour évaluer correctement ces différentes questions cliniques, il est nécessaire d'inclure des études de différents types. Une nouvelle hiérarchie a été élaborée par le NHMRC afin d'attribuer aux études des niveaux de preuve selon le type de question clinique considéré (intervention, diagnostic, pronostic, étiologie et dépistage) ; cette hiérarchie reconnaît l'importance d'un type d'étude adapté à un type de question (tableau 25) (à titre de comparaison, voir la hiérarchie des niveaux de preuve de l'*Oxford Centre for Evidence-based Medicine* annexe 6).

Tableau 25. Hiérarchie des niveaux de preuve d'après le NHMRC, 2009 : niveaux de preuve pour les études d'interventions et les études portant sur la validité diagnostique des tests (7)

Niveau	Domaine				
	Intervention	Diagnostic*	Pronostic	Étiologie	Intervention de dépistage
I	Une revue systématique d'études de niveau II.	Une revue systématique d'études de niveau II.	Une revue systématique d'études de niveau II.	Une revue systématique d'études de niveau II.	Une revue systématique d'études de niveau II.
II†	Un essai contrôlé randomisé.	Une étude portant sur la validité diagnostique avec : une comparaison indépendante, en insu, avec un examen de référence standard‡ fiable, parmi des personnes consécutives ayant un tableau clinique défini.	Une étude de cohorte prospective.	Une étude de cohorte prospective.	Un essai contrôlé randomisé.
III-1	Un essai contrôlé pseudo-randomisé (allocation alternée ou autre méthode).	Une étude portant sur la validité diagnostique avec : une comparaison indépendante, en insu, avec un examen de référence standard fiable, parmi des personnes non consécutives ayant un tableau clinique défini.	Tous ou aucun des sujets ayant le facteur de risque présentent la maladie.	Tous ou aucun des sujets ayant le facteur de risque présentent la maladie.	Un essai contrôlé pseudo-randomisé (allocation alternée ou autre méthode).
III-2	Une étude comparative avec des témoins concomitants : - essai contrôlé expérimental non randomisé ; - étude de cohorte ; - étude cas-témoins ; - série temporelle interrompue avec un groupe témoins.	Une comparaison avec un examen de référence standard qui ne remplit pas les critères requis pour les niveaux II et III-1.	Analyse des facteurs pronostiques parmi les sujets d'un seul bras d'un essai contrôlé randomisé.	Une étude de cohorte rétrospective.	Une étude comparative avec des témoins concomitants : - essai contrôlé expérimental non randomisé ; - étude de cohorte ; - étude cas-témoins.
III-3	Une étude comparative sans témoins concomitants :	Étude cas-témoins portant sur la validité diagnostique d'un test.	Une étude de cohorte rétrospective.	Une étude cas-témoins.	Une étude comparative sans témoins concomitants :

	Domaine				
	- étude contrôle historique ; - étude à deux bras ou plus ; - série temporelle interrompue sans groupe témoins parallèle.				- étude contrôle historique ; - étude à deux bras ou plus.
IV	Série de cas avec des résultats soit post-test soit pré-test/post-test.	Étude du rendement diagnostique (sans examen de référence) ‖.	Série de cas, ou étude de cohorte de personnes à différentes phases de la maladie.	Une étude transversale ou série de cas.	Série de cas.

* : ces niveaux s'appliquent uniquement aux études portant sur la validité diagnostique du test. Pour évaluer l'efficacité d'un test diagnostique en population, il est nécessaire de prendre en considération l'impact du test sur la prise en charge du patient et les résultats cliniques.

† : une revue systématique aura le niveau de preuve des études qu'elle contient, excepté celles incluant des études de niveau II. Une revue systématique doit inclure au minimum 2 études.

‡ : la validité de l'examen de référence standard doit être déterminée dans le contexte de la maladie considérée. Les critères pour déterminer la validité de l'examen de référence doivent être pré-spécifiés.

‖ : des études sur le rendement diagnostique fournissent la proportion de patients diagnostiqués sans confirmation de la valeur de ce diagnostic par un examen de référence standard.

► Grade des recommandations

Le niveau de preuve d'une étude, qui reflète le risque de biais d'une étude liés à sa conception n'est qu'une petite partie de l'évaluation des données scientifiques pour une recommandation. Il est nécessaire de prendre en compte d'autres aspects : la qualité de l'étude et la probabilité que les résultats aient été affectés par des biais ; la cohérence des résultats avec ceux des autres études ; l'impact clinique des résultats ; la généralisabilité des résultats à la population cible des recommandations ; l'applicabilité des résultats au système de soins australien.

► Évaluation d'une étude

Chaque étude incluse dans une revue systématique est évaluée dans trois dimensions :

- Force des données scientifiques :
  - Niveau de preuve : chaque étude est évaluée selon sa place dans la hiérarchie des niveaux de preuve (tableau 25). Cette hiérarchie reflète la capacité de chaque étude ou de chaque revue systématique, incluses dans la revue systématique soutenant la recommandation, à répondre de façon adéquate à une question de recherche particulière (intervention, diagnostic ou autre). Elle est fondée sur la probabilité que la conception de l'étude minimise l'impact des biais sur les résultats.
  - Qualité des données scientifiques (risque de biais) : évalue la façon dont les biais, les facteurs de confusion et/ou le hasard ont pu influencer les résultats de l'étude.
  - Précision statistique : évalue si l'effet est réel ou dû au hasard (degré de significativité statistique de p et/ou intervalle de confiance de l'estimation de l'effet).
- Taille de l'effet :

Elle est utile pour évaluer l'importance clinique des résultats de chaque étude (mesure de l'effet ou de l'estimation).

- Pertinence des données scientifiques :

Cette dimension est en rapport avec la traduction des données scientifiques dans la pratique courante. Elle comporte deux éléments clés :

- pertinence des critères de jugement pour les patients ;
- pertinence de la question de l'étude (les éléments de la question de l'étude [selon PICO] correspondent-ils aux éléments de la question considérée dans la recommandation ?).

► Évaluation de l'ensemble des données scientifiques et formulation des recommandations

L'ensemble des données scientifiques pour chaque recommandation est examiné dans cinq dimensions (tableau 26).

- **Les études sources des données :** en termes de quantité, de niveau de preuve et de qualité (risque de biais) des études incluses :
  - Quantité : nombre d'études qui ont été incluses comme sources de données pour chaque recommandation (et listées dans un tableau de synthèse des données scientifiques ou dans le texte). Elle prend aussi en compte la puissance statistique de ces études.
  - Niveau de preuve, qui reflète les meilleurs types d'études pour un type de question spécifique. Chaque type d'étude est évalué selon sa place dans la hiérarchie des niveaux de preuve (tableau 25). Le type d'étude le plus adapté pour répondre à chaque question clinique (intervention, test diagnostique, pronostic, étiologie) est le niveau II. Le niveau I correspond à des revues systématiques d'études appropriées de niveau II dans chaque cas.
  - Qualité des études : reflète la manière dont les études ont été menées pour éliminer des biais (incluant la sélection des sujets, l'assignation des sujets à chaque groupe, la prise en charge et le suivi, et la façon dont les résultats de l'étude ont été mesurés).

- **La cohérence** de l'ensemble des données scientifiques : évalue la cohérence des résultats entre les études incluses (réalisées à partir de diverses populations et correspondant à divers types d'étude), pour s'assurer du fait que les résultats sont probablement reproductibles ou surviennent dans certaines conditions (analyse de l'hétérogénéité pour une méta-analyse d'essais contrôlés randomisés).
- **L'impact clinique** qui mesure le bénéfice potentiel d'une application de la recommandation à une population. Il prend en compte :
  - la pertinence des données scientifiques pour répondre à la question clinique, le degré de significativité (ou la précision de l'estimation de l'effet indiquée par l'intervalle de confiance) et la taille de l'effet (incluant l'importance clinique) et la pertinence de l'effet pour les patients comparé à une autre option de prise en charge ;
  - la durée du traitement pour obtenir l'effet ;
  - le rapport bénéfices-risques (en prenant en compte la taille de la population concernée).
- **La généralisabilité** évalue dans quelle mesure les sujets et le cadre de réalisation des études correspond à celui de la recommandation de bonne pratique, en particulier la population cible de la RBP et le contexte clinique dans lequel les recommandations seront mises en œuvre. Les caractéristiques de la population qui pourraient intervenir sont le sexe, l'âge, l'origine ethnique, le risque de base, le niveau de soins. Fondamentalement, une évaluation de la généralisabilité vise à déterminer si l'ensemble des données scientifiques disponibles répond à la question clinique posée.
  Dans les études portant sur la validité diagnostique des tests, il est nécessaire de prendre en compte d'autres critères incluant le stade de la maladie, la durée de la maladie, la prévalence de la maladie dans la population de l'étude comparée à la prévalence de la maladie dans la population cible de la RBP.
- **L'applicabilité** évalue la pertinence des données scientifiques pour l'ensemble du système de soins australien ou dans un cadre plus local pour des recommandations spécifiques. Les facteurs pouvant restreindre l'application directe des résultats d'une étude sont des facteurs organisationnels (disponibilité d'une équipe entraînée, équipement spécialisé, tests, autres ressources) et des facteurs culturels (attitudes à l'égard des questions de santé, en particulier celles qui affectent l'observance des recommandations).

Ces cinq éléments sont cotés avec la matrice des données scientifiques ci-dessous (tableau 26).

Tableau 26. Matrice de l'ensemble des données scientifiques d'après le NHMRC 2009 (7)

Éléments	A	B	C	D
	Excellent	Bon	Satisfaisant	Médiocre
Études sources des données Quantité Niveau de preuve* Qualité des études (risque de biais)	Au moins une étude de niveau I avec un risque de biais faible ou plusieurs études de niveau II avec un risque de biais faible.	Une ou deux études de niveau II avec un risque de biais faible ou une revue systématique de plusieurs études de niveau III avec un risque de biais faible.	Une ou deux études de niveau III avec un risque de biais faible ou des études de niveau I ou II avec un risque de biais modéré.	Études de niveau IV ou études de niveau I à III/revues systématiques avec un risque de biais important.
Cohérence†	Toutes les études sont cohérentes.	La plupart des études sont cohérentes, et l'incohérence peut être expliquée.	Incohérence reflétant une véritable incertitude autour de la question clinique.	Les données scientifiques sont incohérentes.
Impact clinique	Très large.	Important.	Modéré.	Léger ou restreint.

	A	B	C	D
Généralisabilité	La (les) population(s) étudiée(s) dans l'ensemble des données scientifiques est(sont) la(les) mêmes que la population cible de la recommandation.	La(les) populations étudiée(s) dans l'ensemble des données scientifiques est(sont) similaire(s) à la population cible de la recommandation.	La(les) population(s) étudiée(s) dans l'ensemble des données scientifiques diffère(nt) de la population cible de la recommandation, mais il est raisonnable cliniquement d'appliquer ces données à la population cible.	La(les) population(s) étudiée(s) dans l'ensemble des données scientifiques diffère(nt) de la population cible, et il est difficile de juger s'il est raisonnable de généraliser à la population cible.
Applicabilité	Directement applicable au contexte du système de soins australien.	Applicable au système de soins avec quelques mises en garde.	Probablement applicable au système de soins avec des mises en garde.	Non applicable au contexte du système de soins.

* : niveau de preuve du NHMRC (voir tableau 25) ; † : si une seule étude est disponible, coter cet élément non applicable.

Il est recommandé que la formulation de la recommandation reflète la force de l'ensemble des données scientifiques. « Doit » ou « il est recommandé » sont utilisés quand les données scientifiques venant à l'appui de la recommandation sont fortes, et « pourrait » est utilisé quand ces données sont plus faibles.

Le grade de la recommandation est fondé sur la somme des cotations de chacun des cinq éléments : études sources des données (quantité, niveau de preuve, qualité des études), cohérence, impact clinique, généralisabilité, applicabilité. Pour qu'une recommandation soit gradée A ou B, les données scientifiques et la cohérence des données scientifiques doivent être l'un et l'autre gradés A ou B.

Les grades des recommandations ont pour objectif d'indiquer la force de l'ensemble des données scientifiques qui sous-tendent la recommandation (tableau 27). Les recommandations de grade A ou B sont généralement fondées sur un ensemble de données scientifiques en lesquelles on peut avoir confiance pour guider la pratique, tandis que les recommandations de grade C ou D doivent être appliquées avec circonspection selon les circonstances cliniques individuelles et organisationnelles, et devraient être interprétées avec soin.

Tableau 27. Gradation des recommandations d'après le NHMRC 2009 (7)

Grade	Description
A	On peut se fier à l'ensemble des données scientifiques pour guider la pratique.
B	On peut se fier à l'ensemble des données scientifiques pour guider la pratique dans la plupart des situations.
C	L'ensemble des données scientifiques fournit des justifications pour la recommandation mais il faut être attentif lors de sa mise en pratique.
D	L'ensemble des données scientifiques est faible et la recommandation doit être appliquée avec précaution.

La cotation de l'ensemble des données scientifiques dans chacune des cinq dimensions et la recommandation et son grade sont reportés dans le formulaire d'énoncé des données scientifiques (tableau 28).

Tableau 28. Formulaire d'énoncé des données scientifiques d'après le NHMRC 2009 (7)

Question :		
1. Études sources des données scientifiques (nombre d'études, niveau de preuve et risque de biais dans les études incluses)		
	A	Au moins une étude de niveau I avec un risque de biais faible ou plusieurs études de niveau II avec un risque de biais faible.
	B	Une ou deux études de niveau II, avec un risque de biais faible ou une revue systématique de plusieurs études de niveau III, avec un risque de biais faible.
	C	Une ou deux études de niveau III, avec un risque de biais faible ou des études de niveau I ou II, avec un risque de biais modéré.
	D	Études de niveau IV ou études de niveau I à III/revues systématiques avec un risque de biais important.
2. Cohérence (si une seule étude est disponible, coter cet élément non applicable)		
	A	Toutes les études sont cohérentes.
	B	La plupart des études sont cohérentes, et l'incohérence peut être expliquée.
	C	Incohérence reflétant une véritable incertitude autour de la question clinique.
	D	Les données scientifiques sont incohérentes.
	NA	Non applicable (une seule étude).
3. Impact clinique (indiquez si les résultats de l'étude varient en fonction d'un facteur inconnu [pas seulement la qualité de l'étude ou la taille de l'échantillon] **et ainsi l'impact clinique de l'intervention ne pourrait pas être déterminé)**		
	A	Très grand.
	B	Modéré.
	C	Léger.
	D	Restreint.
4. Généralisabilité (comment l'ensemble des données scientifiques correspond à la population et au contexte des soins cibles de la RBP ?)		
	A	Données scientifiques directement généralisables à l a population cible.
	B	Données scientifiques directement généralisables à la population cible avec quelques mises en garde.
	C	Données scientifiques non directement généralisables à la population cible, mais pourrait être appliquées de façon judicieuse.
	D	Données scientifiques non directement généralisables à la population cible, et il est difficile de juger s'il est judicieux de les appliquer.
5. Applicabilité *(*Est-ce que l'ensemble des données scientifiques est pertinent pour le système de soins australien/prestation des soins et facteurs culturels ?)		
	A	Directement applicable au contexte du système de soins australien.
	B	Applicable au système de soins avec quelques mises en garde.
	C	Probablement applicable au système de soins avec des mises en garde.
	D	Non applicable au contexte du système de soins.

Question :		
Autres facteurs (indiquez ici tout autre facteur que vous prenez en compte pour évaluer l'ensemble des données scientifiques [par exemple, résultats qui pourraient pousser le groupe à augmenter ou à diminuer le grade de la recommandation])		
MATRICE DE L'ÉNONCÉ DES DONNÉES SCIENTIFIQUES *Résumez la synthèse élaborée par le groupe des données scientifiques relatives à cette question, en prenant en compte tous les facteurs ci-dessus*		
Composant	Cotation	Description
1. Études sources des données scientifiques		
2. Cohérence		
3. Impact clinique		
4. Généralisabilité		
5. Applicabilité		
Indiquez toute opinion différente dans le groupe		
RECOMMANDATION *Quelles recommandations le groupe de travail doit-il déduire de ces données scientifiques ?*		**GRADE**

1.10 *National Institute for Health and Clinical Excellence*

► Interventions

► Analyse critique des études

Le NICE (8) a commencé à utiliser l'approche GRADE (voir § 1.5 système GRADE) pour des questions concernant des interventions dans ses RBP. Les principales différences entre l'approche du NICE et l'approche GRADE sont que le NICE :

- inclut aussi une revue de la qualité des études coût-efficacité ;
- n'a pas d'échelle ordinale pour la qualité des données scientifiques ou la force des recommandations ;
- utilise une formulation des recommandations qui reflète la force de la recommandation (voir infra).

