anses
agence nationale de sécurité sanitaire
alimentation, environnement, travail

Connaître, évaluer, protéger

Évaluation du poids des preuves à l'Anses : revue critique de la littérature et recommandations à l'étape d'identification des dangers

Rapport d'étape

Avis de l'Anses
Rapport d'expertise collective

Juillet 2016 Édition scientifique

anses
agence nationale de sécurité sanitaire
alimentation, environnement, travail

Connaître, évaluer, protéger

Évaluation du poids des preuves à l'Anses : revue critique de la littérature et recommandations à l'étape d'identification des dangers

Rapport d'étape

Avis de l'Anses
Rapport d'expertise collective

Juillet 2016 Édition scientifique

Le directeur général

Maisons-Alfort, le 25 juillet 2016,

AVIS
de l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail

relatif au rapport d'étape sur l'évaluation du poids des preuves à l'Anses : revue critique de la littérature et recommandations à l'étape d'identification des dangers

L'Anses met en œuvre une expertise scientifique indépendante et pluraliste.
L'Anses contribue principalement à assurer la sécurité sanitaire dans les domaines de l'environnement, du travail et de l'alimentation et à évaluer les risques sanitaires qu'ils peuvent comporter.
Elle contribue également à assurer d'une part la protection de la santé et du bien-être des animaux et de la santé des végétaux et d'autre part l'évaluation des propriétés nutritionnelles des aliments.

Elle fournit aux autorités compétentes toutes les informations sur ces risques ainsi que l'expertise et l'appui scientifique technique nécessaires à l'élaboration des dispositions législatives et réglementaires et à la mise en œuvre des mesures de gestion du risque (article L.1313-1 du code de la santé publique).

Ses avis sont rendus publics.

L'Anses s'est autosaisie le 31 mars 2015 pour la réalisation d'une analyse critique sur les approches d'évaluation des niveaux de preuve à l'étape d'identification des dangers.

1. CONTEXTE ET OBJET DE LA SAISINE

1.1 Contexte

L'Anses contribue à la sécurité sanitaire humaine dans les domaines de l'environnement, du travail, de l'alimentation, à la protection de la santé et du bien-être des animaux, ainsi qu'à la protection de la santé des végétaux. L'agence est chargée de mettre en œuvre une expertise scientifique indépendante et pluraliste, afin d'évaluer les risques et de proposer aux autorités compétentes toute mesure de nature à préserver la santé publique.

L'agence identifie et caractérise les dangers biologiques, chimiques ou physiques en mobilisant l'ensemble des données scientifiques disponibles, et évalue les risques pour les populations humaines, animales ou végétales en prenant en compte la diversité des types d'exposition des individus aux différents dangers. L'Anses évalue également des produits (phytosanitaires, biocides, médicaments vétérinaires, nouveaux aliments, produits chimiques, etc.) et procédés (traitement des eaux ou des aliments, etc.) au regard d'exigences réglementaires, en vue d'une décision de mise sur le marché ou d'autorisation d'utilisation par les ministères en charge de la réglementation.

Les questions instruites par l'Agence sont traitées dans le cadre du processus d'évaluation des risques sanitaires, constitué des quatre étapes suivantes : identification des dangers, caractérisation des dangers (incluant l'établissement de la relation dose-réponse), évaluation de l'exposition et caractérisation des risques. Ces questions peuvent être de différents types ; par exemple : "La consommation de fibres alimentaires influence-t-elle le risque de cancer colorectal ?", "Les maladies à prions sont-elles transmissibles à l'Homme ?".

Les données utilisées pour répondre à ces questions proviennent de sources multiples (bases de données de la littérature scientifique, expertises de professionnels), sont de différentes natures (*in vitro*, *in vivo*, toxicologiques, épidémiologiques, etc.) et peuvent être contradictoires. Il convient de les analyser de façon approfondie pour en évaluer la pertinence et la validité, puis de les combiner pour répondre aux questions posées. Dans la littérature, la méthodologie permettant d'évaluer la pertinence et la qualité des données, puis de combiner des données hétérogènes est appelée « évaluation du poids des preuves[1] » (*Weight of Evidence* en anglais). La littérature fait principalement référence à l'évaluation du poids des preuves à l'étape d'identification du danger. Cette étape a pour objectif d'identifier le type et la nature des effets néfastes qu'un agent (biologique, chimique ou physique) peut causer sur un organisme, un système ou une population (IPCS 2004).

1.2 Identification des enjeux et besoins

Pour orienter l'action publique, les pouvoirs publics et les parties prenantes sont en demande de preuves de différentes natures pour un large éventail de domaines (NRC 2014, OCDE 2015, US EPA 2014). Récemment, les divergences entre les conclusions apportées par différents organismes sur la toxicité ou sur le risque lié à l'exposition à diverses substances (par exemple, entre l'EFSA et l'IARC sur le glyphosate[2] ; entre l'EFSA et l'Anses sur le Bisphénol A[3]) ont mis en avant l'impact des systèmes de classification des effets sanitaires et des modèles conceptuels explicatifs de l'apparition d'un effet sanitaire sur le résultat final de l'évaluation. L'évaluation des preuves dans le cadre de l'expertise scientifique fait donc l'objet d'un intérêt croissant de la part d'agences sanitaires, au niveau national et international, notamment en vue d'améliorer la robustesse et la transparence des travaux d'expertise (Hardy et al. 2015, OHAT 2015). Cette transparence est indispensable pour assurer la crédibilité et la confiance de la communauté scientifique comme des autres parties prenantes auprès des pouvoirs publics.

Bien que fréquemment utilisé, le concept de « poids des preuves » est ambigu, souvent mal défini, et ne recouvre pas toujours une méthodologie présentée de manière transparente et cohérente (NRC 2014, Weed 2005). Historiquement, l'évaluation du poids des preuves s'est développée dans le secteur médical comme un outil d'aide à la décision clinique de hiérarchisation des connaissances de la recherche médicale (Sackett et al. 1996). Les méthodes étaient alors focalisées principalement sur la revue critique de la littérature. Après avoir été utilisées dans le cadre de la médecine factuelle[4], ces méthodes ont été adaptées à l'environnement et à la santé environnementale (Mandrioli et Silbergeld 2015, Krimsky 2005). Par la suite, les méthodes ont évoluées pour combiner différentes sources d'information de façon transparente et systématique, notamment dans le cadre des expertises collectives.

Le rapport sur l'état des lieux des pratiques de l'Anses a montré que certaines saisines avaient mobilisé des méthodes d'évaluation du poids des preuves. Ce rapport révèle que les pratiques des collectifs d'experts diffèrent selon les thématiques (risques physico-chimiques, biologiques, liés à

[1] Le Centre National de Ressource et de Traitement de la Langue donne une définition générale du terme « preuve » : « Fait, témoignage, raisonnement susceptible d'établir de manière irréfutable la vérité ou la réalité de (quelque chose).»

[2] Voir http://www.efsa.europa.eu/en/press/news/160113

[3] (Anses 2015b)

[4] Traduction française de « evidence based medecine »

la nutrition, etc.) et au sein d'une même thématique. En particulier, le niveau de formalisation des méthodes utilisées et leur expression au sein des rapports varient selon les collectifs d'experts.

1.3 Objet de la saisine

Dans l'objectif d'améliorer la transparence du processus d'expertise requise pour un organisme certifié ISO 9001, l'Anses a confié au Groupe de Travail « Méthodologie de l'Évaluation des Risques » (GT MER) la conduite d'une analyse critique sur les approches d'évaluation des niveaux de preuve à l'étape d'identification des dangers.

Les objectifs de cette autosaisine sont de :

- Décrire les pratiques actuelles de l'Anses et les comparer avec celles d'autres organismes/agences sanitaires.
- Réaliser une revue critique des approches sur les niveaux de preuve.
- Proposer des procédures harmonisées pour évaluer la qualité des études et des données disponibles, ainsi que des niveaux de preuve par rapport aux questions ou hypothèses avancées.
- Proposer des procédures harmonisées pour évaluer et communiquer le niveau de preuve global associé à l'ensemble des données et études disponibles.
- Démontrer l'applicabilité des recommandations grâce à des études de cas.

Les résultats de cette saisine contribueront à améliorer la transparence et la reproductibilité de l'évaluation du poids des preuves à l'Anses.

2. ORGANISATION DE L'EXPERTISE

2.1 Modalités de traitement de la saisine et objectif du rapport

L'instruction de cette autosaisine est réalisée en trois étapes :

- Réalisation d'un état des lieux des pratiques actuelles de l'Anses,
- Revue de la littérature sur le poids des preuves et formulation de recommandations visant à harmoniser les procédures de l'Anses,
- Évaluation des recommandations à travers des études de cas en interaction avec les collectifs de l'Anses et rédaction d'un guide.

La première étape a été réalisée par l'équipe d'action (EA) « État des lieux » mise en place par le GT MER. Cette EA a établi un état des lieux des pratiques de l'Anses sur l'analyse de l'incertitude et l'évaluation du poids des preuves qui a été présenté et discuté au conseil scientifique le 22 septembre 2015.

Ce rapport présente les résultats de la seconde étape. Il a été réalisé par une autre EA « Poids des preuves ». Ce rapport détaille la revue de la littérature réalisée par cet EA et formule une série de recommandations visant à harmoniser les pratiques de l'Anses. Ce rapport d'étape a été validé par l'ensemble du GT MER.

La troisième étape mentionnée ci-dessus fera l'objet d'un travail spécifique, réalisé en interaction avec les collectifs d'experts de l'Anses en 2016-2017. Un guide méthodologique sera rédigé à l'issue de cette dernière étape. Il proposera des méthodes adaptées à différentes situations pratiques pour évaluer la qualité des études et des données disponibles, et pour évaluer et communiquer le poids des preuves.

Le travail d'expertise présenté dans ce rapport a été conduit par un collectif d'experts intervenant dans les différents domaines de l'Agence et ayant des compétences dans les méthodes d'évaluation des risques.

Ce rapport a été soumis au conseil scientifique pour commentaires lors d'une réunion spécifique le 29 février, puis le 1er mars 2016. Le GT MER a élaboré une nouvelle version du rapport en prenant en compte les remarques du conseil scientifique et en répondant aux questionnements posés. Cette dernière version a été transmise aux membres du conseil scientifique et validée en collège scientifique le 1er avril 2016.

L'expertise a été réalisée dans le respect de la norme NF X 50-110 « Qualité en expertise – prescriptions générales de compétence pour une expertise (Mai 2003) » (AFNOR mai 2003).

2.2 Prévention des risques de conflits d'intérêts.

L'Anses analyse les liens d'intérêts déclarés par les experts avant leur nomination et tout au long des travaux, afin d'éviter les risques de conflits d'intérêts au regard des points traités dans le cadre de l'expertise.

Les déclarations d'intérêts des experts sont rendues publiques *via* le site internet de l'Anses (www.anses.fr).

3. ANALYSE ET CONCLUSIONS DU CONSEIL SCIENTIFIQUE

Le conseil scientifique reprend les conclusions du rapport d'expertise collective du GT MER :

3.1 Définitions

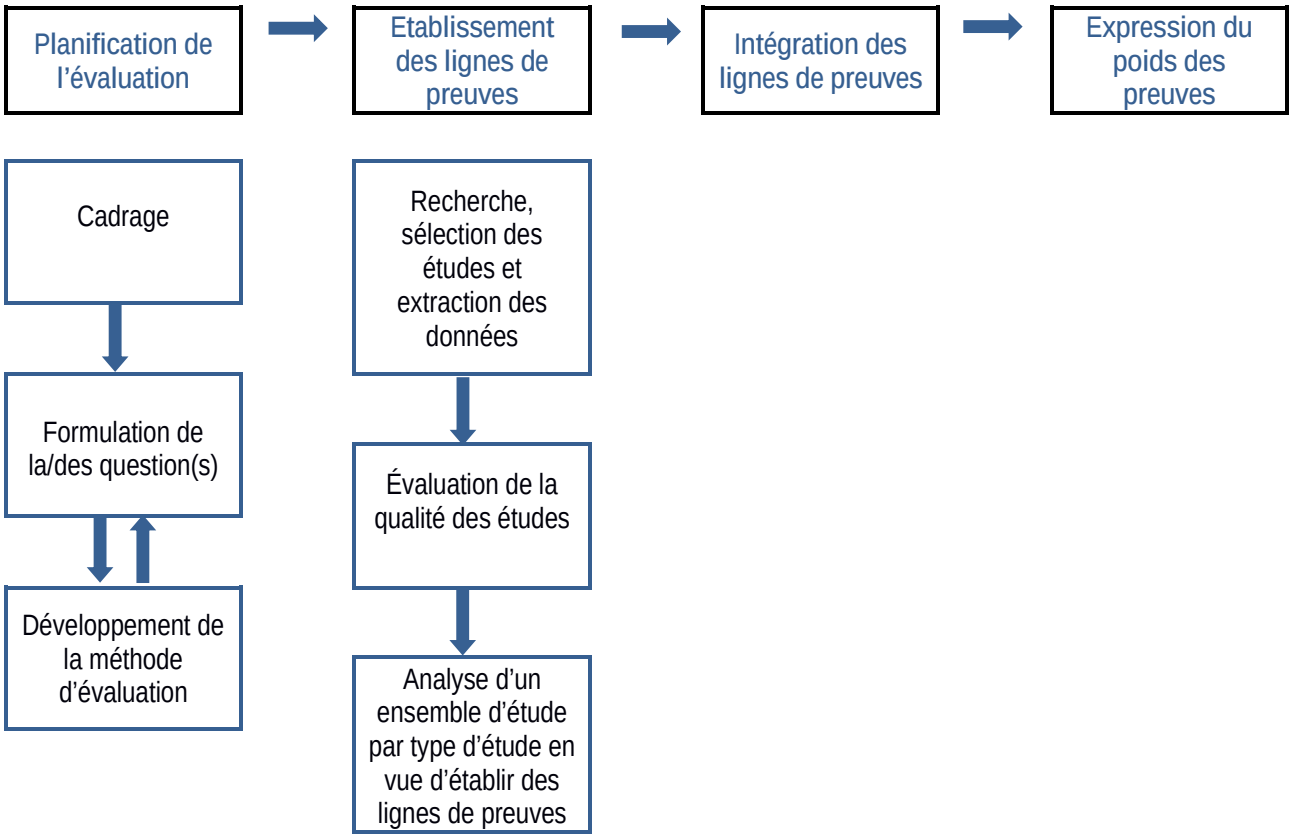
Dans ce rapport, le GT MER propose des définitions pour trois termes clés : poids des preuves, ligne de preuves et revue systématique.

> « Le **poids des preuves** est une synthèse formalisée de lignes de preuves, éventuellement de qualités hétérogènes, dans le but de déterminer le niveau de plausibilité d'hypothèses. »
>
> « Une **ligne de preuves** est un ensemble d'informations de même nature, intégrées pour évaluer une hypothèse».
>
> « Une **revue systématique** de la littérature scientifique consiste à assembler, évaluer et synthétiser de manière exhaustive toutes les études pertinentes, parfois contradictoire, qui abordent une question précise. Une revue systématique est basée sur la rédaction d'un protocole détaillé au préalable favorisant la transparence de la démarche et sa reproductibilité. »

3.2 Les étapes de l'évaluation du poids des preuves

Le GT MER recommande de structurer le processus d'évaluation du poids des preuves en quatre étapes principales (Figure 1).

Figure 1 : Démarche d'évaluation du poids des preuves



La **planification de l'évaluation** (étape 1) a pour objectif de définir le périmètre de l'évaluation du risque, de décrire les enjeux et d'identifier les méthodes à mettre en œuvre, incluant celles pour l'évaluation du poids des preuves. Cette première étape comporte trois sous-étapes opérationnelles :

- cadrage,
- formulation de la ou des questions à évaluer,
- développement de la méthode d'évaluation.

L'**établissement des lignes de preuves** (étape 2) comporte également trois sous-étapes opérationnelles :

- recherche, sélection des études (individuelles ou de synthèse) et extraction des données,
- évaluation de la qualité des études (individuelles ou de synthèse),
- analyse d'un ensemble d'études par type d'étude (épidémiologique, toxicologique,…) en vue d'établir des lignes de preuves.

L'**intégration des lignes de preuves** pour établir le poids des preuves (étape 3) a pour objectif de synthétiser les lignes de preuves disponibles dans le but de déterminer le niveau de validité d'hypothèses ou d'estimer des quantités d'intérêt.

L'étape **d'expression des conclusions sur le poids des preuves** (étape 4) permet d'exprimer le poids des preuves sous forme d'un résultat clair et explicite pour l'aide à la décision.

3.3 Analyse critique des méthodes recensées dans la littérature scientifique

Le GT MER a recensé 25 méthodes d'évaluation du poids des preuves dans des articles scientifiques ou des guides méthodologiques. Les méthodes ont été analysées au moyen de grilles de lecture pour chacune des étapes décrites ci-dessus. Les méthodes ont été comparées entre elles au moyen de trois critères :

- Le **caractère directif** : degré de formalisation de la méthode rendant l'expertise transparente et reproductible.
- La **pertinence** : degré de cohérence de la méthode en regard de l'objectif attendu. La pertinence est à positionner dans le contexte de la planification de l'évaluation (au regard de la question posée).
- La **faisabilité** : degré de mobilisation de temps, de ressources matérielles et humaines, ainsi que de compétence méthodologique spécifique (modélisation, statistique, etc.).

Les résultats de cette évaluation ont ensuite été utilisés pour formuler des recommandations.

3.4 Recommandations

Les méthodes recensées lors de la revue bibliographique ont été comparées à celles inventoriées dans l'état des lieux sur l'analyse de l'incertitude et l'évaluation du poids des preuves à l'Anses, afin de définir des pistes de progrès. Les documents du système qualité de l'Anses en lien avec l'évaluation du poids des preuves ont également été considérés pour évaluer l'intérêt d'inscrire l'évaluation du poids des preuves dans le processus d'expertise.

Le rapport formule des recommandations méthodologiques en tenant compte à la fois du caractère directif des méthodes, de leur pertinence, et de la faisabilité pour leur mise en œuvre.

Recommandations pour l'étape de planification de l'évaluation

A l'Anses, la planification de l'évaluation est réalisée lors de la rédaction du document de cadrage. Il n'existe pas à l'Anses de document précisant les méthodes d'évaluation et les formulations des questions pour les collectifs d'experts. Le GT MER recommande la mise en place d'un processus complet pour la planification de l'évaluation en trois sous-étapes, avec pour chacune un formulaire approprié.

La sous-étape 1 (cadrage) est réalisée par les agents de l'Anses. Les deux sous-étapes suivantes mobilisent les CES et les experts impliqués dans le traitement de la saisine. Le GT MER recommande d'adopter une démarche rétroactive entre les sous-étapes 2 (formulation de la ou des questions à évaluer) et 3 (développement de la méthode d'évaluation) de manière à affiner la formulation des questions en garantissant que celles-ci puissent être correctement traitées.

Le GT MER recommande à l'aide d'une revue préliminaire et si besoin d'auditions :

- D'identifier les enjeux sanitaires, environnementaux, sociétaux, économiques en collaboration avec les décideurs (et, dans l'idéal, les parties prenantes).
- D'évaluer l'étendue du corpus des données disponibles.
- De formaliser la question au moyen d'une structure de description présentée dans la littérature (PICO, PECO, etc.).

- De formaliser les sous-questions à traiter par le collectif d'experts, si possible sous forme de modèle conceptuel.
- De rédiger un plan de développement de la méthode d'évaluation en préalable à sa conduite.
- De définir les modalités de communication de la conclusion dans le document du cadrage.

Dans le cadre de l'évaluation du poids des preuves, pour la sous-étape 3, le GT MER recommande de spécifier au minimum :

- Le choix du type de revue de la littérature : revue systématique ou revue approfondie pour traiter les questions définies lors de la sous-étape 2. Ce choix doit tenir compte des éléments issus du document de cadrage, de l'analyse socio-économique de l'Anses et des critères proposés par l'EFSA : impact potentiel du résultat de la revue systématique, quantité et qualité des données, source et confidentialité des données, transparence nécessaire et controverse du sujet de la saisine, ressources à mettre en œuvre.
- Les critères utilisés pour évaluer la qualité des études par type d'étude.
- La méthode retenue pour évaluer un ensemble d'étude : méta-analyse, analyses multicritères, ou approches qualitatives.
- La méthode retenue pour intégrer des lignes de preuves : modélisation statistique avec mise en place ou non d'élicitation de dires d'experts, analyses multicritères, ou approches qualitatives.
- Les modalités d'expression des conclusions concernant le poids des preuves.

Recommandations pour l'étape d'établissement des lignes de preuves

L'Anses, dans le cadre de ses avis, a déjà conduit des revues systématiques de la littérature. Cependant, la réalisation de ces revues reste occasionnelle au sein de l'agence. Plusieurs collectifs de l'Anses mobilisent des grilles de lecture pour sélectionner les études et extraire des données, mais ces grilles sont insuffisamment partagées. Suite à ce constat, le GT MER formule les recommandations suivantes :

- Développer une grille de lecture, voire un tableau d'extraction des données, adaptable en fonction de la question de recherche et du type d'étude scientifique. Cette grille et ce tableau devront être élaborés et testés sur des études de cas en coordination avec les CES. Ces grilles devraient inclure des éléments concernant la pertinence des études par rapport à la question posée, des éléments descriptifs de l'étude et des critères d'évaluation de la qualité des études.
- Le recours à au moins deux experts est souhaitable pour remplir les grilles de lecture.
- Lorsqu'une revue systématique est réalisée, certains grands principes doivent être appliqués : l'utilisation d'au moins deux bases de données, la sélection des études par deux personnes indépendantes, et la définition des critères de sélection et d'exclusion des études en amont.
- Dans le cas où une revue systématique n'est pas réalisée, le GT MER recommande que la procédure de recherche, de sélection et d'extraction soit décrite de manière aussi précise que possible dans le rapport d'expertise, en se rapprochant autant que possible des pratiques d'une revue systématique.

Pour l'évaluation de la qualité des études, le GT MER recommande l'utilisation de listes de critères formalisés par type d'études (épidémiologiques, toxicologiques,…) de manière à assurer la transparence du processus d'évaluation. Ces listes devront être élaborées en coordination avec les CES, en exploitant par exemple les critères mentionnés dans la méthode GRADE. Lorsque des études sont exclues, les critères d'exclusion doivent être explicites et définis par le collectif d'experts. Lorsque des études de synthèse sont considérées dans le processus d'expertise, le GT MER recommande l'utilisation de la méthode AMSTAR ou R-AMSTAR pour évaluer leur qualité.

Pour l'évaluation d'un ensemble d'études par type d'étude en vue d'établir des lignes de preuves, le GT MER recommande :

- Pour les méthodes qualitatives, de définir aussi précisément que possible les critères et la signification des notes attribuées à ces critères par les experts.
- D'utiliser la méta-analyse pour établir des lignes de preuves sur des sujets stratégiques définis en tenant compte des enjeux en termes de risques sanitaires et des ressources disponibles (ressources humaines, disponibilité des données). Ces méta-analyses devraient être menées en s'inspirant des documents guides disponibles, notamment du document guide Cochrane.
- D'analyser la sensibilité des résultats aux paramètres d'entrée lorsque des méthodes quantitatives sont utilisées.
- De tester l'intérêt des approches multicritères pour évaluer leur utilité dans le contexte de l'Anses.

Recommandations pour l'étape d'intégration des lignes de preuves

Le GT MER recommande :

- D'encourager les experts à décrire et expliciter les choix réalisés afin d'assurer un niveau de transparence aussi élevé que possible. Certaines méthodes permettant de rendre le processus d'expertise transparent pourraient être utilisées dans ce but (ex : méthodes d'élicitation d'experts, intégration des avis d'experts dans les modèles prédictifs de type QSAR).
- D'utiliser les méthodes qualitatives telles que celles proposées par le CIRC ou le WCRF pour la combinaison des lignes de preuves, en incluant explicitement une liste permettant de vérifier la prise en compte des critères de Hill.
- Lorsque cela est réalisable, d'utiliser la modélisation statistique pour combiner différentes lignes de preuves. Ces approches nécessitant des compétences spécifiques et un investissement en temps relativement important, la faisabilité de leur utilisation devra cependant être évaluée sur des exemples concrets.
- D'analyser la sensibilité des résultats aux paramètres d'entrée lors de l'utilisation de méthodes quantitatives.
- De tester l'intérêt des méthodes d'analyse multicritères sur des cas concrets en coordination avec les CES.
- D'utiliser la méthode d'évaluation comparative du poids des preuves dans le domaine spécifique de l'évaluation du mode d'action.

Recommandations pour l'étape d'expression des conclusions sur le poids des preuves

Le GT MER recommande :

- D'associer un texte explicatif aux résultats numériques lorsque le poids des preuves est analysé avec une méthode quantitative.
- D'exprimer les conclusions concernant le poids des preuves selon une classification en quatre niveaux correspondant à des niveaux de preuve croissants, lorsque le poids des preuves est analysé avec une méthode qualitative. Une classe supplémentaire « évaluation impossible » peut également être considérée. Le GT MER recommande que chaque niveau soit défini précisément dans les rapports d'expertise. Une classification sur une échelle numérique pourrait être élaborée et testée sur des études de cas en coordination avec les CES.
- De structurer le contenu de la conclusion en adaptant les travaux de la collaboration Cochrane et de la méthode GRADE aux domaines de l'agence. Un rapprochement avec la Collaboration Cochrane pourrait être envisagé afin de bénéficier de son expérience.
- De caractériser l'incertitude dans la conclusion soit sous forme qualitative, soit sous forme quantitative, selon la méthode d'analyse du poids des preuves utilisée.

Recommandations opérationnelles

Afin de faciliter l'appropriation des concepts et méthodes d'évaluation du poids des preuves par les collectifs d'experts, le GT MER recommande :

- De consulter les fiches méthodes et les références présentées dans ce rapport.
- De mettre à disposition des collectifs d'experts un soutien méthodologique pour faciliter la mise en œuvre des méthodes d'évaluation du poids des preuves. Des référents méthodologiques pourraient être identifiés pour aider des collectifs d'experts à effectuer les sous étapes 2 et 3 de la planification de l'évaluation, à réaliser les revues systématiques et à appliquer les méthodes quantitatives (méta-analyse, modélisation statistique).
- De mettre en place un système d'information pour capitaliser les travaux d'expertise passés. Ce système d'information pourrait inclure le contenu des grilles de lecture des revues bibliographiques et un descriptif de la méthode d'évaluation du poids des preuves mise en œuvre dans l'expertise.

Si elles sont adoptées par les collectifs d'experts, ces recommandations permettront d'harmoniser les procédures de l'agence concernant le poids des preuves. Cependant, dans le but de vérifier leur pertinence et de faciliter leur diffusion, le GT MER propose d'évaluer ces recommandations en réalisant des études de cas en collaboration avec les collectifs de l'Anses.

3.5 Conclusions et recommandations du conseil scientifique

Le conseil scientifique souligne le travail important réalisé par le GT « Méthodologie de l'évaluation des risques » et endosse ce rapport d'étape. Il insiste sur la nécessité de mettre en œuvre des études de cas avec les différents collectifs d'experts de l'Anses afin de tester la faisabilité des recommandations, et préconise le développement des outils méthodologiques pour les collectifs d'experts (tels que les grilles de lecture) ainsi que la mise à disposition des ressources nécessaires.

4. CONCLUSIONS ET RECOMMANDATIONS DE L'AGENCE

L'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail adopte les conclusions du Conseil Scientifique et du GT « Méthodologie de l'Évaluation des Risques» et recommande une évolution du système qualité dans le cadre de l'amélioration continue.

Le Directeur Général

MOTS-CLES

Poids des preuves, ligne de preuves, planification de l'évaluation, processus d'expertise, revue systématique, transparence

Évaluation du poids des preuves à l'Anses : revue critique de la littérature et recommandations à l'étape d'identification des dangers

Rapport d'étape

Saisine « n°2015-SA-0089 »

RAPPORT
d'expertise collective

Groupe de travail « Méthodologie d'évaluation des risques »

Mai 2016

Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail,
14 rue Pierre et Marie Curie, 94701 Maisons-Alfort Cedex
Téléphone : + 33 (0)1 49 77 13 50 - Télécopie : + 33 (0)1 49 77 26 26 - www.anses.fr

Mots clés

Poids des preuves, ligne de preuves, planification de l'évaluation, processus d'expertise, revue systématique, transparence

Key Words

Weight of evidence, line of evidence, scoping, expertise process, systematic review, transparency

Présentation des intervenants

PRÉAMBULE : Les experts externes, membres de comités d'experts spécialisés, de groupes de travail ou désignés rapporteurs sont tous nommés à titre personnel, *intuitu personae,* et ne représentent pas leur organisme d'appartenance.

GROUPE DE TRAVAIL « METHODOLOGIE D'EVALUATION DES RISQUES » (GT MER)

Président

M. David MAKOWSKI – Directeur de recherche à l'INRA – Évaluation des risques, statistique, modélisation

Membres

Mme Isabelle ALBERT – Chargée de recherche à l'INRA – Évaluation des risques alimentaires, statistique, microbiologie, modélisation

Mme Nathalie BONVALLOT – Enseignante chercheuse à l'EHESP – Évaluation des risques, toxicologie

Mme Soraya BOUDIA – Professeure à l'Université Paris Descartes – Histoire et sociologie des risques

Mme Céline BROCHOT – Chercheuse à l'INERIS – Modélisation, toxicocinétique, statistique

M. Olivier BRUYERE – Professeur à l'Université de Liège – Méthodologie, épidémiologie, santé publique

M. Philippe GLORENNEC – Enseignant chercheur à l'EHESP et à l'IRSET – UMR Inserm 1085 – Méthodologie d'évaluation des expositions et des risques sanitaires environnementaux

M. Pierre MARTIN – Chercheur au CIRAD – Santé végétale, knowledge management, modélisation informatique

Mme Bette MEEK – Directrice associée à l'Université d'Ottawa – Évaluation des risques des produits chimiques, toxicologie

M. Claude SAEGERMAN – Professeur à la Faculté de médecine vétérinaire, Université de Liège – Épidémiologie et analyse de risques appliquées aux sciences vétérinaires

Mme Mathilde TOUVIER – Chercheuse à l'INSERM – Épidémiologie nutritionnelle

Mme Jessica TRESSOU – Chargée de recherche à l'INRA – Statistique, modélisation, évaluation du risque alimentaire

Mme Laurence WATIER – Chargée de recherche à l'INSERM – Biostatistique, épidémiologie, santé publique

Participation Anses

Mme Claire BLADIER – Chef de projet scientifique – Direction de l'évaluation des risques

Mme Eve FEINBLATT – Chargée de projet en sciences humaines et sociales – Direction de l'information, de la communication et du dialogue avec la société

Mme Sandrine FRAIZE-FRONTIER – Chef de projet scientifique – Direction de l'évaluation des risques

M. Moez SANAA – Conseiller scientifique pour l'évaluation des risques - Direction de l'évaluation des risques

Secrétariat administratif

Mme Virginie SADE – Anses

ÉQUIPE D'ACTION « PREUVES »

Président

M. Pierre MARTIN – Chercheur au CIRAD – Santé des plantes, knowledge management, modélisation informatique

Membres

M. Olivier BRUYERE – Professeur à l'Université de Liège – Méthodologie, épidémiologie, santé publique

M. David MAKOWSKI – Directeur de recherche à l'INRA – Évaluation des risques, statistique, modélisation

Mme Bette MEEK – Directrice associée à l'Université d'Ottawa – Évaluation des risques des produits chimiques, toxicologie

Mme Mathilde TOUVIER – Chercheuse à l'INSERM – Épidémiologie nutritionnelle

Mme Laurence WATIER – Chargée de recherche à l'INSERM – Biostatistique, épidémiologie, santé publique

Coordination et contribution scientifique Anses

Mme Claire BLADIER – Chef de projet scientifique – Direction de l'évaluation des risques

Mme Eve FEINBLATT – Chargée de projet en sciences humaines et sociales – Direction de l'information, de la communication et du dialogue avec la société

RAPPORTEURS DU CONSEIL SCIENTIFIQUE

M. Michel GERIN – Professeur honoraire, professeur associé, Département de santé environnementale et santé au travail, école de santé publique, université de Montréal

M. Alfred BERNARD – Professeur à l'université catholique de Louvain

CONSEIL SCIENTIFIQUE

Les travaux, objets du présent rapport, ont été suivis et adoptés par le Conseil Scientifique :

Président

M. Paul FRIMAT – Professeur de médecine du travail à l'université Lille-II

Membres de droit

la présidente du conseil scientifique de l'Agence nationale de sécurité du médicament et des produits de santé (ANSM), représentée par M. Robert BAROUKI, professeur des universités, directeur d'unité à l'Institut national de la santé et de la recherche médicale (INSERM)

le président du conseil scientifique de l'Institut de veille sanitaire (InVS), représenté par M. Christian DUCROT, directeur de recherche à l'Institut national de la recherche agronomique (INRA)

Personnalités scientifiques compétentes

Mme Geneviève ABADIA-BENOIST – Responsable du département Études et assistance médicales à l'Institut national de recherche et de sécurité pour la prévention des accidents du travail et des maladies professionnelles

M. Alfred BERNARD – Professeur à l'université catholique de Louvain

M. Xavier BIGARD – Professeur agrégé du Val-de-Grâce

M. Olivier BORRAZ – Directeur de recherche au Centre national de la recherche scientifique, directeur du Centre de sociologie des organisations CNRS-Institut d'études politiques de Paris

Mme Soraya BOUDIA – Professeure de sociologie à l'Université Paris Descartes

M. Thierry CANDRESSE – Directeur de recherche à l'Institut national de la recherche agronomique

Mme Claude CASELLAS – Professeure à la faculté de pharmacie de l'université Montpellier-I

Mme Véronique COXAM – Directrice de recherche à l'Institut national de la recherche agronomique (INRA)

M. Jean-Pierre CRAVEDI – Directeur de recherche à l'Institut national de la recherche agronomique (INRA)

M. Joseph DOMENECH – Inspecteur général honoraire de la santé publique vétérinaire, chargé de mission à l'Organisation mondiale de la santé animale (OIE)

Mme Brigitte ENRIQUEZ – Professeure à l'Ecole nationale vétérinaire d'Alfort

M. Tony FLETCHER – Épidémiologiste à la London School of Hygiene and Tropical Medicine et au Public Health England

Mme Jeanne GARRIC – Directrice de recherche à l'Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture (IRSTEA)

M. Michel GERIN – Professeur honoraire, professeur associé, département de santé environnementale et santé au travail, école de santé publique, université de Montréal

M. Philippe GIORDANENGO – Professeur des universités à l'Institut Sophia Agrobiotech

M. Pierre KERKHOFS – Directeur général du Centre d'études et de recherches vétérinaires et agrochimiques (CODA-CERVA)

Mme Bette MEEK – Directrice associée d'évaluation des risques des produits chimiques à l'université d'Ottawa

Mme Isabelle MOMAS – Professeure des universités, directrice du département santé publique et biostatistique à la faculté des sciences pharmaceutiques et biologiques de l'université Paris-Descartes

M. Christophe NGUYEN-THE – Directeur de recherche à l'Institut national de la recherche agronomique (INRA)

M. François PAQUET – Professeur, responsable de la mission programmes et stratégie scientifique à l'Institut de radioprotection et de sûreté nucléaire (IRSN)

Mme Sylvia RICHARDSON – Professeure à l'université de Cambridge, directrice de l'unité de biostatistique du Medical Research Council

M. Noël TORDO – Chef de l'unité stratégies antivirales à l'Institut Pasteur de Paris

AUDITION DE PERSONNALITÉS EXTÉRIEURES

M. Idesbald BOONE **–** Bundesinstitut für Risikobewertung – Épidémiologie vétérinaire, analyse des risques et diagnostic, recherche et développement

Mme Myriam MERAD **–** Directeur de recherche INERIS **–** Direction des risques chroniques modélisation environnementale et décision

M. Lorenz RHOMBERG **–** Principal chez Gradient (cabinet de conseil) - Massachusetts, USA **–** Évaluation quantitative des risques

SOMMAIRE

Présentation des intervenants 3
Sigles et abréviations 8
Liste des tableaux 9
Liste des figures 9
1. Contexte, objet de la saisine et du rapport 10
1.1 Contexte 10
1.2 Identification des enjeux et besoins 10
1.3 Objet de la saisine 11
1.4 Modalités de traitement de la saisine et objectif du rapport 12
1.5 Prévention des risques de conflits d'intérêts. 12
2. Organisation de l'expertise 13
2.1 Déroulement de l'expertise 13
2.2 Domaines couverts par les documents retenus 16
3. Revue critique des documents guides et de la littérature scientifique 17
3.1 Définitions 17
3.1.1 Définitions du poids des preuves 17
3.1.2 Définition de ligne de preuves 18
3.2 Les étapes de l'évaluation du poids des preuves 19
3.2.1 Revue de la littérature 19
3.2.2 Proposition du groupe de travail 21
3.2.3 Place de la revue systématique dans l'évaluation du poids des preuves 22
3.3 Analyse critique des démarches et méthodes recensées dans la littérature scientifique 24
3.3.1 Planification de l'évaluation 26
3.3.2 Établissement des lignes de preuves 29
3.3.3 L'intégration des lignes de preuves pour établir le poids des preuves 33
3.3.4 L'expression des conclusions concernant le poids des preuves 36
4. Revue des pratiques actuelles de l'Anses 38
4.1 Le processus d'expertise à l'Anses 38
4.2 Résultats du rapport « État des lieux sur l'analyse de l'incertitude et l'évaluation du poids des preuves à l'Anses » 40
4.3 Exemples de classification pour l'expression des conclusions 41
5. Recommandations 42
5.1 Planification de l'évaluation 42
5.2 Établissement des lignes de preuves 43
5.2.1 Recherche, sélection des études et extraction des données 43
5.2.2 Évaluation de la qualité des études 43
5.2.3 Évaluation d'un ensemble d'études par type d'étude en vue d'établir des lignes de preuve 44
5.3 Intégration des lignes de preuves 44
5.4 Expression des conclusions concernant le poids des preuves 45
5.5 Intégration dans le processus d'expertise 45
5.6 Classification des méthodes recommandées par le GT MER en fonction de leur niveau de faisabilité 46
6. Conclusions du groupe de travail 47
7 Bibliographie 49

