Université de Montréal

Revue systématique de l'évaluation de la qualité et du contenu scientifique des guides de pratique

par
François Désy

École de santé publique de l'Université de Montréal
Département de gestion, d'évaluation et de politique de santé

Mémoire présenté à l'École de santé publique
en vue de l'obtention du grade de M.Sc.
en évaluation des technologies de la santé et gestion

avril, 2018

Résumé

Contexte : Les guides de pratique cliniques (GPC) sont des énoncés incluant des recommandations destinées à optimiser les soins au patient reposant sur une revue des données probantes dont les liens avec le contenu des recommandations devraient être explicites et constants. Nombreux sont les guides de pratique qui souffrent d'un manque de rigueur dans leur développement et dont la qualité est sous optimale. En outre, il peut y avoir discordance entre le contenu des recommandations du guide et les données des études primaires sur lesquelles elles reposent de telle sorte que des recommandations peuvent s'avérer inappropriées. Le défi est donc de parvenir à déterminer le niveau de qualité d'un guide de pratique, non seulement de la façon dont le guide de pratique a été développé, mais également de la valeur du contenu ou de la validité des recommandations qu'il comporte.

Objectif : L'objectif est de caractériser les outils disponibles pour évaluer la qualité des GPC et notamment la validité du contenu des recommandations et potentiellement proposer une amélioration à cet égard.

Méthodes : Une première revue systématique a été entreprise dans le but d'identifier les revues existantes ayant déjà procédé à la recension des instruments d'évaluation des GPC. Les bases de données Pubmed et Embase ont été consultées de janvier 1990 à décembre 2015. Puis, cette revue a été complétée par une seconde revue systématique qui ne portait, elle, que sur des études primaires, soit les instruments ayant été développés depuis la recherche documentaire de la revue la plus récente. Les bases de données Pubmed, Embase et Evidence Based Medecine ont été consultées de janvier 2011 à décembre 2015. Une veille scientifique utilisant la même stratégie de recherche a été maintenue jusqu'en décembre 2017.

Résultats : La première revue systématique a permis de repérer 3 revues dénombrant elles-mêmes 47 instruments d'évaluation. La seconde revue systématique a permis d'ajouter 5 instruments de plus. Il apparaît que l'instrument AGREEII est en quelque sorte considéré comme l'étalon d'or des outils d'évaluation. Il est complet sans être excessif et permet une évaluation spécifique par domaine tout comme une évaluation globale en plus d'avoir été validé. Une réflexion sur des situations dans lesquelles un contenu scientifique inapproprié pourrait

passer inaperçu a mené à la suggestion d'évaluer spécifiquement pour chacune des recommandations clés:

- Si le lien est explicite entre les recommandations et les données probantes sur lesquelles elles reposent;
- Si les données probantes utilisées demeurent à jour;
- Si les données probantes utilisées s'appliquent toujours aux patients actuels.

Conclusion : L'ajout suggéré à l'outil AGREEII est susceptible d'améliorer la capacité d'avoir un aperçu de la valeur du contenu scientifique des recommandations d'un GPC.

Mots-clés : Guide de pratique clinique, lignes directrices, évaluation, AGREE

Abstract

Background: Clinical practice guidelines (CPG) are statements that include recommendations for optimizing patient care based on a thorough review of evidence that should be explicitly and consistently linked to the content of the recommendations. Many practice guides suffer from a lack of rigor in their development and their quality is suboptimal. In addition, there may be some discrepancy between the content of the guide's recommendations and the data from the primary studies on which they are based so that some recommendations may be inappropriate. The challenge, therefore, is to determine the level of quality of a practice guide, not only how the practice guide was developed, but also the value of the content or validity of the recommendations it contains.

Objective: The objective is to characterize the tools available to assess the quality of practice guidelines and in particular the validity of the content of the recommendations and potentially propose an improvement in this respect.

Methods: A first systematic review was undertaken in order to identify existing reviews that have already reviewed practice guidelines assessment instruments. Pubmed and Embase databases were consulted from January 1990 to December 2015. Then, this review was supplemented by a second systematic review which only covered primary studies, i.e. instruments that had been developed since the most recent review. Pubmed, Embase and Evidence Based Medecine databases were consulted from January 2011 to December 2015. A scientific watch using the same research strategy was maintained until December 2017.

Results: The first systematic review identified 3 systematic reviews counting 47 assessment instruments. The second systematic review added 5 more instruments. It appears that the AGREEII instrument is somehow considered as the gold standard for evaluation tools. It is complete without being excessive and allows a specific evaluation by domain as well as a global evaluation in addition to having been validated. Reflecting on situations where inappropriate scientific content may go unnoticed has led to the suggestion of specifically assess for each of the key recommendations:

- Whether there is an explicit link between the recommendation and the supporting evidence;
- Whether the evidence used is still up to date;

• Whether the evidence used still applies to current patients.