► Synthèse et présentation des résultats

Les données caractéristiques d'une étude sont incluses dans un tableau de synthèse des données scientifiques (tableau 29).

Tableau 29. Tableau de synthèse des données scientifiques pour des études d'intervention d'après le NICE, 2009 (8)

Question :										
Référence	Type d'étude	Qualité de l'étude	Nombre de patients	Caractéristiques des patients	Intervention	Comparaison	Durée du suivi	Critères de jugement et taille de	Source de financement	Commentaires

Question :										
								l'effet		

L'ensemble des données scientifiques sur une question donnée est présenté dans un profil des données scientifiques (*evidence profile*) ou un tableau de résumé des résultats (*Summary of findings* – SoF) (voir § 1.5 système GRADE).

► Diagnostic

► Analyse critique des études

Les études portant sur la validité diagnostique des tests sont évaluées en utilisant la check-list QUADAS[5,6] (*Quality Assessment of Diagnostic Accuracy Studies*).

Il est nécessaire de déterminer, avant de débuter l'analyse critique des études, quels critères (issus de la check-list QUADAS) sont probablement les indicateurs de qualité les plus importants pour la question posée. Ces critères sont utiles par la suite pour décider de la qualité globale d'une étude, pour exclure certaines études, et pour la présentation des résultats.

Remarque : les études « *Teste et traite* » (dans ces études, les résultats de patients chez lesquels on réalise un nouveau test diagnostique [en combinaison avec une stratégie de prise en charge] sont comparés aux résultats de patients chez lesquels on réalise le test diagnostique habituel et une stratégie de prise en charge) devraient être abordées de la même façon que les études d'intervention.

► Synthèse et présentation des résultats

Un résumé de la qualité des données est réalisé en se fondant sur les critères d'évaluation de la qualité de la check-list QUADAS définis préalablement.

Les données chiffrées sur la validité diagnostique des tests sont présentées dans des tableaux de synthèse des données scientifiques (tableau 30).

Tableau 30. Tableau de synthèse des données scientifiques pour des études portant sur la validité diagnostique des tests d'après le NICE, 2009 (31)

Question :											
Référence	Type d'étude	Qualité de l'étude	Nombre de patients	Prévalence	Caractéristiques des patients	Type de test	Référence standard	Sensibilité et spécificité	VPP VPN	Source de financement	Commentaires
VPN : valeur prédictive négative ; VPP : valeur prédictive positive.											

[5] Une version révisée a été publiée en 2011, QUADAS-2 qui comporte 4 domaines : sélection des patients, test évalué, test de référence, flux de patients et timing des tests. Chaque domaine est évalué par rapport au risque de biais, et l'applicabilité est évaluée pour les trois premiers domaines (29).

[6] La check-list QUADAS est également recommandée par la *Cochrane Collaboration* comme point de départ pour évaluer la qualité des études incluses dans les revues systématiques *Cochrane* sur les tests diagnostiques. Onze items sur les 14 items de QUADAS ont été repris dans l'outil Cochrane (exclusion des items 2, 7 et 8 relatifs à la description plus qu'à la méthode). QUADAS a été conçue pour évaluer les études transversales (30).

► Développement et formulation des recommandations

► Interprétation des données scientifiques pour rédiger les recommandations

Dans l'argumentaire, le but est de montrer clairement comment le groupe de travail passe des données scientifiques à la recommandation. Ceci est sous-tendu par le concept de force des recommandations (voir § 1.5 système GRADE).

Le système GRADE attribue des symboles pour représenter la force des recommandations, tandis que le NICE a choisi de refléter la force des recommandations dans leur formulation. La force des recommandations est dégagée des discussions du groupe de travail sur les principaux points qui sont :

- la valeur relative des résultats considérés ;
- un arbitrage entre les bénéfices et les inconvénients d'une intervention ;
- un arbitrage entre les bénéfices de santé et l'utilisation des ressources ;
- la qualité des données scientifiques.

► Formulation des recommandations

La formulation des recommandations doit :

- être focalisée sur les mesures que les acteurs sont amenés à prendre ;
- inclure ce que les lecteurs ont besoin de savoir ;
- refléter la force des recommandations (voir tableau 31) ;
- insister sur l'implication du patient (et/ou de leur soignants) dans les décisions thérapeutiques ;
- suivre les procédures standard du NICE pour les recommandations sur les médicaments, les délais d'attente et de recours, et les interventions inefficaces.

Tableau 31. Formulation des recommandations d'après le NICE 2009 (8)

Niveau de certitude	Formulation
Recommandations pour des interventions qui **doivent ou ne doivent pas** être utilisées *	- Habituellement utilisée seulement s'il existe une obligation légale à appliquer la recommandation (dans ce cas, donner la référence légale). - Occasionnellement si le fait de ne pas suivre une recommandation peut avoir des conséquences sérieuses sur l'état de santé.
Recommandations pour des interventions **recommandées ou non recommandées** †	Le groupe de travail est sûr que, pour la grande majorité des patients, l'intervention fera plus de bien que de mal et sera coût-efficace Dans la mesure du possible formuler les recommandations comme des instructions directes. Utiliser les verbes tels « offrir », « conseiller », « discuter ».
Recommandations pour des interventions qui **peuvent être utilisées** ‡	Le groupe de travail est sûr que, pour la grande majorité des patients, l'intervention fera plus de bien que de mal et sera coût-efficace. Cependant d'autres options sont également coût-efficaces. Il est probable que le choix de l'intervention (ou la décision d'une intervention) varie selon les valeurs et les préférences d'une personne. Dans la mesure du possible, formuler les recommandations comme des instructions directes. Ajouter « considérer » avant le verbe pour indiquer que la recommandation est moins forte que pour une intervention recommandée.

** : recommendations for interventions that must (or must not) be used.*

† : *recommendations for interventions that should (or should not) be used.*

‡ : *recommendations for interventions that could be used.*

1.11 *American College of Physicians' system*

Le système de l'ACP pour évaluer la qualité des données scientifiques et la force des recommandations est dérivé du système GRADE (voir § 1.5) (12). Bien que le système GRADE convienne mieux pour les recommandations sur les interventions, il peut être utilisé pour grader la qualité des données scientifiques et la force des recommandations pour les études sur les tests ou les stratégies diagnostiques.

► Gradation de la qualité des données scientifiques

La gradation de la qualité des données scientifiques est présentée dans le tableau 32.

Tableau 32. Système gradation de la qualité des données scientifiques de l'*American college of physicians'* d'après Qaseem *et al*., 2010 (12)

Qualité	Définition
Élevée	Au moins 1 ECR bien conçu et bien mené apportant des résultats cohérents et directement applicables. Il est très improbable que des recherches futures changent notre confiance en l'estimation de l'effet.
Moyenne	ECRs ayant des limites importantes (par exemple : évaluation biaisée de l'effet du traitement, nombreux perdus de vue, absence d'insu, hétérogénéité inexpliquée [même si elle est issue d'ECRs rigoureux], données scientifiques indirectes issues d'une population d'intérêt similaire [mais non identique] et ECRs ayant un très petit nombre de participants ou d'événements observés). De plus, les données scientifiques issues d'essais contrôlés bien menés mais sans randomisation, d'études de cohortes bien menées ou d'études cas-témoins, et les séries temporelles multiples avec ou sans intervention sont dans cette catégorie. Il est probable que des recherches futures aient un effet important sur notre confiance en l'estimation de l'effet, et que l'estimation de l'effet puisse changer.
Faible	Obtenue à partir d'études observationnelles avec un risque de biais. Il est très probable que des recherches futures aient un effet important sur notre confiance en l'estimation de l'effet. Il est probable que l'estimation de l'effet change.
	Cependant, la qualité des données scientifiques peut être cotée moyenne ou élevée, selon les conditions dans lesquelles les données scientifiques sont obtenues à partir des études observationnelles. Les facteurs pouvant contribuer à augmenter la qualité des données scientifiques incluent une grande taille de l'effet observé, une relation dose-réponse, ou la présence d'un effet observé quand tous les facteurs confondants possibles réduiraient cet effet.
Données scientifiques insuffisantes pour déterminer des bénéfices nets ou des risques nets	Quand les données scientifiques sont insuffisantes pour se positionner pour ou contre une procédure en pratique courante. Les données scientifiques peuvent être contradictoires, de qualité médiocre ou absente, et le rapport bénéfices-risques ne peut pas être déterminé. Il n'y a aucune estimation de l'effet qui est très incertain, les données scientifiques soit étant indisponibles, soit ne permettant pas une conclusion.

ECR : essai contrôlé randomisé.

► Gradation des recommandations

Une recommandation forte signifie que les bénéfices l'emportent nettement sur les risques et la lourdeur du traitement et *vice versa* (tableau 33).

Une recommandation est classée faible quand les bénéfices d'une part et les risques et la lourdeur du traitement d'autre part s'équilibrent ou qu'il existe une incertitude sur l'ampleur des bénéfices et des risques. Les préférences du patient peuvent influencer fortement le traitement.

Tableau 33. Système de gradation des recommandations de l'*American College of Physicians'* d'après Qaseem *et al*., 2010 (12)

Qualité des données scientifiques	Force des recommandations	
	Les bénéfices l'emportent nettement sur les risques et la lourdeur du traitement ou les risques et la lourdeur du traitement l'emportent nettement sur les bénéfices	Les bénéfices d'une part les risques et la lourdeur du traitement d'autre part s'équilibrent
Élevée	Forte	Faible
Moyenne	Forte	Faible
Faible	Forte	Faible
Les données scientifiques sont insuffisantes pour déterminer des bénéfices ou des risques nets		

L'interprétation des données scientifiques et son lien avec les recommandations est présentée dans le tableau 34.

Tableau 34. Interprétation des données scientifiques et son lien avec les recommandations d'après Qaseem *et al*., 2010 (12)

Grade des recommandations	Bénéfices *versus* risques et lourdeur du traitement	Qualité des données scientifiques	Interprétation	Implications
Recommandation forte ; données scientifiques de qualité élevée (1A)	Les bénéfices l'emportent nettement sur les risques et la lourdeur du traitement ou les risques et la lourdeur du traitement l'emportent nettement sur les bénéfices.	Essais contrôlés randomisés sans limites importantes ou données scientifiques de première importance issues d'études observationnelles.	Recommandation forte. Peut être appliquée à la plupart des patients dans la plupart des situations sans réserve.	Pour les patients : la plupart choisiraient l'intervention recommandée, et seule une faible proportion d'entre eux ne la choisiraient pas.
Recommandation forte ; données scientifiques de qualité moyenne (1B)	Les bénéfices l'emportent nettement sur les risques et la lourdeur du traitement ou les risques et la lourdeur du traitement l'emportent nettement sur les bénéfices.	Essais contrôlés randomisés avec des limites importantes (résultats discordants, biais méthodologiques, données scientifiques indirectes ou imprécises) ou exceptionnellement données scientifiques fortes issues d'études observationnelles.		Une personne devrait solliciter une discussion si l'intervention n'était pas proposée. Pour les médecins : la plupart des patients devraient recevoir l'intervention recommandée.

Grade des recommandations	Bénéfices *versus* risques et lourdeur du traitement	Qualité des données scientifiques	Interprétation	Implications
Recommandation forte ; données scientifiques de qualité faible (1C)	Les bénéfices l'emportent nettement sur les risques et la lourdeur du traitement ou les risques et la lourdeur du traitement l'emportent nettement sur les bénéfices.	Études observationnelles ou séries de cas.	Recommandation forte, mais peut changer quand des données d'une qualité supérieure seront disponibles.	Pour les décideurs : la recommandation peut être adoptée comme politique de soins dans la plupart des situations.
Recommandation faible ; données scientifiques de qualité élevée (2A)	Les bénéfices d'une part les risques et la lourdeur du traitement d'autre part s'équilibrent.	Essais contrôlés randomisés sans limites importantes ou données scientifiques de première importance issues d'études observationnelles.	Recommandation faible. La meilleure action à entreprendre peut différer selon les circonstances ou les préférences du patient ou de la société.	Pour les patients : la plupart choisiraient l'intervention recommandée, mais certains ne la choisiraient pas – la décision pourra dépendre des situations individuelles. Pour les médecins : différents choix seront adaptés pour différents patients, et il faudra parvenir à une décision cohérente avec les préférences des patients et les circonstances. Pour les décideurs : la décision pourra nécessiter un débat et une implication des parties prenantes.
Recommandation faible ; données scientifiques de qualité moyenne (2B)	Les bénéfices d'une part les risques et la lourdeur du traitement d'autre part s'équilibrent.	Essais contrôlés randomisés avec des limites importantes (résultats discordants, biais méthodologiques, données scientifiques indirectes ou imprécises) ou exceptionnellement données scientifiques fortes issues d'études observationnelles.		
Recommandation faible ; données scientifiques de qualité faible (2C)	Incertitude sur les estimations des bénéfices, des risques et de la lourdeur du traitement. Les bénéfices d'une part les risques et la lourdeur du traitement d'autre part peuvent s'équilibrer.	Études observationnelles ou séries de cas.	Recommandation faible. D'autres procédures peuvent être également raisonnables.	
Insuffisant	Le rapport bénéfices-risques ne peut pas être déterminé.	Les données sont discordantes, de qualité faible, ou absentes.	Données insuffisantes pour recommander ou non la procédure en pratique	Pour les patients, les médecins et les décideurs, des décisions ne peuvent pas être

Grade des recommandations	Bénéfices *versus* risques et lourdeur du traitement	Qualité des données scientifiques	Interprétation	Implications
			courante.	prises sur les données scientifiques.

1.12 Groupe d'étude canadien sur les soins de santé préventifs

Le Groupe d'étude canadien sur les soins de santé préventifs (auparavant Groupe d'étude canadien sur l'examen médical périodique) a adopté le système GRADE (32).

1.13 Synthèse

Une synthèse des niveaux de preuve et gradations des principaux systèmes est présentée en annexe 7.

Comment les trois éléments suivants, une étude, un ensemble d'études, une recommandation, sont-ils appréciés dans les principaux systèmes pour répondre à une question ?

► Étude

Selon les systèmes, une étude est considérée sous l'angle de son niveau de preuve ou de sa qualité.

Pour la HAS, le niveau de preuve d'une étude caractérise la capacité de l'étude à répondre à la question posée (2). Cette capacité est jugée sur la correspondance de l'étude au cadre du travail (question, population, critères de jugement) et sur les caractéristiques suivantes :

- l'adéquation du protocole d'étude à la question posée (selon le domaine exploré (diagnostic, pronostic, dépistage, traitement, etc.), un fort niveau de preuve peut être donné par des études dont le type de protocole sera différent) ;
- l'existence ou non de biais importants dans la réalisation de l'étude ;
- l'adaptation de l'analyse statistique aux objectifs de l'étude ;
- la puissance de l'étude et en particulier la taille de l'échantillon.

La taille de l'effet n'est pas prise en compte.

Le NZGG et l'AAP (3,11) utilisent le terme « qualité d'une étude » pour désigner ce qui correspond dans le système de la HAS au niveau de preuve d'une étude. Le système du NZGG prend en compte les mesures prises pour diminuer les biais, la taille de l'effet et la précision, l'applicabilité et la généralisabilité.

Pour le NHMRC, le niveau de preuve reflète la capacité de chaque étude ou de chaque revue systématique à répondre de façon adéquate à une question de recherche particulière (intervention, diagnostic, autre), fondé sur la probabilité que son type a minimisé l'impact des biais sur les résultats (7). Il permet d'évaluer avec la qualité de l'étude et la précision statistique, la force des données scientifiques.

Pour le SIGN, le niveau de preuve d'une étude est décidé sur le code attribué à la qualité de la méthode de l'étude (codée ++, +, -) couplé au type de l'étude (4).

Pour l'USPSTF, Le niveau de preuve d'une étude est évalué sur la validité interne d'une étude avec une liste de critères minimum adaptée à chaque type d'étude et sur la généralisabilité (6).