7.1 Publications 49
7.2 Normes 56
Annexe 1 : Décision d'autosaisine 58
..... 59
Annexe 2 : Grille de lecture 60
Annexe 3 : Liste des organismes consultés pour le recensement des guides existants sur l'évaluation du poids des preuves 61
Annexe 4 : Documents de référence 64
Annexe 5 : Définitions WOE issues de la recherche bibliographique 66
Annexe 6 : Définitions de « revue systématique » dans la littérature 67
Annexe 7 : Fiches méthodes 68
Annexe 8 : Exemple de modèles conceptuels 98
Annexe 9 : Saisines Anses utilisant des méthodes d'évaluation du poids des preuves analysées par le GT MER 99

Sigles et abréviations

AFNOR	Association française de normalisation
Anses	Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail
BfR	Bundesinstitut für Risikobewertung (Institut fédéral d'évaluation des risques)
CES	Comité d'experts spécialisés
CIRC	Centre international de Recherche sur le Cancer
CS	Conseil scientifique
EA	Équipe d'action
ECHA	Agence européenne des produits chimiques
EFSA	European Food Safety Authority (Autorité européenne de sécurité des aliments)
EPICOT	Evidence ; Population ; Intervention ; Comparison ; Outcome ; Time stamp (preuve ; population ; intervention ; comparaison ; effet; temps)
ERS	Évaluation des risques sanitaires
SGH	Système général harmonisé
GRADE	Grading of Recommendations Assessment, Development and Evaluation
GECU	Groupe d'expertise collective d'urgence
GT	Groupe de travail
GT MER	Groupe de travail « méthodologie de l'évaluation des risques »
IARC	International Agency for Research on Cancer (Centre international de recherche sur le cancer)
ISO	International Organization for Standardization (Organisation internationale de normalisation)
MOA	Mode of action (mode d'action)
OCDE	Organisation de Coopération et de Développement Économiques
OHAT	Office of Health Assessment and Translation (office d'évaluation sanitaire et de traduction)
PDP	Poids des preuves
PECO	Patient, problem or population ; Exposure ; Comparison, control or comparator ; Outcomes (patient, problème ou population ; exposition ; comparaison, contrôle ou comparateur ; effets)
PECOTS	Populations, Exposures, Comparators, Outcomes, Timings, Settings of interest (populations ; expositions ; comparateurs ; effets ; temps; localisation d'intérêt)
PICO	Patient, problem or population ; Intervention ; Comparison, control or comparator ; Outcomes (patient, problème ou population ; intervention ; comparaison, contrôle ou comparateur ; effets)
QSAR	Quantitative structure-activity relationship (relation quantitative structure-activité)
RA	Revue approfondie de la littérature
REACH	Registration, Evaluation and Authorization and Restriction of Chemicals
RS	Revue systématique de la littérature
SAR	Structure-activity relationship (relation structure-activité)
SCENIHR	Scientific Committee on Emerging and Newly Identified Health Risks (comité scientifique européen sur les risques émergents et nouvellement identifiés pour la santé)
SGH	Système général harmonisé
VLEP	Valeurs limites d'exposition professionnelle
VTR	Valeurs toxicologiques de référence
WCRF	World Cancer Research Fund (Fonds mondial de recherche contre le cancer)
WOE	Weight of evidence

Liste des tableaux

Tableau 1 : Domaines couverts par les publications sur le poids des preuves† 16
Tableau 2 : Exemples de seuils du poids des preuves 17
Tableau 3 : Etapes de l'évaluation du poids des preuves issues de la recherche bibliographique....20
Tableau 4 : Étapes du processus d'évaluation du PDP couvertes par les méthodes recensées dans la revue de littérature....25
Tableau 5 : Domaines couverts par les méthodes recensées dans la revue de littérature....26
Tableau 6 : Notes comparatives des méthodes pour l'étape de planification de l'évaluation 29
Tableau 7: Notes comparatives des méthodes pour l'étape de l'établissement des lignes de preuves 29
Tableau 8 : Exemple de notation de lignes de preuve à partir de scores obtenus pour chaque critère et de leurs poids associés (Linkov et al. 2011)....32
Tableau 9 : Notes comparatives des méthodes pour l'étape de l'intégration des lignes de preuves....33
Tableau 10 : Exemple d'évaluation comparative du poids des preuves pour le mode d'action mutagénique du 2,3-trichloropropane (Meek et al, 2013) 35
Tableau 11 : Exemples d'expression du poids des preuves 36
Tableau 12 : Notes comparatives des méthodes pour l'étape de l'expression des conclusions sur le poids des preuves à l'étape d'identification du danger 37
Tableau 13 : Formulaires du processus d'expertise en lien avec le poids des preuves 38
Tableau 14 : Méthodes utilisées dans les saisines† Anses analysées par le GT MER....40
Tableau 15 : Exemples de classification pour l'expression des conclusions dans les avis 41
Tableau 16 : Classification des méthodes† recommandées par le GT MER en fonction de leur niveau de faisabilité (ressources). 46
Tableau 17 : Documents retenus suite à la recherche bibliographique 64
Tableau 18 : Guides supplémentaires retenus suite au questionnaire envoyé aux agences 65

Liste des figures

Figure 1 : Stratégie de sélection des documents....14
Figure 2 : Démarche d'évaluation du poids des preuves....22
Figure 3 : Modèle conceptuel décrivant les voies environnementales qui résultent en un risque accru d'infection humaine et animale par des bactéries résistantes aux antibiotiques (Ashbolt et al. 2013) 27
Figure 4 : Eléments de protocole de la revue systématique (NRC 2014)....28
Figure 5 : *Forest plot* – Méta-analyse dose-réponse du WCRF sur la relation entre indice de masse corporelle (IMC) et risque de cancer de l'ovaire (WCRF/AICR 2014) 31
Figure 6 : Arbre de décision utilisé dans le cadre de la saisine relative à l'évaluation des risques du bisphénol A (BPA) pour la santé humaine (Anses 2013a) 35

1. Contexte, objet de la saisine et du rapport

1.1 Contexte

L'Anses contribue à la sécurité sanitaire humaine dans les domaines de l'environnement, du travail, de l'alimentation, à la protection de la santé et du bien-être des animaux ainsi qu'à la protection de la santé des végétaux. L'agence est chargée de mettre en œuvre une expertise scientifique indépendante et pluraliste afin d'évaluer les risques et de proposer aux autorités compétentes toute mesure de nature à préserver la santé publique.

L'agence identifie et caractérise les dangers biologiques, chimiques ou physiques au travers de l'ensemble des données scientifiques disponibles, et évalue les risques pour les populations humaines, animales ou végétales en prenant en compte les différents types d'exposition des individus aux différents dangers. L'Anses évalue également des produits (phytosanitaires, biocides, médicaments vétérinaires, nouveaux aliments, produits chimiques, etc.) et procédés (traitement des eaux ou des aliments, etc.) au regard d'exigences réglementaires, en vue d'une décision de mise sur le marché ou d'autorisation d'utilisation par les ministères en charge de la réglementation.

Les questions instruites par l'Agence sont traitées dans le cadre du processus d'évaluation des risques sanitaires, constitué des quatre étapes suivantes : l'identification des dangers, la caractérisation des dangers (incluant l'établissement de la relation dose-réponse), l'évaluation de l'exposition et la caractérisation des risques. Ces questions peuvent être de différents types, par exemples : "La consommation de fibres alimentaires influence-t-elle le risque de cancer colorectal ?", "Les maladies à prions sont-elles transmissibles à l'homme ?". Les informations nécessaires pour répondre aux questions instruites proviennent de sources multiples (bases de données de la littérature scientifique, expertises de professionnels). Les données à mobiliser sont de différentes natures (in vitro, in vivo, toxicologiques, épidémiologiques, etc.) et peuvent être contradictoires. Il convient alors d'analyser ces informations de façon approfondie pour en évaluer la pertinence et la validité, et de les combiner afin d'apporter une réponse à la question posée. Dans la littérature, la méthodologie permettant d'évaluer la pertinence et la qualité des données puis de combiner des données hétérogènes est appelée « évaluation du poids des preuves[1] » (Weight of Evidence en anglais). La littérature fait principalement référence à l'évaluation du poids des preuves à l'étape d'identification du danger. Cette étape a pour objectif d'identifier le type et la nature des effets néfastes qu'un agent (biologique, chimique ou physique) peut causer sur un organisme, un système ou une population (IPCS 2004).

1.2 Identification des enjeux et besoins

Pour orienter l'action publique, les pouvoirs publics et les parties prenantes sont en demande de preuves de différentes natures pour un large éventail de domaines (NRC 2014, OCDE 2015, US EPA 2014). Récemment, les divergences entre les conclusions apportées par différents organismes sur la toxicité ou sur le risque lié à l'exposition à diverses substances (par exemple, entre les conclusions de l'EFSA et de l'IARC sur le glyphosate[2] ; entre les conclusions de l'EFSA et de

1 Le Centre National de Ressource et de Traitement de la Langue donne une définition générale du terme « preuve » : « Fait, témoignage, raisonnement susceptible d'établir de manière irréfutable la vérité ou la réalité de (quelque chose).»

2 Voir http://www.efsa.europa.eu/en/press/news/160113

l'Anses sur le Bisphénol A[3]) ont mis en avant l'impact des systèmes de classification des effets sanitaires et des modèles conceptuels explicatifs de l'apparition d'un effet sanitaire sur le résultat final de l'évaluation. L'évaluation des preuves scientifiques dans le cadre de l'expertise scientifique fait donc l'objet d'un intérêt croissant de la part d'agences sanitaires, au niveau national et international notamment en vue d'améliorer la robustesse et la transparence des travaux d'expertise (Hardy et al. 2015, OHAT 2015). Cette transparence est indispensable pour assurer la crédibilité et la confiance de la communauté scientifique comme des autres parties prenantes auprès des pouvoirs publics.

Bien que fréquemment utilisé, le concept de « poids des preuves » est ambigu, souvent mal défini, et ne recouvre pas une méthodologie transparente et cohérente (NRC 2014, Weed 2005). Historiquement, l'évaluation du poids des preuves s'est développée dans le secteur médical comme un outil d'aide à la décision clinique de hiérarchisation des connaissances de la recherche médicale (Sackett et al. 1996). Les méthodes étaient alors focalisées principalement sur la revue critique de la littérature, d'abord dans le domaine de la médecine factuelle[4] puis adaptées à l'environnement et à la santé environnementale (Mandrioli et Silbergeld 2015, Krimsky 2005). Par la suite, des développements méthodologiques ont été réalisés pour combiner différentes sources d'information de façon transparente et systématique, notamment dans le cadre des expertises collectives.

Au sein de l'Anses, un état des lieux des pratiques sur l'analyse de l'incertitude et l'évaluation du poids des preuves a montré que certaines saisines avaient mobilisé des méthodes d'évaluation du poids des preuves. Ce rapport révèle que les pratiques des collectifs d'experts diffèrent selon les thématiques (risques physico-chimiques, biologiques, liés à la nutrition, etc.) et au sein d'une même thématique. En particulier, le niveau de formalisation des méthodes utilisées et leur expression au sein des rapports varient selon les collectifs d'experts.

1.3 Objet de la saisine

Dans l'objectif d'améliorer la transparence du processus d'expertise requise pour un organisme certifié ISO 9001, l'Anses a confié au Groupe de Travail « Méthodologie de l'Évaluation des Risques » (GT MER) la conduite d'une analyse critique sur les approches d'évaluation des niveaux de preuve mobilisables pour l'étape d'identification des dangers.

Par décision en date du 31 mars 2015, l'Anses s'est donc autosaisie afin de conduire une analyse critique sur les approches d'évaluation des niveaux de preuve pour cette étape (Annexe 1).

Les objectifs de cette autosaisine sont de :

- Décrire les pratiques actuelles de l'Anses et les comparer avec celles d'autres organismes/agences sanitaires.
- Faire une revue critique des approches sur les niveaux des preuves.
- Proposer des procédures harmonisées pour évaluer la qualité des études et des données disponibles, ainsi que des niveaux de preuve par rapport aux questions ou hypothèses avancées.
- Proposer des procédures harmonisées pour évaluer et communiquer le niveau de preuve global associé à l'ensemble des données et études disponibles.
- Démontrer l'applicabilité des recommandations grâce à des études de cas.

3 (Anses 2015c)

4 Traduction française de « evidence based medecine »

Les résultats de cette saisine contribueront à améliorer la transparence et la reproductibilité de l'évaluation du poids des preuves à l'Anses.

1.4 Modalités de traitement de la saisine et objectif du rapport

L'instruction de cette autosaisine est réalisée en trois étapes :

- Réalisation d'un état des lieux des pratiques actuelles de l'Anses,
- Revue de la littérature sur le poids des preuves et formulation de recommandations visant à harmoniser les procédures de l'Anses,
- Évaluation des recommandations à travers des études de cas en interaction avec les collectifs de l'Anses et rédaction d'un guide.

La première étape a été réalisée par l'équipe d'action (EA) « État des lieux » mise en place par le GT MER. Cette EA a établi un état des lieux des pratiques de l'Anses sur l'analyse de l'incertitude et l'évaluation du poids des preuves qui a été présenté et discuté au conseil scientifique le 22 septembre 2015 (Anses 2015b).

Le présent rapport présente les résultats de la seconde étape. Il a été réalisé par l'équipe d'action « Poids des preuves ». Ce rapport détaille la revue de la littérature sur le poids des preuves et formule une série de recommandations visant à harmoniser les pratiques de l'Anses.

La troisième étape mentionnée ci-dessus fera l'objet d'un travail spécifique, réalisé en interaction avec les collectifs d'experts de l'Anses en 2016-2017. Un guide méthodologique sera rédigé à l'issue de cette dernière étape. Il proposera des méthodes adaptées à différentes situations pratiques pour évaluer la qualité des études et des données disponibles, et pour évaluer et communiquer le poids des preuves.

Le travail d'expertise présenté dans ce rapport a été conduit par un collectif d'experts intervenant dans les différents domaines de l'Agence et ayant des compétences dans les méthodes d'évaluation des risques.

Le présent rapport a été validé par le GT MER. Il a été soumis au conseil scientifique (CS) et tient compte des observations et éléments complémentaires transmis par les membres du CS.

L'expertise a été réalisée dans le respect de la norme NF X 50-110 « Qualité en expertise – prescriptions générales de compétence pour une expertise (Mai 2003) » (AFNOR 2003).

1.5 Prévention des risques de conflits d'intérêts.

L'Anses analyse les liens d'intérêts déclarés par les experts avant leur nomination et tout au long des travaux, afin d'éviter les risques de conflits d'intérêts au regard des points traités dans le cadre de l'expertise.

Les déclarations d'intérêts des experts sont rendues publiques *via* le site internet de l'Anses (www.anses.fr).

2. Organisation de l'expertise

2.1 Déroulement de l'expertise

L'expertise s'est déroulée en six phases.

Phase 1 : **Revue bibliographique préliminaire**

Une revue bibliographique préliminaire a été réalisée par les experts de l'équipe d'action EA « Poids des preuves » et la coordination scientifique de l'Anses. Cette revue a permis d'identifier un corpus initial de 22 documents, en particulier des guides méthodologiques d'agences. Cinq de ces documents ont été exclus après lecture des résumés car ils ne traitaient pas d'évaluation du poids des preuves. Une grille de lecture permettant de disposer d'une vision synthétique des méthodes utilisées a été établie puis testée sur les 17 documents. Cette grille incluait notamment les informations suivantes : auteur, domaine concerné, problématique traitée, définitions des concepts étudiés, codage du niveau de preuve, limites, etc. (Annexe 2).

Cette revue bibliographique préliminaire a montré que le poids des preuves était considéré pour les différentes étapes de l'évaluation des risques. Par exemple, un rapport du comité scientifique des risques sanitaires émergents et nouveaux (SCENIHR 2012) préconise d'évaluer le poids des preuves relatives à des données d'exposition. Dans le domaine de l'évaluation des risques écologiques, Hope et Clarkson (2014) prennent en considération le poids des preuves concernant les mesures d'effets adverses mais aussi d'exposition comme la concentration d'une substance chimique dans un milieu. Ce constat a conduit l'équipe d'action EA « Poids des preuves » à ne pas restreindre la revue bibliographique à l'étape d'identification des dangers, mais à considérer l'ensemble des étapes de l'évaluation des risques pour le choix des termes de la requête bibliographique.

Phase 2 : **État des lieux des guides des agences et des organismes internationaux sur l'évaluation du poids des preuves**

Un courrier a été envoyé à 63 institutions nationales et internationales dans le domaine de l'évaluation des risques afin de recenser les guides existants sur l'évaluation du poids des preuves (Annexe 3).

Les 36 réponses reçues ont permis d'identifier trois guides méthodologiques en plus de ceux identifiés lors de la phase 1 (Annexe 4, Tableau 18).

Phase 3 : **Recherche approfondie de la littérature sur l'évaluation du poids des preuves, sélection des documents et extraction des données**

Une recherche approfondie de la littérature (cf. partie 3) a été réalisée dans les bases de données Scopus et Pubmed, sur la base de critères et termes définis par l'équipe d'action. Les critères retenus pour la recherche bibliographique étaient les suivants :

- Type : Review
- Année : à partir de 2010
- Langue : Français et Anglais
- Mots de la requête : dans titre, abstract, mots clés

Les différents termes de la requête retenus pour la recherche bibliographique étaient les suivants :

- "weight of evidence"/"weight-of-evidence"/"weighing of evidence"
- "scoring method"
- "quality criteria"
- "data integration"
- "lines of evidence"/"line of evidence"

- "level of evidence"/"levels of evidence"
- "strength of evidence"/"strengths of evidence"
- "quality of evidence"
- "evidence integration/integration of evidence"

chacun de ces termes ayant été complété par "risk management", "risk assessment" ou "risk analysis".

Les publications ont été retenues à partir de 2010 dans le but d'identifier les méthodes les plus récentes qui n'auraient pas été incluses dans les revues de la littérature de Rhomberg et al (2013) et de la HAS (2013), complétant celles listées suite à la revue bibliographique préliminaire.

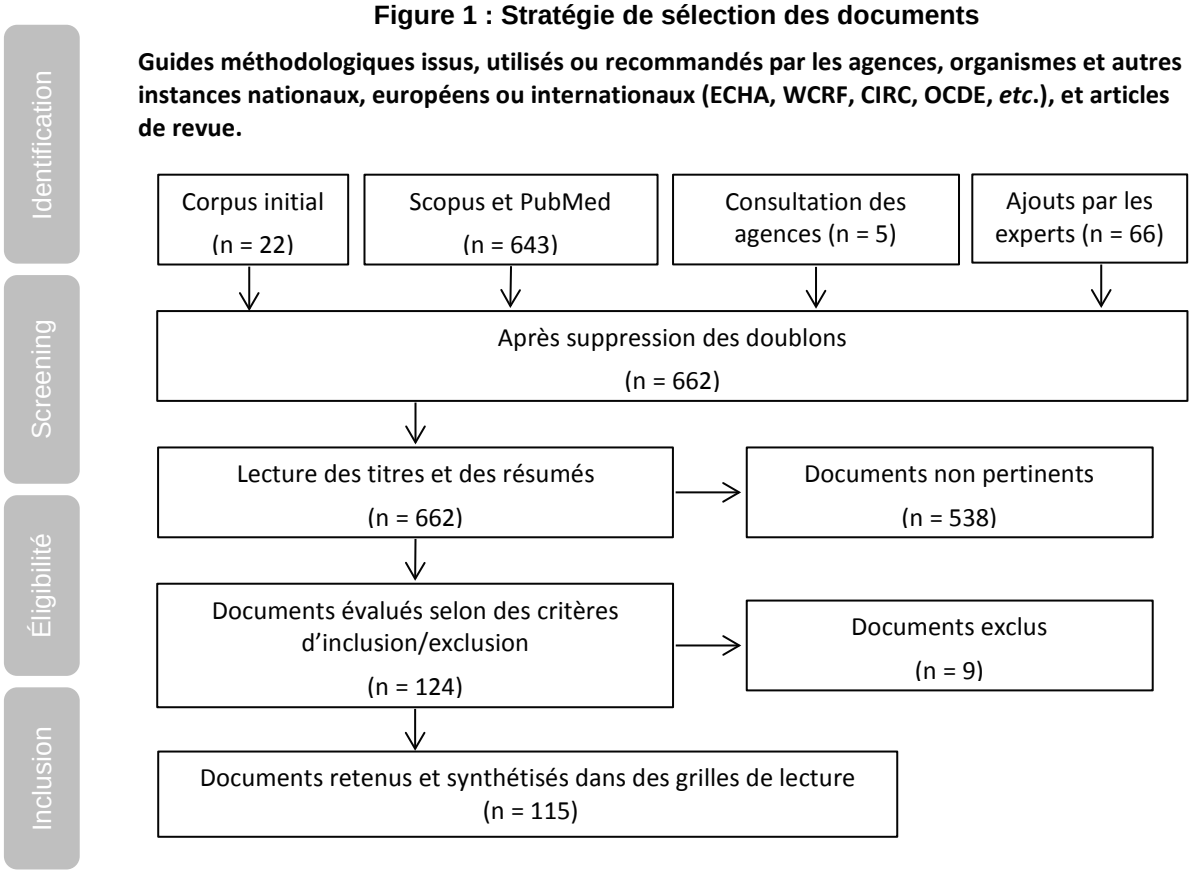
Ces requêtes ont conduit à recueillir 643 documents (571 après exclusion des doublons). Afin de sélectionner ceux portant sur les principes de l'analyse du poids des preuves ou sur une méthode spécifique, chaque titre a été revu par deux membres de l'équipe d'action. En cas d'indécision, les résumés ont été consultés. Trente-un ont ainsi été sélectionnés et décrits à l'aide de la grille de lecture (Annexe 2).

Soixante-six documents supplémentaires ont été identifiés par les membres de l'EA « Poids des preuves » à partir de leur expertise et des références bibliographiques des documents retenus, notamment pour préciser des éléments autour des méthodes identifiées dans la recherche bibliographique.

Au terme de cette phase, les 124 documents sélectionnés (issus du corpus initial, de la recherche bibliographique approfondie, de la consultation des agences et des ajouts par les experts) ont fait l'objet d'une analyse critique et collective. Après analyse, 115 documents de référence ont finalement été retenus pour la suite du travail, dont 22 issus de la recherche bibliographique approfondie (Annexe 4).

La stratégie de sélection des documents est résumée dans la **Figure 1**, inspirée du schéma PRISMA (Moher et al. 2015).

Figure 1 : Stratégie de sélection des documents



Phase 4 : **Auditions complémentaires**

Pour éclaircir certains points de la revue de la littérature, des auditions ont été conduites :

- auprès de l'Institut fédéral d'évaluation des risques allemand (BfR), sur l'évaluation de la qualité des données et des hypothèses dans le domaine du risque microbien,
- auprès de Lorentz Rhomberg de la société de conseil Gradient pour sa revue des cadres de travail dédiés à l'évaluation du poids des preuves (Rhomberg et al. 2013),
- auprès de Myriam Merad de l'INERIS sur le processus d'aide à la décision et son impact sur la formulation du problème et l'expression des conclusions.

Phase 5 : **Revue des méthodes et démarches**

Les caractéristiques des méthodes et démarches utilisées dans le cadre de l'évaluation du poids des preuves ont été comparées au moyen des grilles de lecture pour chacune des étapes d'évaluation du poids des preuves. L'EA « Poids des Preuves » a défini trois critères pour évaluer les méthodes et démarches :

- **Caractère directif** : une méthode reçoit une note élevée pour ce critère si elle propose un cadre précis rendant l'expertise transparente et reproductible.
- **Pertinence** : une méthode reçoit une note élevée pour ce critère si elle permet de répondre à l'objectif défini.
- **Faisabilité** : une méthode reçoit une note élevée pour ce critère si elle mobilise des ressources en temps et en ressources humaines peu importantes, et si elle ne requiert pas de compétence méthodologique spécifique (modélisation, statistique).

Ces critères ont été évalués par l'EA « Poids des Preuves » en fixant une note de 1 à 4, d'abord de manière individuelle, puis en groupe afin d'aboutir à un consensus.

L'EA « Poids des Preuves » s'est focalisée sur le classement relatif des méthodes, plus que sur leur évaluation individuelle.

Phase 6 : **Comparaison entre la littérature et les pratiques à l'Anses – Recommandations**

Les méthodes recensées lors de la revue bibliographique ont été comparées à celles inventoriées dans le rapport « État des lieux Anses » afin de définir des pistes de progrès. Les documents du système qualité de l'Anses ont également été considérés pour évaluer l'inscription de l'évaluation du poids des preuves dans le processus d'expertise.

2.2 Domaines couverts par les documents retenus

Les documents de référence couvrent plusieurs domaines de compétences de l'agence, mais de façon inégale (

Tableau **1**). La plupart des publications concernent les domaines Santé-Environnement et Santé-Alimentation. Pour ce dernier, la majorité des documents concernent les contaminants chimiques dans l'alimentation. Une moindre part concerne la nutrition, la santé au travail ou la santé et la protection des végétaux. Seuls trois documents retenus couvrent la microbiologie des aliments, et deux la santé et le bien-être des animaux.

En dehors des champs de compétence de l'agence, une grande proportion des documents (17) fait référence au domaine médical. De plus, huit documents sont spécifiques au domaine de l'écologie et l'environnement. Toutefois, les méthodes qui y sont développées peuvent être étendues à d'autres domaines.

Tableau 1 : Domaines couverts par les publications sur le poids des preuves[†]

Domaine	Nombre de documents
Santé-Travail	10
Santé-Alimentation	**27** (majoritairement Chimie)
Alimentation et santé animale	2
Santé et protection du végétal	7
Santé-Environnement	**31**
Produits phytosanitaire, biocides et fertilisants	6
Domaine hors Anses : Médical	*17*
Écologie – Environnement	*8*

[†]un document peut couvrir plusieurs domaines

3. Revue critique des documents guides et de la littérature scientifique

3.1 Définitions

Différentes définitions du poids des preuves sont proposées dans la littérature. Une de ces définitions est probabiliste, les autres sont qualitatives.

3.1.1 Définitions du poids des preuves

- **Définition probabiliste**

Historiquement, la définition du poids des preuves est probabiliste. Elle a été proposée par Good sur la base des travaux de Turing (Good 1979, 1985). Good définit le poids W des preuves E associées à l'hypothèse H de la façon suivante :

$$W(H:E) = log\left(\frac{P(E|H)}{P(E|\bar{H})}\right)$$

où $P(E|H)$ correspond à la probabilité conditionnelle et $\bar{H}$ correspond à l'hypothèse alternative. Cette expression correspond au logarithme du rapport des vraisemblances, rapport également appelé « Facteur de Bayes » dans l'approche bayésienne, c'est à dire au logarithme du rapport des probabilités d'observer les preuves sachant que l'hypothèse est vraie ou fausse.

Dans certaines situations, le calcul de $W(H:E)$ est simple. Par exemple, dans le cas où E est le résultat d'un test de diagnostic médical et où H et $\bar{H}$ correspondent respectivement au fait d'être malade et non malade, la valeur de W est égale au logarithme du ratio de la sensibilité du test de diagnostic sur le taux de faux positifs (1 – spécificité du test).

Certains auteurs (Jeffreys 1961) ont proposé des classes de valeur arbitraire de $W(H:E)$ avec leur interprétation (Tableau 2).

Tableau 2 : Exemples de seuils du poids des preuves

Valeur (log en base 10) de $W(H:E)$	Poids des preuves selon Jeffreys
<0	Négatif
0 à 0.5	Faible
0.5 à 1	Substantiel
1 à 1.5	Fort
1.5 à 2	Très fort
> 2	Décisif

Ces seuils correspondent à des conventions, utilisées notamment en sciences cognitives (Tenenbaum et al. 2011). Pour illustrer cette convention, lorsque $log_{10}W(H:E)$ est supérieur à 1, cela indique un poids des preuves fort en faveur de l'hypothèse H.

La valeur de $W(H:E)$ est souvent difficile à calculer, car elle nécessite un modèle spécifique. Les approches mobilisées dans l'analyse du poids des preuves étant actuellement, et pour la majorité, qualitatives, la valeur de $W(H:E)$ est rarement calculée pour les domaines d'intérêt de l'Anses.

- **Définitions qualitatives**

Dans les documents publiés recouvrant les domaines de l'Anses, les définitions du poids des preuves sont qualitatives. La recherche bibliographique a permis de recenser 16 définitions (Annexe 5), à partir desquelles trois notions distinctes du concept de poids des preuves ont été identifiées :

- Cadre de travail transparent pour tirer des conclusions ;
- Intégration de différentes lignes de preuve pour tester une hypothèse (Dans certains exemples, le poids des preuves correspond à une seule ligne de preuves) ;
- Processus pour prendre en considération les forces et les faiblesses de différents éléments d'information pour tester une hypothèse.

Dans ces notions, on retrouve l'expression de l'hypothèse[5] ou de la conclusion, de même que la relation entre des preuves et les hypothèses de la définition de Turing-Good, mais exprimée de manière qualitative.

Les définitions de « poids des preuves » se distinguent par ailleurs de celles de « force de la preuve ». D'après Weed (2005), "strength of evidence" est moins général que "weight of evidence" et renvoie à l'usage d'un sous-ensemble de preuves (parfois issu d'une seule étude) pour identifier un effet statistiquement significatif. D'après Krimsky (2005), "strength of evidence" est associé à l'intensité et à la pertinence des informations au regard d'indicateurs spécifiques, tels que le nombre de tumeurs produites chez les animaux.

Dans le but de disposer d'une définition intégrative du poids des preuves, le GT MER propose la définition suivante :

> **« Le poids des preuves (PDP) est une synthèse formalisée de lignes de preuves, éventuellement de qualités hétérogènes, dans le but de déterminer le niveau de plausibilité d'hypothèses. »**

La traduction anglaise de cette définition est : "Weight of evidence (WOE) is the structured synthesis of lines of evidence, possibly of varying quality, to determine the extent of support for hypotheses".

3.1.2 Définition de ligne de preuves

Deux définitions de "Line of evidence" ont été identifiées dans la bibliographie :

"Set of information used to evaluate endpoint. Lines of evidence are not all equally important in making the overall conclusion" (Hristozov, Zabeo, et al. 2014).

"Line of evidence is a measure associated with a specific risk hypothesis. Multiple lines of evidence can be associated with a single risk hypothesis" (Hope et Clarkson 2014).

Dans ces définitions, la ligne de preuves se réfère à trois notions distinctes :

- Un ensemble d'informations
- Une mesure

5 Le Larousse propose la définition suivante pour le terme "hypothèse": "Proposition visant à fournir une explication vraisemblable d'un ensemble de faits, et qui doit être soumise au contrôle de l'expérience ou vérifiée dans ses conséquences".

  - Utilisation d'informations de même nature (selon la discipline, par exemple : essais contrôlés randomisés, études épidémiologiques, études cliniques, études expérimentales, études centrées sur le mode d'action, etc.) pour évaluer un effet critique.

- Dans l'objectif de disposer d'une définition de ligne de preuves qui s'applique quelle que soit la méthode d'évaluation utilisée, le GT MER propose de combiner ces notions dans la définition suivante :

« Une ligne de preuves est un ensemble d'informations de même nature, intégrées pour évaluer une hypothèse».

- Par exemple, dans le domaine de l'étude des relations entre un facteur nutritionnel donné et le risque de cancer pour une localisation spécifique, le WCRF (WCRF/AICR 2014) établit une « ligne de preuves épidémiologiques » d'une part (réalisant le plus souvent une revue systématique et une méta-analyse dose-réponse des données disponibles), et une « ligne de preuves expérimentales » d'autre part (intégrant les résultats des études sur les modèles animaux concernant la plausibilité mécanistique et le mode d'action du facteur nutritionnel dans la carcinogenèse).

Exprimée en anglais, cette définition est : "A line of evidence is a set of relevant information of similar type grouped to assess a hypothesis".

Dans certains cas, la ligne de preuves est constituée d'une seule information.

3.2 Les étapes de l'évaluation du poids des preuves

3.2.1 Revue de la littérature

Le nombre de guides traitant du processus d'évaluation du poids des preuves est faible et beaucoup d'entre eux se focalisent sur la revue systématique. Cependant, cinq documents (produits par des organismes publiques d'évaluation des risques ou par des sociétés de conseil[6]) proposent une démarche générale pour l'évaluation du poids des preuves (Hope et Clarkson 2014, NRC 2014, OHAT 2015, Rhomberg et al. 2013, SCENIHR 2012). Ils appréhendent l'évaluation du poids des preuves selon un processus structuré en plusieurs étapes clés (Tableau 3). Ces cinq documents présentent des points communs mais également des différences parfois importantes, notamment sur le nombre d'étapes considérées, leurs définitions et le niveau de détails de chacune de ces étapes.

Les étapes de planification, cadrage, formulation du problème et développement de protocole proposées par le NRC (2014), Hope et Clarkson (2014) et l'OHAT (2015) comportent des éléments de planification de l'évaluation. Ces éléments permettent de préparer la conduite de l'évaluation des risques, comme le propose l'EPA (US EPA 2014). Par la suite, les cinq documents distinguent différentes étapes pour établir les lignes de preuves : la recherche et la sélection des études, l'extraction des données, l'évaluation de la qualité des études (avec application de critères), l'évaluation du poids/niveau de preuve de chaque ligne de preuves. Enfin, le poids des preuves est évalué en combinant/intégrant les lignes de preuve, à partir duquel les conclusions sont développées. Le SCENIHR (2012) étaye la conclusion en intégrant l'expression de l'incertitude au sein du processus d'évaluation du poids des preuves. Dans le contexte de l'évaluation des risques écologiques, Hope et Clarkson (2014) proposent d'associer une estimation du risque au poids des preuves.

6 Gradient pour l'American Chemistry Council (Rhomberg et al. 2013) ; CH2M et Ramboll Environ pour l'EPA (Hope et Clarkson 2014)

Tableau 3 : Etapes de l'évaluation du poids des preuves issues de la recherche bibliographique

Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR 2012)	Rhomberg et al (2013)	Hope et Clarkson (2014)	National Research Council (NRC 2014)	Office of Health Assessment and Translation (OHAT 2015)
		Planification et cadrage (Planning and scoping)	Cadrage (scoping)	
Identification des sources potentielles des données et du manque de données en relation avec l'objectif de l'ERS (Identification of the possible sources of data and data gaps in relation to the aim of the assessment)	Formulation des questions concernant la causalité et développement de critères pour la sélection des études (Define causal question and develop criteria for study selection)	Formulation du problème (Problem formulation)	Formulation du problème (Problem formulation) Développement du protocole (Protocol development)	Formulation du problème (avec « scoping » intégré) et développement du protocole (Formulate problem and develop protocol)
Analyse préliminaire de ces sources de données pour identifier celles qui sont pertinentes pour la question posée. (Initial screening of these data sources to identify those that are relevant to address the question posed)		Établissement des lignes de preuves (Lines of evidence)	Recherche et identification des études pertinentes pour la question posée (Identify evidence)	Recherche et sélection des études pertinentes pour la question posée (Search for and select studies for inclusion) Extraction des données et informations des études (Extract data from studies)
Évaluation de la qualité des études individuelles (Assessment of individual data sources)	Application des critères pour l'évaluation des études individuelles (Develop and apply criteria for review of individual studies)		Évaluation de la qualité des études individuelles (Evaluate studies)	Évaluation de la qualité des études individuelles (Assess internal validity of individual studies)
Évaluation du poids de chaque ligne de preuves (Weighing of the individual lines of evidence)	Intégration des preuves et évaluation globale du poids des preuves (Integrate and evaluate evidence)		Intégration des lignes de preuve (Integrate evidence)	Évaluation de la confiance accordée à chaque ligne de preuves (Synthesize evidence and rate confidence in body of evidence)
Évaluation globale du poids des preuves (Weighing of the totality of evidence)		Évaluation du poids des preuves pour chaque groupe de preuve / combinaison des poids des différentes lignes de preuve (Weighing of lines of evidence within evidence groups)		Évaluation du niveau de preuve pour un effet sanitaire, pour chaque ligne de preuves (Translate confidence ratings into level of evidence for health effect) Intégration des preuves et développement des conclusions pour l'identification du danger (Integrate evidence to develop hazard identification conclusions)
	Développement des conclusions basées sur les inférences (Draw conclusions based on inferences)	Développement des conclusions en croisant les poids des preuves et les estimations du risque (Risk matrix : risk estimates and evidence groups)		
Expression de l'incertitude (Expression of uncertainty)				

3.2.2 Proposition du groupe de travail

Sur la base des étapes décrites dans le paragraphe ci-dessus, le groupe de travail propose de structurer la démarche d'évaluation du poids des preuves selon quatre étapes clés (Figure 2)**.** Le terme « Formulation du problème » étant interprété de différentes manières dans la littérature, le GT MER propose de regrouper les opérations précédant la collecte des données sous le terme « Planification de l'évaluation ».

