HAUTE AUTORITE DE SANTE

Etat des lieux - Niveau de preuve et gradation des recommandations de bonne pratique

(Avril 2013)

Extraits concernant les études observationnelles et leur niveau de preuve évalué selon plusieurs méthodes internationales

---

Pages 6, 7 et 8

Guide d'analyse de la littérature et gradation des recommandations publié par l'Anaes en 2000 (2)[1].

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 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.

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

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) ;
	- 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 '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.

Pages 13, 14 et 15

1.4 *American academy of pediatrics*

La procédure d'élaboration de RBP fondées sur des données scientifiques comporte trois étapes (11)[2] :

- 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	

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

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

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.

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

Qualité des données scientifiques	Prépondérance des bénéfices ou des inconvénients	Équilibre des bénéfices et des inconvénients
A. Essais contrôlés randomisés ou études diagnostiques bien conçues sur des populations pertinentes	Recommandation forte	Option
B. Essais contrôlés randomisés ou études diagnostiques avec des limitations mineures ; données scientifiques cohérentes à la grande majorité	Recommandation	
C. Études observationnelles (étude cas-témoins, étude de cohorte)		
D. Avis d'experts, observations, raisonnement à partir des principes physiopathologiques de base	Option	Pas de recommandation

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	Recommandation forte / Recommandation

Pages 15 à 18

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 oeuvre (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)

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).

► 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.

Page 43

American College of Physicians' system

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.