Le niveau de preuve d'une étude se limite à son type pour le NHMRC, il englobe le type d'étude et la qualité de l'étude pour le SIGN. Pour l'USPSTF, le niveau de preuve d'une étude est évalué sur

la validité interne et la généralisabilité. De façon analogue pour la HAS, le niveau de preuve est jugé non seulement sur le type et la qualité de l'étude, mais aussi sur des éléments ayant trait au caractère direct des données scientifiques (correspondance de l'étude au cadre du travail). Le jugement de ces différents éléments pour parvenir au niveau de preuve est laissé à l'appréciation du rédacteur de l'argumentaire (34).

► Ensemble d'études

Dans le système HAS, l'évidence scientifique (*body of evidence* des Anglo-Saxons, ensemble des données scientifiques) est appréciée lors de la synthèse des résultats de l'ensemble des études sélectionnées. Elle constitue la conclusion des tableaux de synthèse de la littérature. D'autres systèmes passent par cette phase de conclusion avec la rédaction d'un énoncé sommaire résumant les données scientifiques des études relatives à la question posée (SIGN, NZGG).

Les éléments pris en compte pour évaluer le niveau de preuve ou la qualité des données scientifiques issues d'un ensemble d'études sont présentés dans le tableau 35. Tous les systèmes considèrent :

- la quantité (nombre d'études, taille de l'échantillon) ;
- le type d'étude ;
- la qualité des études ;
- la cohérence des résultats.

Le rapport bénéfice risque est pris en compte à cette étape dans les systèmes du NZGG, du SIGN et du NHMRC.

Le système du NHMRC cote chacune des cinq dimensions (quantité-type d'étude-qualité, cohérence, impact clinique, généralisabilité, applicabilité).

Le système de l'USPSTF prend en compte le niveau de preuve des données scientifiques pour une question (à partir de l'analyse critique des études – méthode, validité interne, généralisabilité, nombre des études et taille de l'échantillon, homogénéité des résultats), et détermine la certitude du bénéfice net d'une intervention de prévention dans sa globalité (à travers les différentes questions) (6).

Parmi les neuf différents systèmes présentés, sept ont une échelle de niveau de preuve ou de qualité d'un ensemble de données scientifiques (2,4,7,11,12,19,33). Cette échelle comporte quatre niveaux.

L'accord d'experts est intégré dans l'échelle du SIGN (il s'agit du niveau 4) (4). Il est regroupé avec les observations et les raisonnements à partir de principes physiopathologiques de base (niveau D) dans le système de l'AAP (11).

Le système de l'ACP comporte un niveau pour les données scientifiques insuffisantes (12). Le système de l'AAP comporte un niveau spécial pour les situations exceptionnelles (dans lesquelles les études de validation ne peuvent pas être réalisées) (11).

Seul le système de la HAS inclut l'analyse de décision fondée sur des études bien menées (niveau de preuve 1).

Dans le système GRADE et le système de l'ACP qui en découle, l'évaluation est centrée sur chaque résultat important à travers toutes les études (« et non sur l'évaluation de tous les résultats de chaque étude »). Cependant, classiquement, le résultat à considérer dans chaque étude est le résultat sur le critère de jugement principal. Les niveaux de qualité des données scientifiques sont fondés sur la confiance dans l'estimation de l'effet. Ils sont déterminés pour chaque résultat important. L'approche du GRADE *working group* pour coter la qualité des données scientifiques, commence par le type d'étude (essai contrôlé randomisé ou étude observationnelle), puis évalue cinq facteurs pouvant diminuer le niveau de qualité des données scientifiques provenant d'études observationnelles et d'essais contrôlés randomisés et trois facteurs pouvant augmenter le niveau de

qualité des données scientifiques provenant d'études observationnelles. Le système GRADE offre une approche systématique pour prendre en compte et rapporter chacun de ces facteurs. Néanmoins, selon le GRADE *working group*, l'évaluation de la qualité des données scientifiques est un processus fondamentalement subjectif (19).

La gradation de l'ensemble des données scientifiques proposée par le système de la HAS ne s'applique qu'aux études d'intervention.

Les systèmes du NZGG, du SIGN, et NHMRC ont une seule échelle de niveaux de preuve utilisée à la fois pour les études d'interventions et les études sur les tests diagnostiques.

Le système de l'AAP indique sur quels éléments évaluer la qualité des études diagnostiques. Le système du GRADE *working group* précise l'évaluation des études portant sur la stratégie diagnostique des tests et sur la validité diagnostique des tests.

Tableau 35. Éléments pris en compte par les systèmes (autres que GRADE et dérivés) pour évaluer le niveau de preuve ou la qualité des données scientifiques issues d'un ensemble d'études

	HAS (2)	INCa (13)	AAP (11)	SIGN (4)	NHMRC (7)
	Gradation de l'évidence scientifique	Classification des niveaux de preuves	Évaluation de la qualité des données scientifiques	Niveau de preuve de l'énoncé des données scientifiques	Évaluation de l'ensemble des données scientifiques
Quantité (nombre d'études, taille des échantillons ou puissance statistique)	Existence de données, NP études	X	X	X	X
Type d'étude(s)	NP études	X	X	NP études	NP études
Qualité des études	NP études	X	X	NP études	X
Homogénéité des résultats	X	X	X	X	X
Impact clinique (pertinence des données scientifiques [critères de jugement : bénéfices et effets indésirables], degré de significativité et taille de l'effet)	Critères de jugement (correspondance de l'étude avec le cadre du travail)	-	X	X (et rapport bénéfices-risques)	X (et rapport bénéfices-risques)
Généralisabilité (population, cadre de l'étude : environnement des soins, professionnels)	NP études (correspondance de l'étude avec le cadre du travail)	-	X	X	X
Applicabilité (au système de soins)	-	-	-	X	X

NP : niveau de preuve ; X : élément pris en compte ; - : élément non pris en compte.

► Recommandations

Selon les systèmes, une recommandation est considérée sous l'angle de son grade et /ou de sa force.

► Gradation des recommandations

Pour la HAS, la gradation des recommandations est fondée sur le niveau de preuve des études venant à l'appui des recommandations (2).

Pour les systèmes du NHMRC, du NZGG, du SIGN, la gradation des recommandations indique la force de l'ensemble des données scientifiques sur lesquelles la recommandation est fondée.

En plus du niveau de preuve d'ensemble, le SIGN prend en compte l'impact des données scientifiques.

L'ancien système des SOR classe les interventions en standard, option ou recommandation en se fondant sur leurs bénéfices et leurs inconvénients.

Pour le système de l'USPSTF, le grade des recommandations reflète à la fois la certitude et l'ampleur du bénéfice net d'une intervention de prévention (6).

► Gradation des recommandations pour les études diagnostiques

La gradation proposée par le système de la HAS est la même que les recommandations soient d'ordre thérapeutique, diagnostique, idem pour le système du SIGN.

Les gradations du NZGG et du NHMRC qui considèrent les données scientifiques sans précision sont utilisables pour les questions diagnostiques.

► Force des recommandations

Pour la HAS, la force des recommandations est appréciée sur le niveau de preuve scientifique et sur l'interprétation des experts. Ainsi, la gradation des recommandations ne présume pas obligatoirement de la force des recommandations. Il n'est pas précisé comment la force de la recommandation est exprimée.

Le SIGN distingue l'importance de la recommandation qui reflète dans quelle mesure le groupe croit que la recommandation aura un impact sur l'état de santé ou la qualité de vie des patients concernés. Les recommandations clés sont identifiées comme telles dans le texte publié.

Les systèmes du GRADE *working group*, de l'AAP et de l'ACP, fondent la force de la recommandation en partie ou totalement sur le rapport bénéfices inconvénients. Les autres facteurs pris en compte sont :

- pour le système du GRADE *working group* la qualité des données scientifiques, l'incertitude sur la variabilité des valeurs et des préférences et le coût ;
- pour l'AAP, le coût attendu de l'adhésion à une recommandation.

La force des recommandations est repérée sur une échelle à deux niveaux pour le système du GRADE *working group*, trois niveaux pour l'ACP et quatre niveaux pour l'AAP.

Le NICE a adopté l'approche GRADE, à ceci près qu'il n'utilise pas d'échelle ordinale pour la force des recommandations. La force est reflétée par la formulation des recommandations.

Les catégories de recommandations de l'AAP, du GRADE *working group* et de l'ACP sont utilisables pour les études d'intervention et pour les études diagnostiques.

2. Comparaisons des systèmes et retours d'expérience

► Comparaisons

Le système GRADE a été comparé à d'autres systèmes de gradation des données scientifiques et des recommandations, en particulier quand celles-ci sont élaborées par des sociétés savantes (projet du *Royal College of Physicians)* (35).

La recherche pour les sociétés savantes comporte un grand éventail de domaines de recherche et de type d'études. Cependant, dans la majorité des hiérarchies de gradation, le type d'étude de référence est l'essai contrôlé randomisé. Il en résulte que les recommandations fondées sur les données d'une recherche composée d'études autres que des ECR, et gradées avec des systèmes de gradation classiques, sont souvent faibles, avec un risque de légitimité réduite.

Les auteurs ont réalisé dans un premier temps une revue des forces et des faiblesses des principaux systèmes de gradation actuels dans le contexte de leur utilisation par des sociétés savantes (35). La revue a porté sur les systèmes du *Scottish Intercollegiate Guidelines Network* (SIGN) 2008 (4), du *Grading of Recommandations Assessement Development and Evaluation* (GRADE) *working group* (18), et du *National Service Framework for Long-Term Conditions grading system* (NSF-LTC) (annexe 8). Les auteurs ont conclu que le système optimal dépend du domaine de recherche auquel se rapporte la question de la recommandation (par exemple pour les études d'interventions, ils suggèrent d'utiliser le système du SIGN ou GRADE ; pour les études portant sur la validité diagnostique des tests, ils suggèrent GRADE ou le système du NSF-LTC.

Dans un deuxième temps, les auteurs ont réalisé une étude dans l'objectif d'évaluer les forces et les faiblesses de la méthode des trois systèmes de gradation (SIGN, GRADE, NSF-LTC), ainsi que leur facilité d'utilisation et leur applicabilité dans le contexte des sociétés savantes (36).

Douze évaluateurs (quatre paires et quatre individuels) reflétant la composition d'un groupe de travail ont été inclus dans l'étude. Huit questions, issues de huit RBP publiées, leur ont été attribuées. Le domaine de recherche était varié : diagnostic [2], qualitative, pronostic, intervention [2], dépistage, étiologie. Les évaluateurs ont eu trois mois pour grader vingt articles et une recommandation en utilisant les trois systèmes de gradation. Ils ont rempli une matrice contenant les grades qu'ils ont attribués à chaque article et à la recommandation avec chacun des trois systèmes, et ils ont complété un questionnaire semi-structuré.

Les résultats de l'analyse qualitative des questionnaires sont présentés dans le tableau 36.

Tableau 36. Forces et faiblesses des systèmes de gradation d'après Baker 2010 (35,36)

Système	Nombre évaluateurs	Temps pour grader un article	Entraînement nécessaire	Forces	Faiblesses
GRADE	10*	30 à 60 min (7/10) > 120 min (2/10)	Oui (8/10)	Le plus rigoureux et le mieux détaillé en terme d'analyse critique centré sur les essais contrôlés randomisés et les méta-analyses.	Complexe et nécessité d'un entraînement avant de l'utiliser avec temps nécessaire pour analyser les articles.
NSF-LTC	12	< 30 min (11/12)	Non (10/12)	Simple rapide et facile à utiliser (11/12) Utilise les mêmes questions pour	Excessivement simpliste (9/11) (insuffisamment détaillé, ou discriminant ou

Système	Nombre évaluateurs	Temps pour grader un article	Entraînement nécessaire	Forces	Faiblesses
				évaluer différentes modalités de recherche (1/11).	rigoureux pour grader les données scientifiques). Manque de clarté, de précision et de différenciation entre qualité et validité des études.
SIGN	12	< 30 min (7/12) 30 à 60 min (5/12)	Non (9/12)	Intérêt de l'éventail des check-lists pour analyser les différents types d'études quantitatives, y compris les études observationnelles. Facile et rapide à utiliser (8/12).	Limité pour l'évaluation des études descriptives et qualitatives. Pour les études observationnelles, ce n'est pas toujours facile de déterminer quelle est la check-list à utiliser. Ne donne pas la possibilité de rétrograder les études (en particulier les essais contrôlés randomisés). Nécessite une reproduction inutile des données sur la check-list.

* : 2/10 évaluateurs n'ont pas utilisé GRADE qui leur paraissait trop complexe.

Avant de sélectionner un système d'analyse critique des articles et de gradation des recommandations, les auteurs recommandent que les sociétés savantes considèrent le type d'études à analyser et l'expérience des évaluateurs (tableau 37). Les évaluateurs devraient suivre un entraînement en analyse de littérature et gradation des recommandations correspondant au système qui sera employé pour la recommandation.

Tableau 37. Choix d'un système de gradation des recommandations d'après Baker 2010 (36)

Type d'études	Système envisagé	Expérience des évaluateurs
Probablement pas des essais contrôlés randomisés.	SIGN	Fournit les outils pour faire face à un éventail de types d'études, et les check-lists de validité interne globale sont faciles à utiliser pour ceux qui ont une expérience limitée en analyse critique des articles.
Essentiellement des essais contrôlés	GRADE	Mieux utilisé par ceux qui ont une expérience dans l'analyse critique des

Type d'études	Système envisagé	Expérience des évaluateurs
randomisés et des méta-analyses.		articles et une bonne connaissance de l'épidémiologie.
Mélange de types d'études incluant des études qualitatives, sociologiques, opinions d'auteurs, et quantitatives.	NSF-LTC	-

Selon un rapport du *Health Services Assessment Collaboration* de Nouvelle-Zélande (2009) les systèmes les plus fréquemment utilisés et très bien cotés dans la littérature sont par ordre alpha-bétique, les systèmes du GRADE *working group*, du NICE, de l'*Oxford Centre for Evidence-based Medicine* (CEBM), du SIGN (37).

► Retours d'expérience de l'utilisation du système GRADE

Le système GRADE a été conçu dans l'optique d'uniformiser la gradation de la qualité des données scientifiques et des recommandations (10,38). Il permet d'élaborer des recommandations fondées sur une évaluation explicite des données de la littérature. Il convient pour :

- résumer les données scientifiques extraites de revues systématiques et de méta-analyses dans les tableaux de résumés des résultats (GRADE n'est pas un outil pour réaliser des revues systématiques et des méta-analyses) ;
- grader la qualité des données scientifiques résumées dans les tableaux de résultats ;
- grader la force des recommandations.

L'expérience de l'utilisation de GRADE a été rapportée dans deux études (39,40).

L'OMS a utilisé le système GRADE pour élaborer des recommandations concernant des troubles mentaux, neurologiques et liés à l'utilisation de substances psycho-actives (39). Les auteurs de l'article rapportant cette expérience indiquent que le système GRADE est une structure utile pour synthétiser et présenter les données scientifiques concernant l'efficacité réelle des interventions cliniques. Il aide à révéler la subjectivité implicite puisqu'il requiert une approche systématique et explicite pour interpréter les données scientifiques.

Cependant, le processus pourrait être amélioré dans plusieurs domaines (tableau 38) : inclusion des données scientifiques issues d'études non randomisées ou qui ne peuvent pas faire l'objet d'une méta-analyse ; reproductibilité et cohérence interne et choix d'un parmi plusieurs critères de jugement pour chaque résultat. L'élaboration de recommandations implique non seulement une revue systématique de la littérature et une évaluation de la qualité des données scientifiques et du rapport bénéfices inconvénients, mais aussi la prise en compte explicite d'autres aspects tels que des jugements de valeurs, l'utilisation des ressources et la faisabilité.