Etape 1. Planification de l'évaluation
Etape 2. Établissement des lignes de preuves
Etape 3. Intégration des lignes de preuves pour établir le poids des preuves
Etape 4. Expression des conclusions sur le poids des preuves

La planification de l'évaluation (étape 1) a pour objectif de définir le périmètre de l'évaluation du risque, de décrire les enjeux et d'identifier les méthodes à mettre en œuvre, incluant celles pour l'évaluation du poids des preuves. Cette première étape comporte trois sous-étapes opérationnelles :

- **Cadrage** : définition du contexte de conduite de l'étude notamment en termes de ressources et de compétences
- **Formulation de la ou des question(s) à évaluer**
- **Développement de la méthode d'évaluation**

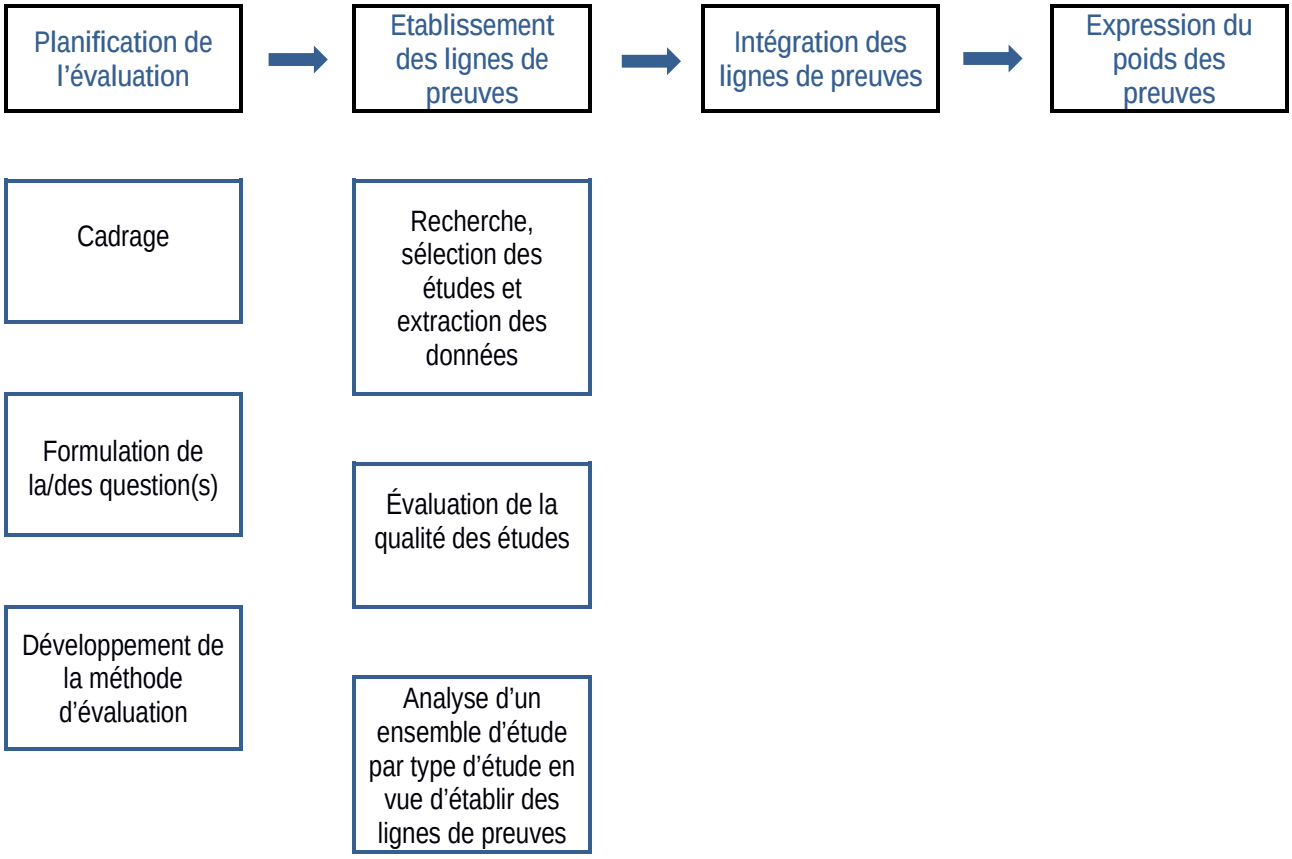
L'établissement des lignes de preuve (étape 2) comporte également trois sous-étapes opérationnelles :

- **Recherche, sélection des études (individuelles ou de synthèse) et extraction des données** : la recherche a pour objectif de récupérer l'ensemble des études pertinentes pour la question posée.
- **Évaluation de la qualité des études (individuelles ou de synthèse) :** cette sous-étape vise à évaluer la qualité des études sélectionnées en utilisant une série de critères explicites en fonction des objectifs, comme par exemple le protocole d'échantillonnage, la présence de biais méthodologiques ou le traitement statistique.
- **Analyse d'un ensemble d'études par type d'étude en vue d'établir des lignes de preuve :** il s'agit ici d'intégrer les résultats de plusieurs études d'un même type (ex : études expérimentales *in vitro*, études épidémiologiques, études toxicologiques) dans le but d'estimer une quantité d'intérêt (ex : taux d'incidence d'une pathologie au sein d'une population, valeur de référence) ou de tester une hypothèse (ex : existence d'une relation causale).

L'intégration des lignes de preuves pour établir le poids des preuves (étape 3) a pour objectif de synthétiser les lignes de preuves disponibles dans le but de déterminer le niveau de validité d'hypothèses ou d'estimer des quantités d'intérêt.

L'étape d'**expression des conclusions sur le poids des preuves (étape 4)** permet d'exprimer le poids des preuves sous forme d'un résultat clair et explicite pour l'aide à la prise de décision.

Figure 2 : Démarche d'évaluation du poids des preuves



3.2.3 Place de la revue systématique dans l'évaluation du poids des preuves

La revue systématique s'inscrit dans les deux premières étapes de la démarche d'évaluation du poids des preuves, c'est à dire la planification de l'évaluation et l'établissement des lignes de preuve. Ce recouvrement d'étapes produit parfois une certaine confusion entre « revue systématique » et « évaluation du poids des preuves ». En général, ces deux concepts sont distincts. Cependant, lorsque la planification de l'évaluation se réduit à une unique question, dont les recherches conduisent à des études d'un même type (épidémiologique, par exemple), les deux concepts sont similaires.

Initialement, les principes de la revue systématique ont été établis dans la recherche médicale, pour la prise de décision médicale (médecine factuelle ou "evidence based medicine"). La collaboration Cochrane, une association à but non lucratif qui a pour but de regrouper des données scientifiquement validées de manière accessible et résumée, a développé une méthode pour effectuer des revues systématiques d'essais randomisés contrôlés d'intervention en santé qui est présentée dans "Cochrane Handbook for Systematic Reviews of Interventions", mis à jour régulièrement (Higgins et Green 2011). La dernière version, V5.1.0, a été mise à jour en mars 2011 et est consultable en ligne. En 2010, l'EFSA a publié un guide spécifique à ses champs d'application (EFSA 2010). En 2015, l'OHAT a publié le "Handbook for Conducting a Litterature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration" (OHAT 2015).

La recherche bibliographique a permis de recenser dix définitions de la revue systématique (Annexe 5). La revue systématique est définie par l'EFSA (2010) comme : "an overview of existing evidence pertinent to a clearly formulated question, which uses pre-specified and standardized methods to identify and critically appraise relevant research, and to collect, report and analyse data from the studies that are included in the

review"[7]. Le GT MER a décidé de ne pas inclure cette dernière étape (l'analyse de données) dans sa définition car, dans ce rapport, il s'agit d'une étape d'analyse statistique quantitative qui n'est réalisée que dans le cadre d'une méta-analyse. La définition suivante est proposée :

> **« Une revue systématique de la littérature scientifique consiste à assembler, évaluer et synthétiser de manière exhaustive toutes les études pertinentes, parfois contradictoire, qui abordent une question précise. Une revue systématique est basée sur la rédaction d'un protocole détaillé au préalable favorisant la transparence de la démarche et sa reproductibilité. »**

D'après l'EFSA (2010), les éléments importants d'une revue de la littérature (systématique ou non) sont les suivants :

- Formulation de la question
- Développement de la méthode de travail
- Recherche, sélection des études, extraction des données
- Évaluation de la qualité des études
- Rédaction d'une synthèse

Les deux premiers éléments font partie de l'étape « planification de l'évaluation », les trois derniers de l'étape « établissement des lignes de preuves ».

La revue systématique se distingue de la revue narrative qui ne dispose pas de processus méthodologique systématique explicite pour collecter et analyser les articles. La revue approfondie de la littérature est une méthode intermédiaire entre la revue narrative et la revue systématique. Elle utilise des règles moins strictes que celle mises en place lors d'une revue systématique (EFSA 2010). A l'Anses, la mise à jour de l'expertise « Radiofréquence et santé » (2011-SA-0150) a été réalisée à partir d'une revue systématique de la littérature telle que définie ici. Dans le cadre de la saisine « Etude des liens entre facteurs de croissance, consommation de lait et de produits laitiers et cancer » (saisine 2009-SA-0261), une revue approfondie de la littérature a été réalisée. Elle ne pouvait être qualifiée de revue systématique car certains critères n'étaient pas remplis (exemple : un seul lecteur par article).

La planification d'une revue systématique nécessite : l'élaboration d'un protocole, la mise en place d'une équipe d'évaluation multidisciplinaire, et la spécification du calendrier et du budget. Le protocole rappelle la question et l'objectif de la revue et mentionne les critères d'inclusion ou d'exclusion des études. Il décrit également les méthodes de recherche et de sélection des études, de collecte des données, d'évaluation de leur qualité méthodologique et enfin de production des synthèses. Une description précise du protocole réduit le risque de biais, de subjectivité, limite les critiques et augmente la reproductibilité.

Les facteurs présentés ci-dessous doivent être pris en compte avant de décider la mise en œuvre d'une revue systématique :

- Impact potentiel du résultat de la revue systématique
- Quantité et qualité des données *a priori*
- Source et confidentialité des données *a priori*
- Transparence nécessaire et controverse du sujet de la saisine

[7]Proposition de traduction de la définition de la revue systématique par l'EFSA (2010) : une vue d'ensemble des éléments de preuve existants pertinents pour une question clairement formulée, qui utilise des méthodes pré-spécifiées et normalisées pour identifier et évaluer de façon critique des recherches pertinentes, et de collecter, rapporter et analyser les données provenant des études qui sont inclus dans la revue.

- Ressources à mettre en œuvre

Il est parfois difficile d'avoir une vision précise de la quantité et qualité des données disponibles *a priori*, ainsi que du niveau de confidentialité de ces données. Une revue de la littérature préliminaire (sans collecte d'information) et la consultation d'experts du domaine concerné peuvent permettre d'évaluer approximativement les informations disponibles. Des outils de *text-mining* peuvent aussi être mobilisés (OHAT 2015).

3.3 Analyse critique des démarches et méthodes recensées dans la littérature scientifique

Le GT MER a recensé 25 méthodes adressant une ou plusieurs étapes de l'évaluation du poids des preuves (Tableau 4). Certaines de ces méthodes (par exemple, la méthode proposée par l'OHAT) concernent l'ensemble des étapes du processus d'évaluation du poids des preuves.

Les méthodes recensées ont été développées principalement pour l'évaluation du poids des preuves à l'étape d'identification du danger. Cependant, certaines abordent également l'évaluation du poids des preuves dans d'autres étapes de l'évaluation des risques (Hope et Clarkson 2014, NRC 2014, OHAT 2015, Rhomberg et al. 2013, SCENIHR 2012). L'évaluation de la qualité des études pour sélectionner des données d'expositions n'est pas abordée dans ce document : par exemple Vlaanderen et al. (2008) pour les études épidémiologiques.

Les méthodes recensées couvrent plusieurs domaines de l'Anses, mais concernent majoritairement le domaine santé-environnement (Tableau 5). Le descriptif des méthodes est présenté en Annexe 7.

Les méthodes ont été qualifiées et comparées entre elles au moyen de trois critères :

- Le **caractère directif** : degré de formalisation de la méthode rendant l'expertise transparente et reproductible.
- La **pertinence :** degré de cohérence de la méthode en regard de l'objectif attendu. La pertinence est à positionner dans le contexte de la planification de l'évaluation (au regard de la question posée).
- La **faisabilité** : degré de mobilisation de temps, de ressources matérielles et humaines, ainsi que de compétence méthodologique spécifique (modélisation, statistique, etc.). Lorsque la méthode nécessite une compétence particulière, celle-ci est précisée.

Chaque critère a été noté entre 1 (plus mauvaise note) et 4 (meilleure note). Les méthodes directives, pertinentes et faisables ont une note élevée sur les trois critères. Les méthodes peu directives, peu pertinentes, peu faisables ont des notes faibles sur les trois critères. Seules les principales conclusions de l'analyse critique des méthodes sont présentées ci-dessous.

Les méthodes recensées sont présentées ci-après (Tableaux 4-5), pour chaque étape d'évaluation du poids des preuves. La plupart de ces méthodes sont qualitatives, avec ou sans notation. On appelle notation le résultat d'un processus explicite permettant d'attribuer un niveau.

Tableau 4 : Étapes du processus d'évaluation du PDP couvertes par les méthodes recensées dans la revue de littérature

		Etablissement des lignes de preuve				
MÉTHODE	Planification de l'évaluation	Recherche, sélection des études et extraction des données	Évaluation de la qualité des études	Analyse d'un ensemble d'études par type d'étude en vue d'établir des lignes de preuve	Intégration des lignes de preuves	Expression des conclusions sur le poids des preuves
(R-)AMSTAR			X			
Analyse multicritères			X	X	X	X
Arbre de décision					X	
B. Hill initial				X	X	
B. Hill pondéré				X	X	
CIRC			X	X	X	X
Epid-Tox			X	X	X	X
Évaluation comparative du PDP					X	X
Évaluation du PDP fondée sur les hypothèses					X	X
FDA (2009)		X	X	X		
GRADE	X		X	X		X
Hope et Clarkson	X		X	X	X	X
ILSI 2010			X	X		
Klimisch			X			
Méta-analyse				X		
Méthode bayésienne					X	X
Navigation Guide		X	X	X	X	
NRC (2014)	X					X
OHAT	X	X	X	X	X	X
PRISMA			X			
RS-Cochrane	X	X	X			X
RS-EFSA	X	X				X
SCENIHR			X	X	X	X
STROBE			X			
WCRF			X	X	X	

Tableau 5 : Domaines couverts par les méthodes recensées dans la revue de littérature

MÉTHODE	Santé-Travail	Alimentation et nutrition humaine			Alimentation et santé animale	Santé et protection du végétal	Santé-environnement	Produits phytosanitaires biocides et fertilisants	Médical	Ecologie-environnement
		Microbiologie des aliments	Chimie des aliments	Nutrition						
(R-)AMSTAR									X	
Analyse multicritères	X						X			X
Arbre de décision		X		X			X			X
B. Hill initial	X			X			X		X	
B. Hill pondéré	X		X	X			X			
CIRC	X		X	X			X	X	X	
Epid-Tox			X				X		X	
Évaluation comparative du PDP							X			
Évaluation du PDP fondée sur les hypothèses							X			
FDA (2009)			X	X						
GRADE				X			X		X	
Hope et Clarkson										X
ILSI 2010				X						
Klimisch			X				X	X		
Méta-analyse				X			X		X	
Méthode bayésienne							X	X		
Navigation Guide							X			
NRC (2014)			X				X			
OHAT							X			
PRISMA	X			X			X			
RS-Cochrane									X	
RS-EFSA		X	X	X	X					
SCENIHR							X			
STROBE				X			X		X	
WCRF			X	X						

3.3.1 Planification de l'évaluation

Hope et Clarkson (2014), NRC (2014) et OHAT (2015) inscrivent la planification de l'évaluation au sein de leur processus de conduite d'expertise (Tableau 3) et appréhendent différemment les trois sous-étapes opérationnelles. D'autres études dans la littérature ne font pas cette différenciation en sous-étapes (Meek et al, 2014, cf. fiche méthode 3.10 évaluation comparative du poids des preuves, annexe 7). L'étape de planification de l'évaluation peut s'appuyer sur une recherche bibliographique préliminaire et sur des auditions. Plusieurs documents traitant de la revue systématique proposent également des méthodes pour la planification de l'évaluation (EFSA 2010, Higgins et Green 2011).

- **Cadrage**

Pour Hope et Clarkson (2014), l'objectif du cadrage est de demander aux gestionnaires de risque, décideurs et parties prenantes de définir les objectifs de gestion écologique en des termes neutres, précis et mesurables. Pour le NRC (2014), l'objectif est de comprendre les besoins du client et des offices régionaux concernant l'évaluation de produits ou de processus chimiques. Enfin, pour l'OHAT (2015), l'objectif est de sélectionner les participants, d'évaluer l'impact de conduite d'une évaluation et d'identifier les activités en cours et connexes de l'étude à conduire.

- **Formulation de la question scientifique**

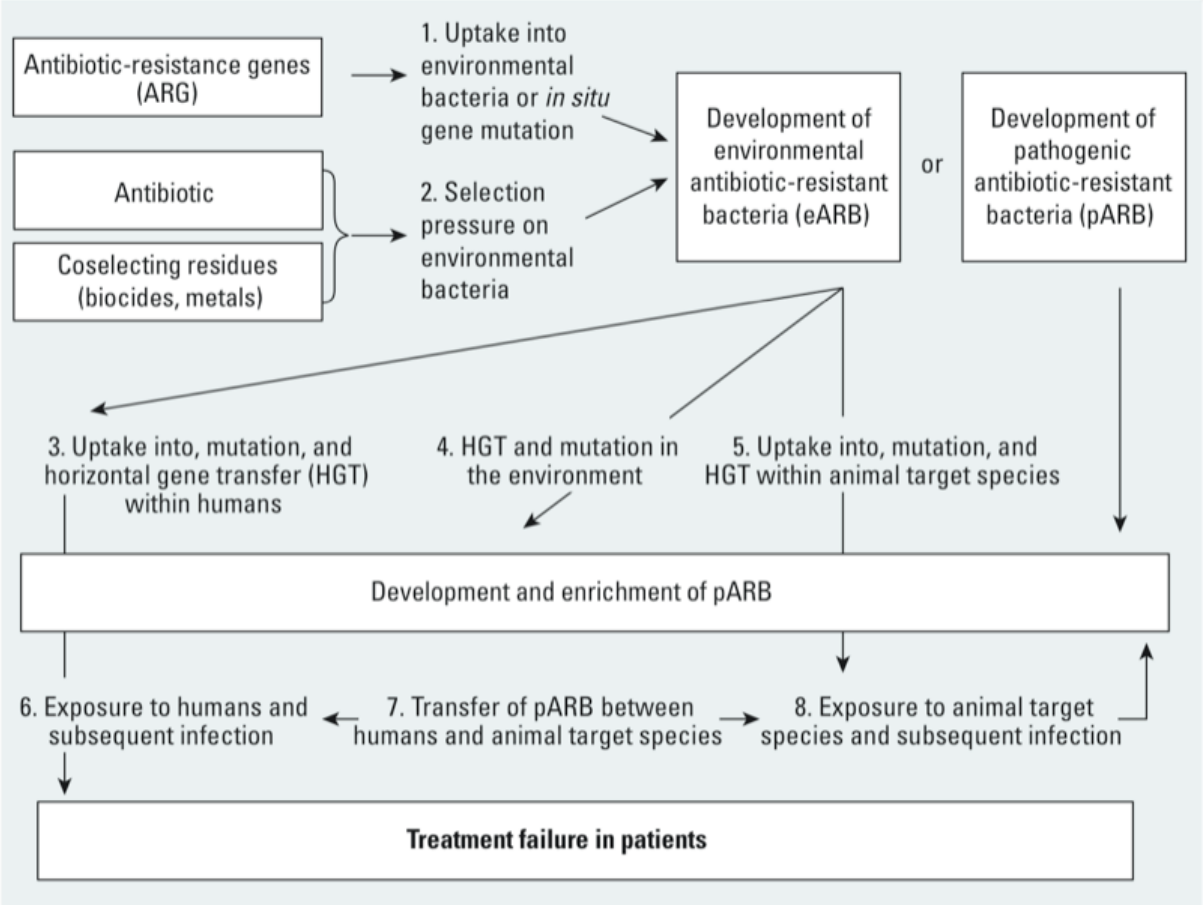
Pour Hope et Clarkson (2014), cette sous-étape consiste à exprimer chaque question sous forme de modèle conceptuel. Utilisé historiquement dans l'évaluation des risques écologiques, mais à présent proposé dans le domaine de l'évaluation des risques sanitaires, le modèle conceptuel consiste en un diagramme dans lequel sont identifiés les agents d'intérêt, les populations exposées, les effets sanitaires qui vont être abordés et les relations entre ces différents éléments décrivant les hypothèses concernant l'effet de l'exposition à un agent (chimiques, microbiologiques, physiques) sur les paramètres de sortie de l'évaluation (*assessment endpoint)* pour différentes populations exposées (US EPA 1998, 2014).

Dans les domaines de compétence de l'Anses, des modèles conceptuels ont été proposés (appelés également « schéma conceptuel » ou « schéma évènementiel ») pour formuler la question scientifique.

Figure 3 présente un exemple de modèle conceptuel pour une évaluation des risques sanitaires examinant le rôle de l'environnement dans l'échec des traitements antibiotiques causé par des pathogènes antibio-résistants (Ashbolt et al. 2013).

Un autre exemple de modèle conceptuel est présenté en Annexe 8.

Figure 3 : Modèle conceptuel décrivant les voies environnementales qui résultent en un risque accru d'infection humaine et animale par des bactéries résistantes aux antibiotiques (Ashbolt et al. 2013)



Pour le NRC (2014), la question est formulée à partir d'un tableau renseignant les effets associés à la substance chimique (ou mixture) à étudier, par type d'études (in vivo, in vitro, etc.). Pour l'OHAT (2015), la formulation de la question consiste à remplir la structure d'information PECO (Population, Exposition, Comparateur, et Effet). La structure PECO est issue de PICO (Patient, Problème ou Population ; Intervention ; Comparaison, Contrôle ou Comparateur ; issues ou effets), initialement utilisée pour répondre à une question clinique. En outre, le NRC et l'OHAT proposent tous les deux de collaborer avec un spécialiste de conduite de revue systématique pour cette sous-étape.

Les autres documents qui s'intéressent à la formulation de la question demandent de remplir une structure d'information adaptée à leur situation d'étude. C'est le cas de la Collaboration Cochrane (Higgins et Green 2011) et Bilotta et al. (2014) qui ont adaptés PICO à leur besoin, ou de Rooney et al. (2014) qui a développé PECOTS (ajoutant ainsi les notions de temps T et de localisation S). Par ailleurs, outre l'utilisation de PICO, GRADE propose une méthode de classification des effets (l'O de PICO) par ordre d'importance, utile *in fine* pour l'établissement des recommandations dans la conclusion (Guyatt, Oxman, Kunz, Atkins, et al. 2011). Outre PICO et PECO, EFSA (2010) recommande d'autres structures spécifiques pour l'évaluation de la précision d'un test et la quantification d'une situation d'intérêt (prévalence par exemple). EFSA (2010) propose également une méthode de remplissage des structures à partir de la littérature.

- **Plan de développement de la méthode d'évaluation**

Pour Hope et Clarkson (2014), le plan de développement a pour objet de recenser les méthodes de traitement des lignes de preuve. Le NRC (2014), présente un protocole rassemblant les éléments à fournir pour conduire une revue systématique (critères d'inclusion et d'exclusion, les types d'études, d'exposition et d'effets, méthodes de collection de données, etc.) comme indiqué sur la

Figure 4. Enfin, pour OHAT (2015) c'est un plan détaillé de l'analyse qui rappelle le cadrage de l'évaluation, présente la structure PECO et demande de fournir le descriptif des toutes les méthodes utilisées dans le processus d'analyse, depuis la localisation des preuves jusqu'à la formulation des conclusions. De la même façon que pour l'OHAT, EFSA (2010) propose de répondre à une série de question, rassemblées dans un document, pour définir la stratégie de la revue bibliographique (revue systématique ou revue approfondie de la littérature).

Figure 4 : Eléments de protocole de la revue systématique (NRC 2014)

BOX 3-1 Systematic-Review Protocol Elements

A. Systematic review question (for example, is benzo[a]pyrene exposure of adult animals associated with neurotoxic effects?)
B. Methods
   1. Inclusion and exclusion criteria for studies:
      a. Types of studies or participants (for example, experimental animal, observational human, or in vitro mechanistic).
      b. Types of exposures (for example, oral or inhalation).
      c. Types of outcome (for example, neurotoxic or developmental).
   2. Search methods for identification of studies.
   3. Assessment of risk of bias and other methodologic characteristics of included studies.
   4. Data-collection methods.
   5. Analysis.

Toutes les méthodes ont été jugées très directives (tableau 6), à l'exception de celle proposée par Hope et Clarkson (2014). Pour l'étape de planification de l'évaluation, les méthodes recensées sont plutôt pertinentes et plutôt faisables (EFSA 2010, Guyatt, Oxman, Kunz, Atkins, et al. 2011, Higgins et Green 2011, NRC 2014, OHAT 2015). Les méthodes proposant des plans de rédaction pour une des trois sous-étapes (EFSA 2010, OHAT 2015) sont jugées plutôt pertinentes dans la mesure où elles favorisent la transparence de l'étape.

Tableau 6 : Notes comparatives des méthodes pour l'étape de planification de l'évaluation

	Planification de l'évaluation		
MÉTHODE	Caractère directif	Pertinence	Faisabilité
GRADE	4	3	3
Hope et Clarkson (2014)	2	2	3
NRC (2014)	4	3	3
OHAT (2015)	4	3	3
Revue systématique – Cochrane (2011)	4	3	3
Revue systématique – EFSA (2010)	4	3	3

3.3.2 Établissement des lignes de preuves

Les notes des méthodes pour l'étape de l'établissement des lignes de preuves sont présentées dans le tableau 7.

- **Recherche, sélection des études (individuelles ou de synthèse) et l'extraction des données**

La littérature, relative à la recherche et la sélection des études, porte principalement sur la revue systématique. Les principes partagés par la Collaboration Cochrane (Higgins et Green 2011), l'EFSA (2010) et l'OHAT (2015) sont : l'utilisation d'au moins deux bases de données, la sélection des études par deux personnes indépendantes, et la définition des critères de sélection en amont. L'extraction des données se fait au moyen d'une grille dont le format est définit en préalable de la revue.

Le GT MER a jugé ces trois méthodes plutôt directives, plutôt pertinentes, mais peu faisables (car plutôt coûteux en ressources humaines).

Tableau 7: Notes comparatives des méthodes pour l'étape de l'établissement des lignes de preuves

	Etablissement des lignes de preuve								
	Sélection des études individuelles et de synthèse et extraction des données			Analyse de la qualité des études individuelles et des études de synthèse			Analyse d'un ensemble d'études		
MÉTHODE	**D[1]**	**P[2]**	**F[3]**	**D[1]**	**P[2]**	**F[3]**	**D[1]**	**P[2]**	**F[3]**
(R-)AMSTAR				4	3	4			
Analyse multicritères				2	4	3	2	4	3
Bradford Hill initial							2	4	4
Bradford Hill pondéré							3	3	3
CIRC				2	4	4	2	3	4
Epid-Tox				2	4	4	2	4	3
FDA (2009)				3	4	4	2	3	3
GRADE				4	3	3	2	3	4
Hope et Clarkson (2014)				2	3	3	2	3	3
ILSI 2010							3	2	3
Klimisch				2	3	4			
Méta-analyse							4	4	1
Navigation Guide	1	3	2	1	3	4	1	3	3
OHAT	3	3	2	3	3	4	2	3	3
PRISMA				4	1	4			
RS – Cochrane	3	3	2	2	4	4			
RS – EFSA	3	3	2						
SCENIHR				2	3	4	1	3	4
STROBE				4	1	4			
WCRF (+ métaanalyse)				2	4	4	4	4	2

[1]Caractère DIRECTIF ; [2]PERTINENCE ; [3]FAISABILITÉ

- **Évaluation de la qualité des études (individuelles ou de synthèse)**

L'évaluation de la qualité des études individuelles

D'une manière générale, la qualité d'une étude s'évalue en analysant les biais méthodologiques potentiels, tels que les biais d'information, de sélection ou de confusion, ou en évaluant le caractère indirect des données scientifiques[8] ou l'imprécision des données. Certains documents recommandent que d'autres critères, comme les biais de publication ou les liens d'intérêt, fassent partie de l'évaluation de la qualité des études. Deux types de méthodes pour évaluer la qualité d'une étude se distinguent dans la littérature : celles sans notation et celles utilisant une notation.

Méthodes sans notation

Des méthodes sans notation ont été proposées par le CIRC (IARC 2006), le WCRF (WCRF/AICR 2014), la collaboration Cochrane (Higgins et Green 2011) et par la FDA pour l'évaluation des allégations santé dans le domaine de l'alimentation (FDA 2009). Ces méthodes proposent des listes de critères relevant des bonnes pratiques de recherche dans chaque domaine (épidémiologie, toxicologie, etc.). Par exemple, la méthode décrite par le WCRF pour juger de la qualité des études épidémiologiques individuelles intègre notamment les concepts méthodologiques clés en épidémiologie, comme la gradation de la qualité des études en fonction de leur design (étude écologique < étude individuelle transversale < étude individuelle prospective < essai contrôlé randomisé). Adami et al (2011) proposent, dans le cadre de la méthode Epid-Tox d'évaluer les études toxicologiques avec les critères de l'EPA (2001) et d'évaluer les études épidémiologiques avec les critères de l'ECETOC (2009). Ces auteurs proposent de classer les études en trois catégories : « acceptable » (limites minimes, à retenir pour l'évaluation du poids des preuves) ; « supplémentaires » (limites modérées, à retenir pour l'évaluation du poids des preuves) et « inacceptables » (limites sévères, à exclure de l'évaluation du poids des preuves).

Toutes ces méthodes ont été jugées par le GT MER peu directives, très pertinentes et très faisables. Compte-tenu de leur caractère peu directif, les résultats de ces méthodes sont susceptibles d'être expert-dépendant, et donc d'inclure une part de subjectivité. Par ailleurs, du point de vue méthodologique, l'absence de notation peut conduire à l'exclusion des études jugées de mauvaise qualité. Dans certains cas, il peut être préférable de les conserver tout en leur donnant un poids relativement faible. Une telle démarche n'est pas toujours facile à mettre en œuvre.

Méthodes avec notation

Les méthodes d'analyse multicritères (Linkov et al. 2011, Linkov et al. 2009), de Hope et Clarkson (2014), GRADE (Balshem et al. 2011), OHAT (2015) et la méthode Klimisch (Klimisch et al. 1997) permettent d'attribuer des scores aux études individuelles en tenant compte de leur qualité, même si cette dernière ne propose qu'un critère binaire. Les outils proposés par GRADE et par l'OHAT proposent de qualifier la qualité des études sur une échelle qualitative et listent une série de questions dans ce but. Par exemple, l'outil de risque de biais de l'OHAT comporte 11 questions relevant des bonnes pratiques de recherche à considérer selon le type d'étude. Pour chaque question, une réponse est à évaluer en termes de risque de biais (bas, probablement bas, probablement élevé, élevé).

Les méthodes proposées par Linkov et al. (2011) et par Hope et Clarkson (2014) permettent de coter les études sur une échelle numérique selon des critères d'évaluations spécifiés. Ces méthodes permettent d'attribuer des scores aux études individuelles, et donc de donner un poids faible aux études de qualité inférieure. Aucune de ces méthodes ne propose une valeur seuil qui permettrait d'exclure certaines études.

[8] Selon HAS (2013), le caractère indirect des données scientifiques correspond à des données scientifiques obtenues par des comparaisons indirectes ou à des différences entre la population, l'intervention, l'intervention de comparaison, les résultats d'intérêt et ceux des études sélectionnées pour la question donnée.

Du fait de critères moins précis que pour les approches de GRADE et de l'OHAT, les méthodes de Klimisch (1997), de Hope et Clarkson (2014) et d'analyse multicritère (Linkov et al. 2011) ont été jugées moins directives. Toutes les méthodes ont été jugées plutôt pertinentes et plutôt faisables par le GT MER.

L'évaluation de la qualité des études de synthèse

Une seule méthode est disponible pour évaluer la qualité des études de synthèse. Il s'agit de la méthode AMSTAR - et sa version révisée R-AMSTAR - (Pieper et al. 2015, Shea, Grimshaw, et al. 2007, Shea et al. 2009, Kung et al. 2010). Le GT MER les juge très directives, plutôt pertinentes et très faisables.

- **Évaluation d'un ensemble d'études par type d'étude en vue d'établir des lignes de preuve**

Plusieurs approches sont présentées dans la littérature scientifique et dans des documents guides :

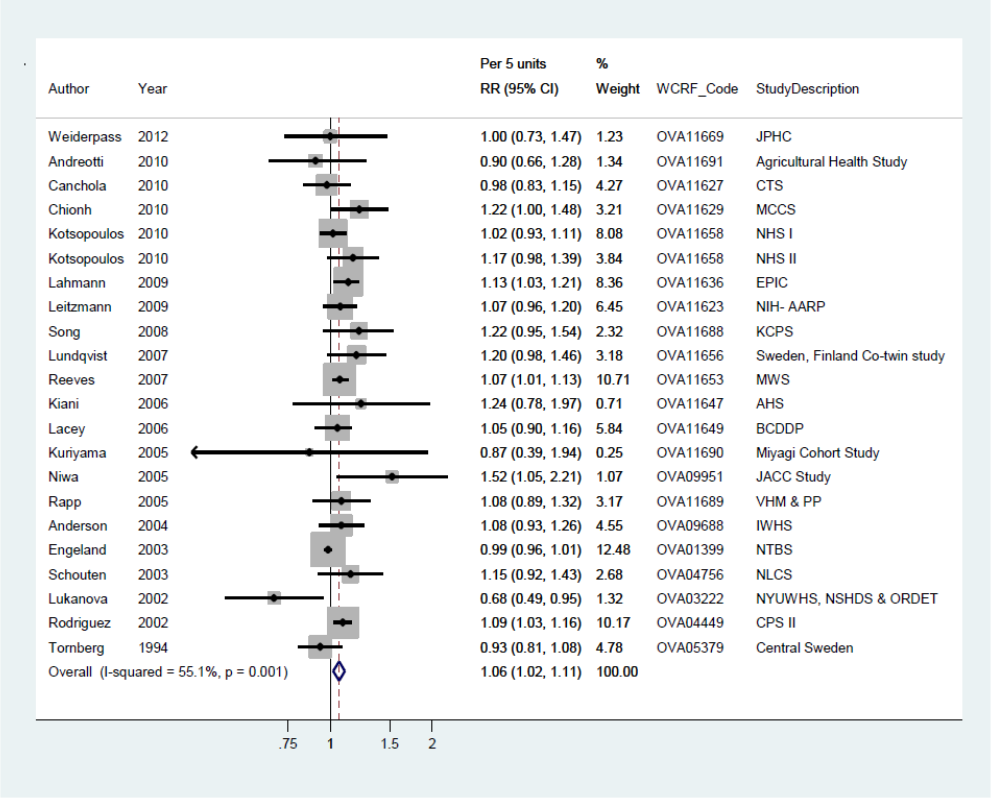
- Méta-analyse,
- Analyse multicritères,
- Méthodes qualitatives avec ou sans notation.

Il existe également des méthodes hybrides, utilisant la méta-analyse pour certains types d'études et des approches qualitatives pour d'autres, ou des méthodes proposant de pondérer les critères de Bradford Hill.

Méta-analyse

La méta-analyse est une démarche statistique combinant les résultats d'une série d'études indépendantes sur un problème donné. Plusieurs articles et documents guides (EFSA 2010, Murad et al. 2014) soulignent l'intérêt de réaliser des méta-analyses pour évaluer un ensemble d'études et établir des lignes de preuves. Le WCRF a récemment conduit une méta-analyse basée sur une revue systématique de la littérature (données publiées) sur la relation entre indice de masse corporelle (IMC) et risque de cancer de l'ovaire (WCRF/AICR 2014). Cette méta-analyse (dont le *forest plot* est présenté Figure 5) inclue 22 études et conclue à une augmentation significative du risque de 6%, pour chaque augmentation de 5 unités d'IMC.

Figure 5 : *Forest plot* – Méta-analyse dose-réponse du WCRF sur la relation entre indice de masse corporelle (IMC) et risque de cancer de l'ovaire (WCRF/AICR 2014)



La méta-analyse et la méthode du WCRF basée en partie sur la méta-analyse (WCRF/AICR 2014) ont reçu des notes élevées pour les critères « Caractère directif » et « Pertinence », mais une note faible pour le critère « Faisabilité ». Les notes élevées attribuées aux critères « Caractère directif » et « Pertinence » sont dues au fait que la méta-analyse est une procédure transparente et répétable qui permet de synthétiser de manière quantitative des études de même nature pour estimer des quantités d'intérêt et tester des hypothèses.

Par ailleurs, la réalisation d'une méta-analyse est coûteuse en temps car, outre les nombreuses analyses statistiques devant être réalisées, elle inclut, comme dans une revue systématique, une étape de sélection des études disponibles sur un sujet donné ainsi qu'une étape d'extraction des données expérimentales publiées dans les études sélectionnées. Il n'est par ailleurs pas recommandé de réaliser de méta-analyse lorsque les données sont peu nombreuses ou lorsqu'elles sont issues d'études trop dissemblables. Dans ce cadre, les résultats de la méta-analyse doivent être interprétés avec prudence. Ces limites expliquent la note faible attribuée au critère « Faisabilité ».

Analyse multicritères

Linkov et al. (2009, 2011) proposent une méthode d'analyse multicritères, appliquée au cas d'une analyse des risques écologiques (sélection de différentes options de restauration d'un site aquatique contaminé). L'analyse d'un ensemble d'études en vue d'établir des lignes de preuve est obtenue en calculant une moyenne pondérée des cotations (un poids est associé à chaque critère). La valeur de chaque poids est déterminée par les experts. Le tableau 8 présente un exemple de notes par critères pour trois lignes de preuves (contamination / toxicité / altération). Dans cet exemple, les poids (appelés « importance ») attribués à chaque critère sont identiques.

Tableau 8 : Exemple de notation de lignes de preuve à partir de scores obtenus pour chaque critère et de leurs poids associés (Linkov et al. 2011)

Table II. SQT Studies Assessed by Scientific Merit

	Importance	Contamination	Toxicity	Alteration
Soundness	0.2	8	6	3
Applicability and utility	0.2	8	5	3
Clarity and completeness	0.2	6	6	3
Uncertainty and variability	0.2	4	2	2
Evaluation and review	0.2	5	5	1
Weight of each line of evidence		6.2	4.8	2.4

La méthode d'analyse multicritères proposée par Linkov et al. (2011) a reçu une note élevée pour le critère « Pertinence » et « Faisabilité » mais une note faible pour le critère « Caractère directif ». La mise en œuvre de cette méthode est en principe assez simple mais nécessite la mobilisation d'experts pour définir les critères et leur pondération relative. Cette sensibilité à l'expertise a conduit le GT MER à juger cette méthode peu directive.

Approches qualitatives avec ou sans notation

Plusieurs approches qualitatives ont été proposées dans la littérature scientifique et dans des documents guides d'agences sanitaires pour analyser un ensemble d'études (Adami et al. 2011, FDA 2009, Guyatt et al. , Hill 1965, Hope et Clarkson 2014, IARC 2006, OHAT 2015, SCENIHR 2012, van Bilsen et al. 2011, INCa 2015). La plupart d'entre-elles attribuent des notes correspondant à un niveau de preuve, de confiance, d'utilité ou de « consistance » selon des échelles dont le nombre de niveaux est variable. Les critères de Bradford Hill sont mobilisés dans ces approches pour la ligne de preuves correspondant aux études épidémiologiques. La méthode Epid-tox (Adami et al. 2011) propose, pour établir la ligne de preuves correspondant aux études toxicologiques, de passer par l'analyse du mode d'action au travers d'un arbre de

décision organisé autour de trois questions[9]. Ces approches ont reçu des notes 3 et 4 pour les critères « Pertinence » et « Faisabilité », sauf GRADE qui a reçu une note de 2 pour le critère « Pertinence ». Pour ces méthodes, l'attribution de notes faibles au critère « Caractère directif », 1 ou 2, repose sur la forte dépendance des résultats, obtenus avec ces méthodes, aux experts mobilisés (caractère subjectif du choix des scores). Ce faisant, le degré de répétabilité de ces méthodes est faible.

Swaen et van Amelssvoort (2009) pondèrent les critères de Bradford Hill afin de rendre la démarche plus systématique et transparente.

3.3.3 L'intégration des lignes de preuves pour établir le poids des preuves

Les notations des méthodes pour l'intégration des lignes de preuves afin d'établir le poids des preuves sont présentées dans le Tableau 9.

Trois catégories de méthodes ont été identifiées pour agréger des lignes de preuve :

- Modélisation statistique (exemple : modèle bayésien)
- Méthodes qualitatives avec ou sans notation (exemple : CIRC) et méthodes semi-quantitatives (exemple : évaluation comparative du poids des preuves)
- Représentation de l'expertise avec des arbres de décision ou par analyse multicritères

Tableau 9 : Notes comparatives des méthodes pour l'étape de l'intégration des lignes de preuves

	Etablissement du poids des preuves		
MÉTHODE	**Caractère directif**	**Pertinence**	**Faisabilité**
Analyse multicritères	2	4	3
Arbre de décision	1	3	3
Bradford Hill initial	2	4	4
Bradford Hill pondéré	3	4	4
CIRC	3	3	4
Epid-Tox	2	4	3
Évaluation comparative du PDP	3	3	3
Évaluation du PDP fondée sur les hypothèses	2	3	3
Hope et Clarkson (2014)	3	3	3
Méthode bayésienne	3	4	2
Navigation Guide	1	3	3
OHAT	3	3	4
SCENIHR	2	3	4
WCRF 2014	3	3	4 (si résultats méta-analyses disponibles)

Modélisation statistique

La modélisation statistique est mentionnée dans la littérature (Gosling et al. 2013, Schleier Iii et al. 2015). Elle peut être appliquée à une grande diversité de problèmes et intégrer des informations de nature différente, notamment par l'utilisation de méthodes bayésiennes (données expérimentales de différentes natures, avis d'experts). Par exemple, dans le domaine des allergies cutanées en réaction à un composé chimique, Gosling et al (2013) proposent d'intégrer des connaissances d'experts avec des données d'études toxicologiques au moyen d'une méthode bayésienne. Les connaissances d'experts sont résumées par des espérances et des variances dont les valeurs sont mises à jour à l'aide des données expérimentales disponibles.
La modélisation dépasse le cadre de l'évaluation des risques et est utilisée dans de nombreux domaines. Cette approche a reçu une note de 3, 4 et 2 pour respectivement le critère « Caractère directif », « Perti-

[9] Le poids des preuves est-il suffisant pour établir un mode d'action chez l'animal ? Le mode d'action animal est-il plausible chez l'humain ? En prenant en compte les facteurs cinétiques et dynamiques, le mode d'action animal est-il plausible chez l'humain ?

nence » et « Faisabilité ». La note assez faible attribuée au critère « Faisabilité » est due à la difficulté posée par la combinaison de données expérimentales avec des avis d'experts, et aux problèmes que peut engendrer l'élicitation des connaissances des experts.

Les approches faisant appel à de la modélisation statistique requièrent une compétence en modélisation statistique.

Approches qualitatives avec ou sans notation et méthodes semi-quantitatives

Ce groupe de méthodes inclut :

- les approches qualitatives avec ou sans notation (exemples : CIRC, WCRF, OHAT, Critères de Bradford Hill, évaluation du poids des preuves fondée sur des hypothèses)
- les méthodes semi-quantitatives (exemples : évaluation comparative du poids des preuves ; Hope et Clarkson 2014).

Les méthodes de cette catégorie ont été bien notées en termes de « pertinence » et « faisabilité » (note de 3 ou 4).

La majorité de ces méthodes ont été jugées relativement directives (note de 3) pour cette étape. Celles jugées faiblement directives (note de 1 ou 2) demandent de prendre des précautions particulières lors de leur utilisation. C'est le cas par exemple des critères de Bradford Hill (Hill 1965) qui présentent un risque accru de subjectivité et de non-reproductibilité selon les experts impliqués. Dans le cadre de l'évaluation comparative du poids des preuves, les critères de Bradford Hill ont été précisés et pondérés pour tenir compte de leur importance relative ; cette importance est basée sur une analyse de l'expérience acquise (Meek, Palermo, et al. 2014). Cette formalisation rend l'approche plus directive (note de 3) et contribue à sa reproductibilité. Le Tableau 10 montre un exemple d'évaluation comparative du poids des preuves.

Les méthodes qualitatives proposées par le CIRC (IARC 2006) et le WCRF (WCRF/AICR 2014) pour l'établissement final du niveau de preuve sont établies selon des principes relativement proches. De même, la matrice proposée par l'OHAT (OHAT 2015) pour intégrer les lignes de preuves humaines et animales est proche de celle proposée par le CIRC. De manière générale, ces méthodes (CIRC, WCRF, etc.) intègrent implicitement les critères initiaux de Bradford Hill dans leur démarche, sans les formaliser spécifiquement. Ces méthodes vont cependant plus loin que la simple prise en compte de ces critères en proposant des classifications sous forme de notation permettant d'aboutir à des catégories de poids des preuves. A titre d'exemple, dans le domaine de la nutrition et du risque de cancer en prévention primaire, la méthode du WCRF/AICR (2014) a permis de statuer sur un poids des preuves (convaincant, probable, suggéré, non concluant ou improbable) pour les relations entre plus de 20 facteurs nutritionnels et plus de 15 localisations de cancer. Dans le domaine du cancer (IARC 2006), cette méthode a été appliquée pour établir un poids des preuves (1, 2A, 2B, 3, 4) concernant le caractère cancérogène de nombreux facteurs environnementaux (ex: radiations non ionisantes) ou nutritionnels (ex: alcool, viandes et charcuteries).

Tableau 10 : Exemple d'évaluation comparative du poids des preuves pour le mode d'action mutagénique du 2,3-trichloropropane (Meek et al, 2013)

Table 6. Comparative weight of evidence analysis for 1,2,3-trichloropropane: mutagenic MOA

Evolved Bradford Hill considerations		Supporting data[a]	Inconsistent data[a]	Missing data[b]
1. Biological concordance		Genotoxic MOA is well established for chemically mediated carcinogenicity		
2. Essentiality of key events		Inducers/inhibitors of metabolism alter amount of DNA binding		Evidence for adduct conversion to genetic damage
3. Concordance of empirical observation	Dose–response	Dose-related formation of DNA-reactive metabolite, DNA adduct formation, tumor formation and time to tumor.		
	Temporality	Metabolism to reactive intermediate occurs within hours of exposure, adducts appear within hours and days of exposure, and tumors first appear after ≈ 9 months.		
	Incidence			No data to assess whether adduct formation frequency different from tumor frequency.
4. Consistency		Mutagenic effects *in vitro* accompanied by limited evidence of *in vivo* mutagenicity.	Adducts occur in tissues where no neoplastic effects were reported (spleen, liver and glandular stomach). Negative results from *in vivo* genotoxicity assessments (dominant lethal and micronucleus).	
5. Analogy		Other halogenated aliphatic chemicals (1,2,-dibromoethane and 1,2-dibromo-3-chloropropane) are mutagenic carcinogens. Other genotoxic chemicals are multisite and multispecies carcinogens.		

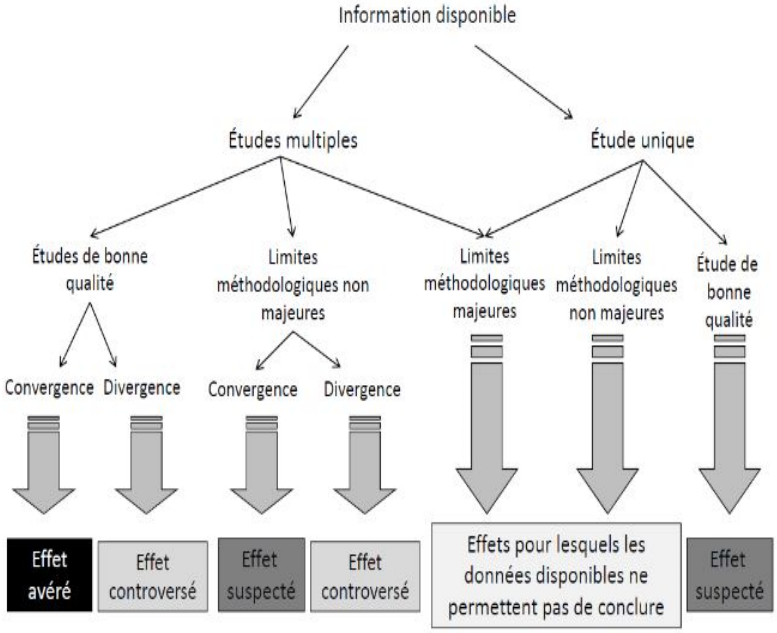
MOA, mode of action.
[a]All conclusions in the above tables were extracted from the original US EPA toxicology review on 1,2,3-trichloropropane (USEPA, 2009).
[b]The IRIS assessment did not comment on missing data; the information here represents the authors' views.

Représentation de l'expertise avec des arbres de décision ou par analyse multicritères

Un arbre de décision représente graphiquement un ensemble de règles de classification organisées de manière arborescente. Les sorties d'un arbre de décision peuvent être quantitatives ou qualitatives. Les règles de classification peuvent être basées sur des informations diverses, notamment sur des études expérimentales, des observations, des sorties de modèles ou sur des avis d'experts. La

Figure 6 présente un exemple d'arbre de décision relatif à l'évaluation des risques liés à l'exposition au bisphénol A.

Figure 6 : Arbre de décision utilisé dans le cadre de la saisine relative à l'évaluation des risques du bisphénol A (BPA) pour la santé humaine (Anses 2013a)



La méthode d'analyse multicritères de Linkov et al. (2011) permet de combiner tout type de données, qu'elles soient qualitatives ou quantitatives. Le caractère générique de cette méthode permet de la mobiliser pour évaluer tous types d'études épidémiologiques et expérimentales. Cette méthode nécessite de réaliser plusieurs choix à dire d'experts, notamment le choix des critères d'évaluation et le choix des poids associés à chaque critère.

Cette méthode et la représentation par arbre de décision sont deux approches ayant reçu des notes élevées pour les critères « Faisabilité » et « Pertinence » (3 à 4). L'attribution de notes faibles pour le critère « Caractère directif » (1 à 2) provient de l'absence de règles d'évaluation précises.

3.3.4 L'expression des conclusions concernant le poids des preuves

Dans la majorité des documents étudiés, le poids des preuves est exprimé à travers une classification dont le nombre de classes et leurs intitulés varient selon les auteurs (Tableau 11). Jeffreys (1961) propose six classes, définies par le logarithme en base 10 du rapport de vraisemblance (cf. paragraphe 3.1.1 du rapport). OHAT (2015) définit cinq classes pour évaluer les interventions en Santé Humaine. Pour GRADE, les quatre classes définies résultent de la composition de deux critères (Force et Orientation), chacun disposant de deux valeurs possibles Weak/Strong pour le premier et Against/For pour le second. Le recours à une cinquième classe permet de ne pas proposer de recommandation (Andrews, Guyatt, et al. 2013). SCENIHR (2012) utilise cinq classes de même que Hope et Clarkson (2014) mais avec des extrêmes différents. Enfin, la collaboration Cochrane (Higgins et Green 2011) utilise quatre classes. La majorité des documents analysés proposent donc quatre classes pour qualifier le poids des preuves, plus une qui permet d'indiquer l'impossibilité d'évaluation. La variante proposée par GRADE, consistant à indiquer l'orientation de la recommandation (against *vs* for), offre un intérêt certain pour notifier un positionnement. Pour chacun des 5 critères de Bradford Hill retenus dans le cadre de l'évaluation comparative du poids des preuves pour le mode d'action, les catégories de poids des preuves faible, modéré et élevé sont précisées et illustrées par les exemples (Meek, Palermo, et al. 2014).

Tableau 11 : Exemples d'expression du poids des preuves

MÉTHODE	Nombre de classes	INTITULÉ DES CLASSES
Jeffreys (1961)	6	Négatif, faible, substantiel, fort, très fort, décisif
CIRC (2006)	5	Groupe 1: l'agent est cancérogène pour l'homme Groupe 2A: l'agent est probablement cancérogène pour l'homme Groupe 2B: l'agent est possiblement cancérogène pour l'homme Groupe 3: l'agent n'est pas classifiable concernant son potentiel cancérogène pour l'homme Groupe 4: l'agent n'est probablement pas cancérogène pour l'homme
Évaluation comparative du PDP[1]	3	Faible, modéré, élevé
GRADE[2]	4	Strong Against, Weak Against, Weak For, Strong For
Hope et Clarkson (2014)	5	Weak, Non précisé, Non précisé, Non précisé, Strong
OHAT (2015)	5	Non identifié dangereux pour l'humain, non classable, suspecté, présumé, avéré
RS – Cochrane[3]	4	Très faible, Faible, Modéré, Fort
SCENIHR (2012)	5	Weighting not possible, Uncertain, Weak, Moderate, Strong

[1](Meek, Palermo, et al. 2014) ; [2](Andrews, Guyatt, et al. 2013) ; [3](Higgins et Green 2011)

Différentes techniques ont été proposées pour étayer les conclusions et analyser les incertitudes associées. Ainsi, OHAT (2015) propose de représenter graphiquement les résultats intermédiaires, NRC (2014) d'effectuer une analyse d'incertitude, SCENIHR (2012) d'effectuer une analyse d'incertitude dont le résultat est ensuite exprimé en regard de classes (certitude avérée, probable, confiant, possible et incertain). Hristozov et al. (2014) propose de conduire une analyse d'incertitude sur les données et sur les jugements

des experts. Enfin, dans le cadre de l'analyse multicritères, Linkov et al. (2011) proposent de conduire une analyse de sensibilité sur les poids et sur certaines données d'entrée. L'intention visée par ces méthodes est de favoriser la transparence de la formulation de la conclusion.

Certains des documents analysés présentent également des supports permettant de structurer la conclusion. Ainsi, EFSA (2010) indiquent les thèmes à traiter dans la discussion et la conclusion. La collaboration Cochrane propose une structure du document avec des thématiques de discussion et de conclusion prédéfinis, dont le contexte d'application de la recommandation. Elle propose également l'utilisation de la structure EPICOT pour traiter des conséquences de la conclusion en matière de besoin de recherche (Higgins et Green 2011). Cette structure est le pendant de la structure PICO utilisée dans la formulation de la question, permettant aux lecteurs d'évaluer la distance entre la question formulée et le résultat obtenu. Pour structurer la discussion et la conclusion, GRADE propose de rappeler la question, la population cible, l'intervention et les ressources disponibles. Deux solutions sont proposées pour éviter les erreurs d'interprétation de la part des destinataires de la conclusion : (1) un système d'étiquetage, élaboré dans le même esprit que celui du système général harmonisé (SGH) pour harmoniser la classification et les éléments de communication du danger des produits chimiques (Nations Unies 2013), pour indiquer la classe de la conclusion, et (2) des expressions à préférer ou éviter. L'intérêt offert par ces supports est de guider la rédaction des conclusions avec le souci de couvrir l'ensemble des thématiques environnant la question.

Les méthodes qui définissent des classes – GRADE, Hope et Clarkson, OHAT, SCENIHR, évaluation comparative du poids des preuves – sont notées assez directives (3), et celles demandant en plus la rédaction d'un document dont la structure est spécifiée – la Collaboration Cochrane, EFSA – sont notées très directives (4) dans le tableau 12. La plupart des méthodes ont reçu des notes de pertinence élevées. L'évaluation comparative du poids des preuves a reçu une note de pertinence moindre du fait que la méthode s'applique principalement dans le domaine du mode d'action. Pour le critère de faisabilité, les méthodes Linkov et al. (2011), NRC (2014), du SCENIHR (2012) et l'évaluation comparative du poids des preuves (Meek, Palermo, et al. 2014) ont été jugés plutôt faisable (3). Les méthodes de l'OHAT (2015) et GRADE (Andrews, Guyatt, et al. 2013) ont reçu la note de 4 « très faisable » car, contrairement aux méthodes précédentes, elles ne nécessitent aucune compétence particulière.

Tableau 12 : Notes comparatives des méthodes pour l'étape de l'expression des conclusions sur le poids des preuves à l'étape d'identification du danger

	Expression des conclusions		
MÉTHODE	**Caractère directif**	**Pertinence**	**Faisabilité**
Analyse multicritères	3	3	3
CIRC	3	4	4
Epid-Tox	3	4	4
Évaluation comparative du PDP	3	2	3
Évaluation du PDP fondée sur les hypothèses	3	4	3
GRADE	3	4	4
Hope et Clarkson (2014)	3	3	3
Méthode bayésienne	3	4	2
NRC (2014)	3	4	3
OHAT	3	4	4
RS – Cochrane	4	4	3
RS – EFSA	4	4	3
SCENIHR	3	4	3

4. Revue des pratiques actuelles de l'Anses

4.1 Le processus d'expertise à l'Anses

L'Anses est certifiée depuis septembre 2013 selon la norme ISO 9001 (ISO 2008), associée à la norme NF X 50-110 pour les processus d'expertise (AFNOR 2003). Les documents du système qualité de l'expertise à l'Anses répondent aux prescriptions de compétence pour une expertise décrites dans la norme NF X 50-110. Ces prescriptions recouvrent le management des ressources de l'organisme d'expertise, les prescriptions techniques pour une expertise (exemples : planification de l'expertise, revue des exigences du client et contrat d'expertise, conception et validation de la méthode d'expertise, réalisation et contenu du produit de l'expertise) et un dispositif d'amélioration continue.

La norme NF X 50-110 précise que le processus d'expertise s'entend de la question posée à la remise du produit de l'expertise conformément au contrat d'expertise. La norme identifie les différents points critiques du processus d'expertise : l'évaluation de la question posée, la sélection des experts, la méthode d'expertise, la réalisation de l'expertise, l'analyse critique des données et des actions menées, la restitution de l'expertise au client, le suivi du produit de l'expertise. L'Anses prend en compte ces différents points critiques dans la procédure ANSES/PR1/9/01 « organisation de la réalisation d'une expertise en réponse à une saisine ou une auto-saisine » qui prévoit certaines des étapes décrites dans la figure 2. L'Anses dispose d'un certain nombre de formulaires en appui au processus d'expertise et en lien avec le poids des preuves (Tableau 13).

Tableau 13 : Formulaires du processus d'expertise en lien avec le poids des preuves

Principales étapes de la réalisation d'une expertise en réponse à une saisine	Formulaires Anses en appui	Etapes d'évaluation du poids des preuves traités dans les formulaires
Cadrage interne – établissement contrat	• Analyse de la saisine et cadrage interne de l'expertise • Grille d'analyse sociologique • Contrat d'expertise	Planification de l'évaluation (cadrage, formulation de la/des question(s))
Collecte des données	• Formulaire de profil de recherche bibliographique • Convention de recherche et développement	Etablissement des lignes de preuves (recherche, sélection des études et extraction des données)
Réalisation de l'expertise	• Rapport d'expertise collective • Expertise collective : synthèse et conclusion	
Elaboration avis de l'agence	• Avis de l'Anses	Expression des conclusions

Le contenu des documents Anses présentés dans le Tableau 13 est brièvement décrit ci-dessous.

Formulaire : « Analyse de la saisine et cadrage interne de l'expertise » :

Ce formulaire constitue un support au cadrage et à la formulation de la question de l'étape de planification de l'évaluation. Il permet notamment de définir le contexte sociétal, règlementaire et international, les questions scientifiques posées, ainsi que les compétences, les moyens et ressources nécessaires à leur traitement, les modalités d'organisation de l'expertise, la planification prévisionnelle et les modalités de

communication. En revanche, le document n'aborde pas les méthodes d'expertise.
Le **« contrat d'expertise »** reprend les principaux éléments de ce formulaire.

Document : **« Grille de questionnement sociologique – Juin 2012 »**

Cette grille constitue un support au cadrage et à la formulation de la question de l'étape de planification de l'évaluation. Elle aborde le contexte institutionnel, le contexte socio-économique, les pratiques des différents groupes sociaux, la construction du problème, les formes de savoirs et les inégalités. Le but de cette grille préliminaire est l'identification des enjeux importants dans le contexte de la saisine (les controverses, les risques institutionnels, l'incertitude scientifique, etc.).

« Formulaire de recherche de profil bibliographique »

Ce formulaire constitue un support à la recherche des études, de l'étape établissement des lignes de preuve. Il prévoit de cibler la requête notamment en fonction des populations cibles, des types d'études et de substances ou sujets concernés, des bases de données à consulter.
Un formulaire est prévu pour réaliser une **« convention de recherche et développement** » en vue d'acquérir des données en appui aux expertises.

Document du système qualité Anses : « Anses – Rapport d'expertise collective »

Ce document est une trame de format pour les rapports d'expertise mais il n'est pas prescriptif. Les titres de sous-sections ne sont pas spécifiés sauf la partie « Contexte, Objet, et Modalités de Traitement » et « Conclusions ».

Document : « Expertise collective : synthèse et conclusion »

Ce document est une trame optionnelle de format de synthèses et conclusions. Il aborde la question posée, le contexte scientifique, l'organisation de l'expertise, la description de la méthode (sans spécification de critères) pour introduire les résultats et les conclusions de l'expertise. Il ne prévoit pas d'exprimer le poids des preuves dans la conclusion.

Document : « Avis de l'Anses »

Ce document est une trame de format des avis de l'Anses. Il contient des sections relatives au « Contexte et Objet de la Saisine », « Organisation de l'Expertise », « Analyse et Conclusions du CES (ou groupe approprié) » et les « Conclusions et Recommandations de l'Agence ». La dernière section prévoit de « donner les éléments et conclusions relatifs à l'expertise mettant en évidence la limite de validité des résultats et le degré d'incertitude, les compléments demandés, les réserves et opinions divergentes argumentées ». Il n'est pas explicitement demandé d'expression du poids des preuves.

En conclusion, les formulaires du processus d'expertise mentionnent les trois sous-étapes de la planification de l'évaluation. Seules les sous-étapes de « cadrage » et de « formulation de la/des question(s) » sont développées dans certains documents. Celle concernant le « développement de la méthode d'évaluation » n'est jamais indiquée. Le choix de la méthode d'expertise est mentionné dans la procédure ANSES/PR1/9/01 « organisation de la réalisation d'une expertise en réponse à une saisine ou une auto-saisine » dans la réalisation de l'expertise en indiquant que ce point doit être explicité dans le compte-rendu de la réunion et le produit de l'expertise. Cette démarche n'est pas intégrée dès la planification de l'évaluation. Le document « Avis de l'Anses » demande de préciser les limites de validité des résultats sans expliciter le lien éventuel avec le poids des preuves. **Les documents n'intègrent pas explicitement une référence au poids des preuves.**

4.2 Résultats du rapport « État des lieux sur l'analyse de l'incertitude et l'évaluation du poids des preuves à l'Anses »

Seize saisines ont été analysées par le GT MER dans l'état des lieux réalisé sur l'analyse de l'incertitude et l'évaluation du poids des preuves à l'Anses (Anses 2015b). Les méthodes utilisées dans ces saisines pour analyser le poids des preuves sont indiquées dans le Tableau 14. La plupart de ces méthodes concernent l'analyse de la qualité des études individuelles et des études de synthèse ainsi que l'analyse d'un ensemble d'étude. Certaines saisines utilisent également des méthodes d'évaluation du poids des preuves qui permettent de combiner des études de différentes natures. A l'exception de deux méthodes, les méthodes mobilisées dans les 16 saisines de l'Anses ont toutes été identifiées lors la revue bibliographique réalisée pour le présent rapport (critères de Bradford Hill, méthode du CIRC, arbres de décision, critères de Klimisch). Les deux autres méthodes ont été publiées en dehors de la période couverte par la revue (2010-2015) : Calabrese et al. (1997) et Lewandowski et Rhomberg (2005). Le score de Calabrese est une approche quantitative utilisée pour hiérarchiser des substances chimiques pouvant potentiellement être des perturbateurs endocriniens. Elle repose sur une notation d'études expérimentales *in vitro* et *in vivo*. L'approche décrite par Lewandovski et Rhomberg (2005) est semi-quantitative, et conduit à la sélection d'une étude parmi un groupe d'études candidates. Elle tient compte des critères de Bradford Hill, ainsi que de la validité interne et externe des études. L'Anses a adapté cette approche pour sélectionner les études utilisées pour établir les Valeurs Toxicologiques de Référence (VTR) et les Valeurs Limites d'Exposition Professionnelle (VLEP).

Des méthodes plus ou moins sophistiquées ont été utilisées à l'Anses selon le caractère plus ou moins sensible de la saisine, le profil des experts sollicités et les informations disponibles.

Tableau 14 : Méthodes utilisées dans les saisines† Anses analysées par le GT MER

	Etablissement des lignes de preuve			
Méthode	Sélection des études et extraction des données	Analyse de la qualité des études individuelles et des études de synthèse	Analyse d'un ensemble d'études	**Etablissement poids des preuves**
Bradford Hill			Facteurs de croissance (Anses 2012)	
CIRC			Radiofréquences (Anses 2013b)	Radiofréquences (Anses 2013b)
Arbre de décision		BPA (Anses 2013a)	BPA (Anses 2013a)	BPA (Anses 2013a)
Klimisch		BPA (Anses 2013a)		
Score de Calabrese		BPA (Anses 2013a)		
Score (qualitatif)			DEP (Anses 2014a) Abeilles (Anses 2015a)	
Critères pour sélectionner des études (Lewandowski et Rhomberg 2005, …)		Cobalt (Anses 2014f), n-hexane (Anses 2014d), Acétaldéhyde (Anses 2014e), BPA (Anses 2013a) DEP (Anses 2014a)		

†Les saisines sont identifiées par un « mot clé » et leur référence abrégée, les références complètes sont listées en Annexe 9. Certaines saisines utilisent plusieurs méthodes.

4.3 Exemples de classification pour l'expression des conclusions

Des exemples de classification pour l'expression des conclusions, issus de saisines de l'Anses, sont regroupés dans le Tableau 15.

Tableau 15 : Exemples de classification pour l'expression des conclusions dans les avis

Évaluation des risques du bisphénol A (BPA) pour la santé humaine (Anses 2013a)
Les conclusions sont narratives (pas de schéma, couleurs). Deux entrées pour les conclusions sont présentées : - Conclusions sur les effets sanitaires du BPA par niveau de preuve (4 classes) et organisme (homme ou animal) (effets avérés ; effets controversés ; effets suspectés ; effets dont les données ne permettent pas de conclure) - Conclusions par organe/système avec détermination des effets à retenir pour l'évaluation des risques (pour chaque organe/système, conclusions concernant les effets chez l'homme et/ou l'animal et si ceux si sont avérés ; controversés ; suspectés ou si les données ne permettent pas de conclure) Conclusion narrative par étape de l'évaluation de risque (caractérisation du danger ; de l'exposition ; des risques) avec conclusion générale en gras.
Radiofréquences et Santé (Anses 2013b)
Les conclusions sont narratives, concernent les études par catégories d'effets (non cancérogène sur le SNC ; autres effets non cancérogènes ; effets cancérogènes potentiels) et indiquent le niveau de preuve selon 5 classes (effet avéré pour l'Homme ; effet probable pour l'Homme ; effet possible pour l'Homme ; Niveau de preuve insuffisant pour conclure à un effet ; Probablement pas d'effet chez l'Homme)
Évaluation des risques liés aux nanomatériaux (Anses 2014c)
Conclusion narrative

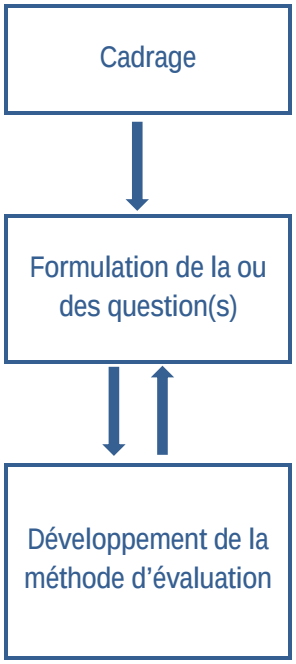
5. Recommandations

Sur la base de la revue critique de la littérature et de la revue des pratiques de l'Anses, des recommandations méthodologiques sont proposées ci-dessous. Pour chaque étape du processus, les recommandations sont classées des plus générales aux plus spécifiques.

5.1 Planification de l'évaluation

A l'Anses, la planification de l'évaluation est réalisée lors de la rédaction du document de cadrage. Il n'existe pas de documents précisant les méthodes d'évaluation et les formulations des questions pour les collectifs d'experts. Le GT MER recommande la mise en place d'un processus complet pour la planification de l'évaluation en 3 sous-étapes (figure 7), avec pour chacune un formulaire approprié.

Figure 7 : Relation entre les sous-étapes de planification de l'évaluation



La sous-étape « cadrage » est réalisée par les agents de l'Anses.

Les deux sous-étapes suivantes mobilisent les CES et les experts impliqués dans le traitement de la saisine.

Le GT MER recommande d'adopter une démarche rétroactive entre les sous-étapes 2 et 3 de manière à affiner la formulation des questions en garantissant que celles-ci puissent être correctement traitées.

Le GT MER recommande **à l'aide d'une revue préliminaire et si besoin d'auditions** :

- D'identifier les enjeux (sanitaires, environnementaux, sociétaux, économiques avec les décideurs (et, dans l'idéal, les parties prenantes).
- D'évaluer l'étendue du corpus des données disponibles.
- **De formaliser la ou les question(s) au moyen d'une structure de description présentée dans la littérature (PICO, PECO, etc.).**
- De formaliser les sous-questions à traiter par le collectif d'experts, si possible sous forme de **modèle conceptuel.**
- De **rédiger un plan de développement de la méthode d'évaluation** en préalable à sa conduite.
- De définir les modalités de communication de la conclusion dans le document du cadrage.

Dans le cadre de l'évaluation du poids des preuves, le GT MER recommande, pour la sous-étape 3, de spécifier au minimum :

- **Le choix du type de revue de la littérature** : revue systématique ou revue approfondie pour traiter la ou les questions définies lors de la sous-étape 2. Ce choix doit tenir compte des éléments issus du document de cadrage, de l'analyse socio-économique de l'Anses et des critères proposés par l'EFSA : impact potentiel du résultat de la revue systématique, quantité et qualité des données, source et confidentialité des données, transparence nécessaire et controverse du sujet de la saisine, ressources à mettre en œuvre.
- Les critères utilisés pour évaluer la qualité des études par type d'étude.
- La méthode retenue pour évaluer un ensemble d'étude : méta-analyse, approche multicritères, ou approches qualitatives.
- La méthode retenue pour intégrer les lignes de preuves : modélisation statistique avec mise en place d'élicitation ou non de dires d'experts, analyse multicritères, ou approches qualitatives.
- Les modalités d'expression des conclusions concernant le poids des preuves.

5.2 Établissement des lignes de preuves

5.2.1 Recherche, sélection des études et extraction des données

L'Anses, dans le cadre de ses avis, a déjà conduit des revues systématiques de la littérature, mais leur réalisation reste occasionnelle au sein de l'agence.

Plusieurs collectifs de l'Anses mobilisent des grilles de lecture pour sélectionner les études et extraire des données, mais ces grilles sont insuffisamment partagées.

Le GT MER formule les recommandations suivantes :

- ➔ **Développer une grille de lecture,** voire un tableau d'extraction des données, adaptable en fonction de la question de recherche et du type d'étude scientifique. Cette grille et ce tableau devront être élaborés et testés sur des études de cas en coordination avec les CES. Ces grilles pourront inclure trois types d'éléments concernant la pertinence des études par rapport à la question posée, des éléments descriptifs de l'étude et des critères d'évaluation de la qualité des études.
- ➔ **Le recours à au moins deux experts est souhaitable pour remplir les grilles de lecture.**
- ➔ **Lorsqu'une revue systématique est réalisée** (Annexe 7, fiche méthode « Revue systématique »), **certains grands principes doivent être appliqués :** utilisation d'au moins deux bases de données, la sélection des études par deux personnes indépendantes, et la définition des critères de sélection et d'exclusion des études en amont. Les collectifs d'experts pourront s'appuyer sur le document guide EFSA (2010).
- ➔ Dans le cas où une revue systématique n'est pas réalisée (en cas de ressources insuffisantes par exemple), le GT MER recommande que la **procédure** de recherche, de sélection et d'extraction soit décrite de manière aussi **précise** que possible dans le rapport d'expertise, en se rapprochant autant que possible des pratiques d'une revue systématique.

5.2.2 Évaluation de la qualité des études

- ➔ Le GT MER recommande l'utilisation de **listes de critères formalisés** par type d'études (épidémiologiques, toxicologiques, etc.) de manière à assurer la **transparence** du processus d'évaluation. Ces listes devront être élaborées en coordination avec les CES, en exploitant par exemple les critères mentionnés dans la méthode GRADE. Lorsque des études sont exclues, les **critères d'exclusion** doivent être explicites et définis par le collectif d'experts.
- ➔ Lorsque des études de synthèse sont considérées dans le processus d'expertise, le GT MER recommande l'utilisation de la méthode **AMSTAR** et **R-AMSTAR** (Kung et al. 2010) pour évaluer leur qualité.

5.2.3 Évaluation d'un ensemble d'études par type d'étude en vue d'établir des lignes de preuve

- ➔ Lors de l'utilisation de méthodes qualitatives avec ou sans notation (ex : critères de causalité Bradford Hill, méthode du CIRC, méthode dérivée de l'approche de Lewandovski et Rhomberg 2005) et dans le but d'assurer un niveau de transparence élevé au processus d'expertise, le GT MER recommande de **définir aussi précisément** que possible **les critères utilisés et la signification des notes attribuées** à ces critères par les experts.
- ➔ Le GT MER estime que **la méta-analyse est une méthode à privilégier** et recommande à l'Anses d'en réaliser sur des sujets stratégiques, en tenant compte des enjeux en termes de risques sanitaires et des ressources disponibles (ressources humaines, disponibilité des données). Ces méta-analyses devraient être menées en s'inspirant des documents guides disponibles, notamment du document guide Cochrane.
- ➔ Lors de l'utilisation de méthodes quantitative, le GT MER recommande **d'analyser la sensibilité des résultats** aux paramètres d'entrée.
- ➔ Les **approches multicritères** semblent potentiellement intéressantes pour formaliser l'évaluation d'un ensemble d'études par un groupe d'experts. Toutefois, nous ne disposons pas d'un recul suffisant pour juger précisément de leur utilité dans le contexte de l'Anses. Le GT MER recommande de les tester sur des cas concrets variés en coordination avec les CES.

5.3 Intégration des lignes de preuves

- ➔ Il est important d'encourager les experts à **décrire et expliciter les choix réalisés à chaque étape** afin d'assurer un niveau de transparence aussi élevé que possible. Certaines méthodes permettant de rendre le processus d'expertise transparent pourraient être utilisées dans ce but (ex : méthodes d'élicitation d'experts, intégration des avis d'experts dans les modèles prédictifs de type QSAR).
- ➔ Le GT MER recommande l'utilisation des méthodes **qualitatives proposées par le CIRC ou le WCRF** pour la combinaison des lignes de preuves, en incluant explicitement une liste permettant de vérifier la prise en compte des critères de Hill (annexe 7, fiches méthodes « OMS-CIRC », « WCRF/AICR » et « Critères de Bradford Hill »).
- ➔ Le GT MER recommande l'utilisation de la **modélisation statistique** pour combiner différentes lignes de preuve (notamment de l'approche Bayésienne, annexe 7, fiches méthodes « Statistique Bayésienne »). Ces approches nécessitant des compétences spécifiques et un investissement en temps relativement important, la faisabilité de leur utilisation devra cependant être évaluée sur des exemples concrets.
- ➔ Lorsque des méthodes quantitatives sont mobilisées, le GT MER recommande **d'analyser la sensibilité des résultats** aux paramètres d'entrée.
- ➔ Les méthodes **d'analyse multicritères** telles que celle proposée par Linkov et al. (2011) (annexe 7, fiche « Analyse multicritères ») sont potentiellement intéressantes, mais le GT MER ne dispose pas d'un recul suffisant pour déterminer si elle pourrait être utile dans le contexte de l'Anses. Le GT MER recommande de la tester sur des cas concrets en coordination avec les CES.
- ➔ Dans le domaine spécifique de l'évaluation du mode d'action, la **méthode d'évaluation comparative du poids des preuves** pourrait être utilisée (annexe 7, fiche méthode « Évaluation comparative du poids des preuves »).

Le groupe de travail a recensé les méthodes existantes et a identifié pour cette étape l'opportunité de développer des méthodes spécifiques (ou des combinaisons de ces méthodes) pour l'intégration des lignes de preuves à l'Anses dans le cadre d'une convention de recherche et développement.

5.4 Expression des conclusions concernant le poids des preuves

- ➔ Lorsque le poids des preuves est analysé avec une méthode quantitative, les **résultats numériques** doivent être associés à un **texte explicatif** répondant à la question posée.
- ➔ Lorsque le poids des preuves est analysé avec une méthode qualitative, le GT MER recommande d'exprimer les conclusions concernant le poids des preuves selon une **classification en 4 niveaux correspondant à des niveaux de preuve croissants. Une classe supplémentaire « évaluation impossible » peut également être considérée**. Le GT MER recommande que chaque niveau soit **défini précisément** dans les rapports d'expertise. Une classification sur une échelle numérique pourrait être élaborée et testée sur des études de cas en coordination avec les CES.
- ➔ Le GT MER recommande de structurer le contenu de la conclusion en adaptant les travaux de la collaboration Cochrane et de la méthode GRADE aux domaines de l'agence. Un rapprochement avec la Collaboration Cochrane pourrait être envisagé afin de bénéficier de son expérience.
- ➔ Le GT MER recommande de **caractériser l'incertitude** dans la conclusion soit sous forme qualitative, soit sous forme quantitative, selon la méthode d'analyse du poids des preuves utilisées. Cette recommandation est également formulée dans le rapport « Traitement de l'incertitude dans le processus d'évaluation des risques sanitaires des substances chimiques » rédigé par le GT « Perturbateurs Endocriniens et reprotoxiques de catégorie 2 ».