Tableau 38. Défis à relever pour développer des recommandations selon le système GRADE d'après l'expérience de l'OMS (39)

Étape d'élaboration de la RBP	Problèmes	Solutions
De la délimitation des questions aux profils des données scientifiques		
Identification des résultats importants et décisifs	L'identification des résultats importants et décisifs est cruciale pour la suite du processus. Souvent les essais cliniques rapportent les résultats sur des résultats qui peuvent être facilement mesurés*,	Durant la cotation des résultats (en décisifs, importants, non importants), il est demandé aux membres du groupe de travail d'apprécier le choix et la valeur des résultats qui devraient être

Étape d'élaboration de la RBP	Problèmes	Solutions
	et omettent d'autres résultats clés qui sont plus difficiles à quantifier.	mesurés plutôt que ceux qui ont été mesurés.
Sélectionner la mesure qui décrit au mieux chaque critère de jugement	Souvent plusieurs mesures sont disponibles pour chaque critère de jugement. L'absence d'instructions pour guider la sélection d'une mesure pour un critère de jugement expose à une hétérogénéité et un risque de biais (les rédacteurs pouvant choisir des mesures qui répondent à leurs opinions *a priori*).	Instructions développées pour aider les rédacteurs à utiliser la même logique pour choisir la mesure d'un critère de jugement. Les rédacteurs ont la possibilité de recourir à une autre logique en la justifiant.
Reproductibilité et cohérence	Des facteurs (limites des études et caractère indirect des données) impliquent un jugement sur la qualité d'un groupe d'essais cliniques, tandis que d'autres facteurs (i.e. : hétérogénéité, imprécision et biais de publication) impliquent un jugement sur le processus de méta-analyse des données. Pour le facteur limite des études, les rédacteurs ont besoin de juger le risque de biais de chaque essai individuel inclus dans la revue systématique sélectionnée, puis de juger si la proportion d'essais à risque de biais est à l'origine d'un risque de biais pour l'ensemble des études. Ceci pose un problème de faisabilité (pour accéder à tous les essais cliniques) quand une revue inclut de nombreux essais, et aussi de reproductibilité et de cohérence. Pour le processus de méta-analyse, un seul jugement est nécessaire, mais il est probable que différents rédacteurs appliqueront différents critères, et que le même rédacteur pourra appliquer différents critères dans différentes situations.	Instructions développées pour grader la qualité des données scientifiques quand le champ des recommandations est vaste et que les rédacteurs sont multiples (afin d'augmenter la reproductibilité du processus et la cohérence des jugements).
Biais de publication	Les instructions de GRADE pour faire un tel jugement sont inadéquates et les méthodes actuellement disponibles pour détecter un biais de publication sont défaillantes (*funnel plot* controversé et pas toujours rapporté dans la revue ; il est demandé de vérifier si les auteurs de la revue ont inclus des données non publiées mais la méthode de l'étude est généralement non disponible et la faisabilité est illusoire si les essais sont très nombreux).	-
Données scientifiques issues d'essais non randomisés	Les études épidémiologiques et les études qualitatives ne peuvent pas être décrites facilement dans les tableaux	L'inclusion d'une description narrative des données scientifiques complémentaires dans le texte

Étape d'élaboration de la RBP	Problèmes	Solutions
	GRADE. Le risque est d'omettre la contribution des études non randomisées dans les profils des données scientifiques.	correspondant au profil des données scientifiques n'est pas satisfaisante. Cela expose à un déséquilibre entre les données scientifiques de qualité élevée et celles de qualité faible, avec une moindre considération donnée à ce qui est présenté en dehors des profils des données scientifiques.
Des données scientifiques aux recommandations		
Données scientifiques *versus* valeurs, préférences et questions de faisabilité	La méthode pour prendre en considération les aspects relatifs aux valeurs, préférences et faisabilité est moins développée que la méthode pour synthétiser et pour porter un jugement sur les données scientifiques.	Développement d'une check-list et d'un modèle pour donner une visibilité à la fois aux données scientifiques et aux valeurs-préférences-faisabilité. Pour chaque question : 1- Synthèse des données scientifiques, de la cotation de leur qualité et description narrative du rapport entre effets souhaitables et indésirables = rédaction de la version 0 de la recommandation. 2- Prise en considération des valeurs-préférences-faisabilité = version modifiée de la recommandation.
Données scientifiques *versus* valeurs, préférences et questions de faisabilité	-	-

* : et pour lesquels on peut espérer mettre en évidence des changements ou des différences sur un intervalle de temps relativement court (41).

Le NICE a démarré, en 2007, une étude pilote d'utilisation de GRADE dans trois RBP portant sur des questions d'interventions. Par la suite, cette utilisation a été étendue à l'ensemble du programme RBP du NICE. Cette expérience a été rapportée dans un article décrivant les défis rencontrés et les solutions qui ont été développées (40).

Le système GRADE encourage une évaluation de la confiance en l'estimation de l'effet pour chaque résultat, et un arbitrage entre effets souhaitables, effets indésirables et coûts amenant à un jugement sur la force de la recommandation.

Le système GRADE a été utilisé dans le but de supprimer la correspondance directe parfois inappropriée entre le type d'étude (*via* la hiérarchie du niveau de preuve) et la force des recommandations, en permettant la séparation des jugements sur la confiance en l'estimation de l'effet d'une part, des jugements sur la force des recommandations d'autre part.

Le système GRADE implique un changement dans la façon de penser par rapport aux précédentes méthodes d'évaluation d'un ensemble de données scientifiques. D'autres problèmes ont été rencontrés, incluant la définition des résultats, la définition des différences importantes minimales et l'imprécision, l'intégration des données scientifiques médico-économiques, l'évaluation des données scientifiques concernant les comparaisons multiples (par exemple les méta-analyses en réseau) (tableau 39). Des travaux sont nécessaires pour améliorer l'utilisation de GRADE pour les études non randomisées et les études sur la validité des tests diagnostiques.

Tableau 39. Problèmes rencontrés et solutions trouvées pour développer des recommandations selon le système GRADE d'après l'expérience du NICE (40)

Problèmes		Solutions
Changer le mode de pensée	Changement de mentalité partant des précédentes méthodes d'évaluation des données scientifiques vers une approche plus structurée et transparente de prise de décision.	Formation à GRADE pour toute l'équipe technique. Les problèmes pratiques sont discutés dans des sessions de formation continue et existence d'une « foire aux questions ». Tous les groupes de travail reçoivent un entraînement aux principes et à la pratique de GRADE en début d'élaboration d'une RBP.
Spécifier les résultats importants	Le GRADE *working group* suggère de limiter le nombre de résultats à 7. Persuader le groupe de travail de limiter le nombre de résultats importants en particulier pour les maladies systémiques ou chroniques.	Le NICE n'a pas imposé de limite rigide, mais a spécifié une liste de résultats par avance dans les protocoles de revue pour les RBP.
Imprécision	Les jugements sur l'importance des effets (définition du seuil de décision clinique pour recommander ou ne pas recommander le traitement) devraient être faits avant la revue des données scientifiques. Mais il a été difficile pour les groupes de travail et les équipes techniques de spécifier les différences de résultats importantes ou significatives.	Les *national collaborating centers* (NCCs) et les groupes de travail ont utilisé diverses approches en fonction de la question considérée. Question en discussion.
Intégrer une analyse économique	Le système GRADE permet d'intégrer l'utilisation des ressources comme résultat dans les profils des données scientifiques, mais la restriction du nombre de résultats pourrait rendre difficile ou impossible d'intégrer toutes les ressources pertinentes impliquées. Les modélisations ne peuvent pas être présentées. Les analyses coût-efficacité sont fondées sur les résultats de méthodes complexes de synthèse des données scientifiques (par exemple : les méta-analyses en réseau) qui ne peuvent pas être intégrées actuellement.	Conception de profils des données économiques inspirés des profils des données scientifiques de GRADE – en cours de développement.
Comparer des interventions multiples	Les analyses dans ces situations sont habituellement faites par paires, avec un profil des données scientifiques pour chaque comparaison, sans matrice pour synthétiser l'interprétation globale de ces profils.	L'adaptation des profils existants pour convenir à des comparaisons thérapeutiques multiples est en cours.
Catégoriser des recommandations	GRADE suggère deux catégorises de recommandations « forte » ou « faible »	Reflète le concept de force dans la formulation de la recommandation en préférant utiliser : - « offrir » pour les recommandations dont nous sommes certains de la force et de la qualité des données

Problèmes		Solutions
		scientifiques, et pour lesquelles nous avons un petit doute sur le fait que les effets souhaitables l'emportent sur les effets indésirables ; - « considérer » pour les recommandations dont nous sommes moins sûrs ; - écrire une recommandation de recherche clinique quand l'incertitude est grande.
Utiliser le logiciel GRADEpro pour réaliser les profils des données scientifiques	Au début du projet GRADEpro n'était pas fiable et ne permettait pas d'évaluer les études observationnelles.	Les équipes techniques ont dû trouver d'autres manières de présenter les profils des données scientifiques.

D'autres aspects du système GRADE suscitent des préoccupations (42) :

- Le système GRADE demande de coter la qualité des données scientifiques en partant du type d'étude, et d'augmenter ou de diminuer le niveau de qualité en prenant en compte divers facteurs. Mais il s'agit de combiner des éléments disparates. Ces facteurs sont fondamentalement différents et ne peuvent pas être additionnés ou soustraits.
- Le système GRADE considère que les options de prise en charge associées à des recommandations fortes peuvent servir de base pour des critères d'évaluation de la qualité et, que pour les décideurs, elles peuvent être adoptées comme mesure de santé. Savoir quelles recommandations sont les meilleures (ou sont valides) n'est pas simple. Il est arrivé que des recommandations fortes aient été plus tard invalidées.

3. Recommandations pour l'élaboration de recommandations

3.1 Institute of Medicine

En 2011, l'*Institute of Medicine* (IOM) a publié des recommandations pour élaborer des recommandations fiables. L'IOM a donné une nouvelle définition des RBP : **« Les RBP sont des propositions qui incluent des recommandations destinées à optimiser les soins qui sont fondées sur une revue systématique de la littérature et sur une évaluation des bénéfices et des inconvénients des différentes options des soins**[7] » (43).

Selon l'IOM, les RBP devraient apporter une explication claire des relations logiques entre les différentes options des soins et les résultats cliniques, et présenter une cotation de la qualité des données scientifiques et de la force des recommandations.

L'IOM a proposé des standards pour développer des RBP fiables (43). Les RBP devraient être fondées sur des bonnes revues systématiques de la littérature (qui devraient remplir les standards IOM des revues systématiques de la recherche comparative sur l'efficacité clinique des traitements médicaux ou des interventions chirurgicales[8]) (43) (tableau 40).

Tableau 40. Recommandations de l'*Institute of Medicine,* 2011 (43)

Préciser les données scientifiques sur lesquelles sont fondées les recommandations et coter la force des recommandations

L'IOM recommande de fournir pour chaque recommandation :

- une explication du raisonnement qui sous-tend la recommandation, incluant :
  - une description claire des bénéfices et des inconvénients potentiels,
  - une synthèse des données scientifiques pertinentes disponibles (et les lacunes dans les données), une description de la qualité (incluant l'applicabilité), la quantité (incluant l'exhaustivité), et l'homogénéité de l'ensemble des résultats disponibles,
  - une explication de la part jouée par les valeurs, l'opinion, la théorie, et l'expérience clinique dans la rédaction de la recommandation ;
- une cotation du niveau de confiance (concernant la certitude) dans les données scientifiques sur lesquelles la recommandation est fondée ;
- une cotation de la force de la recommandation en considérant les points précédents ;
- une description et une explication de toute différence d'opinion par rapport à la recommandation.

Formuler des recommandations

Les recommandations devraient être formulées de manière standardisée en détaillant précisément l'action recommandée, et dans quelles circonstances elle doit être réalisée.

Les recommandations fortes devraient être formulées de telle sorte que le respect de la recommandation puisse être évalué.

[7] *Clinical practice guidelines are statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options (43).*

[8] http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews/Standards.aspx.

L'élaboration des RBP nécessite habituellement une interprétation des données scientifiques concernant différentes questions.

Les recommandations vont au-delà d'une simple revue systématique de la littérature, et reflètent des jugements de valeurs concernant toutes les questions pertinentes pour la décision clinique. Les recommandations n'émergent pas toujours directement des données empiriques revues par un groupe de travail. Les facteurs qui interviennent dans la divergence entre les données scientifiques et les recommandations incluent (43) :

- la nature des données scientifiques de qualité faible ou non applicables parce que dans ce cas, le jugement a de fortes chances d'entrer en jeu ;
- l'expérience clinique ;
- les préférences des patients ;
- la disponibilité du traitement ;
- les valeurs des cliniciens.

3.2 Guidelines International Network

Le *Guidelines International Network* (GIN) a listé les composants clés pour l'élaboration de RBP de bonne qualité et fiables (44). En ce qui concerne la revue de la littérature, le fondement des recommandations, la cotation des données scientifiques et des recommandations, le GIN propose que :

- **ceux qui élaborent des RBP utilisent les méthodes des revues systématiques de la littérature pour identifier et évaluer des données scientifiques relatives au thème de la RBP :** les revues systématiques font appel à des méthodes rigoureuses pour identifier les questions cliniques, les critères d'inclusion et d'exclusion, et à des méthodes pour coter la qualité des données scientifiques disponibles. Le GIN a développé des tableaux de résumé standardisé des études pour les questions relatives aux interventions et au diagnostic (annexe 9) ;
- **une recommandation soit énoncée clairement et fondée sur des données scientifiques pour les bénéfices, les inconvénients et si possible les coûts.** Les recommandations relatives aux interventions devraient être rédigées dans un langage direct et non ambigu de manière à refléter la force des données scientifiques. Ceux qui élaborent des recommandations devraient s'efforcer d'utiliser des termes comme « recommande », et éviter des mots et des phrases vagues comme « peut » ou « considère », à moins qu'il n'existe une incertitude réelle sur les données scientifiques d'efficacité en population, ces termes n'étant pas d'un grand secours pour la mise en œuvre en pratique ;
- **ceux qui élaborent des RBP utilisent un système de cotation pour informer de la qualité et de la reproductibilité à la fois des données scientifiques et de la force de leurs recommandations.**

  La force des recommandations devrait être attribuée sur la base d'une évaluation :

  - des résultats pour les bénéfices et les inconvénients,
  - de l'homogénéité,
  - de l'effet clinique,
  - de la généralisabilité et de l'applicabilité,
  - et sur les préférences des patients.

Plusieurs systèmes de gradation sont actuellement disponibles.

4. Retour sur le niveau de preuve

Un article a été consacré au niveau de preuve dans le cadre de l'évaluation thérapeutique (45). L'objectif a été de déterminer dans un premier temps les éléments, dont dépend le niveau de preuve d'un essai clinique et dans un deuxième temps si les échelles connues prennent en compte ces éléments.

La pratique de l'*Evidence based medicine* nécessite un outil pour évaluer et discerner parmi les données disponibles celles fondées sur des bases objectives et ainsi, faciliter l'accès à des informations fiables (45). Le niveau de preuve permet de comparer les résultats issus de multiples études testant une même hypothèse.

4.1 Niveau de preuve d'un essai clinique

Le concept de niveau de preuve (*level of evidence* ou *strength of evidence*) est souvent confondu avec le concept de qualité (45). Mais le niveau de preuve recouvre plus que la qualité d'une étude. La notion de qualité reflète généralement une propriété qui n'est pas directement liée à l'objectif de l'étude, tandis que le niveau de preuve des résultats d'une étude inclut à la fois l'objectif et la nature de l'étude. L'objectif d'une étude est associé à la formulation d'une hypothèse tandis que la qualité d'une étude peut rester complètement indépendante de l'hypothèse à tester.

Ainsi, il est possible que les résultats d'une étude qui a rempli tous les critères d'un essai contrôlé randomisé n'offrent qu'un niveau de preuve modeste parce que les résultats de l'étude et l'hypothèse testée ne correspondent pas à la question qu'un utilisateur (selon son point de vue soit de chercheur, soit de clinicien ou autre) cherche à évaluer.

Les trois dimensions clés d'une échelle de niveaux de preuve sont :

- la première **relative à la méthode** de l'étude **:** il s'agit de la capacité du protocole expérimental à minimiser les biais ;
- la deuxième est **qualitative :** il s'agit de la manière dont l'étude a été réalisée en pratique et des mesures prises pour minimiser les biais de sélection et/ou le nombre de patients perdus de vue, qui renseigne sur la force des données sur lesquelles la taille de l'effet est estimée ;
- la troisième, spécifique aux études d'intervention, est liée à la **pertinence clinique de l'hypothèse testée,** donc des résultats attendus de l'étude et plus précisément du bénéfice clinique qu'ils représentent pour les patients ainsi que du contexte dans lequel l'intervention est réalisée. Cela conduit à prendre en considération :
  - la nature des critères de jugement cliniques choisis,
  - la nature de l'intervention,
  - l'intervention de référence utilisée comme contrôle,
  - pour un traitement, les traitements adjuvants autorisés,
  - la pertinence de la durée de suivi,
  - le choix de la population à l'étude.