5.5 Intégration dans le processus d'expertise

L'objectif recherché par l'évaluation du poids des preuves est de rendre transparent les résultats de l'expertise pour en assurer la crédibilité et la confiance de la part de la communauté scientifique comme des parties prenantes. La mise en place généralisée de cette évaluation par tous les collectifs d'experts pourrait alourdir sensiblement la charge de travail. Afin de faciliter l'appropriation des concepts et méthodes d'évaluation du poids des preuves par les collectifs d'experts :

- ➔ Le GT MER recommande de mettre à disposition des collectifs d'experts un **soutien méthodologique** pour faciliter la mise en œuvre des méthodes d'évaluation du poids des preuves. Des référents méthodologiques pourraient être identifiés pour aider des collectifs d'experts à réaliser les sous étapes 2 et 3 de la planification de l'évaluation, à réaliser les revues systématiques et à appliquer les méthodes quantitatives (méta-analyse, modélisation statistique).
- ➔ Le GT MER recommande de mettre en place **un système d'information** pour capitaliser les travaux d'expertise passés. Ce système d'information inclurait le contenu des grilles de lecture des revues bibliographiques et un descriptif de la méthode d'évaluation du poids des preuves mise en œuvre dans l'expertise.

Les documents du système qualité Anses ont été développés en regard d'une norme (NF X-50-110) publiée en 2003. La prise en compte de l'évaluation du poids des preuves dans le processus d'expertise demande de revoir certains documents. Ainsi,

- ➔ Le GT MER recommande d'inscrire le **processus d'évaluation** du poids des preuves au sein du processus d'expertise de l'Anses.
- ➔ Le GT MER recommande d'intégrer explicitement les éléments relatifs à l'évaluation du poids des preuves dans le **document de cadrage interne**, dans le rapport d'expertise collective, dans le document de synthèse et conclusion de l'expertise collective, et dans l'avis de l'Anses de façon à accroître la transparence des travaux d'expertise.
- ➔ Le GT MER recommande de développer **un formulaire opérationnel de description du plan de développement de la méthode d'évaluation** par le collectif d'experts. Pour cela, le GT MER recommande d'adapter le plan détaillé proposé par l'OHAT aux besoins de l'agence (Annexe 7, fiche méthode « Processus d'évaluation du poids des preuves OHAT »).

➔ Le GT MER recommande d'adapter **le formulaire opérationnel de profil de recherche bibliographique** de façon à permettre la prise en compte des informations transmises dans les structures de description de la question posée (PICO, PECO, etc.).

5.6 Classification des méthodes recommandées par le GT MER en fonction de leur niveau de faisabilité

Le tableau ci-dessous présente l'ensemble de méthodes recommandées par le GT MER (Tableau 16). Les méthodes recommandées ont reçu des notes élevées pour le niveau de « Pertinence » et le « Caractère directif ». Celles-ci sont classées en deux groupes correspondant à des niveaux de faisabilité (temps et ressources) contrastés.

Tableau 16 : Classification des méthodes† recommandées par le GT MER en fonction de leur niveau de faisabilité (ressources).

Etape du processus d'évaluation du poids des preuves	Méthodes de niveau I (Faisable)	Méthodes de niveau II (Moins faisable)
1 - Planification de l'évaluation		
Cadrage	Document structuré	
Formulation de la/des question(s)	Définition d'un modèle conceptuel Structure d'informations (PICO, etc.)	
Plan de travail	Revue préliminaire Plan de développement de la méthode d'évaluation	
2 - Etablissement des lignes de preuve		
Recherche, sélection, extraction	Revue approfondie de la littérature - Grille de lecture pour la sélection des études - Description de la procédure utilisée	Revue Systématique de la littérature
Évaluation de la qualité des études	Grille de lecture incluant des critères de qualité AMSTAR ou R-AMSTAR	
Évaluation d'un ensemble d'étude	Méthodes qualitatives (ex : CIRC, Bradford Hill)	Méta-analyse
3 - Évaluation du poids des preuves		
	Méthodes qualitatives et semi-quantitatives (ex : CIRC, WCRF, méthodes multicritères)	Modélisation statistique
4 - Expression de la conclusion		
	Classification en 4 niveaux Structuration du contenu de la conclusion	

†Les méthodes listées ci-dessous ont reçu des notes élevées pour leur niveau de pertinence et leur caractère directif

6. Conclusions du groupe de travail

Ce rapport d'expertise collective est basé sur une revue critique de la littérature et sur une analyse des pratiques actuelles de l'Anses. Il propose des définitions pour trois termes clés (poids des preuves, ligne de preuves et revue systématique) et formule des recommandations visant à harmoniser les procédures utilisées par l'Anses pour l'étape d'identification des dangers.

Les définitions du poids des preuves proposées dans la littérature sont hétérogènes. Certaines considèrent que le poids des preuves correspond à un cadre de travail transparent permettant de tirer des conclusions, d'autres insistent sur l'idée d'intégration de différentes lignes de preuve et, finalement, certaines définitions considèrent que le poids des preuves correspond à un processus prenant en considération les forces et les faiblesses de différents éléments d'information. Dans le but de disposer d'une définition intégrative du poids des preuves, GT MER propose de définir le poids des preuves de la façon suivante : « Une synthèse formalisée de lignes de preuves, éventuellement de qualités hétérogènes, dans le but de déterminer le niveau de plausibilité d'hypothèses ». Le GT MER a également combiné différentes définitions de « ligne de preuves » présentées dans la littérature pour aboutir à la définition intégrative suivante : « Une ligne de preuves est un ensemble d'informations de même nature, intégrées pour évaluer une hypothèse».

Le GT MER recommande de structurer le processus d'évaluation du poids des preuves en quatre étapes : (1) Planification de l'évaluation ; (2) Établissement des lignes de preuves ; (3) Intégration des lignes de preuve pour établir le poids des preuves ; (4) Expression des conclusions sur le poids des preuves.

Pour chacune de ces étapes, les méthodes recensées lors de la revue bibliographique ont été comparées par le GT MER à celles inventoriées dans l'état des lieux sur l'analyse de l'incertitude et l'évaluation du poids des preuves à l'Anses afin de définir des pistes de progrès. Les documents du système qualité de l'Anses en lien avec l'évaluation du poids des preuves ont également été considérés pour évaluer l'inscription de l'évaluation du poids des preuves dans le processus d'expertise.

Le GT MER formule des recommandations méthodologiques pour chaque étape du processus, en tenant compte du caractère directif des méthodes, de leur pertinence et de la faisabilité pour leur mise en œuvre. Les méthodes jugées comme étant les plus directives et les plus pertinentes sont privilégiées et sont classées selon leur niveau de faisabilité par rapport aux ressources disponibles.

Concernant la planification de l'évaluation (étape 1), le GT MER recommande la mise en place d'un processus complet et itératif en trois sous-étapes, impliquant à la fois l'Anses et les experts mobilisés pour le traitement de la saisine. Ces trois sous-étapes sont : le cadrage, la formulation des questions et le développement de la méthode d'évaluation du poids des preuves. Pour la dernière sous-étape, le GT MER recommande de spécifier au minimum le choix du type de revue de littérature, les critères d'évaluation de la qualité des études, les méthodes retenues pour évaluer un ensemble d'étude et pour intégrer les lignes de preuves ainsi que les modalités d'expression du poids des preuves.

L'établissement des lignes de preuves (étape 2) repose en partie sur la revue de la littérature. La qualité de cette revue est déterminante. Lorsqu'une revue systématique est réalisée, certains grands principes doivent être appliqués : utilisation d'au moins deux bases de données, sélection des études par deux personnes indépendantes et, si possible, définition des critères de sélection et d'exclusion en amont avec les parties prenantes. Dans le cas où une revue systématique ne peut pas être réalisée, le GT MER recommande que la procédure de recherche, de sélection et d'extraction soit décrite de manière aussi précise que possible. Pour évaluer la qualité des études, des grilles détaillées, sous forme de listes, formalisées par type d'études individuelles (épidémiologiques, toxicologiques,…) devraient être utilisées de manière à assurer la transparence du processus d'évaluation. Pour établir des lignes de preuves sur des sujets jugées stratégiques par l'Anses, le GT MER recommande de réaliser des méta-analyses pour synthétiser de manière quantitative les données disponibles. Lorsque les lignes de preuves sont élaborées en utilisant une méthode qualitative, le GT MER recommande de définir aussi précisément que possible les critères utilisés

et la signification des notes attribuées à ces critères par les experts dans le but d'assurer un niveau de transparence élevé.

Le GT MER recommande l'utilisation de modèles statistiques pour intégrer différentes lignes de preuve (étape 3). Leur développement et leur utilisation nécessitent des compétences spécifiques. Lorsque de tels modèles sont mobilisés, le GT MER recommande d'analyser la sensibilité des résultats aux paramètres d'entrée. Lorsqu'une méthode qualitative est utilisée pour combiner différentes lignes de preuves, il est important de décrire et d'expliciter les choix réalisés à chaque étape de manière à rendre le processus transparent et reproductible.

Le groupe de travail recommande d'exprimer les conclusions concernant le poids des preuves (étape 4) selon une classification en quatre niveaux. Lorsqu'une méthode quantitative est appliquée, ces niveaux peuvent être définis directement à partir des quantités calculées par le modèle utilisé. Dans le cas contraire, des définitions aussi précises que possibles doivent être établies par les collectifs d'experts en fonction du problème traité.

Des fiches méthodes présentées dans le rapport visent à faciliter l'appropriation des concepts et méthodes d'évaluation du poids des preuves par les collectifs d'experts.

Enfin, le groupe de travail recommande de mettre en place un système d'information permettant de capitaliser les travaux d'expertise passés, pouvant inclure le contenu des grilles de lecture des revues bibliographiques et un descriptif de la méthode d'évaluation du poids des preuves mise en œuvre dans l'expertise.

Bien que ce rapport se focalise sur l'évaluation du poids des preuves à l'étape d'identification des dangers, la démarche d'évaluation du poids des preuves recommandée peut également être mobilisée pour d'autres étapes de l'évaluation des risques. En effet, la définition élaborée par le GT MER pour le poids des preuves concerne le niveau de plausibilité d'hypothèses sans se restreindre aux hypothèses de causalité entre l'exposition à un agent et le type et la nature des effets néfastes qu'il peut engendrer.

Dans le but de vérifier la pertinence et de faciliter la diffusion de ces recommandations, le GT MER propose de les évaluer en réalisant des études de cas en collaboration avec les collectifs de l'Anses. Un guide méthodologique sera rédigé à l'issue de cette dernière étape et proposera, en fonction de différentes situations, des méthodes pour évaluer la qualité des études et des données disponibles, ainsi que pour évaluer et communiquer le poids des preuves. Si elles sont adoptées par les collectifs d'experts, ces recommandations permettront d'harmoniser les procédures de l'Agence concernant le poids des preuves à l'étape d'identification du danger.

Date de validation du rapport d'expertise collective par le GT MER : 24 mars 2016

7 Bibliographie

7.1 Publications

Adami, Hans-Olov, Sir Colin L. Berry, Charles B. Breckenridge, Lewis L. Smith, James A. Swenberg, Dimitrios Trichopoulos, Noel S. Weiss, et Timothy P. Pastoor. 2011. "Toxicology and Epidemiology: Improving the Science with a Framework for Combining Toxicological and Epidemiological Evidence to Establish Causal Inference." *Toxicological Sciences* 122 (2):223-234. doi: 10.1093/toxsci/kfr113.

Akl, E. A., N. Maroun, G. Guyatt, A. D. Oxman, P. Alonso-Coello, G. E. Vist, P. J. Devereaux, V. M. Montori, et H. J. Schunemann. 2007. "Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial." *J Clin Epidemiol* 60 (12):1298-305. doi: 10.1016/j.jclinepi.2007.03.011.

Andrews, J. C., H. J. Schunemann, A. D. Oxman, K. Pottie, J. J. Meerpohl, P. A. Coello, D. Rind, V. M. Montori, J. P. Brito, S. Norris, M. Elbarbary, P. Post, M. Nasser, V. Shukla, R. Jaeschke, J. Brozek, B. Djulbegovic, et G. Guyatt. 2013. "GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength." *J Clin Epidemiol* 66 (7):726-35. doi: 10.1016/j.jclinepi.2013.02.003.

Andrews, Jeff, Gordon Guyatt, A. D. Oxman, Phil Alderson, Philipp Dahm, Yngve Falck-Ytter, Mona Nasser, Joerg Meerpohl, P. N. Post, Regina Kunz, Jan Brozek, Gunn Vist, David Rind, E. A. Akl, et H. J. Schünemann. 2013. "GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations." *Journal of Clinical Epidemiology* 66 (7):719-725.

Anses. 2012. Avis et rapport de l'Anses relatif à l'Étude des liens entre facteurs de croissance, consommation de lait et de produits laitiers et cancers. Maisons-Alfort: Anses.

Anses. 2013a. Avis et rapport de l'Anses relatif à l'Évaluation des risques du bisphénol A (BPA) pour la santé humaine. - Tome 1 : Évaluation des risques du bisphénol A (BPA) pour la santé humaine et aux données toxicologiques et d'usage des bisphénols S, F, M, B, AP, AF, et BADGE. Maisons-Alfort: Anses.

Anses. 2013b. Avis et rapport de l'Anses relatif à la mise à jour de l'expertise « Radiofréquences et santé ». Maisons-Alfort: Anses.

Anses. 2013c. Évaluation des risques du bisphénol A (BPA) pour la santé humaine - Tome 1. Maisons-Alfort, France: Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail.

Anses. 2013d. Radiofréquences et santé - Mise à jour de l'expertise. Maisons-Alfort, France: Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail.

Anses. 2014a. Avis de l'Anses relatif au risque d'émergence de la diarrhée épidémique porcine (DEP) due à un nouveau variant du virus de la DEP en France. Maisons-Alfort: Anses.

Anses. 2014b. Avis de l'Anses relatif au risque d'émergence de la diarrhée épidémique porcine (DEP) en Europe par le biais de l'alimentation animale. Maisons-Alfort: Anses.

Anses. 2014c. Avis et rapport de l'Anses relatif à l'Évaluation des risques liés aux nanomatériaux. Enjeux et mise à jour des connaissances. Maisons-Alfort: Anses.

Anses. 2014d. Avis et rapport de l'Anses relatif à la Valeur toxicologique de référence chronique par voie respiratoire pour le n-hexane. Maisons-Alfort: Anses.

Anses. 2014e. Avis et rapport de l'Anses relatif à une Proposition de valeurs guides de qualité d'air intérieur. L'acétaldéhyde. Maisons-Alfort: Anses.

Anses. 2014f. Avis et rapport de l'Anses relatif à la proposition de valeurs limites d'exposition à des agents chimiques en milieu professionnel - Evaluation des effets sur la santé et des

méthodes de mesure des niveaux d'exposition sur le lieu de travail pour le cobalt et de ses composés à l'exception du cobalt associé au carbure de tungstène. Maisons-Alfort: Anses.

Anses. 2015a. Avis de l'Anses relatif à la hiérarchisation des dangers sanitaires exotiques ou présents en France métropolitaine chez les abeilles. Maisons-Alfort: Anses.

Anses. 2015b. Etat des lieux sur l'analyse de l'incertitude et l'évaluation du poids des preuves - Rapport interne. Maisons-Alfort, France: Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail.

Anses. 2015c. Note d'appui scientifique et technique de l'Anses relatif à la demande d'avis relatif à l'évaluation des risques pour la santé humaine du bisphénol A. Maisons-Alfort: Anses.

Ashbolt, NJ., A. Amézquita, T. Backhaus, P. Borriello, KK. Brandt, P. Collignon, A. Coors, R. Finley, WH. Gaze, T. Heberer, JR. Lawrence, DG. Larsson, SA. McEwen, JJ. Ryan, J. Schönfeld, P. Silley, JR. Snape, C. Van den Eede, et E. Topp. 2013. "Human health risk assessment (HHRA) for environmental development and transfer of antibiotic resistance." *Environ Health Perspect* (121):993-1001. doi: http://dx.doi.org/10.1289/ehp.1206316.

Bailey, L. A., M. A. Nascarella, L. E. Kerper, et L. R. Rhomberg. 2016. "Hypothesis-based weight-of-evidence evaluation and risk assessment for naphthalene carcinogenesis." *Crit Rev Toxicol* 46 (1):1-42. doi: 10.3109/10408444.2015.1061477.

Balshem, Howard, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris, et Gordon H. Guyatt. 2011. "GRADE guidelines: 3. Rating the quality of evidence." *Journal of Clinical Epidemiology* 64 (4):401-406. doi: 10.1016/j.jclinepi.2010.07.015.

Bergman, A., G. Becher, B. Blumberg, P. Bjerregaard, R. Bornman, I. Brandt, S. C. Casey, H. Frouin, L. C. Giudice, J. J. Heindel, T. Iguchi, S. Jobling, K. A. Kidd, A. Kortenkamp, P. M. Lind, D. Muir, R. Ochieng, E. Ropstad, P. S. Ross, N. E. Skakkebaek, J. Toppari, L. N. Vandenberg, T. J. Woodruff, et R. T. Zoeller. 2015. "Manufacturing doubt about endocrine disrupter science - A rebuttal of industry-sponsored critical comments on the UNEP/WHO report "State of the Science of Endocrine Disrupting Chemicals 2012"." *Regul Toxicol Pharmacol* 73 (3):1007-17. doi: 10.1016/j.yrtph.2015.07.026.

Berkman, ND, KN Lohr, M Ansari, M McDonagh, E Balk, E Whitlock, SC Morton, M Viswanathan, P Sista, et S Chang. 2012. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality : An Update. In *Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290 - 2007 - 10056 - I)*. Rockville, MD: Agency for Healthcare Research and Quality.

Bilotta, Gary S., Alice M. Milner, et Ian Boyd. 2014. "On the use of systematic reviews to inform environmental policies." *Environmental Science & Policy* 42:67-77. doi: http://dx.doi.org/10.1016/j.envsci.2014.05.010.

Boobis, A. R., S. M. Cohen, V. Dellarco, D. McGregor, M. E. Meek, C. Vickers, D. Willcocks, et W. Farland. 2006. "IPCS framework for analyzing the relevance of a cancer mode of action for humans." *Crit Rev Toxicol* 36 (10):781-92. doi: 10.1080/10408440600977677.

Boobis, A. R., J. E. Doe, B. Heinrich-Hirsch, M. E. Meek, S. Munn, M. Ruchirawat, J. Schlatter, J. Seed, et C. Vickers. 2008. "IPCS framework for analyzing the relevance of a noncancer mode of action for humans." *Crit Rev Toxicol* 38 (2):87-96. doi: 10.1080/10408440701749421.

Brown, Polly, Klara Brunnhuber, Kalipso Chalkidou, Iain Chalmers, Mike Clarke, Mark Fenton, Carol Forbes, Julie Glanville, Nicholas J. Hicks, Janet Moody, Sara Twaddle, Hazim Timimi, et Pamela Young. 2006. "How to formulate research recommendations." *BMJ : British Medical Journal* 333 (7572):804-806. doi: 10.1136/bmj.38987.492014.94.

Calabrese, E. J., E. J. Stanek, R. C. James, et S. M. Roberts. 1997. "Soil ingestion: a concern for acute toxicity in children." *Environmental Health Perspectives* 105 (12):1354-1358.

Chalmers, I., L. V. Hedges, et H. Cooper. 2002. "A brief history of research synthesis." *Eval Health Prof* 25 (1):12-37.

ECETOC. 2009. Framework for the Integration of Human and Animal Data in Chemical Risk Assessment. In *Technical Report*. Brussels, Belgium: European Centre for Ecotoxicology and Toxicology of Chemicals.

ECHA. 2011. Guidance on information requirements and chemical safety assessment. Chapter R.4 Evaluation of available information. Helsinki, Finland: European Chemicals Agency.

EFSA. 2010. "Application of systematic review methodology to food and feed safety assessments to support decision making." *EFSA Journal* 8 (6). doi: 10.2903/j.efsa.2010.1637.

EFSA. 2014. "Scientific Opinion on the risk of Phyllosticta citricarpa (Guignardia citricarpa) for the EU territory with identification and evaluation of risk reduction options." *EFSA Journal* (12):3355.

FAO/OMS. 2001. Evaluation of allergenicity of genetically modified foods - Report of a Joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology. Rome, Italy: Food and Agriculture Organization of the United Nations.

FDA. 2009. Guidance for Industry: Evidence-Based Review System for the Scientific Evaluation of Health Claims - Final. Washington D.C: U.S. Food and Drug Administration.

Good, I. J. 1979. "Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II." *Biometrika* 66 (2):393-6.

Good, I. J. 1985. "Weight of Evidence: A Brief Survey." *Bayesian Statistics* 2:249-270.

Goodman, M., K. Squibb, E. Youngstrom, L. G. Anthony, L. Kenworthy, P. H. Lipkin, D. R. Mattison, et J. S. Lakind. 2010. "Using systematic reviews and meta-analyses to support regulatory decision making for neurotoxicants: lessons learned from a case study of PCBs." *Environ Health Perspect* 118 (6):727-34. doi: 10.1289/ehp.0901835.

Gosling, J P , Andy Hart, H Owen, M David, J Li, et C MacKay. 2013. "A Bayes linear approach to weight-of-evidence risk assessment for skin allergy." *Bayesian Analysis* 8 (1):169-186.

Groupe BioBayes. 2015. *Initiation à la statistique bayésienne*: Ellipses.

Guha, N., A. Roy, L. Kopylev, J. Fox, M. Spassova, et P. White. 2013. "Nonparametric Bayesian methods for benchmark dose estimation." *Risk Anal* 33 (9):1608-19. doi: 10.1111/risa.12004.

Guyatt, G. H., A. D. Oxman, R. Kunz, D. Atkins, J. Brozek, G. Vist, P. Alderson, P. Glasziou, Y. Falck-Ytter, et H. J. Schunemann. 2011. "GRADE guidelines: 2. Framing the question and deciding on important outcomes." *J Clin Epidemiol* 64 (4):395-400. doi: 10.1016/j.jclinepi.2010.09.012.

Guyatt, G. H., A. D. Oxman, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, David Rind, P. J. Devereaux, V. M. Montori, Bo Freyschuss, Gunn Vist, Roman Jaeschke, J. W. Williams, Jr., M. H. Murad, David Sinclair, Yngve Falck-Ytter, Joerg Meerpohl, Craig Whittington, Kristian Thorlund, Jeff Andrews, et H. J. Schünemann. 2011. "GRADE guidelines 6. Rating the quality of evidence - imprecision." *Journal of Clinical Epidemiology* 64 (12):1283-1293. doi: 10.1016/j.jclinepi.2011.01.012.

Guyatt, G. H., A. D. Oxman, Regina Kunz, James Woodcock, Jan Brozek, Mark Helfand, Pablo Alonso-Coello, Paul Glasziou, Roman Jaeschke, E. A. Akl, Susan Norris, Gunn Vist, Philipp Dahm, V. K. Shukla, Julian Higgins, Yngve Falck-Ytter, et H. J. Schünemann. 2011. "GRADE guidelines: 7. Rating the quality of evidence - inconsistency." *Journal of Clinical Epidemiology* 64 (12):1294-1302. doi: 10.1016/j.jclinepi.2011.03.017.

Guyatt, G. H., A. D. Oxman, Victor Montori, Gunn Vist, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, Ben Djulbegovic, David Atkins, Yngve Falck-Ytter, J. W. Williams, Jr., Joerg Meerpohl, S. L. Norris, E. A. Akl, et H. J. Schünemann. 2011. "GRADE guidelines: 5. Rating the quality of evidence - publication bias." *Journal of Clinical Epidemiology* 64 (12):1277-1282. doi: 10.1016/j.jclinepi.2011.01.011.

Guyatt, G. H., A. D. Oxman, Shahnaz Sultan, Paul Glasziou, E. A. Akl, Pablo Alonso-Coello, David Atkins, Regina Kunz, Jan Brozek, Victor Montori, Roman Jaeschke, David Rind, Philipp Dahm, Joerg Meerpohl, Gunn Vist, Elise Berliner, Susan Norris, Yngve Falck-Ytter, M. H. Murad, et H. J. Schünemann. 2011. "GRADE guidelines: 9. Rating up the quality of evidence." *Journal of Clinical Epidemiology* 64 (12):1311-1316. doi: 10.1016/j.jclinepi.2011.06.004.

Guyatt, G. H., A. D. Oxman, Gunn Vist, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, Victor Montori, E. A. Akl, Ben Djulbegovic, Yngve Falck-Ytter, S. L. Norris, John Williams, Jr, David Atkins, Joerg Meerpohl, et H. J. Schünemann. 2011. "GRADE guidelines: 4. Rating the quality of evidence - study limitations (risk of bias)." *Journal of Clinical Epidemiology* 64 (4):407-415. doi: 10.1016/j.jclinepi.2010.07.017.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, James Woodcock, Jan Brozek, Mark Helfand, Pablo Alonso-Coello, Yngve Falck-Ytter, Roman Jaeschke, Gunn Vist, Elie A. Akl, Piet N. Post, Susan Norris, Joerg Meerpohl, Vijay K. Shukla, Mona Nasser, et Holger J. Schünemann. 2011. "GRADE guidelines: 8. Rating the quality of evidence - indirectness." *Journal of Clinical Epidemiology* 64 (12):1303-1310. doi: 10.1016/j.jclinepi.2011.04.014.

Guyatt, Gordon, A. D. Oxman, Shahnaz Sultan, Jan Brozek, Paul Glasziou, Pablo Alonso-Coello, David Atkins, Regina Kunz, Victor Montori, Roman Jaeschke, David Rind, Philipp Dahm, E. A. Akl, Joerg Meerpohl, Gunn Vist, Elise Berliner, Susan Norris, Yngve Falck-Ytter, et H. J. Schünemann. "GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes." *Journal of Clinical Epidemiolog* 66 (2):151-157.

Guzelian, P. S., M. S. Victoroff, N. C. Halmes, R. C. James, et C. P. Guzelian. 2005. "Evidence-based toxicology: a comprehensive framework for causation." *Hum Exp Toxicol* 24 (4):161-201.

Hardy, Anthony, Jean-Louis Dorne, Elisa Aiassa, Jan Alexander, Bernard Bottex, Qasim Chaudhry, Andrea Germini, Birgit Nørrung, Josef Schlatter, Didier Verloo, et Tobin Robinson. 2015. "Editorial: Increasing robustness, transparency and openness of scientific assessments." *EFSA Journal* 13 (3):3. doi: 10.2903/j.efsa.2015.e13031.

HAS. 2013. Niveau de preuve et gradation des recommandations de bonne pratique Haute Autorité de santé.

Higgins, J P T, et S Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*: The Cochrane Collaboration.

Hill, Austin Bradford. 1965. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58 (5):295-300.

Hope, B. K., et J. R. Clarkson. 2014. "A Strategy for Using Weight-of-Evidence Methods in Ecological Risk Assessments." *Human and Ecological Risk Assessment* 20 (2):290-315. doi: 10.1080/10807039.2013.781849.

Howard, BE, R Shah, K Walker, K Pelch, S Holmgren, et K Thayer. 2014. "Use of text-mining and machine learning to prioritize the results of a complex literature search." Society of Toxicology (SOT). 53rd annual meeting., Phoenix, AZ.

Hristozov, D. R., S. Gottardo, M. Cinelli, P. Isigonis, A. Zabeo, A. Critto, M. Van Tongeren, L. Tran, et A. Marcomini. 2014. "Application of a quantitative weight of evidence approach for ranking and prioritising occupational exposure scenarios for titanium dioxide and carbon nanomaterials." *Nanotoxicology* 8 (2):117-31. doi: 10.3109/17435390.2012.760013.

Hristozov, D. R., A. Zabeo, C. Foran, P. Isigonis, A. Critto, A. Marcomini, et I. Linkov. 2014. "A weight of evidence approach for hazard screening of engineered nanomaterials." *Nanotoxicology* 8 (1):72-87. doi: 10.3109/17435390.2012.750695.

IARC. 2006. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans - Preamble. Lyon, France: World Health Organization International Agency for Research on Cancer.

INCa. 2015. Nutrition et prévention primaire des cancers : actualisation des données. Institut national du cancer.

IPCS. 2004. IPCS Risk Assessment Terminology. Geneva: World Health Organization.

Jeffreys, Harold. 1961. *Theory of Probability*. Clarendon: Oxford University Press.

Kho, M. E., et M. C. Brouwers. 2012. "The systematic review and bibliometric network analysis (SeBriNA) is a new method to contextualize evidence. Part 1: description." *J Clin Epidemiol* 65 (9):1010-5. doi: 10.1016/j.jclinepi.2012.03.009.

Khosrovyan, A., A. Rodríguez-Romero, M. Antequera Ramos, T. A. DelValls, et I. Riba. 2015. "Comparative analysis of two weight-of-evidence methodologies for integrated sediment quality assessment." *Chemosphere* 120:138-144.

Klimisch, H. J., M. Andreae, et U. Tillmann. 1997. "A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data." *Regul Toxicol Pharmacol* 25 (1):1-5. doi: 10.1006/rtph.1996.1076.

Krimsky, S. 2005. "The weight of scientific evidence in policy and law." *Am J Public Health* 95 Suppl 1:S129-36. doi: 10.2105/ajph.2004.044727.

Kung, J., F. Chiappelli, O. O. Cajulis, R. Avezova, G. Kossan, L. Chew, et C. A. Maida. 2010. "From Systematic Reviews to Clinical Recommendations for Evidence-Based Health Care: Validation of Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) for Grading of Clinical Relevance." *Open Dent J* 4:84-91. doi: 10.2174/1874210601004020084.

Lewandowski, T. A., et L. R. Rhomberg. 2005. "A proposed methodology for selecting a trichloroethylene inhalation unit risk value for use in risk assessment." *Regul Toxicol Pharmacol* 41 (1):39-54. doi: 10.1016/j.yrtph.2004.09.003.

Linkov, I., D. Loney, S. Cormier, F. K. Satterstrom, et T. Bridges. 2009. "Weight-of-evidence evaluation in environmental assessment: review of qualitative and quantitative approaches." *Sci Total Environ* 407 (19):5199-205. doi: 10.1016/j.scitotenv.2009.05.004.

Linkov, I., P. Welle, D. Loney, A. Tkachuk, L. Canis, J. B. Kim, et T. Bridges. 2011. "Use of multicriteria decision analysis to support weight of evidence evaluation." *Risk Anal* 31 (8):1211-25. doi: 10.1111/j.1539-6924.2011.01585.x.

Mandrioli, D, et E K Silbergeld. 2015. "Evidence from Toxicology: The Most Essential Science for Prevention." *Environ Health Perspect*. doi: 10.1289/ehp.1509880.

Marvier, M., C. McCreedy, J. Regetz, et P. Kareiva. 2007. "A meta-analysis of effects of Bt cotton and maize on nontarget invertebrates." *Science* 316 (5830):1475-7. doi: 10.1126/science.1139208.

Marvier, Michelle. 2011. "Using meta-analysis to inform risk assessment and risk management." *Journal für Verbraucherschutz und Lebensmittelsicherheit* 6 (1):113-118. doi: 10.1007/s00003-011-0675-6.

Meek, M. E. 2008. "Recent developments in frameworks to consider human relevance of hypothesized modes of action for tumours in animals." *Environ Mol Mutagen* 49 (2):110-6. doi: 10.1002/em.20369.

Meek, M. E., A. Boobis, I. Cote, V. Dellarco, G. Fotakis, S. Munn, J. Seed, et C. Vickers. 2014. "New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis." *Journal of Applied Toxicology* 34 (1):1-18.

Meek, M. E., J. R. Bucher, S. M. Cohen, V. Dellarco, R. N. Hill, L. D. Lehman-McKeeman, D. G. Longfellow, T. Pastoor, J. Seed, et D. E. Patton. 2003. "A framework for human relevance analysis of information on carcinogenic modes of action." *Crit Rev Toxicol* 33 (6):591-653. doi: 10.1080/713608373.

Meek, M. E., C. M. Palermo, A. N. Bachman, C. M. North, et R. Jeffrey Lewis. 2014. "Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence." *J Appl Toxicol* 34 (6):595-606. doi: 10.1002/jat.2984.

Metcalfe, Dean D. 2005. "Genetically modified crops and allergenicity." *Nature Immunology* 6:857-860.

Moher, D., L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, et LA. Stewart. 2015. "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement." *Syste Rev* 4 (1):1.

Money, C. D., J. A. Tomenson, M. G. Penman, P. J. Boogaard, et R. Jeffrey Lewis. 2013. "A systematic approach for evaluating and scoring human data." *Regul Toxicol Pharmacol* 66 (2):241-7. doi: 10.1016/j.yrtph.2013.03.011.

Murad, M. H., V. M. Montori, J. P. Ioannidis, R. Jaeschke, P. J. Devereaux, K. Prasad, I. Neumann, A. Carrasco-Labra, T. Agoritsas, R. Hatala, M. O. Meade, P. Wyer, D. J. Cook, et G. Guyatt. 2014. "How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature." *Jama* 312 (2):171-9. doi: 10.1001/jama.2014.5559.

Nardo, Michela, Michaela Saisana, Andrea Saltelli, et Stefano Tarantola. 2005. Tools for Composite Indicators Building. Ispra, Italy: Joint Research Center, European Commission.

Nations Unies. 2013. Système général harmonisé de classification et d'étiquetage des produits chimiques (SGH). New York et Genève: Nations Unies.

NRC. 2014. *Review of EPA's Integrated Risk Information System (IRIS) Process*. Washington, DC: The National Academies Press.

O'Connor, Denise, Sally Green, et Julian PT Higgins. 2011. "Chapter 5: Defining the review question and developing criteria for including studies." In *Cochrane Handbook of Systematic Reviews of Intervention. Version 5.1.0*, edited by Julien PT Higgins and Sally Green. The Cochrane Collaboration.

OCDE. 2014. Users' handbook supplement to the guidance document for developing and assessing AOPs. Paris: Organisation de Coopération et de Développement Économiques.

OCDE. 2015. Scientific Advice for Policy Making: The Role and Responsibility of Expert Bodies and Individual Scientists. In *Technology and Industry Policy Papers*. Paris: OECD Publishing.

OHAT. 2015. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Research Triangle Park, NC: OHAT.

OMS. 2012. Daily iron and folic acid supplementation in pregnant women - Guideline. Geneva: World Health Organization.

Pieper, Dawid, Roland Brian Buechter, Lun Li, Barbara Prediger, et Michaela Eikermann. 2015. "Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties." *Journal of Clinical Epidemiology* 68 (5):574-583. doi: 10.1016/j.jclinepi.2014.12.009.

Rhomberg, L. R., J. E. Goodman, L. A. Bailey, R. L. Prueitt, N. B. Beck, C. Bevan, M. Honeycutt, N. E. Kaminski, G. Paoli, L. H. Pottenger, R. W. Scherer, K. C. Wise, et R. A. Becker. 2013. "A survey of frameworks for best practices in weight-of-evidence analyses." *Crit Rev Toxicol* 43 (9):753-84. doi: 10.3109/10408444.2013.832727.

Rhomberg, Lorenz. 2015. "Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology." *Risk Analysis* 35 (6):1114-1124. doi: 10.1111/risa.12206.

Rooney, A A, A L Boyles, M S Wolfe, Bucher J R, et Thayer KA. 2014. "Systematic review and evidence integration for literature-based environmental health science assessments." *Environmental Health Perspectives* (122):711-718. doi: http://dx.doi.org/10.1289/ehp.1307972

Rothman, K. J., et S. Greenland. 2005. "Causation and causal inference in epidemiology." *Am J Public Health* 95 Suppl 1:S144-50. doi: 10.2105/ajph.2004.059204.

Sackett, D. L., W. M. Rosenberg, J. A. Gray, R. B. Haynes, et W. S. Richardson. 1996. "Evidence based medicine: what it is and what it isn't." *Bmj* 312 (7023):71-2.

SCENIHR. 2012. Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty. Brussels: European Commission.

Schleier Iii, J. J., L. A. Marshall, R. S. Davis, et R. K. Peterson. 2015. "A quantitative approach for integrating multiple lines of evidence for the evaluation of environmental health risks." *PeerJ* 3:e730. doi: 10.7717/peerj.730.

Schneider, K., M. Schwarz, I. Burkholder, A. Kopp-Schneider, L. Edler, A. Kinsner-Ovaskainen, T. Hartung, et S. Hoffmann. 2009. ""ToxRTool", a new tool to assess the reliability of toxicological data." *Toxicol Lett* 189 (2):138-44. doi: 10.1016/j.toxlet.2009.05.013.

Schünemann, Holger J., Andrew D. Oxman, Gunn E. Vist, Julian PT. Higgins, Jonathan J. Deeks, P. Glasziou, Gordon H. Guyatt, et on beholf of the Cochrane Applicability and Recommandations Methods Groups. 2011. "Chapter 12: Interpreting results and drawing conclusions." In *Cochrane Handbook for Systematic Reviews of Interventions*, edited by Green S Higgins JPT. The Cochrane Collaboration.

Shea, B. J., L. M. Bouter, J. Peterson, M. Boers, N. Andersson, Z. Ortiz, T. Ramsay, A. Bai, V. K. Shukla, et J. M. Grimshaw. 2007. "External validation of a measurement tool to assess systematic reviews (AMSTAR)." *PLoS One* 2 (12):e1350. doi: 10.1371/journal.pone.0001350.

Shea, B. J., J. M. Grimshaw, G. A. Wells, M. Boers, N. Andersson, C. Hamel, A. C. Porter, P. Tugwell, D. Moher, et L. M. Bouter. 2007. "Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews." *BMC Med Res Methodol* 7:10. doi: 10.1186/1471-2288-7-10.

Shea, B. J., C. Hamel, G. A. Wells, L. M. Bouter, E. Kristjansson, J. Grimshaw, D. A. Henry, et M. Boers. 2009. "AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews." *J Clin Epidemiol* 62 (10):1013-20. doi: 10.1016/j.jclinepi.2008.10.009.

Spiegelhalter, D. J., K. R. Abrams, et J. P. Myles. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Edited by Vic Barnett Stephen Senn.

Swaen, G., et L. van Amelsvoort. 2009. "A weight of evidence approach to causal inference." *J Clin Epidemiol* 62 (3):270-7. doi: 10.1016/j.jclinepi.2008.06.013.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, et N. D. Goodman. 2011. "How to grow a mind: statistics, structure, and abstraction." *Science* 331 (6022):1279-85. doi: 10.1126/science.1192788.

US EPA. 1997. Rules of thumb for superfund remedy selection. Washington DC: U.S. Environmental Protection Agency.

US EPA. 1998. Guidelines for Ecological Risk Assessment. Washington D.C: U.S. Environmental Protection Agency.

US EPA. 2001. HED Standard Operating Porcedure: Executive Summaries for Toxicology Data Evaluation Record (DERs). Washington DC: U.S. Environmental Protection Agency.

US EPA. 2003. A summary of general assessment factors for evaluating the quality of scientific and technical information. Washington, DC: U.S. Environmental Protection Agency

US EPA. 2014. Framework for Human Health Risk Assessment to Inform Decision Making. edited by EPA Risk Assessment Forum. Washington D.C: U.S. Environmental Protection Agency.

van Bilsen, J. H., S. Ronsmans, R. W. Crevel, R. J. Rona, H. Przyrembel, A. H. Penninks, L. Contor, et G. F. Houben. 2011. "Evaluation of scientific criteria for identifying allergenic foods of public health importance." *Regul Toxicol Pharmacol* 60 (3):281-9. doi: 10.1016/j.yrtph.2010.08.024.

Vinken, M. 2013. "The adverse outcome pathway concept: a pragmatic tool in toxicology." *Toxicology* 312:158-65. doi: 10.1016/j.tox.2013.08.011.

Viswanathan, M, MT Ansari, ND Berkman, S Chang, L Hartling, LM McPheeters, PL Santaguida, T Shamliyan, K Singh, A Tsertsvadze, et JR Treadwell. 2012. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. In *Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews*: AHQR.

Vlaanderen, J., R. Vermeulen, D. Heederik, et H. Kromhout. 2008. "Guidelines to evaluate human observational studies for quantitative risk assessment." *Environ Health Perspect* 116 (12):1700-5. doi: 10.1289/ehp.11530.

WCRF/AICR. 2014. Continuous Update Project Report. Food, Nutrition, Physical Activity, and the Prevention of Ovarian Cancer 2014. World Cancer Research Fund, American Institute for Cancer Research.

Weed, D. L. 2005. "Weight of evidence: a review of concept and methods." *Risk Anal* 25 (6):1545-57. doi: 10.1111/j.1539-6924.2005.00699.x.

Williams, M. S., E. D. Ebel, et D. Vose. 2011. "Framework for microbial food-safety risk assessments amenable to Bayesian modeling." *Risk Anal* 31 (4):548-65. doi: 10.1111/j.1539-6924.2010.01532.x.

Woodruff, T. J., et P. Sutton. 2011. "An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences." *Health Aff (Millwood)* 30 (5):931-7. doi: 10.1377/hlthaff.2010.1219.

Woodruff, TJ, et P Sutton. 2014. "The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes." *Environ Health Perspect* (122):1007-1014. doi: http://dx.doi.org/10.1289/ehp.1307175.

7.2 Normes

AFNOR. 2003. NF X 50-110. Qualité en expertise - Prescriptions générales de compétence pour une expertise.

ISO. 2008. Norme ISO 9001 : Système de management de la qualité.

ANNEXES

Annexe 1 : Décision d'autosaisine

anses
agence nationale de sécurité sanitaire
alimentation, environnement, travail

2015 -SA- 0 0 8 9

Décision N° 2015-03-098

AUTOSAISINE

Le directeur général de l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses),

Vu le code de la santé publique, et notamment son article L. 1313-3 conférant à l'Anses la prérogative de se saisir de toute question en vue de l'accomplissement de ses missions,

Décide :

Article 1er.- L'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail se saisit afin de réaliser une expertise dont les caractéristiques sont listées ci-dessous.

1.1 Thématiques et objectifs de l'expertise

L'Anses s'autosaisit afin de conduire une analyse critique sur les approches d'évaluation des niveaux de preuve à l'étape d'identification des dangers.

1.2 Contexte de l'autosaisine

L'identification des dangers est une étape essentielle de l'évaluation des risques, elle a pour objectif de déterminer les effets néfastes potentiels sur la santé humaine liés à l'exposition à un agent biologique, chimique ou physique et doit s'appuyer sur les meilleures preuves scientifiques disponibles au moment de sa mise en œuvre. Il s'agit donc à cette première étape de l'évaluation des risques d'apporter une réponse qualitative (sur la probabilité qu'un effet néfaste se produise dans les populations exposées) fondée sur une analyse approfondie et documentée des données scientifiques pertinentes et extraites des bases de données appropriées, de la littérature scientifique et, le cas échéant, d'autres sources de données non publiées telles que celles provenant des professionnels. Pour ce faire, les experts doivent évaluer l'ensemble des données scientifiques pertinentes sur la base d'échelles de niveau de preuve préexistantes.

Selon les thématiques de travail couvertes par l'Agence (risques physico-chimiques, biologiques, liés à la nutrition, etc.), mais également selon les GT, au sein d'une même thématique, les pratiques actuelles des experts diffèrent en matière d'évaluation des niveaux de preuve laissant une part importante à la subjectivité. Il est donc nécessaire d'objectiver les critères permettant la fixation des niveaux de preuve et, autant que faire se peut, d'harmoniser les pratiques d'expertise à ce niveau.

1.3 Questions sur lesquelles portent les travaux d'expertise à mener

Les objectifs de cette autosaisine sont de :

- Décrire les pratiques actuelles de l'Anses et les comparer avec les pratiques actuelles d'autres organismes/agences sanitaires.
- Faire une revue critique des approches sur les niveaux des preuves.
- Proposer des procédures harmonisées pour évaluer la qualité des études et des données disponibles, ainsi que des niveaux de preuve par rapport aux questions ou hypothèses avancées.
- Proposer des procédures harmonisées pour évaluer et communiquer le niveau de preuve global associé à l'ensemble des données et études disponibles.

- Démontrer l'applicabilité des recommandations grâce à des études de cas.

1.4 Durée prévisionnelle de l'expertise : 2 ans

A l'issue de la première année, le groupe de travail publiera un rapport présentant son analyse et ses recommandations. La deuxième année sera réservée aux études de cas.

Article 2.- Un avis sera émis et publié par l'Agence à l'issue des travaux.

Fait à Maisons-Alfort, le 31 MARS 2015

Marc MORTUREUX
Directeur général

Annexe 2 : Grille de lecture

N° Ordre		2	**Lecteur**	
Année publication		4	**Type de document**	
Pages concernées				
Auteurs/Institution (Pays)				
Titre				
Statut du document				

9 ☐ Document retenu

si autre, précisez:

Domaines couverts

- ☐ Santé - Travail
- ☐ Santé-Environnement
- ☐ Microbiologie (Aliments)
- ☐ Chimie (Aliments)
- ☐ Santé et Bien être Animale
- ☐ Santé- Nutrition
- ☐ Santé et Protection des végétaux
- ☐ Autres

si autre, précisez:

11 Problématiques traitées

- ☐ Evaluation d'une étude individuelle
- ☐ Evaluation d'un ensemble d'études

Copier la définition du concept "niveau de preuve"
Lister les systèmes de cotation développés ou recommandés dans le document (Noms et références)

Types d'études couverts par le document	Oui/Non	Commentaires/descriptifs des critères pris en compte
Etudes expérimentales *in vivo*		
Etudes expérimentales *in vitro*		
Etude d'intervention chez l'homme		
Etudes épidémiologiques d'observation		
Autres		

Citer les cas d'étude jugés pertinents pour les domaines de compétence de l'agence
Limites d'application des recommandations du document
Codage du niveau de preuve
Critères retenus pour fixer chaque niveau de preuve

Commentaires (par exemple, ce qui concerne la communication, le processus, ou autre)

Annexe 3 : Liste des organismes consultés pour le recensement des guides existants sur l'évaluation du poids des preuves

Acteurs internationaux
Organisations - Agences
CIRC (Centre international de recherche sur le cancer) - Programme des Monographies
FAO (Organisation des Nations Unies pour l'alimentation et l'agriculture)
ILO (Organisation internationale du Travail)
OCDE (Organisation de Coopération et de Développement Économiques)
WHO/OMS (Organisation mondiale de la Santé) - Département de Santé publique, environnement et déterminants sociaux de la santé département de sécurité sanitaire et zoonoses
Comités d'experts
GRADE WG (The Grading of Recommendations Assessment, Development, and Evaluation Working Group) http://www.gradeworkinggroup.org/
ECETOC (Centre européen d'écotoxicologie et de toxicologie des produits chimiques)
GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat)
ILSI (International Life Sciences Institut)
ICNIRP (International Commission on Non-Ionizing Radiation Protection)
ICRP (International Commission on Radiological Protection)
IPCS (International Programme on Chemical Safety)
WCRF (World Cancer Research Fund International)
Acteurs européens
Agences
ECHA (Agence européenne des produits chimiques)
EEA (Agence européenne pour l'environnement)
EFSA (Autorité européenne de sécurité des aliments)
EMA (Agence européenne des médicaments)
ECDC (European Centre for Disease prevention and Control)
EU-OSHA (Agence européenne pour la sécurité et la santé au travail)
OEPP (Organisation européenne et méditerranéenne pour la protection des plantes)
Commission européenne
JRC (Joint Research Center)
SCCS (Scientific Committee on Consumer Safety)
SCHER (Scientific Committee on Health and Environmental Risks)
SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks)

Acteurs nationaux en Europe
Allemagne
BAuA (German Federal Agency for Occupational Safety and Health)
BfR (German Federal Institute for Risk Assessment)
BVL (Federal Office of Consumer Protection and Food Safety)
UBA (German Federal Environmental Agency)
UFZ (Helmholtz Centre for Environmental Research)
Autriche
AGES (Austrian Agency for Health and Food Safety)
Belgique
AFSCA (Belgian Federal Agency for the Safety of the Food Chain)
WIV-ISP (Scientific Institute of Public Health)
Bulgarie
Risk Assessment Center (RAC) - Bulgarian Food Safety Agency
Chypre
SGL (State General Laboratory)
Croatie
Croatian Food Agency
Danemark
DTU-Food (National Food Institute DTU)
Danish EPA
Espagne
AECOSAN (The Spanish Agency for Consumer Affairs, Food Safety and Nutrition)
Estonie
Ministry of Agriculture - Food Safety Department
Finlande
Evira (Finnish Food Safety Authority)
THL (National Institute for Health and Welfare)
Grèce
EFET (Hellenic Food Authority)
Hongrie
National Food Chain Safety Office
Islande
The Icelandic Food and Veterinary Authority
Irlande
FSAI
Italie
Istituto Superiore di Sanità (ISS)
Lettonie
Institute of Food Safety, Animal Health and Environment (BIOR)
Lituanie
National Food and Veterinary Risk Assessment Institute
Luxembourg
Ministry of Agriculture ; Ministry of Health
Malte
Malta Competition and Consumer Affairs Authority
Norvège
FHI (The Norwegian Institute of Public Health)
VKM (The Norwegian Scientific Committee for Food Safety)

Acteurs nationaux en Europe (suite)
Pays Bas
NVWA (Food and Consumer Product Safety Authority)
PBL (Netherlands Environmental Assessment Agency)
RIVM (National Institute of Public Health and the Environment) RIVM/MNP (Netherlands Environmental Assessment Agency)
IRAS (Universiteit Utrecht · Institute for Risk Assessment Sciences)
Portugal
ASAE (Portuguese Economy and Food Safety Authority)
République Slovaque
Ministry of Agriculture and Rural Development of the SR
République Tchèque
Ministry of Agriculture of the Czech Republic
Pologne
Polish EFSA Focal Point
Roumanie
National Sanitary Veterinary and Food Safety Authority
Royaume-Uni
FERA (Food and Environment Research Agency)
FSA (UK Food Standards Agency)
Centre for Mathematical Sciences (Cambridge)
Imperial College of London
MRC (Medical Research Council)
University of Durham
Slovénie
Ministry of Agriculture Forestry and Food
Suède
SLV (National Food Agency)
Suisse
FSVO (Federal Food Safety and Veterinary Office)
ETH Zurich (Swiss Federal Institute of Technology in Zurich)
Acteurs nationaux hors de l'Europe
Canada
INSPQ (Institut national de santé publique du Québec)
IRSST (Institut de recherché Robert-Sauvé en santé et en sécurité du travail)
Santé Canada
États-Unis
AHRQ (Agency for Healthcare Research & Quality)
AICR (American Institute for Cancer Research)
ATSDR (Agency for Toxic Substances and Disease Registry)
US FDA (US Food and Drug Administration)
NIOSH (CDC-National Institute for Occupational Safety and Health)
NIEHS (NIH – National Institute of Environmental Health Sciences)
US NRC (United States Nuclear Regulatory Commission)
US EPA (Environmental Protection Agency)
NCEA (US EPA – National center for environmental assessment)
NAS (National Academies – National Academy of Sciences)- IOM (National Academies - Institute of Medicine of the National Academies) – NRC (National Academies – National Research Council)
Nouvelle Zélande
ANZSA

Annexe 4 : Documents de référence

Tableau 17 : Documents retenus suite à la recherche bibliographique

Nom de l'agence /Auteurs	Année	Titre du guide/article
		Recherche bibliographique préliminaire
Efsa	2010	Application of systematic review methodology to food and feed safety assessments to support decision making
Anses	*2016 en cours*	Rapport d'étude : Traitement de l'incertitude dans le processus d'évaluation des risques sanitaires des substances chimiques Groupe de travail « Perturbateurs Endocriniens et reprotoxiques de catégorie 3 »
Anses	2013	Avis + rapport « Radiofréquences et santé »
WHO	2012	Guideline: Daily iron and folic acid supplementation in pregnant women
World Cancer Research Fund (WCRF) / American Institute for Cancer Research (AICR)	2014	Continuous Update Project : Ovarian Cancer 2014 Report, Appendix - Criteria for grading evidence (mais idem pour toutes les expertises du WCRF depuis le rapport de 2007)
Institut National du Cancer (INCa)	*2015*	Nutrition et prévention primaire des cancers : actualisation des données scientifiques
Haute Autorité de Santé	2013	Niveau de preuve et gradation des recommandations de bonne pratique
Office of Health Assessment and Translation (OHAT)	2014	Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration
Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR)	2012	Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty
National Research Council (NRC)	2014	Review of EPA's Integrated Risk Information System (IRIS) Process
Berkman et al. pour l'Agency for Healthcare Research and Quality (AHRQ)	2012	Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update
The Cochrane Collaboration	2012	Cochrane Handbook for Systematic Reviews of Interventions
Meek et al.	2014	Mode of action human relevance (species concordance) framework : Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence
Meek et al.	2014	New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis.
Swaen et al.	2009	Article : A WOE approach to causal inference
Krimsky et al.	2005	Article : The weight of scientific evidence in policy and laws
Hristozov et al.	2014	Article : A weight of evidence approach for hazard screening of engineered nanomaterials
		Recherche bibliographique approfondie
Linkov et al.	2011	Use of multicriteria decision analysis to support weight of evidence evaluation
Vinken	2013	The adverse outcome pathway concept: A pragmatic tool in toxicology
Bilotta, Milner et Boyd	2014	On the use of systematic reviews to inform environmental policies
Marvier	2011	Using meta-analysis to inform risk assessment and risk management
Rhomberg	2014	Hyopthesis-Based Weight of Evidence : An Approach to Assessing Causation and its Application to Regulatory Toxicology
Schleier Iii et al.	2015	A quantitative approach for integrating multiple lines of evidence for the evaluation of environmental health risks
van Bilsen et al.	2011	Evaluation of scientific criteria for identifying allergenic foods of public health importance

Goodman et al.	2010	Using systematic reviews and meta-analyses to support regulatory decision making for neurotoxicants: Lessons learned from a case study of PCBs
Guyatt et al.	2011	GRADE guidelines: 4. Rating the quality of evidence - Study limitations (risk of bias)
Hristozov et al.	2014	Application of a quantitative weight of evidence approach for ranking and prioritising occupational exposure scenarios for titanium dioxide and carbon nanomaterials
Rhomberg	2013	A survey of frameworks for best practices in weight-of-evidence analyses
Kho et Brouwers	2012	Application of the systematic review and bibliometric network analysis (SeBriNA) methodology contextualizes evidence.
Khosrovyan et al.	2015	Comparative analysis of two weight-of-evidence methodologies for integrated sediment quality assessment
Adami et al.	2011	Toxicology and Epidemiology: Improving the Science with a Framework for Combining Toxicological and Epidemiological Evidence to Establish Causal Inference
Woodruff et Sutton	2014	The Navigation Guide. The Navigation Guide Systematic Review Methodology: A Rigorous and Transparent Method for Translating Environmental Health Science into Better Health Outcomes
Gosling et al.	2013	A Bayes Linear Approach to Weight-of-Evidence Risk Assessment for Skin Allergy
Hope et Clarkson	2014	A Strategy for Using Weight of Evidence Methods in Ecological Risk Assessment
Murad et al	2014	How to Read a Systematic Review and Meta-analysis and Apply the Results to Patient Care Users' Guides to the Medical Literature
Pieper et al.	2015	Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties.
CIRC	2006	IARC Monographs on the Evaluation of Carcinogenic Risks to Humans - PREAMBLE
Moher et al.	2015	Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) statement
Kung et al	2010	From Systematic Reviews to Clinical Recommendations for Evidence-Based Health Care: Validation of Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) for Grading of Clinical Relevance.

Tableau 18 : Guides supplémentaires retenus suite au questionnaire envoyé aux agences

(après suppression des doublons)

Nom de l'agence	Année	Titre du guide
Santé Canada	2011	Poids de la preuve : Facteurs à considérer pour la prise de mesures appropriées et en temps opportun dans une situation d'enquête sur une éclosion de maladie d'origine alimentaire
OCDE	2014	New guidance document on an integrated approach on testing and assessment for skin corrosion and irritation - part 2 Weight of evidence analysis
ECHA	2010	How to report weight of evidence

Annexe 5 : Définitions WOE issues de la recherche bibliographique

Définition "Transparent framework for drawing conclusions"	Référence
When we refer to "WoE frameworks," we mean approaches that have been developed for taking the process all the way from scoping of the assessment and initial identification of relevant studies through the drawing of appropriate conclusions.	Rhomberg et al 2013
Walker cites three objectives of a WOE analysis: (1) it provides a "clear and transparent framework" for evaluating the evidence in a risk determination; (2) it offers regulatory agencies a consistent and standardized approach to evaluating toxic substances; (3) it helps to identify the discretionary assumptions in risk determinations from experts.	Krimsky et al., 2005
The term is often used by EPA in the context of a WOE "narrative." In the case of a carcinogenic risk assessment, the narrative consists of a short summary that "explains what is known about an agent's human carcinogenic potential and the conditions that characterize its expression" (EPA 2011). In EPA's Guidelines for Carcinogen Risk Assessment, the WOE narrative "explains the kinds of evidence available and how they fit together in drawing conclusions, and it points out significant issues/strengths/limitations of the data and conclusions" (EPA 2005, p. 1-12).	EPA 2011 ; EPA 2005
A process or method in which all scientific evidence that is relevant to the status of a causal hypothesis is taken into account.	Krimsky et al., 2005

Définition "Process of considering the strengths and weaknesses of various pieces of information"	
The process of considering the strengths and weaknesses of various pieces of information in order to inform a decision being made among competing alternatives	Hope and Clarkson 2014
The present committee found that the phrase weight of evidence has become far too vague as used in practice today and thus is of little scientific use. In some accounts, it is characterized as an oversimplified balance scale on which evidence supporting hazard is placed on one side and evidence refuting hazard on the other. That analogy neglects to account for the total weight on either side (that is, the scope of evidence available) and captures only where the balance stands. Others characterize WOE as a single scale, and different kinds of evidence have different weights. For example, a single human study with low risk of bias might be considered as providing the same evidential weight as three well-conducted animal studies. The weights might be adjusted according to the quality of the study design. This analogy neglects to account for the "weight for" vs the "weight against" hazard. Perhaps the overall idea of the WOE for hazard should combine both characterizations. It is evident, however, that its use in the literature and by scientific agencies, including EPA, is vague and varied.	National Academy of Sciences, 2014

Définition "Integration of different lines of evidences"	
Includes all varieties of evidence, positive and negative, mechanistic and nonmechanistic, in vivo and in vitro, as well as human and animal studies.	Krimsky et al., 2005
Several well-established methods of evidence-based research synthesis:the hierarchy of research study designs, the systematic narrative review, meta-analysis, and application of so-called causal criteria. Our approach to WOE included the idea that all (rather than some) of the evidence would be considered, emphazing (i.e. putting more "weight") studies that tested the scientific hypotheses better than others.	Alexandre et al, 2011
Weight of evidence (WOE) can be defined as a framework for synthesizing individual lines of evidence, using methods that are either qualitative (examining distinguishing attributes) or quantitative (measuring aspects in terms of magnitude) to develop conclusions regarding questions concerned with the degree of impairment or risk. In general, qualitative methods include presentation of individual lines of evidence without an attempt at integration, or integration through a standardized evaluation of individual lines of evidence based on qualitative considerations. Quantitative methods include integration of multiple lines of evidence using weighting, ranking, or indexing as well as structured decision or statistical models.	Linkov et al, 2009
WOE = a methodology with a simple premise: that all available evidence should be examined and interpreted (Weed 2005). Cet article se focalise sur le poids des preuves des études épidémiologiques et considère que approche de type "poids des preuves" n'est possible que si les études considérées utilisent toutes les mêmes méthodes (ou des méthodes voisines).	Goodman et al, 2010
Integration of different lines of evidence (chemical concentrations, toxicological responses, in situ surveys) lies at the basis of the WOE approach	Khosroyan et al., 2015
the concept of weight of evidence (WOE) integrates data from different studies, or lines of evidence (LOEs), that address questions relating to the presence of chemical pollutants, their bioavailability, and the onset of adverse effects at different levels of biological organization, i.e. from a molecular level to organism or community effects (Chapman and Hollert, 2006)	Piva et al., 2011
Sometimes the phrase weight of evidence is invoked when a reviewer has simply drawn her or his own conclusions about a series of studies without any formal analytical tools, whereas on other occasions weight of evidence is used to describe a rigorous quantitative synthesis of effect size from multiple experiments.	Marvier, 2011
In short, a WOE approach is a synthetic process that combines the information content of multiple weighted pieces of evidence" (Suter and Cormier 2011).	Hope and Clarkson 2014 citant Suter and Cormier 2011

Autres définitions	
combining lines of evidence of varying quality in a risk assessment	Gosling et al., 2013
WOE consists of a diverse set of methods, often built for particular applications	Linkov et al, 2011

Annexe 6 : Définitions de « revue systématique » dans la littérature

Definitions de "systematic review"	Auteurs
The application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic. Meta-analysis may be, but is not necessarily, used as part of the process	Chalmers et al, 2002
Rhomberg et al. (2013), in a review article surveying best practices for WOE frameworks or analyses, describe WOE as encompassing all of causal inference […] They then describe the wide array of meanings attached to the phrases systematic review and weight of evidence as follows: Some terms are used differently in different frameworks. In particular, to some practitioners, the term "systematic review" refers specifically to the systematic assembly of evidence (for example, by using explicit inclusion and exclusion criteria or by using standard tabulation and study-by-study quality evaluation), while "weight of evidence" refers to the subsequent integration and interpretation of these assembled selected studies/data as they are brought to bear on the causal questions of interest. To others, "systematic review" refers to the whole process from data assembly through evaluation, interpretation, and drawing of conclusions; for still others, this whole suite of processes is subsumed under WoE. …when we refer to "WoE frameworks," we mean approaches that have been developed for taking the process all the way from scoping of the assessment and initial identification of relevant studies through the drawing of appropriate conclusions.	Rhomberg et al, 2013
Systematic review methodology can be implemented to answer well-formulated specific questions generated by the risk assessment process or by other analytical frameworks developed in food and feed safety in a transparent, reproducible, and rigorous evidence-based way. However, several aspects must be considered in order to decide whether specific questions obtained by simplifying broad food or feed safety policy problems are suitable for systematic review. A useful means of determining whether a question is answerable by SR is to identify the structure of the question. If the question structure can be specified in such a way that a particular primary research study design can be envisaged that would answer the question, then it is likely that a systematic review would be appropriate.	EFSA, 2010
In evidence-based policy-making in the healthcare field, systematic review (SR) processes are used in order to tackle these issues, helping to present a comprehensive, policy-neutral, transparent and reproducible synthesis of the evidence.	Bilotta et al, 2014
Systematic reviews address a specific clinical question using specific methodologies to minimize bias and improve the precision of estimates (i.e., reduce random error).	M. E. Kho and M. C. Brouwers, 2012
A systematic review is a research summary that addresses a focused clinical question in a structured, reproducible manner. It is often, but not always, accompanied by ameta-analysis, which is a statistical poolingor aggregation of results from different studies providing a single estimate of effect.	Murad et al., 2014
A well conducted systematic review addresses a carefully formulated question by analyzing all available evidence. It employs an objective search of the literature, applying predetermined inclusion and exclusion criteria to the literature, critically appraising what is found to be relevant. It then extracts and synthesizes data from the available evidence base to formulate findings	Shea et al., 2007
Systematic Review definition [http://www.nlm.nih.gov/nichsr/hta101/ta101014.html] (National Information Center on Health Services Research and Health Care Technology (NICHSR) – HTA: Health Technology Assessment) : a form of structured literature review that addresses a question that is formulated to be answered by analysis of evidence, and involves objective means of searching the literature, applying predetermined inclusion and exclusion criteria to this literature, critically appraising the relevant literature, and extraction and synthesis of data from evidence base to formulate findings.	Shea et al., 2007
Systematic reviews (SRs) are the cornerstone of evidence based health care as they can provide the highest level of evidence.	Pieper et al., 2015
Systematic reviews are: the reference standard for synthesizing evidence in health care because of their methodological rigor; based on pre-defined eligibility criteria and conducted according to a pre-defined methodological approach as outlined in an associated protocol.	ECHA, 2010

Annexe 7 : Fiches méthodes

1) Documents couvrant l'ensemble des étapes de l'évaluation du poids des preuves .....

1.1 Processus d'évaluation du poids des preuves de l'OHAT (2015) 69

1.2 Processus d'évaluation du poids des preuves Hope et Clarkson (2014) 72

1.3 Processus d'évaluation du poids des preuves du NRC (2014) 73

1.4 Processus d'évaluation du poids des preuves de Rhomberg (2013) 74

1.5 Processus d'évaluation du poids des preuves du SCENIHR (2012) 75

2) Revue systématique et évaluation de la qualité .....

2.1 Revue systématique EFSA (2010) 77

2.2 Revue systématique Collaboration Cochrane (2011) 79

2.3 Revue bibliographique appliquée à l'environnement Bilotta (2014) 81

2.4 Revue systématique « Navigation guide » (Woodruff et Sutton 2011) 82

2.5 Évaluation d'une revue systématique par l'outil AMSTAR (2007) 83

2.6 Evaluation d'une revue systématique du domaine médical Murad et al (2014) 83

2.7 Lignes directrices pour la rédaction d'articles scientifiques 84

3) Méthodes qualitatives, quantitatives ou semi-quantitatives .....

3.1 Critères de Bradford Hill (1965) 85

3.2 GRADE (2011) 86

3.3 OMS-CIRC (WHO-IARC en anglais, IARC 2006) 88

3.4 WCRF/AICR (WCRF/AICR 2007) 90

3.5 Klimisch (1997) 92

3.6 Evaluation du poids des preuves fondée sur les hypothèses (Rhomberg 2014) 93

3.7 Méta-analyse 93

3.8 Analyse multicritères (Linkov et al, 2011) 94

3.9 Méthodes statistiques Bayésiennes 95

3.10 Evaluation comparative du poids des preuves (Meek et al, 2014) 96

3.11 Hristozov et al (2014) 97

3.12 Arbre de décision 97

1.1 Processus d'évaluation du poids des preuves de l'OHAT (2015)

Pour le domaine de la santé-environnement, le document de l'OHAT (2015) propose un processus d'évaluation du poids des preuves en sept étapes : « Formulation du problème et développement du protocole » ; « Recherche et sélection des études en vue de leur inclusion » ; « Extraction des données des études » ; « Évaluation de la qualité des études individuelles » ; « Synthèse des preuves et notation de la confiance » ; « Traduction des niveaux de confiance en niveaux de preuves pour les effets sanitaires donnés » ; « Intégration des preuves pour développer des conclusions d'identification du danger ».

Démarche pour la planification de l'évaluation

Le document de l'OHAT (2015) considère la formulation de la question au sein de son protocole de conduite d'une revue systématique et d'intégration des preuves. En préalable, le protocole demande de spécifier le cadrage de l'étude, processus consistant à solliciter les agences nationales (ou fédérales), le public et les parties prenantes pour définir le cadre de l'évaluation : personnes impliquées, domaine d'action et recherches en cours. La formulation du problème, consistant à renseigner la structure d'information PECO (P : Population, Exposition, Comparateur, et Effet) en collaboration avec un spécialiste du traitement de l'information, permet de définir le processus de revue de la littérature en termes de la stratégie de recherche, de critères d'inclusion/exclusion, de types de données extraites des études, de stratégie de synthèse et de compte-rendu des résultats. Le recours à un outil de *text-mining*, à l'exemple de SWIFT (Howard et al. 2014), pour conduire une revue de littérature préliminaire (sans collecte des informations) et accéder à la densité et profondeur de la bibliographie, permet d'alimenter cette étape. Ces informations sont enfin rassemblées dans un document, appelé « *Concept Document* », constitué des sections suivantes : Objectif général ; Contexte avec historique des saisines et aperçu des données d'exposition humaine et d'effets sanitaires ; Les critères PECO ; Portée du document (utilisation prévue et format) et un résumé des activités de formulation du problème (cadrage ; considération des contributions scientifiques et du public ; considération des évaluations en cours dans d'autres institutions ; considération des problématiques scientifiques clés et des domaines de complexité). OHAT (2015) indique que la difficulté de mise en œuvre de cette partie porte sur les discussions nécessaires à la production du résumé des activités de formulation du problème.

La construction du protocole repose principalement sur deux travaux conduits dans le domaine santé-environnement :

- Rooney et al. (2014) propose d'utiliser les critères PECOTS au lieu de PECO au sein du protocole : populations ; expositions ; comparateurs ; effets ; temps; localisation d'intérêt.
- NRC (2014) propose une matrice de spécification des effets à considérer pour la formulation du problème en lieu et place de PECO. Cette matrice est constituée de trois colonnes (études humaines *in vivo*, études animales in vivo et études mécanistiques in vitro) et dix-huit lignes (Génotoxicité, Oncogenèse, Reproduction, Neurologie, Endocrinien, etc.). Son remplissage consiste à préciser (1) le composé chimique, le processus ou la mixture à évaluer, (2) le type d'étude d'intérêt et (3) les effets d'intérêt ainsi que le système d'organe potentiellement affecté. La limite de cette approche est que son utilisation est circonscrite aux thématiques situées au croisement des lignes et des colonnes.

Démarche pour l'établissement de lignes de preuve

- *Sélection des études (individuelles ou de synthèse) et l'extraction des données*

L'OHAT renvoie aux protocoles génériques de revue systématique pour ce qui est de la méthodologie de la recherche bibliographique. L'OHAT recommande de documenter la stratégie de recherche bibliographique (termes, base de données consultées, date) afin que la recherche bibliographique soit reproductible. Ensuite, l'OHAT prévoit des critères préliminaires d'inclusion et d'exclusion pour la sélection des études par rapport aux questions clés identifiées préalablement au moyen de la structure d'information PECO. C'est-à-dire que si des limites méthodologiques majeures ne permettant pas de répondre aux questions clés sont connues avant l'évaluation de la qualité des études individuelles, ces facteurs peuvent être utilisés comme critères d'exclusion (par exemple, des méthodes ne permettant pas d'évaluer l'exposition ou l'effet sanitaire).

Ces critères, *in fine* (notamment les critères d'inclusion), permettent de diriger la recherche de littérature. Le document liste des exemples de critères d'inclusion pour les différents aspects de PECO (population ; exposition ; comparateurs et effet) et les différents types d'études (humaines, animaux, mécanistiques). L'OHAT recommande que deux *examinateurs* effectuent de manière indépendante le screening des titres et résumés. Enfin, l'OHAT prévoit la consultation des parties prenantes afin d'identifier des études qui ont pu être oubliées. En ce qui concerne l'extraction des données, le but est de résumer le design et les résultats des études afin de faciliter l'évaluation de leur qualité et/ou conduire des analyses statistiques. Les éléments (minimaux) à extraire sont proposés par l'OHAT pour chaque type d'études (humain, animaux in vivo, in vitro). A chaque fois, la source de financement et les conflits d'intérêts déclarés sont précisés. L'OHAT gère l'extraction des données sous des formes structurées dans des bases de données en utilisant des logiciels spécifiques, suivant l'objectif et la complexité du sujet. Il est recommandé que deux examinateurs travaillent indépendamment pour extraire les données de chaque étude : un examinateur extrait les données et le second vérifie l'extraction. Les désaccords sont discutés entre examinateurs et un membre additionnel peut alors intervenir. Lorsque des données sont manquantes, l'OHAT essaie de joindre les auteurs de l'étude.

- *Évaluation de la qualité des études individuelles*

L'OHAT évalue la validité interne, appelée « risque de biais » d'une manière qualitative et propose une notation de chaque étude selon 3 niveaux. Pour cela, l'OHAT développe un outil consistant en 11 questions relevant des bonnes pratiques de recherche à considérer selon le type d'étude. Les questions concernent les biais de sélection, de confusion, de performance, d'attrition/d'exclusion, de détection, de reporting et d'autres biais. Pour chaque question considérée, une réponse est à donner en termes de risque de biais (bas, probablement bas, probablement élevé, élevé). Ces réponses sont néanmoins parfois subjectives. Les études peuvent être classées selon trois niveaux (niveau 1, 2 ou 3) en fonction du nombre de réponse de risque de biais bas ou élevé. Tous les types d'études sont considérés avec cette méthode. L'exclusion des études en fonction de leur qualité (par exemple les études de niveau 3) doit être réalisée au cas par cas.

- *Établissement lignes de preuves*

L'OHAT propose plusieurs méthodes pour analyser un ensemble d'étude par type d'étude :
- la méta-analyse, lorsque c'est pertinent
- l'évaluation de la confiance en un corpus de preuves, suivi de la traduction en niveau de preuves pour chaque ligne de preuve

L'évaluation de la confiance repose sur une version affinée de GRADE (affinée pour prendre en compte les spécificités des études animales in vivo ; in vitro ; et des études humaines observationnelles). Un niveau de confiance initial est évalué sur la base de caractéristiques principales (exposition contrôlées, exposition précédent l'effet sanitaire ; données individuelles d'effet ; utilisation de contrôles), et augmenté ou diminué sur la base d'autres facteurs (diminué si risque de biais, hétérogénéité des résultats, caractère indirect des données scientifiques ou manque d'applicabilité ; imprécision des données, et biais de publication ; et augmenté si association forte ; gradient dose-réponse ; cohérence entre différent types d'études ; et prise en considération de facteurs de confusion ou d'autres facteurs qui augmentent la confiance dans une association ou un effet). La confiance est cotée en 4 niveaux :
- "**high confidence** in the association between exposure to the substance and the outcome. The true effect is highly likely to be reflected in the apparent relationship"
- "**moderate confidence** in the association between exposure to the substance and the outcome. The true effect is may be reflected in the apparent relationship"
- "**low confidence** in the association between exposure to the substance and the outcome. The true effect is highly likely may be different from the apparent relationship"
- "**very low confidence** in the association between exposure to the substance and the outcome. The true effect is highly likely to be different from the apparent relationship."

Ces niveaux de confiance sont ensuite traduits en niveaux de preuve d'un effet sanitaire selon l'existence ou non d'un effet sanitaire. Les niveaux de preuves sont de cinq catégories : "high level of evidence for health effect" ; "moderate level of evidence for health effect" ; "low level of evidence for health effect" ; "inadequate level of evidence for health effect" ; "evidence of no health effects".

Démarche pour l'intégration des lignes de preuves

Pour évaluer le poids des preuves global, l'OHAT propose une matrice croisant les niveaux de preuve pour l'existence d'un effet sanitaire dans les études chez l'humain et les études chez l'animal. Les études mécanistiques peuvent être utilisées pour augmenter ou abaisser la catégorie de poids des preuves. Des facteurs à considérer lors de la prise en compte de ces données mécanistiques sont listées. La matrice proposée est relativement proche de celle utilisée par le CIRC.

Démarche pour l'expression des conclusions

L'OHAT établit la conclusion dans la septième et dernière étape de son processus. La communication des conclusions consiste à inscrire le niveau de preuve dans une des cinq catégories suivantes : « Known to be a hazard to humans » (avérée), « Presumed to be a hazard to humans » (présumée), « Suspected to be a hazard to humans » (suspectée), « Not classifiable as a hazard to humans » (non classable) et « Not identified as a hazard to humans » (non identifiée dangereuse pour l'humain). Un changement de catégorie (catégorie supérieure ou inférieure) peut être effectué dans le cas où des données mécanistiques biologiques soutiendraient fortement la relation entre exposition et effet sur la santé. Cette classification est étayée par des représentations graphiques des résultats intermédiaires.

Le mode de catégorisation adopté par l'OHAT est issu du SGH, le système harmonisé d'étiquetage des matières dangereuses développé par les Nations Unies (Nations Unies 2013). Le SGH considère vingt-neuf classes de matières : seize physiques (explosifs, aérosols, etc.), dix pour la santé humaine (peau, yeux, etc.) et deux pour l'environnement (aquatique et ozone). Chacune de ces classes comporte une classification propre établie en regard de critères semi-quantitatifs et/ou qualitatifs.

1.2 Processus d'évaluation du poids des preuves Hope et Clarkson

Dans le cadre d'une commande de l'EPA (U.S. Environmental Protection Agency), Hope et Clarkson (2014) ont proposé un protocole de conduite d'évaluation du poids des preuves pour l'évaluation prédictive des risques écologiques en cinq étapes. L'évaluation prédictive des risques écologiques consiste à utiliser les informations sur les relations de causes à effet pour estimer la probabilité qu'une action (de gestion) X effectuée mène à un effet adverse Y.

Démarche pour la planification de l'évaluation

La première étape recommandée par Hope et Clarkson, celle de « Planning and Scoping phase », conduite par les gestionnaires de risque, les décideurs et les parties prenantes, a pour objet de définir les objectifs de gestion écologique en des termes neutres, précis et mesurables. La seconde étape « Problem formulation phase », conduite par les évaluateurs, en interaction avec les acteurs de la première phase, a pour objet l'identification des objectifs et des valeurs environnementales à protéger (espèce, ressource écologique ou habitat), la construction d'un modèle conceptuel exprimé en termes statistiques et présentant la façon selon laquelle une succession d'hypothèses peut nuire aux valeurs environnementales, et enfin le développement du plan d'analyse des lignes de preuve. Bien que les contenus, les intervenants et l'interaction soient proposés, ces auteurs ne proposent pas de méthode ou de modèle de formulation des contenus.