Remarque : l'intensité attendue de l'effet, bien qu'elle fasse partie de l'hypothèse et qu'elle soit un des composants de la pertinence clinique, n'est pas prise en compte dans la dimension clinique du niveau de preuve.

La majeure partie des échelles de niveau de preuve existantes pour les essais cliniques est centrée sur la méthode de l'étude. Certaines peuvent inclure la qualité de réalisation, mais souvent partiellement. Quelques-unes prennent en compte la pertinence clinique et seulement de manière incomplète. Les questions de reproductibilité sont rarement considérées.

4.2 Niveau de preuve d'une revue systématique

Les auteurs ont également étudié des échelles de niveau de preuve utilisées pour les revues systématiques (45). Les revues systématiques correspondent à une revue de la littérature à un moment donné, elles fondent leur système de hiérarchie sur la qualité du protocole des essais inclus dans la revue, et elles prennent peu ou pas en considération la méthode de la revue en elle-même. Le niveau de preuve paraît découler d'une « somme » d'études de bonne qualité méthodologique (voir annexe 10).

Ces échelles devraient prendre en compte non seulement la méthode de l'étude, la qualité de réalisation, la pertinence clinique de l'hypothèse testée (cf. essai clinique), mais aussi le processus utilisé pour produire la synthèse (par exemple : méta-analyse) (45).

Dans une revue systématique, la validité finale repose en partie sur la validité de chacun des essais individuels poolés. La nécessité de prendre en compte la validité de chacun des essais poolés rend la cotation du niveau de preuve de la revue systématique encore plus difficile que celle des essais cliniques.

Annexe 1. Recherche documentaire

La recherche documentaire a consisté en :

- L'identification des guides méthodologiques d'élaboration de recommandations de bonne pratique et des systèmes de gradation et de niveaux de preuves ; les sites internet explorés pour cela sont présentés tableau 41.

Tableau 41. Type de protocole préférentiellement proposé pour une question donnée

Sites internet explorés
Agencia de Evaluación de Tecnología e Investigación Médicas de Cataluña
Agencia de Evaluación de Tecnologías Sanitarias de Galicia
Agency for Healthcare Research and Quality
Alberta Heritage Foundation for Medical Research
American College of Physicians
Australian Government - Department of Health and Ageing
Centre fédéral d'expertise des soins de santé
Centre for Evidence-Based Medicine (Oxford)
CISMeF
CMAInfobase
Department of Health UK
Cochrane Library Database
GIN (Guidelines International Network)
Grading of Recommendations Assessment, Development and Evaluation GRADE Working group
Haute Autorité de Santé
Institute for Clinical Systems Improvement
Institute of Medicine (IOM)
Institut National d'Excellence en Santé et en Services Sociaux
National Health and Medical Research Council
National Institute for Health and Clinical Excellence
National Institutes of Health
New Zealand Guidelines Group
Scottish Intercollegiate Guidelines Network
Servicio de Evaluación de Tecnologías Sanitarias OSTEBA
Singapore Ministry of Health
Tripdatabase
US Preventive Services Task Force
World Health Organization

- L'utilisation du fonds documentaire (ouvrages, articles scientifiques) tenu à la HAS depuis 1990, en complément une veille bibliographique a été effectuée par la surveillance des sommaires des revues suivantes : *Annals of Internal Medicine*, *Archives of Internal Medicine*, *British Medical Journal*, *Canadian Medical Association Journal*, JAMA, *Lancet*, *New England Journal of Medicine*, *Journal of Clinical Epidemiology,* etc.

Annexe 2. Glossaire

***Evidence*[9]*-based medicine*.** A été définie par Sackett *et al*, comme l'utilisation consciencieuse, explicite et judicieuse des meilleures données actuelles pour la prise de décision dans les soins à prodiguer à des patients individuels. La pratique de l'EBM signifie l'intégration à l'expérience clinique individuelle des meilleures données objectives disponibles de source externe découlant d'une recherche systématique. Par expérience clinique individuelle, les auteurs désignent les compétences et le jugement que chaque clinicien acquiert par la pratique clinique. Par meilleures données objectives disponibles de source externe, les auteurs veulent dire des données cliniquement pertinentes souvent issues des sciences médicales fondamentales, mais surtout de la recherche clinique centrée sur le patient concernant les tests diagnostiques, les marqueurs pronostiques, l'efficacité et la sécurité des schémas thérapeutiques, de réadaptation et de prévention. Les données objectives de source externe infirment des tests diagnostiques ou des traitements précédemment acceptés et les remplacent par des nouveaux plus puissants, plus précis, plus efficaces et à moindres risques (49).

L'EBM n'est pas limitée aux essais contrôlés randomisés et aux méta-analyses. Elle implique d'identifier les meilleures données de source externe permettant de répondre aux questions cliniques. Pour évaluer la validité diagnostique d'un test, nous avons besoin d'études transversales appropriées de patients chez lesquels on suspecte la maladie en question, et non d'un essai contrôlé randomisé. Pour une question sur le pronostic, nous avons besoin d'études de suivi appropriées de patients inclus à un moment identique précocement dans l'évolution clinique de leur maladie. Quelquefois les données proviendront des sciences fondamentales comme la génétique ou l'immunologie. C'est pour les questions relatives aux traitements que nous devrions éviter les approches non expérimentales, puisqu'elles conduisent régulièrement à des conclusions faussement positives sur l'efficacité...

Dans le Grand dictionnaire terminologique, la définition de l'EBM est la suivante : médecine qui est fondée sur une prise en compte des meilleures données scientifiques actuelles dans la prise de décisions concernant les malades. On trouve en français un grand nombre d'expressions pour désigner cette notion : « médecine factuelle », « médecine fondée sur les preuves », « médecine fondée sur des données probantes », « médecine fondée sur des faits » et « médecine fondée sur des faits prouvés » (50).

Le terme « données actuelles de la science » apparaît dans le droit français en 1946 (Cass. 1re civ., 20 février 1946, Gaz. Pal. 1946, 1, 209). Il correspond à des données admises par l'ensemble du corps médical (51). Il a été repris dans le Code de la santé publique Section 4 : Déontologie des masseurs-kinésithérapeutes, Sous-section 2 : Devoirs envers les patients, Article R4321-80 : « Dès lors qu'il a accepté de répondre à une demande, le masseur-kinésithérapeute s'engage personnellement à assurer au patient des soins consciencieux, attentifs et fondés sur les données actuelles de la science » (version en vigueur au 16 janvier 2012, depuis le 6 novembre 2008.

Généralisabilité. Capacité d'une observation d'être généralisée à une classe d'observations à laquelle elle appartient (dictionnaire Larousse en ligne consulté le 08/01/2013).

Revue systématique. Une revue systématique est un type d'investigation scientifique de la littérature sur un sujet donné dans lequel les « sujets » sont les articles évalués (52). Ce sont des études rétrospectives observationnelles et sont exposées aux erreurs systématiques et dues au hasard. La qualité de la revue dépend des méthodes utilisées pour diminuer les erreurs et les biais. Les revues systématiques sont développées selon un protocole comprenant (53) :

- une question clinique précise, souvent limitée ; formulée selon quatre variables : la population spécifique et le contexte des soins, la maladie d'intérêt, l'exposition à un test ou à un traitement, et un ou plusieurs résultats spécifiques ;
- une recherche documentaire explicite et les sources exploitées ;
- l'identification et la sélection des études selon des critères d'inclusion et d'exclusion ;
- les types de données à extraire de chaque article ;

[9] *Evidence : ground for belief or disbelief ; data on which to base proof or to establish truth or falsehood* (dictionnaire Collins en ligne consulté le 17/01/2012 http://dictionnaire.reverso.net/anglais-definition/evidence).

- la synthèse des données sous forme d'un texte résumé (revue systématique qualitative) ou d'une revue systématique quantitative ou méta-analyse, dans laquelle on utilise des méthodes statistiques pour combiner les résultats de deux études ou plus.

Les RBP sont fondées sur des revues systématiques de la littérature.

Validité interne d'une étude. C'est la qualité de la méthode d'une étude. Elle reflète jusqu'à quel point on peut montrer que tous les aspects de la conception d'une étude et la manière dont l'étude a été menée ont pu protéger vis-à-vis de biais systématiques, de biais non systématiques et d'une erreur inférentielle (54).

Validité externe d'une étude. Elle correspond à la cohérence avec les connaissances et les données qui ne sont pas celles de l'étude (physiopathologique, pharmacologiques, épidémiologiques).

Annexe 3. Type de protocole préférentiellement proposé pour une question donnée

Tableau 42. Type de protocole préférentiellement proposé pour une question donnée

Question	Protocole
THÉRAPEUTIQUE Efficacité	Étude contrôlée randomisée
THÉRAPEUTIQUE Sécurité	Étude contrôlée randomisée ou suivi de cohorte
DIAGNOSTIC Reproductibilité/Variabilité	Transversal comparatif avec répétition de mesure
DIAGNOSTIC Sensibilité/Spécificité	Transversal comparatif avec étalon-or
DIAGNOSTIC Efficacité/Utilité	Étude contrôlée randomisée
DIAGNOSTIC Stratégie	Étude contrôlée randomisée ou arbre décisionnel
CAUSALITÉ Phénomène contrôlable fréquent	Étude contrôlée randomisée
CAUSALITÉ Phénomène non contrôlable fréquent	Suivi de cohorte (exposés/non exposés)
CAUSALITÉ Phénomène rare	Étude cas-témoin
PRONOSTIC Maladie fréquente	Étude contrôlée randomisée ou suivi de cohorte
PRONOSTIC Maladie rare	Étude cas-témoin

Annexe 4. Évaluation de biomarqueurs pronostiques et prédictifs de la réponse aux traitements

Lorsqu'il s'agit d'évaluation de biomarqueurs pronostiques et prédictifs de la réponse aux traitements, d'autres systèmes d'attribution de niveaux de preuve aux données de la littérature existent. Parmi eux, la grille de Hayes (55) et sa mise à jour (celle de Simon (56)) définissent une classification globale de niveaux de preuve, se rapportant à la fois à la validité analytique, à la validité clinique et à l'utilité clinique d'un biomarqueur, qu'il soit mesuré par des techniques biochimiques ou de biologie moléculaire.

- Validité analytique : capacité du test à mesurer d'une manière précise et reproductible le génotype ou la substance à analyser :
  - préanalytique : *timing* et site du prélèvement, mode de prélèvement, mode d'extraction (sous forme protéique ou d'acides nucléiques), acheminement et conservation des échantillons, stabilité des échantillons ;
  - analytique : emploi de techniques internationales et standardisées, calibration standardisée et précise, sensibilité, spécificité, reproductibilité, robustesse, contrôle de qualité interne, participation aux programmes d'évaluation des compétences (contrôle de qualité externe) ;
  - postanalytique (report des résultats) : minimisation des risques des résultats erronés attribuables aux interférences (ex. : anticorps hétérophiles, *turn-over* ou temps de demi-vie du marqueur, cinétique), emploi dans les études de validation cliniques du même test et des mêmes seuils ainsi que d'une population similaire à celle employée dans l'étude expérimentale.
- Validité clinique : capacité du test à identifier d'une manière précise et reproductible un sous-groupe à risque de patients ou de prédire une évolution clinique de patients liée ou non à la contribution d'un traitement ;
- Utilité clinique : rapport bénéfice/risque lorsque le marqueur est employé dans l'objectif de modifier la prise en charge des patients ; l'utilité correspond à la valeur ajoutée/puissance du marqueur par rapport aux autres outils existants.

À noter que cette grille est limitée à l'évaluation du niveau de preuve et ne permet pas, utilisée seule, l'attribution d'un grade.

Figure 1 : grille de Hayes (55)

Hypothèse:
Cible moléculaire potentielle

↓

Phase 1

Objectif : caractérisation et mesure de la cible : définition de la méthode analytique
Acteur(s) : laboratoire de recherche
Méthodes : étude de faisabilité : large étude rétrospective

LOE III

↓

Phase 2

Objectif : standardisation, contrôle qualité
Acteur(s) : laboratoire expert
Méthodes : Analyse* de protocoles cliniques prospectifs où l'étude du marqueur est un objectif secondaire tant sur le plan du descriptif du protocole que de l'étude statistique

LOE II

↓

Phase 3

Objectif : utilité clinique
Acteur(s) : réseau de laboratoires et groupes cliniques
Méthodes : compilation positive et concordante des études de niveau II et III (études rétrospectives) ou méta-analyses positives

↓

Phase 4

Objectif : valeurs pronostique et/ou prédictive
Acteur(s) : groupes cliniques
Méthodes : essai prospectif dédié où l'évaluation du biomarqueur tumoral est l'objectif principal de l'étude dans son descriptif et analyse statistique (inclusion de groupes homogènes de patients sur les plans cliniques et thérapeutiques)

LOE I

* Le marqueur, étudié en tant qu'objectif secondaire dans les essais prospectifs, est analysé en rétrospectif.

Tableau 43. Niveaux de preuve selon Simon (56)

Niveau de preuve	Description des études	Études de validation disponibles
LOE IA	Prospectives.	Non nécessaires
LOE IB	Prospectives-rétrospectives utilisant des échantillons archivés prospectivement dans le cadre d'un essai clinique.	Une étude ou plus avec des résultats concordants. Échantillons provenant d'essais cliniques différents.
LOE IIB	Prospectives-rétrospectives utilisant des échantillons archivés prospectivement dans le cadre d'un essai clinique.	Aucune étude ou plusieurs études avec des résultats non concordants.
LOE IIC	Prospectives-observationnelles (registre).	Deux études ou plus avec des résultats concordants.
LOE IIIC	Prospectives-observationnelles (registre).	Aucune étude ou une étude avec des résultats concordants ou non concordants.
LOE IV-VD	Rétrospectives-observationnelles.	Non applicable.

Particularité : « attribution d'un niveau de preuve LOE IB à des études rétrospectives s'appuyant sur des échantillons archivés d'un biomarqueur qui avaient été collectés prospectivement dans le cadre d'un essai randomisé non dédié à l'étude de ce marqueur ».

- Les résultats doivent être confirmés dans le cadre d'au moins une autre étude similaire à la précédente dont les échantillons proviennent d'un essai différent.
- Les échantillons disponibles doivent être en quantité suffisante permettant d'assurer la représentativité de la population de l'essai, et donc une puissance acceptable de l'étude (au moins 2/3) ou que les patients soient sélectionnés de manière à éviter les biais de sélection par exemple en mimant une randomisation via un schéma mathématique.
- Les données préanalytiques doivent être parfaitement contrôlées et doivent correspondre à la pratique actuelle (procédures opératoires standards) :
  - le test doit être validé en analytique et en préanalytique pour son utilisation sur des échantillons archivés ;
  - la technologie du marqueur doit être précise, reproductible, robuste.
- Le design de l'étude doit être complètement défini et écrit avant la conduite des essais sur des tissus archivés ; et doit être dédié à l'évaluation d'un seul marqueur bien défini.
- Le design et l'analyse de l'étude doivent être adéquats et appropriés à l'étude de l'utilité du marqueur pour une utilisation clinique précise.
- Les patientes ne doivent pas avoir reçu un traitement adjuvant.
- Les données cliniques (*outcome* et traitement) doivent être en aveugle : le test doit être conduit sans que les données cliniques ne soient communiquées.

Annexe 5. Analyse des différents systèmes – GRADE *working group*, 2004

L'étude a eu pour objectif de décrire une analyse critique des six principaux systèmes de gradation du niveau de preuve et de la force des recommandations et les résultats de cette analyse (18).

Les six systèmes analysés ont été ceux des organismes suivants (37) :

- *American College of Chest Physician* (2001) ;
- *Australian National Health and Medical Research Council* (2000) ;
- *Oxford Centre for Evidence-based Medicine* (1998) ;
- *Scottish Intercollegiate Guidelines Network* (2001) ;
- *US Preventive Services Task Force* (2001) ;
- *US Task Force on Community Preventive Services* (USTFCPS, 2000).

Une description de ces systèmes a été donnée à douze personnes collaborant au groupe GRADE qui ont analysé de façon indépendante les six systèmes.