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études individuelles*

La méthode décrite par Hope et Clarkson est qualitative, avec notation. Chaque étude est évaluée suivant 5 ou 10 critères (5 critères pour les études concernant les niveaux d'exposition et les caractéristiques ; 10 pour les études concernant les effets). Chaque critère est étayé de façon plus précise par des questions. Les critères généraux à tout type d'étude sont les suivants : la qualité des études, l'utilisation de méthodes standardisées, la spécificité du site d'étude, la représentation spatiale, la représentation temporelle. Les critères supplémentaires pour les études concernant les effets sont les suivants : l'association entre l'attribut et l'effet, la relation entre l'exposition et l'effet ; la sensibilité de l'agent ; la spécificité de l'agent ; la quantification de la réponse. Chaque critère est noté 0 (échec) ou 1 (passe). Une étude a donc un maximum de 5 ou de 10 points. Un texte narratif justifie le score sur la base des questions associées à chaque critère. Le choix de faire passer ou non un critère revient à l'évaluateur ainsi que celui de retenir *in fine* ou pas une étude selon son « poids », c'est-à-dire la somme des critères passés.

Démarche pour l'évaluation du poids des preuves

Hope et Clarkson résonnent en termes de groupes d'évidence, associés à une hypothèse de risque (hypothèses concernant les effets du développement d'une technologie par exemple). La combinaison des lignes d'évidence pour un effet et pour une exposition forment un groupe d'évidence. Le poids attribué à un groupe d'évidence correspond à la somme des poids des différentes lignes d'évidence retenues (ligne d'évidence pour un effet, une exposition et éventuellement une caractéristique) ramenée sur une échelle de 1 à 5. Chaque groupe d'évidence obtient donc un poids.

Démarche pour l'expression des conclusions

Hope et Clarkson proposent de conclure en croisant les poids des groupes d'évidence avec les estimations de risque pour chaque groupe d'évidence, pour chaque hypothèse. Les poids d'évidence sont répartis sur un continuum : de 1 à 2 (faible); de 2 à 3 ; de 3 à 4 et de 4 à 5 (fort) et les estimations de risques exprimés en terme de probabilités (<5% = faible ; 5-24% ; 25 à 50% ; 51 à 75% ; > 75%). Cela revient à conclure le niveau de preuve d'une certaine probabilité d'effet adverse sur une population du fait d'une hypothèse concernant le facteur en cause et le niveau de preuve d'une certaine probabilité d'effet adverse sur une population du fait d'une autre hypothèse. Cela constitue une aide à la détermination des mesures de gestions appropriées.

1.3 Processus d'évaluation du poids des preuves du NRC (2014)

Pour les domaines de la santé-environnement et des risques chimiques dans les aliments, le NRC (2014) propose, dans la cadre de sa revue du système d'identification des dangers de l'EPA (le "Integrated Risk Information System"), un processus d'évaluation du poids des preuves en six étapes. Le document se concentre essentiellement sur la revue systématique.

Démarche pour la planification de l'évaluation

Le document du NRC prévoit une étape de formulation du problème et une étape de développement du protocole de la revue systématique. Selon le NRC, l'enjeu de la formulation du problème est d'identifier les effets à considérer dans l'étude. Cette sous-étape comporte trois parties : (1) identifier dans la littérature les effets associées à la substance chimique étudiée via une revue systématique large en collaboration avec un spécialiste du traitement de l'information ; (2) construire un tableau des effets à prendre en considération, dans lequel les études identifiées sont placées selon le types d'étude (in vivo, in vitro, etc.) et les effets sanitaires répertoriés (ex : caractère génotoxique ou mutagène, effets sur développement, effets rénaux, effets endocriniens, effets cardiovasculaires…) et (3) au vue des éléments de la matrice, déterminer pour quel(s) effet(s) sanitaire il est pertinent de réaliser une revue systématique et formuler les questions de revue systématique.

Une fois les questions de revue systématique précisées, le NRC recommande de développer un protocole pour chaque revue, précisant le processus et les méthodes utilisées dans la revue systématique. Le protocole inclut généralement les éléments suivants :

- la question de revue systématique,
- les méthodes d'inclusion et d'exclusion des études, en précisant le type d'étude (ex : études observationnelles sur humain, études expérimentales sur animaux, étude mécanistiques in vitro), le type de voie d'exposition (ex : orale ou inhalation) et le type d'effet (ex : neurotoxique, ou toxique pour le développement),
- les méthodes de recherche pour l'identification des études,
- les méthodes pour évaluer le risque de biais et autres caractéristiques méthodologiques des études sélectionnées,
- les méthodes de collecte des données,
- les méthodes d'analyse.

Démarche pour l'établissement de lignes de preuve

- *Sélection des études (individuelles ou de synthèse) et l'extraction des données*

Le NRC recommande de mener des revues systématiques avec des stratégies de recherche et des formats d'extraction standardisés, ainsi que d'examiner l'applicabilité des standards de l'Institut de Médecine (IOM) aux évaluations menées par l'EPA. Le NRC insiste sur la nécessité de disposer d'un spécialiste du traitement de l'information et en encourageant la présence de deux examinateurs indépendant pour rechercher et sélectionner les études.

- *Évaluation de la qualité des études individuelles*

Le NRC recommande d'évaluer les risques de biais (validité interne) pour chaque type d'étude, sans recommander une liste spécifique d'éléments à examiner.

Démarche pour l'expression des conclusions

Le NRC recommande de conduire de façon systématique une analyse d'incertitude.

1.4 Processus d'évaluation du poids des preuves de Rhomberg (2013)

Rhomberg et al. (2013) présente une revue de cadres de travail dédiés à l'évaluation du poids des preuves. Cinquante-quatre cadres de travail ont été recensés, issus d'organisations gouvernementales, d'organisations internationales, de consortiums et autres. Une grille d'analyse indique pour chacun de ces cadres de travail les méthodes proposées pour les 4 phases d'évaluation du poids des preuves proposées par les auteurs : (1) définition de la question causale et développement de critères pour la sélection des études ; (2) développement et application des critères pour la revue des études individuelles ; (3) évaluation et intégration des preuves ; et (4) développement des conclusions sur la base des inférences identifiées. Le croisement du contenu des 54 cadres de travail a permis d'identifier différentes fonctionnalités pour chacune de ces étapes :

- Etape 1 : Définition de la question causale ou de l'hypothèse, identification des critères d'inclusion des études, élaboration du plan de recherche bibliographique, élaboration des stratégies de recherche bibliographique, recherche et sélection des études.
- Etape 2 : Extraction des caractéristiques des études, extraction des données, évaluation de la qualité des études, évaluation de la qualité des études par catégorie, évaluation des résultats des études individuelles, évaluation de la pertinence des catégories d'étude, considération d'autres facteurs.
- Phase 3 : Considération de facteurs clés transversaux aux études, évaluation des effets contradictoires, évaluation du mode d'action, évaluation de l'intérêt du mode d'action vis-à-vis de l'homme, évaluation de la relation dose-réponse pour des évènements clés du mode d'action, intégration des données négatives ou nulles dans l'évaluation, évaluation des données au sein et entre les domaines de preuve, évaluation quantitative du poids des preuves, considération d'observations non publiées, formulation de la conclusion du poids des preuves, proposition d'étapes suivantes.
- Phase 4 : Utilisation d'inférences ou de conclusions pour recommander des pratiques spécifiques, des recherches à mener ou des mesures de gestion, proposition de recommandations pour l'évaluation du risque, et utilisation d'inférence ou de conclusions pour proposer des catégories d'évaluation en regard d'un jeu de critères causaux définis (connu, apprécié, non-apprécié ou incertain).

1.5 Processus d'évaluation du poids des preuves du SCENIHR (2012)

Le document du SCENIHR (2012) aborde l'évaluation du poids des preuves et de l'incertitude conjointement, dans une large variété de domaines couvrant les risques sanitaire et environnementaux émergents ou nouveaux (ex : nanotechnologies ; antibiorésistance ; cancers d'organes endocriniens, bruit, champs éléctromagnétiques, effets cumulés et synergiques).

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études individuelles*

Dans le document du SCENIHR, une matrice est proposée pour évaluer chaque étude : cette matrice croise la pertinence et la qualité. Le document précise que tous les types d'études (humaine ou animales, in vivo et in vitro, expérimentale, observationnelle ou de modélisation…) doivent être évalués de la même manière. Le document liste quatre thématiques clés à évaluer (caractérisation de l'agent en question, bien-fondé et adéquation (à la question posée) de la méthodologie utilisée, reproductibilité des résultats entre différentes études, disponibilité des détails de la méthodologie, pertinence des données pour un effet particulier) à aborder lors de cette évaluation, sans formuler de questions précises. Ainsi, chaque étude est caractérisée selon quatre catégories de qualité (colonnes de la matrice) et trois catégories de pertinence (lignes de la matrice).

Les catégories de qualité sont :

- Good Scientific quality. The study is considered to be appropriately designed, conducted and reported, and to have used valid methodology.
- Adequate/utilizable scientific quality but with significant limitations. The study is scientifically acceptable but there are some important deficiencies in the design and/or conduct and /or the reporting of the experimental findings.
- Inadequate scientific quality. There are serious concerns about the design and/or conduct of the study.
- Not assignable. The study is lacking insufficient detail to make an evaluation.

Les catégories de pertinence sont :

- Direct relevance. The study addresses the specific agent (stressor), model and outcome of interest
- Indirect relevance. The study concerns a related agent (stressor), model or outcome of interest
- Insufficient relevance. The study cannot be used for the purposes of the risk assessment.

- *Établissement des lignes de preuves*

Dans le document du SCENIHR, des critères (questions) sont spécifiés par type d'étude[10] afin d'établir le poids des lignes de preuves. Une matrice est proposée pour évaluer chaque ligne de preuve, croisant l' « utilité » et la « cohérence ».

L'utilité est une combinaison de la qualité et de la pertinence (décrits ci-dessus). Le document propose de remplir dans un premier temps un tableau afin de caractériser le niveau d'utilité (haute, moyenne ou faible). Pour cela, il doit être indiqué, pour chaque ligne de preuve :

- La proportion de publications de « good scientific quality » et « direct relevance » (correspondant à une utilité haute)
- La proportion de publication de « Adequate/utilizable scientific quality » et « direct relevance » (correspondant à une utilité moyenne)
- La proportion de publication de « good scientific quality » et « indirect relevance » (correspondant à une utilité moyenne)
- La proportion de publication de « « Adequate/utilizable scientific quality » et « indirect relevance » (correspondant à une utilité faible).

10 Études épidémiologiques, avec critères spécifiques pour les études sur volontaires humains, les études biomarqueurs humains, les études cliniques, et autres), les études expérimentales sur animaux, les études in vitro, les modèles mathématiques, structure-activité et autres données *in silico*, les études sur les modes/mécanismes d'action, les omics et d'autres méthodes en développement

Dans un second temps, le document propose de caractériser la cohérence entre les différentes études d'un même type. La cohérence est définie comme "the agreement in the results of the analysis between all the individual publications/data sets » et est caractérisée ainsi :

- HIGH – most studies show findings in the same direction;
- MEDIUM – the studies show a mixture of findings in the same direction and those consistent with either outcome;
- LOW – little agreement between studies. This may be due to heterogeneity of results because of particular features of the studies considered or to effect modification, e.g. because of the presence of susceptible subgroups in the study.

Au final, le niveau de chaque ligne de preuve est caractérisée au moyen d'une matrice « utilité/cohérence » : une croix indique la catégorie dans laquelle la ligne de preuve se situe.

Démarche pour intégration des lignes de preuves

Dans le document du SCENIHR, une liste de question à se poser en fonction de l'étape de l'évaluation des risques est dressée, mais il n'est pas précisé comment combiner les différentes lignes de preuves entre elles.

Démarche pour l'expression des conclusions

SCENIHR (2012) communique le résultat de l'évaluation du poids des preuves en le situant au sein d'une des catégories suivantes :

- " Strong overall weight of evidence" : l'ensemble des données, relatives à l'être humain, l'animal et les études mécaniques, fait preuve et est cohérent. Aucune autre donnée n'est contradictoire.
- "Moderate overall weight of evidence" : Preuve avérée des données d'importance sachant que certaines sont manquantes.
- "Weak overall weight of evidence" : Le niveau de preuve des données d'importance est faible.
- "Uncertain overall weight of evidence" : Le niveau de preuve est incertain en raison d'informations contradictoires provenant de sources de données non explicables en termes scientifiques.
- "Weighing of evidence not possible" : Pas de preuve disponible.

Une analyse d'incertitude étaye ensuite la recommandation, dont le résultat est qualifiée en termes de certitude avérée (doute très faible, 1/100 de chance d'être faux), probable (confiance raisonnable, 1/10 de chance d'être faux), confiant (certaine confiance, entre 1/3 et 1/5 chances d'être faux), possible (confiance assez limitée) et incertain (pas de confiance).

2.1 Revue systématique EFSA (2010)

Démarche pour la planification de l'évaluation

Pour la sécurité alimentaire, l'EFSA (2010) considère la formulation de la question dans le cadre de l'élaboration du protocole de conduite d'une revue bibliographique systématique. EFSA (2010) distingue les questions fermées (close-framed), qui comportent l'ensemble des informations clés suffisantes pour conduire la revue systématique, des questions ouvertes (open-framed) où certaines sont manquantes. Il est alors nécessaire de cerner la nature des éléments manquants puis d'élaborer une stratégie spécifique de conduite de la synthèse bibliographique. Trois structures génériques de spécification des informations clés sont proposées, chacune étant associée à un type d'étude.

L'évaluation de l'effet d'une intervention et d'une exposition sur une population s'effectue au moyen des structures PICO et PECO :

- Population (P) : la population d'intérêt (groupe ou communauté de personnes, animaux, plantes, alimentation ou produit alimentaire, un système ou un secteur d'agriculture, un taxon ou une échelle géographique, ou un problème (e.g. vecteur ou insecte des cultures)).
- Intervention (I) ou Exposition (E) : facteur d'exposition de la population.
- Comparator (C) : scénario de référence auquel l'intervention ou l'exposition sera comparée.
- Outcome (O) : propriétés mesurables d'une population qui indiquent les conséquences d'une intervention ou d'une exposition.

La précision d'un test est décrite au moyen de la structure PIT :

- Population : la population d'intérêt
- Index test : le test évalué
- Target conditions : les dégâts, les conditions (présence/absence) ou la quantité recherchée

Enfin, la quantification d'une population en termes de prévalence, d'incidence ou d'occurrence, est obtenue via la structure PO :

- Population : la population, l'organisme ou les conditions dans lesquels la situation d'intérêt est mesurée.
- Outcome : ce qui est précisé ou mesuré au sein de la Population.

Selon EFSA (2010),ces structures permettent la production des revues systématiques, qui nécessitent une formulation ciblée et explicite du problème, les revues narratives ne requérant qu'une définition relativement vague du périmètre.

Le document de l'EFSA propose ainsi de répondre à une série de questions dans le but de définir la stratégie de conduite de recherche bibliographique à développer. La planification en amont de la revue systématique implique : l'élaboration d'un protocole, la mise en place d'une équipe d'évaluation multidisciplinaire, et le calendrier et le budget. Le protocole doit expliquer et définir la question et l'objectif de la revue ainsi que les critères d'inclusion ou d'exclusion des études. Il doit également décrire les méthodes pour chercher et sélectionner les études, collecter les données des études incluses et évaluer leur qualité méthodologique et synthétiser les données des études incluses. Selon l'EFSA, une définition précise du protocole de revue systématique permet de réduire le risque de biais, limite les critiques ultérieurs et améliore le niveau de reproductibilité

Démarche pour l'établissement de lignes de preuve

- *Sélection des études (individuelles ou de synthèse) et l'extraction des données ;évaluation de la qualité des études individuelles ; établissement des lignes de preuve*

L'EFSA (2010) utilise une méthodologie dont les huit étapes sont largement inspirées de Cochrane : *(1)* préparation de la revue : développement du protocole et identification de la logistique (décrite au dessus) ; *(2)* recherche des études ; (3) sélection des études selon les critères d'inclusion ou d'exclusion ; (4) collecte des données dans une grille prédéfinie incluant le recueil des paramètres relatifs à la validité ; (5) évaluation de la qualité méthodologique des études ; (6) synthèse des données avec ou sans méta-analyses ; (7) présentation des données et des résultats ; (8) interprétation des résultats et structuration des conclusions.

Ces étapes doivent être très bien documentées afin de s'assurer de la transparence et de la reproductibilité. A chaque étape, à l'exception de l'étape (2), L'EFSA associe des critères d'évaluation de la qualité de la méthodologie mise en œuvre dans les différentes études analysées. A chaque étape, l'ensemble des études doivent être évaluées de manière indépendantes par plus d'un examinateur afin de limiter l'introduction d'erreurs et de biais individuels. Inclure des experts du domaine étudié est essentiel, cependant, pour réduire les risques d'opinions pouvant biaiser l'évaluation, prendre un expert d'un autre domaine peut être avantageux. Le protocole doit décrire combien d'examinateurs sont responsables de la collecte des données et combien de désaccords ont été résolus. EFSA (2010) considère le conflit d'intérêt dans la sélection des données. Les revues systématiques utilisent au moins deux examinateurs en parallèle pour extraire les données puis les extractions sont comparées. Une approche séquentielle, collecte par un examinateur puis vérification par un autre, peut aussi être utilisée. Les examinateurs peuvent être assignés, à chaque collecte, de manière aléatoire.

Démarche pour l'expression des conclusions

Pour le domaine de l'environnement, EFSA (2010) émet des recommandations concernant le contenu de la discussion des résultats concernant la méta-analyse réalisée à partir de la revue systématique dans l'objectif de favoriser la transparence de l'établissement de la recommandation. Cette discussion doit porter sur la quantité et la qualité des preuves mobilisées, l'interprétation statistique et biologique des résultats appuyée par une analyse de sensibilité, les limitations du processus de revue, et enfin les accords et désaccords appuyée par des études et/ou revues publiées par ailleurs. Les manques de preuves sont enfin mis en évidence, étayée par des recommandations en matière de recherche à conduire.

2.2 Revue systématique Collaboration Cochrane (2011)

La collaboration Cochrane, un réseau international à but non lucratif qui regroupe les données de la recherche médicale scientifiquement validée sous forme accessible et résumée (www://www.cochrane.org/), a développé un document décrivant la méthodologie de la revue systématique de la littérature : Cochrane Handbook for Systematic Reviews of Interventions ***(Higgins et Green 2011).*** L'objet des revues produites par la collaboration Cochrane est d'aider les patients, cliniciens, administrateurs et décisionnaires à la prise de décision pour une intervention (Schünemann et al. 2011).

Démarche pour la planification de l'évaluation

La collaboration Cochrane, propose d'utiliser la structure PICO pour formuler la question et définir les critères descriptifs du contours d'une revue bibliographique (O'Connor et al. 2011). Ces critères sont déclinés de la façon suivante :

- Participants : caractéristiques des patients (âge, sexe, etc.) et/ou du problème qu'il pose (diagnostic, etc.)
- Intervention : nouveau traitement, test diagnostic, etc.
- Comparaisons : intervention servant de témoin, si appropriée (placébo, traitement ou test de référence).
- Outcome : évènements cliniques, survie/mortalité, effets contraires, etc.

Démarche pour l'établissement de lignes de preuve

- *Sélection des études (individuelles ou de synthèse) et l'extraction des données*

La méthode Cochrane permet d'obtenir une méthodologie claire en ce qui concerne la recherche de la littérature (données sources, planification du processus de recherche, développement de stratégie de recherche, management des références bibliographiques, documentation du processus de recherche) mais aussi en ce qui concerne la sélection des études ainsi que l'extraction des données. D'une manière plus spécifique, la collaboration Cochrane recommande l'utilisation de plusieurs bases de données spécifiques ou non (avec PubMed au minimum), l'utilisation de mots clés présents dans le texte ou de thesaurus, une recherche de documentation sans restriction de langues, un contact éventuel avec les auteurs ou avec les industriels. Une recherche privilégiant une grande sensibilité doit être préférée à une recherche privilégiant la spécificité. Pour la sélection des articles scientifiques, la collaboration recommande une sélection initiale sur base du titre et de l'abstract et, ensuite, une sélection des articles présélectionnés sur base de la lecture d*u texte intégral de l'article*. En ce qui concerne l'extraction des données, une liste est proposée. Notons tout de même que cette méthode Cochrane a été développée pour les revues systématiques de la littérature concernant les études épidémiologiques interventionnelles mais que le principe est relativement générique et peut être utilisé pour d'autre type d'étude.

- *Évaluation de la qualité des études individuelles*

La méthode Cochrane, estime, qualitativement, les risques de biais des études incluses dans la revue systématique de la littérature. Une grille avec les biais à évaluer est proposée. Il n'y a pas d'exclusion des données sur base des biais méthodologiques observés. Cependant, la méthode Cochrane propose d'effectuer les analyses principales des résultats en se focalisant uniquement sur les études considérées comme à faible risque de biais. Cette méthode s'intéresse plus particulièrement aux études interventionnelles randomisées mais permet tout de même d'inclure des études non randomisées.

Démarche pour l'expression des conclusions

La communication des résultats (outcome et adverse outcome / effet et effet indésirable) et des conclusions qui en ressortent – en termes d'avantages, d'inconvénients, de charge et de couts – est effectuée dans un rapport dont la structure est prédéfinie. Le rapport comporte une partie discussion et une partie conclusion.

La partie discussion du rapport indique les éléments disponibles mobilisés pour conduire l'étude ainsi que son positionnement dans le contexte approprié. Ce chapitre comporte les cinq sections suivantes : (1) résumé des principaux résultats présentés sous forme de tableaux et discutés en termes d'avantages et inconvénients ; (2) complémentarité et applicabilité des preuves ; (3) qualité des preuves ; (4) biais potentiels du processus de revue ; et (5) accords/désaccords en regard d'autres études ou revues.

La seconde section, portant sur l'applicabilité des preuves, demande de considérer les éléments suivants : le point de vue des auteurs (i.e. assurer la conformité de la revue avec les ressources mobilisées en regard de leur champs d'expertise et confronter les résultats avec des hypothèses formulées à priori), les problèmes de variation biologique incluant les divergences en pathophysiologie et les agents causatifs, les variations induites par le contexte et la culture (sachant que des interventions ont vu leurs effets améliorés dans certaines situations et dégradées dans d'autres), la divergence d'adhésion pour des questions économiques ou comportementales (sachant que le plus fort niveau d'adhésion serait étroitement lié aux essais randomisés) et enfin la considération des valeurs et préférences individuelles pour la gestion de la décision

Dans la troisième section, l'évaluation de la qualité des preuves est effectuée en regard de quatre niveaux de qualité :

- Fort : correspondant aux essais randomisés ou études observationnelles à deux niveaux,
- Modéré : essais randomisés dégradés ou études observationnelles améliorées,
- Faible : essais randomisés doublement dégradés ou études observationnelles,
- Très faible : essais randomisés triplement dégradés, études observationnelles dégradées ou étude de cas.

Ce niveau de qualité peut par la suite être réduit en regard de facteurs dégradants – e.g. hétérogénéité/inconsistance/imprécision des résultats ou preuves indirectes – ou augmentés par d'autres, à l'exemple de l'envergure des effets.

La conclusion du rapport traite dans une première section des conséquences de la mise en pratique de l'intervention en regard de quatre critères : qualité des preuves, ratio avantages/inconvénients, les valeurs et préférences des patients, et enfin les ressources à mobiliser. Dans ce rapport, les auteurs ne doivent pas faire de recommandations, mais peuvent mettre en évidence diverses actions pertinentes en regard de schémas particuliers de valeur et de préférence. La seconde section du chapitre Conclusion traite des conséquences en matière de besoin de recherche.

Celles-ci peuvent être décrites selon le format EPICOT (Brown et al. 2006) :

- E (Evidence) : what is the current evidence?
- P (Population) : diagnosis, disease stage, co-morbidity, risk factor, sex, age ethnic group, specific inclusion or exclusion criteria, clinical setting
- I (Intervention) : type, frequency, dose, duration, prognostic factor
- C (Comparison) : Placebo, routine care, alternative treatment/management
- O (Outcome) : which clinical or patient-related outcomes will the researcher need to measure, improve, influence or accomplish ?
- T (Time stamp) : date of litterature search or recommendation

Cette structure est en réponse à la structure PICO utilisée dans l'étape de planification de l'évaluation pour la formulation de la question.

2.3 Revue bibliographique appliquée à l'environnement Bilotta (2014)

Démarche pour la planification de l'évaluation

Dans l'objectif de répondre aux questions relatives aux politiques environnementales, Bilotta et al. (2014) proposent d'utiliser la structure PICO pour formuler la question et définir les critères descriptifs du contour d'une revue bibliographique. Les critères PICO sont déclinés de la façon suivante :

- Participant : animal, plante, habitat, écosystème ou membre d'une société,
- Intervention : option de gestion de l'environnement (technique agricole ou de mesure de contrôle des maladies des plantes/animaux, etc.),
- Comparateur : études environnementales utilisant des approches avant-et-après, ou intervention versus cas-témoin,
- Outcome : métrique de mesure de l'animal, de la plante, de la santé de l'écosystème ou sa productivité, ou d'un événement social.

Démarche pour l'expression des conclusions

Pour les politiques environnementales, Bilotta et al. (2014)proposent d'utiliser le logiciel développé par la collaboration Cochrane pour interpréter la synthèse bibliographique et de présenter un résumé neutre des évidences.

2.4 Revue systématique « Navigation guide » (Woodruff et Sutton 2011)

La « Navigation Guide » (Woodruff et Sutton 2011, Woodruff et Sutton 2014)est une méthode systématique et transparente de synthèse de résultats de la recherche dans le domaine de la santé-environnement, plus spécifiquement concernant les risques reprotoxiques et développementaux d'un agent chimique.

Le processus proposé de déroule en 4 étapes : (1) préciser la question à étudier (2) Sélectionner les preuves (3) Noter la qualité et la force des preuves (4) Grader la force des recommandations

Démarche pour la fla planification de l'évaluation

Le « Navigation guide systematic review methodology » estime que la rédaction d'une question de recherche pertinente est d'importance capitale. En effet, dans le protocole de recherche, développé en amont de la revue systématique, l'identification et la sélection d'articles intéressants doit se faire sur base des critères PECO (Participants, Exposure, Comparator, Outcome).

Démarche pour l'établissement de lignes de preuves

- *Recherche, Sélection des études et extraction des données*

La stratégie de recherche doit combiner une recherche systématique d'études publiées et non publiées. Aucune autre information pragmatique n'est recommandée ni discutée.

L'inclusion ou l'exclusion des études doit se faire de manière consistante et transparente et tous les critères de jugement doivent être documentés. Aucune autre information pragmatique n'est recommandée.

- *Evaluation de la qualité des études individuelles*

Les auteurs proposent de séparer les critères de qualité de l'étude qui peuvent introduire une erreur systématique dans l'amplitude ou dans la direction des résultats, d'autres erreurs qui n'influencent pas systématiquement les résultats de l'étude. Les auteurs insistent aussi sur l'importance d'analyser les risques de biais dans les études sur les animaux. Les auteurs suggèrent d'inclure les conflits d'intérêts et les biais de publication comme risque de biais. L'outil pour évaluaer le risque de biais est inspiré de ceux développé par la collaboration Cochrane (Higgins et Green 2011) et par l' Agency for Healthcare Research and Quality (Viswanathan et al. 2012) et modifié afin d'évaluer les études toxicologiques.

- *Etablissement des lignes de preuves*

La « force » des lignes de preuves « humaines » et les lignes de preuve « non-humaines » est qualifiée de « suffisante », « limitée », « inadaptée » ou « preuve d'absence de toxicité ».

Démarche pour l'intégration des lignes des preuves

Le poids des preuves global est obtenu à l'aide d'une matrice croisant la force des lignes de preuve « humaines » et « non humaines ». Cinq classes peuvent ainsi être obtenues : « Known to be Toxic to Human Reproduction » ; « Probably

Toxic » ; « Possibly Toxic" ; " Not classificable" et "probably not toxic".

Démarche pour l'expression des conclusions

Les auteurs de ce guide suggèrent de combiner (via une matrice) la la force de l'évidence et le niveau d'exposition (haute, moyenne, faible), pour effectuer une gradation des forces des recommandations (recommandation forte ou facultative). Ces recommandations peuvent aussi être modulées par l'existence d'alternatives moins toxiques et les préférences et de l'avis des patients.

2.5 Évaluation d'une revue systématique par l'outil AMSTAR (2007)

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études de synthèse*

En ce qui concerne les études de synthèse de la littérature, un seul outil d'évaluation de leur qualité est proposé, AMSTAR, développé en 2007 (Shea, Grimshaw, et al. 2007, Shea, Bouter, et al. 2007). Basé sur des questionnaires/listes existants (OQAQ, Sacks), AMSTAR comporte 11 items. Chaque item vaut 1 point, le score total varie de 0 à 11. Aucun seuil critique n'a été défini.

En 2010, une version révisée de AMSTAR est développée : R-AMSTAR (Kung et al. 2010). Pour chaque item, des critères sont définis permettant de donner un score entre 1 à 4. Le score total varie de 11 à 44. Cet outil permet de « classer » plusieurs revues systématiques portant sur un même sujet. Le système de cotation (A-D) est basé sur les percentiles de la distribution des scores obtenus pour chacune des revues (90ème, 80ème ...). Là encore, aucun seuil critique n'a été défini. Ces outils ont été développés et validés pour les revues d'essais cliniques mais sont aussi utilisés pour les revues d'études observationnelles.

2.6 Évaluation d'une revue systématique du domaine médical Murad

Murad et al. (2014) proposent une méthode d'évaluation des revues systématiques, dans le domaine clinique.

Démarche pour la planification de l'évaluation

Murad et al. (2014) dissocie la formulation de la question de l'utilisation de PICO, ce dernier servant de support d'identification des critères d'éligibilité des études. Les auteurs ne proposent cependant pas de structure de formulation de la question ni de contenu.

Démarche pour l'expression des conclusions

Murad et al (2014) proposent deux séries de questions pour évaluer et appliquer les résultats d'une revue systématique conjuguée d'une méta-analyse. La première série concerne l'évaluation de la crédibilité des méthodes de revue systématique mobilisées (« did the review explicitly address a sensible clinical question », « was the search relevant studies exhaustive », « were selection and assessments of studies reproductible », « did the review present results that are ready for clinical application », and « did the review address confidence in estimates of effects ? »).

La seconde série traite du taux de confiance dans l'estimation des effets: « how serious is the risk of bias in the body of evidence? », « are the results consistent across studies? », « how precise are the results? », « is there concern about reporting bias? », et « are there reasons to increase the confidence rating? ».

2.7 Lignes directrices pour la rédaction d'articles scientifiques

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études de synthèse*

Des lignes directrices à la rédaction d'articles scientifiques ont été proposées. Citons notamment STROBE pour les études observationnelles, PRISMA pour les revues systématiques et des méta-analyses (Moher et al. 2015), CONSORT pour les essais contrôlés randomisés, STARD les études sur les tests diagnostiques, COREQ pour la recherche qualitative, ENTREQ pour les études sur la recherche qualitative, REFLECT-LFS pour études expérimentales dans le domaine de la sécurité alimentaire. Notons cependant que ces lignes directrices n'ont pas été développées pour juger directement de la qualité des études mais les articles scientifiques rédigés suivant ces lignes directrices auront toutes les informations nécessaires à leur évaluation qualitative.

3.1 Critères de Bradford Hill (1965)

Démarche pour l'analyse d'un ensemble d'étude et l'évaluation du poids des preuves

Les critères de Bradford Hill (Hill 1965) interviennent principalement dans les étapes « Établissement de lignes de preuve » et « Intégration des lignes de preuves pour établir le poids des preuves ». Ils sont de type qualitatif sans notation. Ils consistent en une liste de critères constituant un argumentaire en faveur de l'inférence causale de la relation étudiée :

1- Force de l'association : plus l'association observée est forte, plus elle a de chance d'être causale.

2- Cohérence entre les études : si plusieurs études dans des populations et/ou pays différents trouvent des résultats similaires, la causalité est renforcée

3- Spécificité : si une substance donnée exerce un effet spécifique, cela constitue un argument en faveur de la causalité. Toutefois, il existe de nombreuses associations causales non spécifiques

4- Temporalité : l'exposition doit précéder la pathologie dans le temps

5- Gradient biologique ou dose-réponse

6- Plausibilité mécanistique au regard des connaissances sur l'étiologie de la maladie et du mode d'action possible de la substance étudiée

7- Cohérence : en référence aux autres effets biologiques observés potentiellement pertinents dans les voies mécanistiques impliquées dans l'étiologie de la maladie (changement histologique dans l'organe cible par exemple)

8- Preuves expérimentales du type parallélisme de la cause et de l'effet (par exemple, si l'incidence de la maladie diminue suite à l'élimination de l'exposition)

9- Analogie : si un agent similaire exerce un effet similaire, ceci constitue également un argument en faveur de la causalité.

En pratique, les critères de Hill ont été très largement cités et utilisés et ont inspiré de nombreux autres systèmes de gradation du niveau de preuve (Swaen et van Amelsvoort 2009, Rothman et Greenland 2005, Bergman et al. 2015, OHAT 2015, Meek, Palermo, et al. 2014, Guzelian et al. 2005, Vinken 2013).Il s'agit d'une méthode générique qui peut être appliquée en santé humaine mais également dans d'autres domaines, et s'applique, à l'origine, principalement aux études épidémiologiques.

Parmi les limites discutées dans la littérature concernant les critères de Hill, on retrouve le fait qu'aucun de ces critères (d'ailleurs plutôt appelés « viewpoints » par Hill) ne suffit à lui seul à établir la causalité et à l'inverse, celle-ci peut être avérée même si tous les critères ne sont pas remplis. Il s'agit réellement d'un faisceau d'arguments en faveur (ou en défaveur) de la causalité (Bergman et al. 2015). La principale limite de cette méthode est son caractère qualitatif. Il s'agit principalement d'une « checklist » d'arguments à examiner mais leur interprétation et leur compilation peut-être sujette à discussion entre différents experts. Dans cette optique, Swaen et van Amelsvoort (2009) ont proposé une méthode permettant de rendre quantitative l'utilisation des critères de Hill. Cette méthode inclut deux étapes : (1) Estimation de la probabilité que le critère soit atteint (pour chaque critère de Hill séparément) ; et (2) Détermination d'un poids pour chacun des neuf critères de Hill, exprimant l'importance relative de chacun dans l'évaluation globale de la causalité (à partir d'une approche par analyse discriminante sur une base de référence issue des travaux du CIRC). Les critères de Hill ainsi utilisés constitueraient une méthode de type « quantitative avec notation ».

Les critères de Bradford Hill ont par ailleurs été considérés dans le domaine spécifique de l'évaluation du mode d'action. Dans ce cadre, ils ont été affinés et une méthodologie semi-quantitative a été proposée (Meek, Palermo, et al. 2014).

3.2 GRADE (2011)

Démarche pour la planification de l'évaluation

Le groupe de travail GRADE (Grading of Recommendations Assessment, Development and Evaluation), impliqué dans la conduite de revue systématique dans le domaine de la santé humaine, appréhende la formulation de la question sous la forme d'identification des questions directrices (Guyatt, Oxman, Kunz, Atkins, et al. 2011). Pour ce faire, la structure PICO (Patient, Intervention, Comparator, Outcome) est utilisée, complétée parfois des conditions dans lesquelles les conclusions seront utilisées à l'exemple du type de population. Dans la mesure où les recommandations produites à partir des effets peuvent être différentes entre les groupes de patients, une méthodologie en trois étapes est proposée pour classer par ordre d'importance les effets :

- Le regroupement préliminaire des effets selon trois classes d'importance vis-à-vis des patients (critique, important mais non critique, peu important) par des experts, des patients et le public.
- La réappréciation de l'importance relative des effets après la revue des preuves, avec la possibilité d'agréger de nouveaux effets en regard de la revue bibliographique.
- Le jugement de la balance entre les effets désirables et non désirables d'une intervention pour produire les recommandations.

Au sein des classes, les effets sont ordonnés en fonction de leur importance relative par attribution d'une note : de 1 à 3 pour les effets peu importants, de 4 à 6 pour les effets importants mais non critique et de 7 à 9 pour effets critiques. Ce positionnement relatif sert de base à l'établissement des recommandations en partant des effets les plus critiques (valeur de 9) au moins critiques (valeur de 4).

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études individuelles*

La méthode GRADE pour évaluer la qualité des études individuelles est qualitative avec notation : les niveaux de qualité peuvent être qualifiés de élevés, modérés, faibles ou très faibles ((Balshem et al. 2011).

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 (les essais cliniques randomisés et contrôlés et les études épidémiologiques observationnelles). Chaque étude est évaluée individuellement mais la définition des niveaux de qualité des données scientifiques pour chaque résultat important se base sur la qualité de l'ensemble des données scientifiques..

Cinq facteurs peuvent diminuer la qualité des données scientifiques : le risque de biais (Guyatt, Oxman, Vist, et al. 2011) (Guyatt, Oxman, Vist, et al. 2011), l'hétérogénéité des résultats (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Glasziou, et al. 2011) le caractère indirect des données scientifiques (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Falck-Ytter, et al. 2011), l'imprécision des données (Guyatt, Oxman, Kunz, Brozek, et al. 2011), les biais de publication (Guyatt, Oxman, Montori, et al. 2011). Par contre, trois facteurs peuvent augmenter la qualité des données scientifiques : la force de l'association , un gradient dose-réponse et la prise en compte de facteurs de confusion qui auraient pu réduire l'effet observé (Guyatt, Oxman, Sultan, et al. 2011).

L'objectif final n'est pas d'exclure les données sur base des biais méthodologiques potentiels mais de limiter leur poids dans l'évaluation de la force des recommandations. Une certaine subjectivité dans l'évaluation des facteurs ne peut cependant pas être exclue. Par ailleurs, pour limiter l'influence de certains articles dont les données sont utilisées à plusieurs reprises par divers articles scientifiques, Kho et Brouwers (2012)proposent de construire le réseau des citations bibliographiques afin de conserver exclusivement les articles originaux pour l'analyse.

Comme discuté précédemment, le système GRADE évalue principalement les essais contrôlés randomisés et les méta-analyses. Remarquons aussi que les biais de publication nécessaires à l'évaluation dans la méthodologie GRADE sont difficilement objectivés sans l'utilisation de statistiques appropriées en effectuant des techniques méta-analytiques.

Démarche pour l'expression des conclusions

Dans le domaine clinique, l'objet de la recommandation émise au moyen de GRADE est d'aider au choix d'une technique d'intervention en regard des preuves disponibles. Pour ce faire, une méthode de catégorisation, d'étiquetage et de formulation des recommandations est proposée. Dans la mesure où toute technique d'intervention présente des effets désirables et non désirables, classés « critiques » ou « importants et peu critiques », GRADE propose de les évaluer au moyen de six critères (Andrews, Guyatt, et al. 2013, Andrews, Schunemann, et al. 2013): l'estimation de l'effet d'une intervention sur les effets, la confidence dans cette estimation, les estimations des valeurs et préférences des patients (ce critère intègre leur perspective, croyance, attente et but pour leur santé et leur vie), la variabilité des estimations des valeurs et préférence des patients, et les ressources à mobiliser (couts, etc.).

La technique d'intervention est ensuite qualifiée en terme de force (weak *vs* strong) et d'orientation (against *vs* for), ce qui donne lieu à la situer au sein d'une des quatre catégories de recommandation :

- « Strong for » : forte confiance dans le fait que les conséquences désirables l'emportent sur les conséquences indésirables
- « Weak for » : faible confiance dans le fait que les conséquences désirables l'emportent sur les conséquences indésirables
- « Weak against » : faible confiance dans le fait que les conséquences indésirables l'emportent sur les conséquences désirables
- « Strong against » : forte confiance dans le fait que les conséquences indésirables l'emportent sur les conséquences désirables

Il est également possible que diverses raisons, e.g. choisir entre des techniques qualifiées « weak for », conduisent à ne proposer aucune recommandation ; ce qui correspondrait à une cinquième catégorie.

La qualification de l'intervention peut être fortement influencée par des effets relatifs à sa mise en œuvre, à l'exemple de sa prévalence, son équité vis-à-vis d'autres interventions, son cout et l'amélioration des soins apportés. Pour l'évaluation de la force de la recommandation, l'impact potentiel sur les parties prenantes est également considéré) : dans la mesure où une recommandation de force « strong » implique l'unicité du choix d'intervention et une recommandation « weak » la pluralité de choix, sa formulation affecte directement : (1) la relation entre patient et clinicien, puisqu'ils doivent décider ensemble de l'intervention à adopter ; et (2) les décisionnaires. En effet pour ces derniers, la recommandation « strong » implique l'adoption systématique d'une même intervention pour tous les cas cliniques, contrairement à la recommandation « weak » qui génèrera de la diversité de traitement entre les individus.

Pour éviter les erreurs d'interprétation, GRADE suggère d'utiliser des symboles pour indiquer la classe de recommandation (Akl et al. 2007) , de même que certaines expressions naturelles pour les formuler, à l'exemple de « we recommend » et « we suggest » pour respectivement les recommandations « strong » et « weak ». Pour favoriser la transparence dans la conduite de la méthode, des documents constitués de tableaux, sont proposés. Pour chaque intervention, l'entête du tableau rappelle la question, la population cible, l'intervention et les ressources disponibles. Le corps du tableau demande de préciser pour les six critères, le jugement sous forme de boite à cocher (oui / non), une explication du jugement et les sous critères expliquant le jugement. Le bas de tableau informe de la recommandation finale, indiquée au moyen de sa catégorie d'appartenance suivie d'une justification. Pour faciliter sa mise en œuvre, GRADE dispose d'un logiciel disponible à l'adresse www.guidelinedevelopment.org.

La méthode GRADE est mobilisée dans de nombreux organismes, par exemple par l'OMS pour ses recommandations concernant la supplémentation en fer et en acide folique des femmes enceintes (OMS 2012) ou par l'AHQR (Berkman et al. 2012).

3.3 OMS-CIRC (WHO-IARC en anglais, IARC 2006)

Dans les domaines santé-nutrition, santé-environnement et santé-travail, la méthode du CIRC est utilisée pour statuer sur le caractère cancérogène chez l'homme de tous les agents auxquels il est susceptible d'être exposé.

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études individuelles*

La méthode décrite par le CIRC (IARC 2006) pour l'évaluation de la qualité des études individuelles est qualitative, sans notation. Pour chaque type de preuve pris en compte dans l'évaluation (études épidémiologiques chez l'homme, études expérimentales chez l'animal, et études sur les mécanismes et modes d'action), la qualité de chaque étude individuelle est évaluée à partir de critères relevant des bonnes pratiques de recherche dans chacun de ces domaines. Pour les études épidémiologiques par exemple, les critères suivants sont examinés pour juger de la qualité de l'étude : design (prospectif ou non, etc.), risque de biais, niveau de prise en compte des facteurs de confusion, possibilité de risque de première espèce (« chance finding »), précision dans la définition de l'exposition et de la maladie, puissance statistique, présence d'une relation dose-effet et/ou de tendances temporelles, etc. Aucune notation n'est proposée à l'issue de cette évaluation des études individuelles et il n'est pas fait mention de grille de cotation pour les études individuelles.

- *Établissement lignes de preuves*

La méthode CIRC est de type qualitative, sans notation**.**

Pour la ligne de preuves correspondant aux études épidémiologiques chez l'homme, la gradation suivante est établie :

- <u>preuve suffisante</u> de cancérogénicité (une relation positive entre l'exposition à l'agent et le cancer est observée, qui ne peut vraisemblablement pas être expliquée par le hasard, un biais ou un phénomène de confusion)
- <u>preuve limitée</u> de cancérogénicité (une relation positive entre l'exposition à l'agent et le cancer est observée mais le hasard, un biais ou un phénomène de confusion ne peuvent pas être exclus avec un niveau de confiance suffisant))
- <u>preuve inadéquate</u> de cancérogénicité (la qualité, la cohérence ou le pouvoir statistique de la littérature disponible pour cette ligne de preuve est insuffisante pour pouvoir conclure sur la présence ou l'absence d'une association causale entre l'exposition à l'agent et le cancer, ou alors aucune donnée humaine sur le cancer n'est disponible)
- <u>preuve suggérant une absence de</u> cancérogénicité (les études disponibles sont suffisamment fiables méthodologiquement pour réfuter l'hypothèse de cancérogénicité pour l'agent et la localisation cancéreuse étudiés).

Pour la ligne de preuves correspondant aux études expérimentales chez l'animal, la gradation suivante est également établie :

- <u>preuve suffisante</u> de cancérogénicité (démontrée chez 2 espèces animales différentes, ou chez une même espèce dans au moins 2 études indépendantes, à 2 temps ou dans 2 laboratoires différents ou avec 2 protocoles différents, ou éventuellement pour les 2 sexes d'une même espèce dans une étude bien conduite - idéalement suivant les Bonne Pratiques de Laboratoire -,ou éventuellement dans une étude chez une espèce et un sexe si les neoplasmes malins apparaissent à un degré inhabituel au regard de l'incident, localisation, type de tumeur ou âge d'apparition, ou lorsque les tumeurs sont nombreuses à de nombreuses localisations).
- <u>preuve limitée</u> de cancérogénicité (une seule expérience ou questions non résolues concernant l'adéquation du design des études, leur conduite ou leur interprétation, ou effet uniquement sur les stades pré-cancéreux ou tumeurs bénignes, ou la preuve de cancérogénicité est restreinte aux études démontrant une activité dans une gamme étroite de tissus ou d'organes) <u>preuve inadéquate</u> de cancérogénicité (qualité

de la littérature disponible pour cette ligne de preuves est insuffisante pour pouvoir conclure du fait de limites méthodologiques ou quantitatives majeures ou de l'absence de données animales disponibles sur le cancer.

- preuve suggérant une absence de cancérogénicité (chez au moins deux espèces animales)

Pour la ligne de preuves correspondant à chaque mécanisme/mode d'action étudié, la gradation suivante est également établie : fort, modéré et faible.

Démarche pour l'intégration des lignes de preuve

La méthode CIRC introduite précédemment (IARC 2006).. Elle consiste à croiser la gradation des éléments de preuve concernant les études épidémiologiques chez l'humain, expérimentales chez l'animales, ainsi que les données relatives aux mécanismes et autres données pertinentes.

Chaque agent pour lequel le caractère cancérogène pour l'homme a été étudié est catégorisé selon le système suivant :

- Groupe 1. L'agent est cancérogène pour l'homme : preuve suffisante de cancérogénicité chez l'homme. Exceptionnellement, un agent peut être classé dans le groupe 1 lorsque la preuve de cancérogénicité chez l'homme est moins que suffisante (« less than sufficient ») mais que la preuve de cancérogénicité chez l'animal est suffisante avec une preuve forte chez les humains exposés que l'agent agit à travers un mécanisme de cancérogénicité pertinent.
- Groupe 2A. L'agent est probablement cancérogène pour l'homme : preuve limitée de cancérogénicité chez l'homme mais suffisante chez l'animal. Dans certains cas,un agent peut être classé dans le groupe 2A lorsque la preuve de cancérogénicité chez est inadéquate chez l'homme mais suffisante chez l'animal, avec une preuve forte que le mécanisme de cancérogénicité a lieu aussi chez les humains. Exceptionnellement, un agent peut être classé dans le groupe 2A avec une preuve limitée de cancérogénicité chez l'homme si l'agent appartient clairement, sur la base de considération mécanistiques, à une classe d'agents dont un ou plusieurs ont été classé groupe 1 ou groupe 2A.
- Groupe 2B. L'agent est peut-être cancérogène pour l'homme : preuve limitée de cancérogénicité chez l'homme et « moins que suffisante » (« less than sufficient ») chez l'animal, ou preuve inadéquate de cancérogénicité chez l'homme mais suffisante chez l'animal. Parfois, un agent peut être classé 2B si la preuve de cancérogénicité est inédéquate chez l'homme et « moins que suffisante » chez l'animal, mais avec des preuves fortes issues de données mécanistiques ou autres. Groupe 3. L'agent est inclassable quant à sa cancérogénicité pour l'Homme : preuve inadéquate de cancérogénicité chez l'homme et inadéquate ou limitée chez l'animal. Exceptionnellement, un agent peut être classé Groupe 3 lorsque la preuve de cancérogénicité est inadéquate chez l'homme et suffisante chez l'animal, avec une preuve forte que les mécanismes en jeu chez l'animal n'opèrent pas chez l'homme. Les agents qui n'entrent pas dans les autres catégories sont classées Groupe 3.En général, des recherches supplémentaires sont nécessaires pour les agents classés dans ce groupe.

Groupe 4 : l'agent n'est probablement pas cancérogène pour l'homme : preuves suggérant une absence de cancérogénicité chez l'homme et chez l'animal ou preuves inadéquates chez l'homme mais preuves suggérant une absence de cancérogénicité chez l'animal, et données mécanistiques ou autres allant dans ce sens. Les limites de la méthodologie décrite dans le document guide général sont la part de subjectivité liée à l'évaluation des différents critères et le fait que le document général transversal corresponde principalement à un exposé des principes scientifiques suivis par le groupe, plus qu'à un système de cotation et/ou des critères détaillés. Cependant, c'est une méthode générique qui peut être appliquée en santé humaine et porte sur tous types d'études épidémiologiques et expérimentales.

3.4 WCRF/AICR (WCRF/AICR 2007)

En pratique, la méthode WCRF/AICR a pour objectif de grader le niveau de preuve de toutes les relations entre facteurs nutritionnels (aliments, nutriments, etc.) et risque de cancer par localisation.

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études individuelles*

La méthode décrite par le WCRF/AICR (WCRF/AICR 2014) est qualitative, sans notation. La qualité de chaque étude individuelle est évaluée à partir de critères classiques, relevant des bonnes pratiques de recherche. Les études épidémiologiques chez l'homme sont au centre du processus d'évaluation du niveau de preuve mais la plausibilité mécanistique en provenance des études expérimentales in vivo ou in vitro est également considérée. La limite de cette méthodologie est que les critères de la grille peuvent ne pas être appliqués en l'état et sont modulés au cas par cas selon différents paramètres faisant l'objet de discussions au sein du groupe d'experts, impliquant une part de subjectivité.

- *Établissement lignes de preuves*

Pour cette étape, la méthode WCRF/AICR (WCRF/AICR 2014) est basée sur une approche quantitative (résultat d'une méta-analyse) pour les études épidémiologiques et qualitative sans notation pour les études expérimentales in vivo et in vitro.

Pour la ligne de preuve correspondant aux études épidémiologiques chez l'homme, le WCRF/AICR réalise des méta-analyses ad hoc pour chaque relation étudiée, sur la base d'une revue systématique de la littérature. Des méta-analyses dose-réponses (linéaires et non-linéaires) et haut versus bas sont réalisées en incluant toutes les études disponibles. Les sources potentielles d'hétérogénéité sont identifiées par méta-régression. Le résultat de cette étape est présenté sous la forme d'un risque relatif résumé (summary RR) assorti de son intervalle de confiance.

Pour la ligne de preuves correspondant aux études expérimentales chez l'animal, aucune gradation particulière n'est imposée. Cette ligne de preuve a pour vocation d'évaluer dans quelle mesure les résultats épidémiologiques (centraux dans le processus) sont supportés ou non par une plausibilité mécanistique.

Comme énoncé précédemment, cette méthode concerne le domaine santé-nutrition, mais est extrapolable aux domaines santé-environnement et santé-travail. Outre la subjectivité liée à l'évaluation de certains critères, les autres limites de cette approche sont celles liées à l'utilisation de méta-analyses. Celles-ci exclues parfois certaines études pour lesquelles toutes les informations statistiques nécessaires ne sont pas publiées. La réalisation d'une revue systématique et d'une méta-analyse pour la ligne de preuve épidémiologique est par ailleurs couteuse en temps et en personnel qualifié.

Démarche pour l'intégration des lignes de preuves

Pour cette étape, la méthode WCRF/AICR) est de type qualitative, avec notation.

Chaque relation facteur nutritionnel – localisation de cancer étudiée est catégorisée selon le système suivant :

- Niveau de preuve convaincant (justifiant de donner lieu à des recommandations: preuves provenant de plusieurs types d'études, au moins deux cohortes indépendantes, faible hétérogénéité dans les méta-analyses, bonne qualité des études épidémiologiques considérées permettant de conclure que le résultat n'est pas dû à des biais de confusion, sélection ou classement, ou au risque de première espèce, présence d'un gradient dose-réponse plausible, linéaire ou non, forte plausibilité mécanistique)
- Niveau de preuve probable (au moins deux cohortes indépendantes ou au moins cinq cas-témoins, faible hétérogénéité, bonne qualité des études épidémiologiques, plausibilité mécanistique satisfaisante)
- Niveau de preuve suggéré (au moins deux cohortes indépendantes ou au moins cinq cas-témoins, hétérogénéité possible, plausibilité mécanistique)
- Niveau de preuve non concluant (peu d'études disponibles ou effets différents entre les études ou études présentant des lacunes méthodologiques. Des recherches additionnelles sont nécessaires pour conclure)

- Niveau de preuve effet peu probable (même exigences que pour le niveau de preuve convaincant, mais montrant une absence d'association)

Comme énoncé précédemment, cette méthode concerne le domaine santé-nutrition, mais est extrapolable aux domaines santé-environnement et santé-travail. Sa limite principale est la subjectivité inhérente à l'évaluation de certains critères, rendant nécessaire la prise de décision et la modulation des règles fixées au cas par cas par le groupe d'experts.

3.5 Klimisch (1997)

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études individuelles*

La méthode Klimisch (Klimisch et al. 1997) est qualitative, de type scoring et a été développée pour évaluer la qualitédes études toxicologiques. La qualité des études est évaluée à partir de critères spécifiques à la toxicologie expérimentale, peu différents des critères promus par les bonnes pratiques de laboratoire et les méthodes standardisées. Ce système repose sur une cotation des études expérimentales en tenant compte de la fiabilité des études (qualité intrinsèque de l'étude évaluée par le score de Klimisch de 1 à 4, respectivement fiable sans restriction, fiable avec restrictions, non fiable, non évaluable) puis de leur pertinence (caractère approprié, représentatif ou non) et de leur adéquation (utilité des données pour la décision finale). Les études de score 1 et 2, considérées comme valides, décrivent avec précision la méthodologie et les résultats observés. Les études de score 3 et 4, considérées comme non valides ou non évaluables, peuvent néanmoins être utilisées comme source d'informations complémentaires. En pratique, la méthode est utilisée pour l'évaluation des substances chimiques dans le domaine santé environnement, tel que dans le contexte réglementaire de REACH (ECHA 2011). Une limite de la méthode est qu'une étude Klimisch de cotation 1 peut être de pertinence nulle. Pour les études scorées 3 ou 4, une méthode complémentaire (analyse de la littérature) est nécessaire pour rendre une conclusion sur l'étude la plus pertinente à prendre en compte. Le manque de précision de certains critères est une autre limite de cette méthode. La fiabilité et l'adéquation des études dépendent de l'évaluateur.Des outils pour aider et préciser la classification Klimisch ont été développés (Schneider et al. 2009). Des adaptations des critères de Klimisch aux études humaines ont aussi été réalisées (Money et al. 2013).

3.6 Evaluation du poids des preuves fondée sur les hypothèses

Rhomberg (2015), Bailey et al. (2016) décrivent un système désigné comme « Poids de la preuve fondé sur des hypothèses » dans lequel le poids des preuves relatives à la toxicité d'un produit chimique est considéré en évaluant la cohérence des explications causales hypothétiques à travers toutes les lignes de preuves disponibles (épidémiologiques, toxicologiques, mécanistiques, toxicocinétiques…). La méthode est qualitative, sans notation. Le système est basé sur l'évaluation de la relative plausibilité des hypothèses alternatives pour prendre en compte les divers résultats des études disponibles, sur la base de questions permettant d'évaluer la logique des hypothèses proposées, pour chaque ligne de preuve et pour l'ensemble des lignes de preuves. En substance, l'objectif est similaire à celui de « l'évaluation comparative du poids des preuves » décrit en 3.10, sans qu'il n'y ait de codification des considérations « *a priori* » basées sur l'expérience collective. La cohérence est ainsi plus faible, dépendant éventuellement du jugement de l'évaluateur.

3.7 Méta-analyse

Démarche pour l'établissement de lignes de preuve

- *Établissement lignes de preuves*

La méta-analyse (Chalmers et al. 2002) est une méthode quantitative qui consiste à réaliser une analyse statistique de données provenant de différentes études, conduites dans différentes conditions mais traitant toutes d'un sujet commun. Les études considérées doivent présenter suffisamment de similarités pour pouvoir être analysées avec des méthodes statistiques (Chalmers et al. 2002).Les données sont extraites d'études individuelles correspondant, souvent, à des publications scientifiques et, parfois également, à de la littérature grise ou à des données brutes.

La méta-analyse combine des estimations individuelles d'une quantité d'intérêt (appelée *taille d'effet*) et produit une estimation de taille d'effet moyenne ainsi qu'un intervalle de confiance décrivant l'incertitude associée à cette estimation moyenne. La taille d'effet moyenne synthétise l'ensemble de données disponibles à travers une valeur unique. Les valeurs des tailles d'effet individuelles ont cependant, elles aussi, un intérêt car elles décrivent la variabilité inter-études de la quantité étudiée. Cette variabilité est due à l'hétérogénéité des conditions expérimentales dans lesquelles les études individuelles ont été réalisées, ainsi qu'aux erreurs de mesure et d'estimation. La taille d'effet moyenne est estimée à l'aide d'une analyse statistique souvent basée sur des modèles mixtes (combinant des effets fixes et des effets aléatoires).

La méta-analyse est une méthode très utile pour synthétiser de manière quantitative l'ensemble des données disponibles sur un sujet donné. Elle a été appliquée dans de nombreux domaines, par exemple :

- Pour analyser l'effet sur la santé humaine d'une exposition au PCB (Goodman et al. 2010),
- Pour étudier les risques environnementaux associés aux cultures OGM (Marvier 2011, Marvier et al. 2007)
- Pour évaluer l'efficacité des pesticides pour contrôler les maladies des cultures (EFSA 2014)

La méta-analyse est souvent délicate à mettre en œuvre et est coûteuse en temps car elle inclut une étape de revue systématique (sélection des études disponibles sur un sujet donné) ainsi qu'une étape d'extraction des données publiées dans les études sélectionnées. Lorsque les données sont peu nombreuses ou lorsqu'elles sont issues d'études trop dissemblables, les résultats de la méta-analyse doivent être interprétés avec grande prudence (Goodman et al. 2010). Les résultats d'une méta-analyse peuvent par ailleurs être biaisés lorsqu'il existe un biais de publication, c'est à dire lorsque les études publiées ne sont pas représentatives de l'ensemble des études réalisées.

3.8 Analyse multicritères (Linkov et al, 2011)

Linkov et al. (2011) ont développé une méthode d'analyse multicritères d'évaluation du poids des preuves dans un contexte de risque écologique permettant de répondre à des objectifs d'analyse pluriels (estimer un risque, produire une classification, etc.). Cette méthode est quantitative. La méthode permet de combiner tout type de données, à l'exemple des études individuelles et d'options managériales (options stratégiques et décisionnelles), qu'elles soient qualitatives ou quantitatives.

Démarche pour l'établissement de lignes de preuve

- *Évaluation de la qualité des études*

L'analyse de la qualité et pertinence des informations s'effectue par notation des informations selon des critères d'évaluation en regard d'une échelle numérique (de 1 à 10 par exemple).

- *Établissement lignes de preuves*

L'analyse d'un ensemble d'étude par type d'étude en vue d'établir des lignes de preuve est obtenue en calculant la moyenne pondérée des notations, sachant qu'un poids est associé à chaque critère. Les valeurs des poids sont déterminées par les experts.

Démarche d'intégration des lignes de preuve pour établir le poids des preuves

L'intégration des différentes lignes de preuve s'effectue en procédant de la même façon que pour l'étape précédente. Par construction, cette méthode autorise autant de niveaux d'intégration des informations que nécessaire avec la possibilité d'adapter la formule d'intégration selon le niveau et le type d'information intégré. La communication du niveau de preuve dans le contexte décisionnel est appuyée au moyen d'une mesure de l'incertitude (Monte Carlo) et/ou d'une analyse de sensibilité, en particulier concernant le poids des critères définis par les experts.

Cette méthode a été utilisée par :

- Linkov et al. (2011) pour la gestion des sédiments d'un lac en regard des recommandations du Comprehensive Environmental Resource Conservation and Liability Act (CERCLA). Deux jeux de critères ont été utilisés pour répondre aux 3 objectifs d'analyse (1- Evaluer des études individuelles relatives à la contamination, la toxicité et l'altération ; 2 - Evaluer les risques liés à la contamination du site et l'utilisation des sédiments ; 3- Evaluer les options d'assainissement) : ceux issus de US EPA (1997) et de US EPA (2003).
- Hristozov, Zabeo, et al. (2014) pour l'identification des risques liés à l'usage des nanomatériaux manufacturés en respect des recommandations de REACH pour le choix des critères.
- Hristozov, Gottardo, et al. (2014) pour le classement des scénarios d'exposition professionnelle au titane et carbone contenu dans les nanomatériaux. Les critères ont été établis à partir de la base NANEX, Development of Exposure Scenarios for Manufactured Nanomaterials (www.Nanex-project.eu).

La généricité de cette méthode permet de l'utiliser pour l'évaluation de tous types d'étude épidémiologique et expérimentale de l'Agence. Cette méthode nécessite de réaliser plusieurs choix à dire d'experts, notamment le choix des critères d'évaluation et le choix des poids associés à chaque critère.

Démarche pour l'expression des conclusions

Dans le cadre d'une analyse multicritères, Linkov et al. (2011) propose de conduire une analyse de sensibilité sur les poids des critères ainsi que pour certaines données d'entrée de façon à évaluer la stabilité de la classification.

3.9 Méthodes statistiques Bayésiennes

La statistique bayésienne constitue une des deux grandes familles de la statistique, l'autre famille étant la statistique dite classique ou fréquentiste. La statistique bayésienne permet d'intégrer à la connaissance que l'on a d'un système, l'information à son propos portée par des données (Groupe BioBayes 2015). Cette intégration se fait dans le cadre de la théorie des probabilités (Groupe BioBayes 2015). Une des spécificités de la statistique bayésienne est qu'elle se réfère à une interprétation subjective des probabilités. Contrairement à la statistique classique, la statistique bayésienne considère qu'une probabilité ne correspond pas à la fréquence de réalisation d'un évènement mais reflète un niveau d'incertitude subjectif. En statistique bayésienne, la probabilité n'est pas une propriété objective de l'évènement mais est liée au contexte de l'étude et dépend de la personne qui l'évalue. Comme la statistique classique, la statistique bayésienne permet d'estimer des quantités d'intérêt et de tester des hypothèses. Mais, contrairement à la statistique classique, la statistique bayésienne n'utilise pas uniquement des données expérimentales ou d'observations mais, également, des informations dites *a priori* qui peuvent venir d'avis d'experts.

Démarche pour l'évaluation du poids de la preuve

La statistique bayésienne permet ainsi d'intégrer des sources d'information diversifiées provenant à la fois d'expérimentation, d'enquêtes et d'avis d'expert, par exemple les données et avis d'expert sur les seuils de sensibilité à des allergènes chez des espèces animales et chez l'homme (Gosling et al. 2013). Ces informations *a priori* sont synthétisées à travers une distribution de probabilité *a priori*. Les données ou observations sont ensuite utilisés pour corriger la distribution *a priori* et pour calculer une distribution de probabilité *a posteriori* à l'aide du théorème de Bayes. Cette distribution *a posteriori* résume l'ensemble des informations disponibles sur une quantité d'intérêt. La statistique bayésienne est une méthode quantitative qui est souvent utilisée en analyse des risques, notamment du fait de sa capacité à décrire l'incertitude et parce qu'elle permet à l'utilisateur de combiner à la fois des données expérimentales et des avis d'experts (Gosling et al. 2013). Elle a été appliquée dans de nombreux autres domaines, par exemple :

- Pour analyser les essais cliniques et évaluer l'efficacité de traitements médicaux (Spiegelhalter et al. 2004)
- Pour analyser les risques de contamination des aliments (Williams et al. 2011)
- En toxicologie, pour estimer des doses de référence (Guha et al. 2013)

La statistique bayésienne est utilisée de plus en plus fréquemment, mais sa mise en œuvre peut poser plusieurs problèmes pratiques. Un de ces problèmes est de définir la distribution *a priori*. Cette étape peut être réalisée en mobilisant des méthodes d'élicitation d'experts, parfois délicate à appliquer. Une deuxième difficulté est de calculer la distribution *a posteriori*. Celle-ci peut rarement être calculée analytiquement et doit être approchée à l'aide d'algorithmes qui requièrent parfois des temps de calcul importants.

3.10 Evaluation comparative du poids des preuves (Meek et al, 2014)

Meek et al (2014) proposent une méthode d'évaluation comparative du poids des preuves dans le domaine de la toxicologie, plus précisément pour l'étude des modes d'action (Meek, Palermo, et al. 2014, Meek, Boobis, et al. 2014).

Démarche pour la planification de l'évaluation

Meek et al (2014) positionne la « formulation du problème » en tête de son protocole de conduite d'une analyse du mode d'action. Cette étape inclut l'identification du périmètre de gestion du risque et des objectifs en regard des scénarios d'exposition potentiels, des ressources disponibles, l'urgence de son évaluation et le niveau d'incertitude acceptable.

Démarche pour l'évaluation du poids des preuves

L'évaluation comparative du poids des preuves est une approche experte semi-qualitative permettant d'intégrer différentes sources de données, telles que des informations toxicologiques, épidémiologiques et mécanistiques, in silico, in vitro et in vivo afin d'envisager des potentiels modes d'action et leur concordance d'espèce.

L'approche est basée sur les critères de Bradford Hill (Hill 1965) modifiés, précisés et pondérés pour tenir compte de leur importance relative, basée sur l'expérience acquise par les experts dans l'analyse du mode d'action. La description et la pondération des critères de Bradford Hill modifiés sont basées sur l'analyse d'un nombre important d'exemples de cas de mode d'action documentés (Meek et al. 2003, Boobis et al. 2006, Boobis et al. 2008). Les critères retenus et modifiés sont :

1. La concordance/plausibilité biologique : Le mode d'action hypothétique s'oppose-t-il à une plus vaste connaissance biologique? A quel point le mode d'action est-il bien établi ?

2. Le caractère essentiel des évènements importants : est-ce que la séquence d'événements est réversible si le dosage est arrêté ou si un événement-clé est bloqué?

3. Le support empirique (concordance des observations empiriques entre des événements clés) sur la relation dose-réponse (les événements clés sont-ils observés aux doses inférieures ou similaires à celles associées aux effets néfastes finaux ?) et le support empirique sur la temporalité (les événements principaux sont-ils observés dans l'ordre hypothétique? L'incidence des effets néfastes finaux est-elle inférieure les événements clés précédents ?)

4. La cohérence (parmi les différents contextes biologiques) : l'ensemble des observations parmi des espèces/organes/systèmes correspond-il à ce qui est attendu selon l'hypothétique mode d'action ?

5. L'analogie (uniformité entre les produits chimiques) : le mode d'action pourrait-il être anticipé sur la base d'une connaissance spécifique de produits chimiques d'une même catégorie ?

En outre, des descriptions pour l'assignation aux catégories de poids de preuve faible, modéré et élevé pour chacune des cinq considérations de Bradford Hill modifiées ont été établies afin d'améliorer la cohérence et la transparence des évaluations du poids des preuves pour le mode d'action et de faciliter une application à plus large échelle, en remplissant des systèmes d'information éléctroniques sur les analyses de mode d'action et de « adverse outcome pathways » (AOPs) (OCDE 2014). L'approche a été adoptée, incorporée dans les lignes directrices et couramment appliquée dans une application réglementaire dans l'évaluation des hypothétiques modes d'action. Meek (2008) liste des exemples représentatifs d'application au sein des organismes règlementaires.

La méthode comparative de poids des preuves sur différents modes d'action hypothétiques est fondée sur la comparaison des poids relatif des éléments de preuve pour un éventail d'options, sur la base d'un examen de l'étendue des données cohérentes et incohérentes pour chacune des considérations Bradford Hill modifiées. Cette analyse comparative fondée sur des considérations « codifiées » reflétant l'expérience acquise dans les cas précédemment documentés permet un examen transparent et cohérent du poids de la preuve pour une série d'hypothèses.

3.11 Hristozov et al (2014)

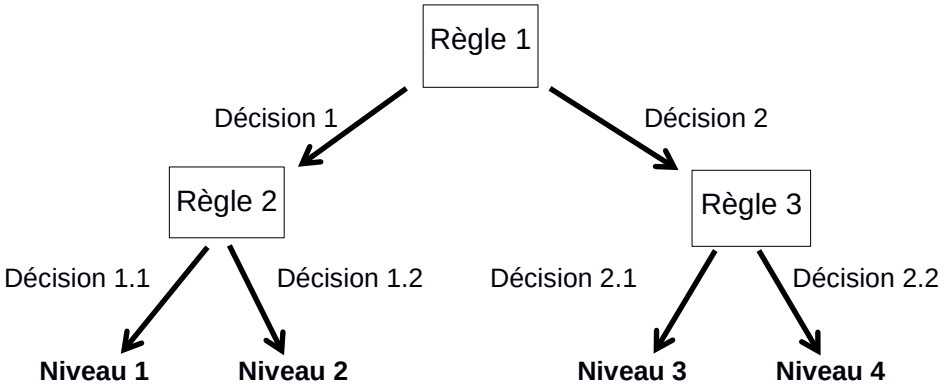
Démarche pour l'expression des conclusions

Pour les nanomatériaux, Hristozov, Gottardo, et al. (2014) propose de comparer les différentes classifications au moyen du modèle de normalisation et de calcul des différences développé par Nardo et al. (2005), puis de conduire une analyse d'incertitude relative à la performance du modèle et au jugement des experts concernant les données manquantes au moyen de la méthode de Monte-Carlo. Les résultats obtenus permettent de l'impact des déductions et hypothèses des experts sur les résultats.

3.12 Arbre de décision

Démarche pour l'intégration des lignes de preuve

Un arbre de décision représente graphiquement un ensemble de règles de classification organisées de manière arborescente. Plusieurs méthodes décrites précédemment peuvent être appliquée, et présentées sous la forme d'un arbre de décision. Cette approche s'applique à l'étape « Intégration des différentes lignes de preuve ». Les sorties d'un arbre de décision peuvent être quantitatives ou qualitatives. Un arbre de décision comporte plusieurs nœuds reliés par des branches. Les nœuds décrivent les différentes règles (ou tests) prises en compte par l'arbre, et les branches décrivent les différentes décisions possibles. Les nœuds terminaux (souvent appelés « feuilles ») indiquent les valeurs (quantitatives ou qualitatives) de variables cibles, comme par exemple des niveaux de risque.



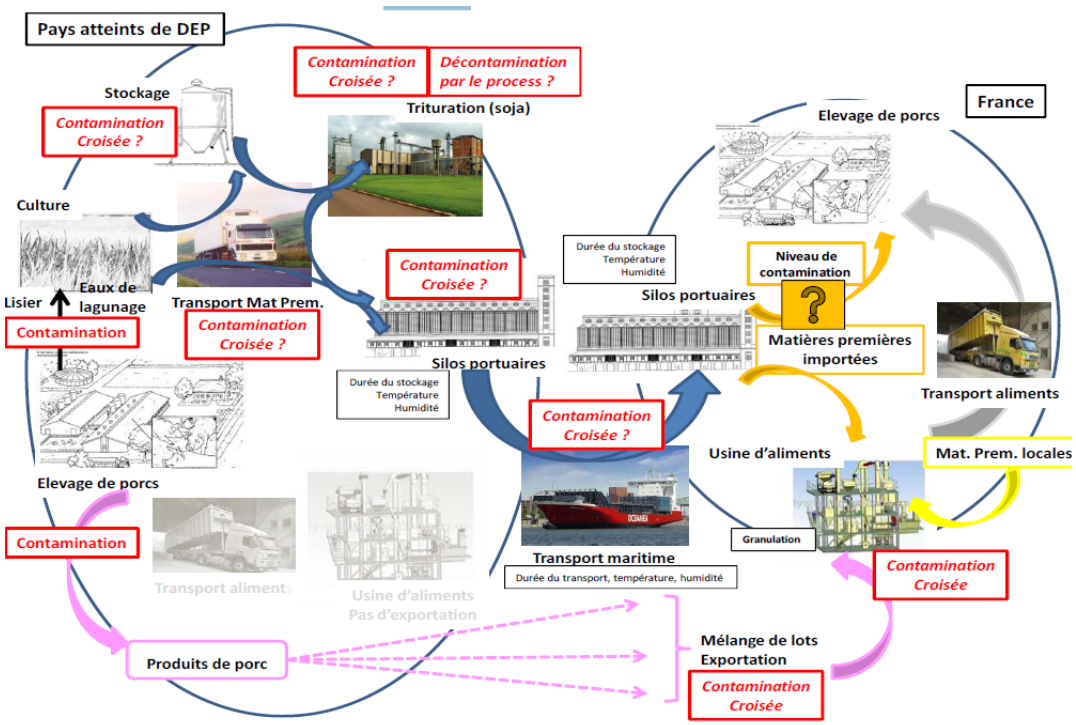
Les règles de classification d'un arbre de décision peuvent être basées sur des informations diverses, notamment sur des études expérimentales, des observations, des sorties de modèles ou sur des avis d'experts. Un arbre permet ainsi de combiner différents types de preuves et de relier ces preuves à des niveaux de risque de manière transparente. Les arbres de décision ont été utilisés dans divers domaines, par exemple :

- Pour analyser les risques associés aux radiofréquences et au bisphénol A, par l'Anses (Anses 2013c, d)
- Pour analyser les risques d'allergies liées aux organismes génétiquement modifiés (FAO/OMS 2001)
- Pour analyser les risques environnementaux liés à l'utilisation de sédiments contaminés (Khosrovyan et al. 2015)

La construction d'un arbre de décision est cependant délicate car elle nécessite d'identifier des règles pertinentes et de pouvoir associer des décisions précises à ces règles. En analyse des risques, les arbres de décision sont généralement construits par des collectifs d'experts (Anses 2013c, d, FAO/OMS 2001) et font parfois l'objet de versions successives (Metcalfe 2005).

Annexe 8 : Exemple de modèles conceptuels

Schéma événementiel des différentes sources potentielles de contamination par le virus à l'origine de la diarrhée épidémique porcine (Anses 2014b)



Ce modèle est élaboré dans le contexte de l'appréciation du risque qu'un aliment composé ne contenant pas de matières premières à base de porc soit contaminé par le virus à l'origine de la diarrhée épidémique porcine (DEP). La figure fait ressortir les différentes sources potentielles de contamination d'un aliment composé fabriqué en France, à partir de matières premières exclusivement végétales, dont certaines importées depuis un pays tiers infecté par la DEP.

Annexe 9 : Saisines Anses utilisant des méthodes d'évaluation du poids des preuves analysées par le GT MER

Mot clé (référence abrégée)	Titre complet du rapport	CES/ GT	Thème
Facteurs de croissance (Anses 2012)	Étude des liens entre facteurs de croissance, consommation de lait de produits laitiers et cancers	CES NUT	Alimentation humaine et nutrition
Cobalt (Anses 2014f)	Proposition de valeurs limites d'exposition à des agents chimiques en milieu professionnel - Evaluation des effets sur la santé des méthodes de mesure des niveaux d'exposition sur le lieu de travail pour le cobalt et de ses composés à l'exception du cobalt associé au carbure de tungstène.	CES VLEP	Santé-travail
Acétaldéhyde (Anses 2014e)	Proposition de valeurs guides de qualité d'air intérieur. L'acétaldéhyde	CES AIR / GT VGA	Santé-environnement
Radiofréquences (Anses 2013b)	Mise à jour de l'expertise « Radiofréquences et santé »	CES AP / GT Radiofréquences et santé	Santé-environnement
n-hexane (Anses 2014d)	Valeur toxicologique de référence chronique par voie respiratoire pou le n-hexane.	CES SUBSTANCES /GT VTR	Santé-environnement
BPA (Anses 2013a)	Évaluation des risques du bisphénol A (BPA) pour la santé humaine - Tome 1 : Évaluation des risques du bisphénol A (BPA) pour la sant humaine et aux données toxicologiques et d'usage des bisphénols S, F, M, B, AP, AF, et BADGE	CES SUBSTANCES /GT PE	Santé-environnement
DEP (Anses 2014b)	Risque d'émergence de la diarrhé épidémique porcine (DEP) en Europe par le biais de l'alimentation animale	GECU DEP	Santé et alimentation animale
Abeilles (Anses 2015a)	Hiérarchisation des dangers sanitaires exotiques ou présents en France métropolitaine chez les abeilles	CES SANT	Santé et alimentation animale

Notes

Agence nationale de sécurité sanitaire
de l'alimentation, de l'environnement et du travail
14 rue Pierre et Marie Curie
94701 Maisons-Alfort Cedex
www.anses.fr
www.anses.fr / @Anses_fr

ISBN 979-10-286-0118-8 - Dépôt légal : septembre 2016 – © Anses Éditions : septembre 2016 – Date de publication : septembre 2016