Les douze évaluations avaient l'expérience d'au moins un système, et la plupart avaient aidé au développement d'un des six systèmes analysés.

La sensibilité des systèmes pour grader les données scientifiques et les recommandations a été évaluée avec douze critères (57) (tableau 44).

Tableau 44. Critères utilisés pour évaluer la sensibilité des systèmes de gradations des données scientifiques et des recommandations par le GRADE *working group* (18)

Critères appliqués pour grader à la fois le niveau de preuve et la force des recommandations	Cotation
1. Dans quelle mesure l'approche est applicable aux différents types de questions efficacité (en population), risques, diagnostic, pronostic ?	Non, pas sûr, oui
2. Dans quelle mesure le système peut-il être utilisé par différents publics : patients, professionnels, décideurs ?	Faible, certaine, large
3. Le système est-il clair et simple ?	Pas très clair, assez clair, très clair
4. Combien de fois une information non habituellement disponible sera-t-elle nécessaire ?	Souvent, quelquefois, rarement
5. Dans quelle mesure les décisions subjectives sont-elles nécessaires ?	Souvent, quelquefois, rarement
6. Y a-t-il des dimensions incluses qui ne sont pas prises en compte dans le niveau de preuve ou la force des recommandations ?	Oui, partiellement, non
7. Y a-t-il des dimensions importantes qui auraient du être incluses et qui ne l'ont pas été ?	Non, partiellement, oui
8. La manière dont les dimensions incluses sont agrégées est-elle claire et simple ?	Non, partiellement, oui
9. La manière dont les dimensions incluses sont agrégées est-elle appropriée ?	Non, partiellement, oui
10. Les catégories sont-elles suffisantes pour faire la distinction entre différents niveaux de preuve et forces des recommandations ?	Non, partiellement, oui
11. Quelle est-la probabilité que le système réussisse à faire la distinction entre un niveau de preuve élevé et faible ou entre une recommandation forte et faible ?	Pas très probable, assez probable, très probable
12. Les évaluations sont-elles reproductibles ?	Probablement pas, pas sûr, probablement

Les résultats ont montré un « faible accord » entre les douze cotateurs sur la sensibilité des six systèmes. Les résultats ont montré que :

- un seul des systèmes, celui de l'OCEBM, convient pour les quatre types de questions (efficacité, risques, diagnostic, pronostic) (accord d'au moins 10 cotateurs sur les 12). Le système de l'ACCP convient pour les questions d'efficacité et de risques. Les systèmes du SIGN et de l'USTFCPS conviennent seulement pour les questions sur l'efficacité ;
- aucun des systèmes ne convient pour être utilisé à la fois par les patients, les professionnels et les décideurs ;
- la plupart des cotateurs ne sont pas sûrs de la reproductibilité des évaluations réalisées avec n'importe lequel de ces systèmes.

Le « faible accord » a été rapporté au fait que certains cotateurs avaient une expérience pratique d'un des systèmes, et leur cotation a pu être biaisée en faveur du système avec lequel ils étaient le plus familier ; les critères ont été appliqués pour grader à la fois le niveau de preuve et la force des recommandations ; quelques-uns des critères n'étaient pas clairs, et ont été appliqués ou interprétés différemment par les cotateurs.

Bien que les auteurs aient limité leur analyse à six systèmes, d'autres systèmes de gradation utilisés par cinquante et une organisations qui financent des RBP (identifiés à partir de la *National Guidelines Clearing House* de l'US AHRQ) ont été comparés aux six systèmes analysés. Ces autres systèmes étaient en substance des variations des six systèmes analysés. De là, les auteurs ont conclu que toutes les approches utilisées pour grader les niveaux de preuve et la force des recommandations avaient d'importants défauts.

Annexe 6. Niveaux de preuves de l'*Oxford Centre for Evidence-Based Medicine*

L'*Oxford Centre for Evidence-based Medicine* a actualisé ses niveaux de preuve en 2011 (58). Ces niveaux de preuve (2011) sont une hiérarchisation de ce que sont probablement les meilleures données scientifiques disponibles (59). Ils permettent de répondre à une série de questions cliniques, concernant le diagnostic, le pronostic, la thérapie, et les effets indésirables.

Ces niveaux de preuve ne visent pas explicitement la rédaction de recommandations définitives.

Ils prennent en compte les revues systématiques. Mais, ils peuvent être utilisés même si il n'y a pas de revue systématique disponible (à la différence de GRADE qui présuppose qu'il y a une revue systématique pour toute question et qui perd son utilité quand il n'en existe pas) (59).

http://www.cebm.net/index.aspx?o=5653

Tableau 45. Niveaux de preuves de l'*Oxford Centre for Evidence-Based Medicine* (version française 2011) (58)

Question	Étape 1 (Niveau 1*)	Étape 2 (Niveau 2*)	Étape 3 (Niveau 3*)	Étape 4 (Niveau 4*)	Étape 5 (Niveau 5*)
Quelle est la fréquence du problème ?	Étude récente et locale sur des échantillons aléatoires (ou recensement).	Revue systématique d'études dont les conditions sont proches, mais non identiques aux conditions locales**.	Étude locale sur des échantillons non aléatoires**.	Série de cas**.	/
Le diagnostic ou le test de contrôle est-il valide ? (Diagnostic)	Revue systématique d'études transversales menées en aveugle et utilisant un standard de référence appliqué de manière constante.	Étude transversale menée en aveugle et utilisant un standard de référence appliqué de manière constante.	Série de cas à recrutement non consécutif ; étude transversale sans standard de référence appliqué de manière constante**.	Étude cas-témoins ; étude avec un standard de référence non indépendant ou de faible qualité**.	Raisonnement déductif fondé sur la pathophysiologie.
Que se passera-t-il si aucun traitement n'est appliqué ? (Pronostic)	Revue systématique d'études de cohortes où les patients sont inclus au début de leur maladie (*inception cohort*).	Étude de cohorte où les patients sont inclus au début de leur maladie (*inception cohort*).	Étude de cohorte ; considération du groupe contrôle (non traité) dans un essai contrôlé randomisé.	Série de cas ; étude cas-témoins ; étude de cohorte pronostique de pauvre qualité**.	/
Cette intervention est-elle bénéfique ? (Bénéfices du traitement)	Revue systématique d'essais contrôlés randomisés ou d'essais de taille 1 (*n-of-1 trials*).	Essai contrôlé randomisé ; étude d'observation avec effet majeur.	Étude de cohorte non randomisée**.	Série de cas ; étude cas-témoins ; étude contrôlée pour laquelle la collecte des données du groupe contrôle a précédé celle du groupe étudié**.	Raisonnement déductif fondé sur la pathophysiologie.
Quels sont les effets indésirables fréquents ? (Effets indésirables du traitement)	Revue systématique d'essais contrôlés randomisés ; revue systématique d'études cas-témoins recrutés dans la population d'une étude de cohorte ; revue systématique d'essais de taille 1 (*n-of-1 trials*) ; revue systématique d'études d'observation avec un effet majeur.	Essai contrôlé randomisé ; (exceptionnellement) étude d'observation avec effet majeur.	Étude de cohorte contrôlée non randomisée (surveillance post-commercialisation) à condition qu'il y ait un nombre suffisant de patients par rapport à la fréquence de l'événement (pour les effets à long terme, la durée du suivi doit être suffisante)**.	Série de cas ; étude cas-témoins ; étude contrôlée pour laquelle la collecte des données du groupe contrôle a précédé celle du groupe étudié**.	Raisonnement déductif fondé sur la pathophysiologie.
Quels sont les effets indésirables rares ? (Effets indésirables du traitement)	Revue systématique d'essais contrôlés randomisés ou d'essais de taille 1 (*n-of-1 trials*).	Essai contrôlé randomisé ; (exceptionnellement) étude d'observation avec effet majeur.			
Ce test (détection précoce) en vaut-il la peine ? (Dépistage)	Essai contrôlé randomisé ; (exceptionnellement) étude d'observation avec effet majeur.	Essai contrôlé randomisé.	Étude de cohorte contrôlée non randomisée**.	Série de cas ; étude cas-témoins ; étude contrôlée pour laquelle la collecte des données du groupe contrôle a précédé celle du groupe étudié**.	Raisonnement déductif fondé sur la pathophysiologie.

* : le niveau de preuve d'une étude peut être rétrogradé en raison des faiblesses intrinsèques de l'étude, d'imprécisions, du caractère indirect de la preuve, à cause de l'incohérence entre études, ou à cause de la taille de l'effet absolu qui est très petite ; le niveau de preuve peut être mieux classé si la taille de l'effet est grande ou très grande ; ** : une revue systématique est généralement meilleure qu'une étude individuelle.

Annexe 7. Synthèse des niveaux de preuve et gradations des principaux systèmes

Tableau 46. Niveau de preuve des études et gradation des recommandations des principaux systèmes : synthèse

Élément	HAS (2)	FNCLCC (9,33)	NZGG (3)
étude	**niveau de preuve :** Capacité de l'étude à répondre à la question posée, jugée sur : - la correspondance de l'étude au cadre de travail (sujet, population, critères de jugement) ; - l'adéquation du protocole d'étude à la question posée - l'existence de biais importants dans la réalisation dont l'adaptation de l'analyse statistique aux objectifs de l'étude ; - la puissance de l'étude (en particulier la taille de l'échantillon). Classification générale : • **fort niveau de preuve,** • **niveau de preuve intermédiaire,** • **faible niveau de preuve.**	-	Évaluation du type d'étude et du **score de qualité :** • **plus (+)** étude robuste dans laquelle tous ou la plupart des critères de validité sont remplis ; • **moins (-)** étude faible sur le plan de la méthode dans laquelle très peu de critères de validité sont remplis et il y a un risque de biais élevé ; • **neutre (∅)** étude pour laquelle tous les critères ne sont pas remplis mais les résultats de l'étude ne sont probablement pas affectés.
ensemble d'études	**Gradation de l'évidence scientifique :** C'est la conclusion des tableaux de synthèse de la littérature. Elle s'appuie sur : - l'existence de données pour répondre aux questions posées ; - le niveau de preuve des études disponibles ; - la cohérence de leurs résultats. Niveau de preuve scientifique fourni par la littérature : • **niveau 1** - essais comparatifs randomisés de forte puissance ; - méta-analyse d'essais comparatifs randomisés ; - analyse de décision fondée sur des études bien menées ; • **niveau 2** - essais comparatifs randomisés de faible puissance ;	**Les niveaux de preuve :** fonction du type et de la qualité des études disponibles ainsi que de l'homogénéité de leurs résultats. • **A** Il existe une (des) méta-analyse(s) « de bonne qualité » ou plusieurs essais randomisés « de bonne qualité » dont les résultats sont cohérents. • **B** Il existe des preuves « de qualité correcte » : essais randomisés (B1) ou études prospectives ou rétrospectives (B2). Les résultats de ces études sont cohérents dans l'ensemble. • **C** Les études disponibles sont critiquables d'un	-

Élément	HAS (2)	FNCLCC (9,33)	NZGG (3)
	- études comparatives non randomisées bien menées ; - études de cohortes ; • **niveau 3** - études cas-témoins ; • **niveau 4** - études comparatives comportant des biais importants ; - études rétrospectives ; - séries de cas ; - études épidémiologiques descriptives (transversale, longitudinale).	point de vue méthodologique ou leurs résultats ne sont pas cohérents dans l'ensemble. • **D** Il n'existe pas de données ou seulement des séries de cas. • **accord d'experts** Il n'existe pas de données pour la méthode concernée, mais l'ensemble des experts est unanime.	
recommandation	**Gradation** fondée sur le niveau de preuve scientifique de la littérature venant à l'appui des recommandations : • **grade A** (niveau de preuve 1**)** ; • **grade B** (niveau de preuve 2) ; • **grade C** (niveau de preuve 3 et 4) ; • accord d'experts. **Force** des recommandations : Appréciée sur le niveau de preuve scientifique et sur l'interprétation des experts. Ne précise pas comment elle est exprimée.	-	**Grade** Fondé sur : - le type et la qualité des études individuelles identifiées pour répondre à la question posée (converties en un énoncé récapitulatif des données scientifiques reflétant l'ensemble des données scientifiques) ; - la quantité, la cohérence, l'applicabilité et l'impact clinique de l'ensemble des données scientifiques ; - le consensus du groupe de travail. • **grade A** les données scientifiques sont issues des résultats d'études de conception robuste pour répondre à la question posée ; • **grade B** - les données scientifiques sont issues des résultats d'études de conception robuste pour répondre à la question posée, mais il existe quelques incertitudes sur la conclusion soit à cause d'une hétérogénéité des résultats des études, soit à cause de biais mineurs ; - ou les données scientifiques sont issues de résultats d'études de conception moins solide pour la question posée, mais les résultats ont été confirmés dans des études différentes, et sont assez cohérentes. Il y a des données scientifiques assez bonnes sur le fait que les bénéfices de la conduite à tenir proposée l'emportent sur les risques ; • **grade C**

Élément	HAS (2)	FNCLCC (9,33)	NZGG (3)
			pour plusieurs résultats, des essais ou des études ne peuvent pas ou n'ont pas pu être réalisées, et la pratique est informée uniquement par l'avis d'experts ; • **grade I** les données scientifiques sont insuffisantes : les données scientifiques manquent, ou sont de qualité médiocre, ou elles sont contradictoires, et le rapport bénéfices risques ne peut pas être déterminé.

Tableau 46 (suite). Niveau de preuve des études et gradation des recommandations des principaux systèmes : synthèse

Élément	AAP (11)	SIGN (4)	USPSTF (6)	NHMRC (7)
étude	Évaluation de la qualité de chaque étude fondée sur le type d'étude et sa réalisation. Niveaux de qualité des études individuelles (d'intervention) : • **élevée** essais contrôlés randomisés bien conçus et bien menés réalisés sur un groupe issu d'une population similaire à la population cible de la RBP ; • **intermédiaire** essais contrôlés randomisés avec des biais non rédhibitoires, ou des limites liées à la méthode ; études de cohortes, études cas témoins ; • **faible** Étude de cas unique, raisonnement issu de principes physiopathologiques, avis d'experts.	**Niveau de preuve** décidé sur le code attribué à la qualité de la méthode de l'étude couplé au type de l'étude : • **++** Tous ou la plupart des critères sont remplis. Il est **très peu probable** que les conclusions de l'étude ou de la revue soient affectées par les critères non remplis. • **+** Plusieurs critères sont remplis. Il est **peu probable** que les conclusions de l'étude soient affectées par les critères non remplis ou non décrits de manière adéquate. • **-** Peu ou aucun des critères ne sont remplis. Il est **probable ou très probable** que les conclusions de l'étude en soient affectées.	Le niveau de preuve d'une étude est évalué sur la validité interne d'une étude avec une liste de critères minimum adaptée à chaque type d'étude et sur la généralisabilité. La **validité interne** est cotée : • **bonne** l'étude remplit tous les critères ; • **moyenne** l'étude ne remplit pas (ou pas clairement) au moins un critère, mais n'a pas de biais majeur ; • **médiocre** l'étude a au moins un biais majeur. La **généralisabilité** est évaluée en considérant la population, l'environnement des soins et les professionnels de l'étude. Elle est cotée : • **bonne** • **moyenne** • **médiocre.**	Chaque étude est évaluée dans trois dimensions : • **force des données scientifiques** - niveau de preuve : capacité de chaque étude à répondre de façon adéquate à une question de recherche particulière intervention, diagnostic, ou autre) ; - qualité (risque de biais) ; - précision statistique (degré de significativité, intervalle de confiance) ; • **taille de l'effet** • **pertinence des données scientifiques** - pertinence des critères de jugement pour les patients ; - pertinence de la question de l'étude (selon PICO).

Élément	AAP (11)	SIGN (4)	USPSTF (6)	NHMRC (7)
ensemble d'études	Considérer : - la cohérence des résultats des études ; - la taille de l'effet estimé dans les études ; - la taille des échantillons de populations individuels et regroupés. Niveaux de qualité des études regroupées : • **A** essais contrôlés randomisés ou études diagnostiques bien conçues sur des populations pertinentes ; • **B** essais contrôlés randomisés ou études diagnostiques avec des limitations mineures ; données scientifiques cohérentes à la grande majorité ; • **C** études observationnelles (étude cas-témoins, étude de cohorte) ; • **D** avis d'experts, observations, raisonnement à partir des principes physiopathologiques de base ; ● **X** situations exceptionnelles dans lesquelles des études de validation ne peuvent pas être réalisées et il y a une nette prépondérance du bénéfice ou des inconvénients.	Niveaux de preuve d'ensemble concernant une question spécifique jugés sur : - la quantité, la qualité et la cohérence des données scientifiques ; - la généralisabilité des résultats ; - l'applicabilité directe des données scientifiques à la population cible. **Liste des niveaux de preuve** pour l'énoncé des données scientifiques : • **1++** méta-analyses de qualité élevée, revues systématiques d'essais contrôlés randomisés, ou essais contrôlés randomisés avec un risque de biais très faible ; • **1+** méta-analyses bien menées, revues systématiques, ou essais contrôlés randomisés avec un risque de biais faible ; • **1-** méta-analyses, revues systématiques, ou essais contrôlés randomisés avec un risque de biais élevé ; • **2++** revues systématiques de qualité élevée d'études cas-témoins ou d'études de cohortes ; études cas-témoins ou études de cohortes avec un faible risque d'effet de facteurs de confusion ou de biais et une probabilité élevée que la relation est causale ;	**Niveau de preuve** des données scientifiques pour une question : • **convaincant,** • **adéquat,** • **inadéquat.** **Niveau de « certitude** » par rapport au bénéfice net de l'intervention : Le niveau de certitude est la probabilité que l'évaluation du bénéfice net d'une intervention de prévention est correcte. • **élevé** Les données scientifiques disponibles incluent en général des résultats cohérents issus d'études bien conçues et bien menées dans des populations des soins de premier recours représentatives. Ces études évaluent les effets de l'intervention de prévention sur des résultats cliniques. Il est peu probable que cette conclusion soit fortement affectée par les résultats d'études futures. • **modéré** Les données scientifiques disponibles sont suffisantes pour déterminer les effets de l'intervention de prévention sur des résultats cliniques, mais la confiance dans l'estimation est limitée par des facteurs tels que : - le nombre, la taille, ou la qualité des études individuelles ; - une hétérogénéité des résultats des études individuelles ; - une généralisabilité limitée des résul-	L'ensemble des données scientifiques pour chaque recommandation est examiné dans cinq dimensions : • études sources des données (en terme de quantité [nombre d'étude, puissance statistique], de niveau de preuve et de qualité [conduite des études]) ; • cohérence des données scientifiques ; • impact clinique (pertinence des données scientifiques pour répondre à la question clinique, degré de significativité [p ou intervalle de confiance] et taille de l'effet, pertinence de l'effet pour les patients) ; • généralisabilité ; • applicabilité. Chacune des cinq dimensions est cotée : • **A : excellent** • **B : bon** • **C : satisfaisant** • **D : médiocre**

Élément	AAP (11)	SIGN (4)	USPSTF (6)	NHMRC (7)
		• **2+** études cas-témoins ou études de cohortes bien menées avec un faible risque d'effet de facteurs de confusion ou de biais et une probabilité modérée que la relation est causale ; • **2-** études cas-témoins ou études de cohortes avec un risque élevé d'effet de facteur de confusion ou de biais et un risque significatif que la relation ne soit pas causale ; • **3** études non analytiques, par exemple séries de cas ; • **4** opinion d'experts.	tats à la pratique courante des soins de premier recours ; - un manque de cohérence dans la chaîne des données scientifiques. Si des informations supplémentaires deviennent disponibles, l'ampleur ou la direction de l'effet observé pourrait changer, et ce changement pourrait être assez grand pour altérer les conclusions. • **faible** Les données scientifiques disponibles sont insuffisantes pour évaluer des effets sur les résultats cliniques à cause : - du nombre limité ou de la taille des études ; - des biais importants dans la conception de l'étude ou des méthodes ; - d'une hétérogénéité des résultats des études individuelles ; - des lacunes dans la chaîne des données scientifiques ; - des résultats non généralisables à la pratique courante des soins de premier recours ; - d'un manque d'information sur des résultats cliniques importants. Davantage d'informations peut permettre une estimation des effets sur des résultats cliniques.	
recommandation	La **force** d'une recommandation indique **l'importance de l'adhésion** à une recommandation. Elle est fondée sur :	Le **grade** se rapporte à la force des données scientifiques sur lesquelles la recommandation est fondée. Il indique aux utilisateurs la probabilité que le résultat prévu soit atteint si la	Les recommandations sont codées pour refléter à la fois la certitude des données scientifiques et l'ampleur du bénéfice net. La formulation des recommandations est	La gradation indique la force de l'ensemble des données scientifiques qui sous-tendent la recommandation. Le grade des recommandations est fondé sur la somme de chacune des cinq

Élément	AAP (11)	SIGN (4)	USPSTF (6)	NHMRC (7)
	- quatre niveaux de qualité des données scientifiques ; - deux catégories du rapport bénéfices inconvénients (bénéfices >> inconvénients ; bénéfices ≈ inconvénients) ; - recommandation dans des situations exceptionnelles. Catégories : • **recommandation forte,** • **recommandation,** • **option,** • **pas de recommandation.**	recommandation est mise en œuvre. Il est fondé sur : - le niveau de preuve d'ensemble ; - l'impact des données scientifiques (inconvénients potentiels associés à la mise en œuvre de la recommandation, impact clinique de la recommandation, possibilité de mise en œuvre pour le système de santé) ; • **grade A** au moins une méta-analyse, une revue systématique, ou un essai contrôlé randomisé coté 1++, et directement applicable à la population cible ; ou un ensemble de données scientifiques composé principalement d'études cotées 1+, directement applicable à la population cible, démontrant une homogénéité globale des résultats ; • **grade B** un ensemble de données scientifiques incluant des études cotées 2++, directement applicable à la population cible, et démontrant une homogénéité globale des résultats ; ou données scientifiques extrapolées d'études cotées 1++ ou 1+ ; • **grade C** un ensemble de données scientifiques incluant des études cotées 2+, directement applicable à la population cible et démontrant une homogénéité globale des résultats, ou données scientifiques extrapolées	standardisée. • **grade A** l'USPSTF recommande l'intervention. Il y a une certitude élevée d'un bénéfice net substantiel ; • **grade B** l'USPSTF recommande l'intervention. Il y a une certitude élevée d'un bénéfice net modéré ou il y a une certitude modérée d'un bénéfice net modéré à substantiel ; • **grade C** les médecins peuvent offrir cette intervention à des patients sélectionnés selon les circonstances. Il y a une certitude élevée ou modérée d'un petit bénéfice net ; • **grade D** l'USPSTF ne recommande pas l'intervention. Il y a une certitude élevée ou modérée de l'absence de bénéfice net ou que les inconvénients l'emportent sur les bénéfices ; • **grade I** les données scientifiques sont insuffisantes pour évaluer le rapport bénéfices inconvénients de l'intervention. Les données scientifiques manquent, sont de qualité faible ou contradictoires, et le rapport bénéfices inconvénient ne peut pas être déterminé.	dimensions de l'évaluation de l'ensemble des données scientifiques. Pour qu'une recommandation soit gradée A ou B, les données scientifiques et la cohérence des données scientifiques doivent être l'un et l'autre gradés A ou B. • **grade A** On peut se fier à l'ensemble des données scientifiques pour guider la pratique. • **grade B** On peut se fier à l'ensemble des données scientifiques pour guider la pratique dans la plupart des situations. • **grade C** L'ensemble des données scientifiques fournit des justifications pour la recommandation, mais il faut être attentif lors de sa mise en pratique. • **grade D** L'ensemble des données scientifiques est faible, et la recommandation doit être appliquée avec précaution.

Élément	AAP (11)	SIGN (4)	USPSTF (6)	NHMRC (7)
		d'études cotées 2++ ; • **grade D** niveau de preuve 3 ou 4, ou données scientifiques extrapolées d'études cotées 2+. L'**importance de la recommandation** n'est pas nécessairement reliée à la force des données scientifiques, mais reflète dans quelle mesure le groupe croit que la recommandation aura un impact sur l'état de santé ou la qualité de vie des patients concernés.		

Tableau 46 (suite). Niveau de preuve des études et gradation des recommandations des principaux systèmes : synthèse

Élément	GRADE (10,18)	NICE (8)	ACP (12)
étude	-	**Études d'intervention :** tableau de résumé standardisé. **Études portant sur la validité diagnostique d'un test :** étude évaluée sur des critères issus de la check-list QUADAS.	-
ensemble d'études	On considère pour chaque **résultat important :** **- Type d'étude :** • Essai contrôlé randomisé (qualité élevée). • Étude observationnelle (qualité faible). **- Cinq facteurs peuvent diminuer la qualité des études observationnelles et des essais contrôlés randomisés :** limites des études, hétérogénéité des résultats, caractère direct des données, imprécision des données, biais de publication. **- Trois facteurs peuvent augmenter la qualité des études observationnelles :** force de l'association,	Pas d'échelle ordinale pour la qualité des données scientifiques. **Études d'intervention :** Profil des données scientifiques ou tableau de résumé des résultats. **Études portant sur la validité diagnostique d'un test :** Tableaux de synthèse des données scientifiques.	**Gradation** de la qualité des données scientifiques : • **élevée** Au moins 1 ECR bien conçu et bien mené apportant des résultats cohérents et directement applicables. • **moyenne** ECRs ayant des limites importantes (par exemple : évaluation biaisée de l'effet du traitement, nombreux perdus de vue, absence d'insu, hétérogénéité inexpliquée [même si elle est issue d'ECRs rigoureux], données scientifiques indirectes issues d'une population d'intérêt similaire [mais non identique]) et ECRs ayant un très petit nombre de participants ou d'événements

Élément	GRADE (10,18)	NICE (8)	ACP (12)
	gradient dose-réponse, présence de facteurs de confusion plausibles qui auraient diminué l'effet observé. **Cotation de la qualité des données scientifiques pour chaque résultat important :** • **élevé** Nous avons une confiance élevée dans l'estimation de l'effet : celle-ci doit être très proche du véritable effet. • **modéré** Nous avons une confiance modérée dans l'estimation de l'effet : celle-ci est probablement proche du véritable effet, mais il est possible qu'elle soit nettement différente. • **faible** Nous avons une confiance limitée dans l'estimation de l'effet : celle-ci peut-être nettement différente du véritable effet. • **très faible** Nous avons très peu confiance dans l'estimation de l'effet : il est probable que celle-ci soit nettement différente du véritable effet. **Cotation de la qualité des données scientifiques dans leur ensemble :** les données scientifiques de qualité la plus faible pour n'importe lequel des **résultats décisifs** devraient fournir la base pour coter la qualité globale des données scientifiques.		observés. De plus, les données scientifiques issues d'essais contrôlés bien menés mais sans randomisation, d'études de cohortes bien menées ou d'études cas-témoins, et les séries temporelles multiples avec ou sans intervention sont dans cette catégorie. • **faible** Obtenue à partir d'études observationnelles avec un risque de biais. Cependant, la qualité des données scientifiques peut être cotée moyenne ou élevée, selon les conditions dans lesquelles les données scientifiques sont obtenues à partir des études observationnelles. • **données insuffisantes pour déterminer des bénéfices nets ou des risques nets** Les données scientifiques peuvent être contradictoires, de qualité médiocre ou absente, et le rapport bénéfices-risques ne peut pas être déterminé. Il n'y a aucune estimation de l'effet qui est très incertain, les données scientifiques soit étant indisponibles, soit ne permettant pas une conclusion.
recommandation	La **force d'une recommandation** reflète la confiance que l'on peut avoir dans le fait que les effets souhaitables d'une intervention l'emportent sur les effets indésirables. **Facteurs qui déterminent la force** d'une recommandation : - rapport bénéfices-inconvénients ; - qualité des données ;	Pas d'échelle ordinale pour la force des recommandations. Force reflétée par la formulation des recommandations.	**Force des recommandations :** • **Recommandation forte :** bénéfices >> risques et lourdeur du traitement et *vice versa.* • **Recommandation faible :** bénéfices ≈ risques et lourdeur du traitement ou incertitude sur l'ampleur des bénéfices et des risques. **Gradation** de la qualité des données scientifiques et de la **force**

Élément	GRADE (10,18)	NICE (8)	ACP (12)
	- incertitude sur la variabilité des valeurs et des préférences ; - coût. Catégories : • **recommandation forte** quand le groupe de travail est confiant dans le fait que les effets souhaitables de l'adhésion à une recommandation l'emportent sur les effets indésirables ; • **recommandation faible** indique que les effets souhaitables de l'adhésion à une recommandation l'emportent probablement sur les effets indésirables, mais le groupe de travail est moins confiant.		**des recommandations :** • recommandation **forte ;** qualité des données scientifiques : **élevée, moyenne, faible ;** • recommandation **faible ;** qualité des données scientifiques : **élevée, moyenne, faible ;** • **insuffisant.**

Annexe 8. *National Service Framework for Long Term Conditions grading system*

Tableau 47. Résumé du système de codage et de gradation des données scientifiques du *National Service Framework for Long Term Conditions.* D'après Turner-Stokes, 2006 (60)

Données d'experts – codé : E			
Type d'experts	Usager, famille, aidant, professionnel, autre partie prenante.		
Processus	Consultation, consensus.		
Données issues de la recherche clinique – codée : R			
Type	primaire		
	P1	Utilisant des approches quantitatives.	
	P2	Utilisant des approches qualitatives.	
	P3	Approche mixte (qualitative et quantitative).	
	secondaire		
	S1	Méta-analyses ou analyse des données actuelles.	
	S2	Analyse secondaire des données actuelles.	
	revues		
	R1		
	R2		
Qualité	1	Les questions/objectifs de la recherche sont-ils clairement décrits ?	0 : non 1 : partiellement 2 : oui
	2	La conception de l'étude est-elle adaptée pour les objectifs et les buts de la recherche ?	
	3	Les méthodes sont-elles clairement décrites ?	
	4	Les données sont-elles adéquates pour supporter les interprétations/conclusions des auteurs ?	
	5	Les résultats sont-ils généralisables ?	
		Total	Score/10
	Qualité élevée : score ≥ 7 ; moyenne : score 4 – 6 ; faible : score ≤ 3		
Applicabilité	directe	Données issues de populations ayant la maladie considérée.	
	indirecte	Données scientifiques extrapolées de populations ayant d'autres maladies.	

Tableau 47 (suite). Gradation des données de la recherche clinique du *National Service Framework for Long Term Conditions*

Données d'experts – codé : E	
Recherche de grade A	Plus d'une étude de qualité élevée et au moins l'une d'elles d'applicabilité directe.
Recherche de grade B	Une étude de qualité élevée ou plus d'une étude de qualité moyenne et au moins l'une d'elles d'applicabilité directe, *Ou* plus d'une étude de qualité élevée d'applicabilité indirecte.
Recherche de grade C	Une étude de qualité moyenne *Ou* des études de qualité faible ou uniquement des études d'applicabilité indirecte.

Annexe 9. Tableau de résumé standardisé des études pour les questions d'interventions

Tableau 48. Résumé standardisé des études pour les questions d'intervention (61)

RUBRIQUE	DESCRIPTION
Référence	Utiliser le style de Vancouver : Auteurs*. Titre. Titre secondaire. Nom du journal. Année de la publication ; volume (issue) : page de début–page de fin.
Type de l'étude	Préciser le type de l'étude (par exemple : essai contrôlé randomisé, essai comparatif non randomisé, étude cas-témoins, série de cas, etc.).
Source de financement	Préciser la source de financement : fonds publics, organisation non gouvernementale, industrie pharmaceutique ou autres (spécifier le nom de l'organisation ou de la compagnie).
MÉTHODE	
Critères d'éligibilité	Citer les critères d'inclusion et d'exclusion.
Cadre et lieu de l'étude	Nombre de centres, le ou les pays concernés, patients ambulatoires ou hospitalisés, zone urbaine, rurale ou mixte, s'il y a lieu.
Interventions	Détailler les interventions pour chaque groupe (incluant par exemple pour les traitements : dose, fréquence, durée du traitement et moment de la prise si pertinent).
Critère de jugement principal	Décrire le critère de jugement principal (habituellement celui utilisé pour le calcul du nombre de sujets nécessaire).
Critère(s) de jugement secondaire(s)	Description brève.
Taille de l'échantillon	Donner le nombre calculé de sujets nécessaire dans chaque groupe et le nombre de patients inclus dans chaque groupe.
Méthode de randomisation	Décrire la méthode de randomisation et celle de l'insu (aveugle) s'il y a lieu.
RÉSULTATS	
Nombre de sujets analysés	Donner le nombre de patients par groupe inclus dans l'analyse, notamment en intention de traiter dans les essais comparatifs.
Durée de l'étude	Donner les dates de début et de fin d'étude, ainsi que les périodes d'inclusion et de suivi.
Caractéristiques des patients et comparabilité des groupes	Décrire les divergences entre les groupes.
Résultats inhérents au critère de jugement principal	Résumer les résultats inhérents au critère de jugement principal dans chaque groupe et entre les groupes en précisant la différence, la valeur de p et l'intervalle de confiance s'ils sont disponibles.
Résultats inhérents au(x) critère(s) de jugement secondaire(s)	Description brève des résultats inhérents au(x) critère(s) de jugement secondaire(s).
Effets secondaires	Décrire les divergences entre les groupes.
Revue critique de la qualité de l'étude	Donner des commentaires détaillés sur : - la validité externe : cadre et lieu de l'étude, critères d'inclusion et d'exclusion, interventions, etc. ; - la validité interne : taille de l'échantillon (risques alpha et bêta utilisés pour le calcul du nombre de sujets nécessaire), randomisation et insu, analyse statistique inappropriée, comparabilité initiale des groupes, etc. ; - commentaires généraux.

*Pour les auteurs, se limiter aux six premiers auteurs et ensuite ajouter *et al*. Dans le cas où il y aurait une société savante, elle compte comme un auteur.

Annexe 10. Revue systématique de l'AHRQ, 2002

Une revue systématique de l'*Agency for Healthcare Research and Quality* (AHRQ), publiée en 2002, a eu pour objectif de décrire des systèmes de gradation de la force des données scientifiques, évaluant notamment la qualité des articles d'une revue de la littérature sur une question spécifique (46).

Les auteurs ont postulé qu'évaluer le niveau de preuve d'un ensemble de données scientifiques est similaire à faire la distinction entre les associations causales et non causales en épidémiologie. Pour évaluer la nature causale d'une association, un certain nombre de critères doivent être utilisés, aucun d'entre eux n'étant suffisant à lui seul (47,48). Deux de ces critères sont directement associés à la cotation de la force d'un ensemble de données scientifiques (46) :

- l'homogénéité (dans quelle mesure différentes approches, par exemple différents types d'études ou populations, pour étudier une association entre un facteur et une maladie fournissent des conclusions similaires ?) ;
- La force de l'association (taille du risque estimé [d'une maladie due à un facteur] et son intervalle de confiance).

Un troisième critère, la cohérence avec ce qui est connu de l'histoire naturelle et de la biologie d'une maladie est plus en rapport avec l'élaboration des RBP.

Les critères d'évaluation ont porté sur trois domaines : la qualité, la quantité, et la cohérence (46) (tableau 48). La qualité représente dans quelle mesure la conception d'une étude, sa réalisation et l'analyse ont minimisé les biais de sélection, de mesure et de confusion. La quantité reflète d'une part l'association entre l'intervention (ou l'exposition) évaluée et le résultat, et d'autre part la quantité d'information supportant cette association. La cohérence reflète dans quelle mesure un ensemble de données scientifiques est cohérent en lui-même (homogénéité) et avec des données externes. Au total, quarante systèmes de gradation de la force d'un ensemble de données scientifiques ont été évalués. Parmi eux, sept ont abordé complètement les trois domaines.

Tableau 49. Domaines et éléments pour estimer la force d'un ensemble de données scientifiques d'après l'AHRQ, 2002 (46)

Domaines	Éléments
Qualité	Regroupement des cotations de la qualité des études individuelles, reflétant dans quelle mesure les biais ont été minimisés.
Quantité	Taille de l'effet. Nombre d'études ayant évalué le sujet donné. Taille de l'échantillon global, largeur de l'intervalle de confiance ou puissance statistique de l'étude.
Homogénéité	Pour un sujet donné, dans quelle mesure des résultats analogues sont rapportés dans des travaux utilisant des études de conception similaire ou différente.

Références

1. Appraisal of guidelines Research and evaluation. Grille d'évaluation de la qualité des recommandations pour la pratique clinique (grille AGREE II). Paris: FNCLCC; 2009.

2. Agence nationale d'accréditation et d'évaluation en santé. Guide d'analyse de la littérature et gradation des recommandations. Paris: ANAES; 2000.

3. New Zealand Guidelines Group. Handbook for the preparation of explicit evidence-based clinical practice guilelines. Wellington: NZGG; 2001.

4. Scottish Intercollegiate Guidelines Network. A guideline developer's handbook. SIGN 50. Edimbough: SIGN; 2008.

5. World Health Organization. WHO Handbook for guideline development. Geneva: WHO; 2010. http://www.who.int/hiv/topics/mtct/grc_handbook_mar2010_1.pdf

6. U.S.Preventive Services Task Force. Grade Definitions 2008. <http://www.uspreventiveservicestaskforce.org/uspstf/grades.htm> [consulté en 07/2012].

7. National Health and Medical Research Council. NHMRC levels of evidence and grades for recommendations for developers of guidelines. Canberra: NHMRC; 2009.

8. National Institute for Health and Clinical Excellence. The guideline manual. London: NHS; 2009.

9. Fervers B, Bonichon F, Demard F, Heron JF, Mathoulin S, Philip T*, et al.* Méthodologie de développement des standards, options et recommandations diagnostiques et thérapeutiques en cancérologie. Bull Cancer 1995;82(10):761-7.

10. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P*, et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-6.

11. American Academy of Pediatrics. Classifying recommendations for clinical practice guidelines. Pediatrics 2004;114(3):874-7.

12. Qaseem A, Snow V, Owens DK, Shekelle P, Clinical Guidelines Committee of the American College of Physicians. The development of clinical practice guidelines and guidance statements of the American College of Physicians: summary of methods. Ann Intern Med 2010;153(3):194-9.

13. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001;323(7308):334-6.

14. Greer N, Mosser G, Logan G, Halaas GW. A practical approach to evidence grading. Jt Comm J Qual Improv 2000;26(12):700-12.

15. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM*, et al.* Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001;20(3 Suppl):21-35.

16. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S*, et al.* Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res 2004;4(1):38.

17. Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, Harbour RT*, et al.* Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. BMC Health Serv Res 2005;5(1):25.

18. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S*, et al.* Grading quality of evidence and strength of recommendations. BMJ 2004;328(7454):1490.

19. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J*, et al.* GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64(4):401-6.

20. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P*, et al.* GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol 2011;64(4):407-15.

21. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M*, et al.* GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol 2011;64(12):1294-302.

22. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M*, et al.* GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol 2011;64(12):1303-10.

23. Guyatt GH, Oxman AD, Kunz R, Brozek J, onso-Coello P, Rind D*, et al.* GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011;64(12):1283-93.

24. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J*, et al.* GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol 2011;64(12):1277-82.

25. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P*, et al.* GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64(12):1311-6.

26. Guyatt GH, Norris SL, Schulman S, Hirsh J, Eckman MH, Akl EA*, et al.* Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141(2 Suppl):53S-70S.

27. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144(11):850-5.

28. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE*, et al.* Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336(7653):1106-10.

29. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB*, et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-36.

30. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeglang MMG, Deeks JJ. Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, ed. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration; 2009.

31. Cancer Care Ontario. Program in evidence-based care handbook 2011. <https://www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=50876> .

32. Birtwhistle R, Pottie K, Shaw E, Dickinson JA, Brauer P, Fortin M*, et al.* Le groupe d'étude canadien sur les soins préventifs. Nous sommes de retour !
964. Can Fam Phys 2012;58:e1-e4.

33. Fervers B, Hardy J, Blanc-Vincent MP, Theobald S, Bataillard A, Farsi F*, et al.* SOR: project methodology. Br J Cancer 2001;84(Suppl 2):8-16.

34. Haute Autorité de Santé. Élaboration de recommandations de bonne pratique. Méthode « Recommandations pour la pratique clinique ». Saint-Denis La Plaine: HAS; 2010.

35. Baker A, Young K, Potter J, Madan I. A review of grading systems for evidence-based guidelines produced by medical specialties. Clin Med 2010;10(4):358-63.

36. Baker A, Potter J, Young K, Madan I. The applicability of grading systems for guidelines. J Eval Clin Pract 2011;17(4):758-62.

37. Health Services Assessment Collaboration, Ali W. What assessment tools are used both in New Zealand and in other countries for grading of evidence? HSAC Report 2009;2(6).

38. Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schunemann H. An emerging consensus on grading recommendations? ACP J Club 2006;144(1):A8-A9.

39. Barbui C, Dua T, van OM, Yasamy MT, Fleischmann A, Clark N*, et al.* Challenges in developing evidence-based recommendations using the GRADE approach: the case of mental, neurological, and substance use disorders. PLoS Med 2010;7(8).

40. Thornton J, Alderson P, Tan T, Turner C, Latchem S, Shaw E*, et al.* Introducing GRADE across the NICE clinical guideline program. J Clin Epidemiol 2012.

41. National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. Handbook series on preparing clinical practice guidelines. Canberra: NHMRC; 2000.

42. Kavanagh BP. The GRADE system for rating clinical guidelines. PLoS Med 2009;6(9):e1000094.

43. Institute of Medicine, Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E. Clinical practice guidelines we can trust. Washington: IOM; 2011. http://www.iom.edu/~/media/Files/Report%20Files/2011/Clinical-Practice-Guidelines-We-Can-Trust/Clinical%20Practice%20Guidelines%202011%20Insert.pdf

44. Qaseem A, Forland F, Macbeth F, Ollenschlager G, Phillips S, van der WP*, et al.* Guidelines International Network: toward international standards for clinical practice guidelines. Ann Intern Med 2012;156(7):525-31.

45. Bossard N, Boissel FH, Boissel JP. Level of evidence and therapeutic evaluation: need for more thoughts. Fundam Clin Pharmacol 2004;18(3):365-72.

46. Agency for Healthcare Research and Quality. Systems to rate the strength of scientific evidence. Rockville: AHRQ; 2002.

47. Surgeon General's Advisory Committee on Smoking and Health, Bayne-Jones,S, Burdette,WJ, Cochran,WG, Farber,E, Fieser,LF, *et al.* Smoking and health. Report of the advisory committee to the Surgeon General of the Public Health Service. Washington: US Department of Health Education and Welfare; 1964.

48. Hill AB. The environment and disease: association or causation ? Proc R Soc Med 1965;58:295-300.

49. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312(7023):71-2.

50. Office québécois de la langue française. Le grand dictionnaire terminologique (GTD) 2011. <http://gdt.oqlf.gouv.qc.ca/> .

51. Beun A. Le principe de précaution en matière de responsabilité médicale [thèse]. Lille: Lille 2 université du droit et de la santé; 2003.

52. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med 1997;126(5):376-80.

53. West S, King V, Carey TS, Lohr KN, Mckoy N, Sutton SF*, et al.* Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ) 2002;(47):1-11.

54. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv 1999;25(9):470-9.

55. Hayes DF, Bast RC, Desch CE, Fritsche H, Kemeny NE, Jessup JM*, et al.* Tumor marker utility grading system: a framework to evaluate

clinical utility of tumor markers
962. J Natl Cancer Inst 1996;88(20):1456-66.

56. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers
963. J Natl Cancer Inst 2009;101(21):1446-52.

57. Feinstein AR. The theory and evaluation of sensibility. In: Clinimetrics. New Haven: Yale University Press; 1987. p. 141-166.

58. Oxford Centre for Evidence-Based Medicine, Durieux N, Pasleau F, Howick J. The Oxford 2011 levels of evidence 2011. <http://www.cebm.net/index.aspx?o=5653> [consulté en 08/2012].

59. Oxford Centre for Evidence-Based Medicine, Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, *et al.* Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document) 2011. <http://www.cebm.net/index.aspx?o=5653> [consulté en 08/2012].

60. Turner-Stokes L, Harding R, Sergeant J, Lupton C, McPherson K. Generating the evidence base for the National Service Framework for Long Term Conditions: a new research typology. Clin Med 2006;6(1):91-7.

61. Mlika-Cabanne N, Harbour R, de BH, Laurence M, Cook R, Twaddle S*, et al.* Sharing hard labour: developing a standard template for data summaries in guideline development. BMJ Qual Saf 2011;20(2):141-5.

Participants

Organismes professionnels

Les organismes professionnels suivants ont été sollicités pour l'élaboration de cet état des lieux :

Collège de médecine générale

Fédération des spécialités médicales*

Institut national du cancer*

(*) Cet organisme a proposé un ou plusieurs experts pour ce projet

Groupe de lecture

Dr Solange Beaumont, représentante d'associations de patients et d'usagers UNAFAM, Paris

Dr Christophe Berkhout, médecin généraliste, Dunkerque

Dr Jacques Birgé, médecin généraliste, Boulay

Pr Jean-Pierre Boissel, pharmacologie clinique, retraité, Lyon

Pr Diane Braguer, pharmacologie, Marseille

Pr Charles Caulin, médecine interne, retraité, Paris

M. Emmanuel Corbillon, HAS, Saint-Denis

M. Guy Cordesse, masseur-kinésithérapeute, La-Ferté-sous-Jouarre

Dr Michel Cucherat, biostatistiques, Lyon

Pr Isabelle Durand-Zalesky, santé publique, Créteil

Pr Alain Durocher, HAS, Saint-Denis

Dr Sylvie Erpeldinger, médecin généraliste, Villeurbanne

M. Jonathan Finzi, Inca, Boulogne-Billancourt

Dr Nicole Garret, pédopsychiatre, Nantes

Pr Bernard Gay, médecin généraliste, La Réole

Dr Gaëtan Gentile, médecin généraliste, Puyricard

Dr Michel Gerson, endocrinologue-diabétologue, Colmar

Mme Françoise Hamers, HAS, Saint-Denis

M. Grégoire Jeanblanc, HAS, Saint-Denis

Diana Kassab-Chahmi, Inca, Boulogne-Billancourt

Mme Virginie Lecaplain, infirmière, assistante qualité, Caen

Dr Marina Martinowsky, HAS, Saint-Denis

M. René Mazars, représentant d'associations de patients et d'usagers – CISS, Luc-la-Primaube

M. François Planchamp, Inca, Boulogne-Billancourt

Dr Patrick Silvestre, médecin généraliste, Sérifontaine

Mme Sophie Stamenkovic, HAS, Saint-Denis

Dr Jean-Michel Thurin, psychiatre, Paris

Mme Laetitia Verdoni, Inca, Boulogne-Billancourt

Mme Laure Zanetti, HAS, Saint-Denis

Dr Philippe Zerr, médecin généraliste, Levallois-Perret

Fiche descriptive

Titre	Titre de la recommandation à compléter
Objectifs	Déterminer s'il est nécessaire d'adopter un système existant ou d'élaborer un nouveau système de niveau de preuve et gradation répondant aux attentes des différents partenaires pour les études d'intervention, études diagnostiques (Gradation HAS, SOR, GRADE, SIGN, etc.) – Partie 1 : État des lieux.
Demandeur	Autosaisine HAS.
Promoteur	Haute Autorité de Santé (HAS), service des bonnes pratiques professionnelles.
Financement	Fonds publics.
Pilotage du projet	Coordination : D[r] Muriel Dhénain et M. Emmanuel Nouyrigat, chefs de projet, service des bonnes pratiques professionnelles de la HAS (chef de service : D[r] Michel Laurence). Secrétariat : M[me] Sladana Praizovic.
Recherche documentaire	Cf. Stratégie de recherche documentaire décrite en annexe 1. Réalisée par M[me] Emmanuelle Blondet, avec l'aide de M[me] Maud Lefèvre (chef du service documentation – information des publics : M[me] Frédérique Pagès).
Auteurs	D[r] Muriel Dhénain, chef de projet, service des bonnes pratiques professionnelles de la HAS.
Participants	Cf. liste des participants.
Validation	Adoption par le Collège de la HAS en avril 2013.
Documents d'accompagnement	Guide d'analyse de littérature et gradation des recommandations (Anaes 2000) – Actualisation en cours.

N°ISBN : 978-2-11-138037-0

Toutes les publications de l’HAS sont téléchargeables sur
www.has-sante.fr