Conclusion: The suggested addition to the AGREEII tool is likely to enhance the ability to gain insight into the value of the scientific content of practice guideline CGP recommendations.

Keywords: clinical practice guidelines, appraisal, evaluation, AGREE

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles

ACG : American College of Gastroenterology

AGA : American Gastroenterological Association

ACC/AHA: American College of Cardiology /American Heart Association

AGREE: *Appraisal of Guidelines for Research and Evaluation*

AMSTAR: *A Measurement Tool to Assess Systematic Reviews*

BSG : British Society of Gastroenterology

COI : *conflict of interest,* conflit d'intérêt

CAG : Canadian Association of Gastroenterology

CEPO : Comité de l'évolution des pratiques en oncologie

CHC : carcinome hépatocellulaire

CPGs : *Clinical practice guidelines*, guides de pratique clinique

EANM: European Association of Nuclear Medicine

ECCO : European Crohn's and Colitis Organisation

ESC/EAS: European Society of Cardiology / European Atherosclerosis Society

GLIA. *GuideLine Implementability Appraisal*

GPC : Guide de pratique clinique

HCC: *hepatocellular carcinoma,* carcinome hépatocellulaire

IAS : International Atherosclerosis Society

IOM: Institute Of Medecine

RPC : recommandations pour la pratique clinique

SNM: Society of Nuclear Medicine

SIGN: *Scottish Intercollegiate Guidelines Network*

TAE: *transarterial embolization,* embolisation transartérielle

TACE: *transarterial chemoembolization,* chimioembolisation transartérielle

Liste des abréviations

vs : versus

À la mémoire de mes parents, Thérèse et Conrad

Remerciements

Merci à Dave Ames, par qui j'ai eu mon tout premier contact avec la discipline de l'évaluation des technologies de la santé.

Merci à Hang Cheng, Claude Warren et Jean Rousseau d'avoir supporté ma candidature au programme Ulysse.

Merci à tous les membres qui formaient cette septième cohorte du programme Ulysse. Tous et toutes possédaient des formations et bagages d'expérience forts différents, mais étaient animés de la même soif d'apprendre dans une atmosphère de camaraderie.

Merci à Lucy Boothroyd et Nicolay Ferrari pour leur collaboration qui s'est traduit en un apport significatif au manuscrit du chapitre 3.

Merci à mon directeur de recherche le Dr Luigi Lepanto pour son ouverture d'esprit, ses judicieux conseils et surtout d'avoir accepté de composer avec les conditions particulières dans lesquelles j'ai effectué ce travail.

Dans ma thèse en 2001, je mentionnais les gazouillis de Xavier, mon fils, qui n'avait que quelques mois. Plusieurs années ont passé, ponctuées d'événements marquants, si bien que je remercie maintenant, 6 garçons : Xavier, Dominic, Christophe, Emmanuel, Félix et Renaud, que j'aime profondément et que j'encourage à poursuivre et atteindre leurs rêves.

Merci à ma conjointe Karine pour son amour, son soutien, ses bons mots toujours opportuns et pour toutes ces périodes au cours desquelles j'ai dû lui fausser compagnie. Merci de m'encourager à me réaliser dans tous les aspects de ma vie. Je t'aime.

CHAPITRE 1 : INTRODUCTION

Les lignes directrices ou guides de pratique clinique (GPC) sont définis de plusieurs façons dans la littérature. Certains aspects sont presqu'invariablement invoqués, notamment la rigueur de leur développement ainsi que leur capacité à éclairer le clinicien quant à l'option de traitement la plus judicieuse à privilégier, à la lumière d'une lecture adéquate des données probantes les plus récentes.

Les travaux de Marilyn J. Field and Kathleen N. Lohr de l'Institute of Medicine (IOM) au début des années 90 (1, 2) constituent des références phares qui ont grandement influencé le développement des GPC et largement contribué à en faire ce qu'ils sont aujourd'hui en traçant les grandes lignes de ce que doivent contenir les GPC et également la façon et la rigueur avec lesquelles ils doivent être élaborés et éventuellement évalués, le tout reposant essentiellement sur la crédibilité et la responsabilité.

1.1 Problématique

Afin de remplir pleinement et adéquatement leurs fonctions, les GPC devraient exposer de façon explicite et constante les liens directs entre le contenu des recommandations et les données probantes scientifiques et cliniques à partir desquelles elles ont été élaborées (2). Toutefois, il est manifeste que, dans la littérature scientifique, tel n'est pas le cas et que les GPC affichent des différences substantielles dans leurs qualités méthodologiques, leur processus de développement, leurs contenus et que plusieurs recommandations sont basées sur des données probantes de faible qualité (3, 4). Cette situation est préoccupante dans la mesure où des GPC peuvent contenir des recommandations invalides et ainsi compromettre l'atteinte du bénéfice anticipé.

C'est d'ailleurs précisément ce qui a conduit à la recommandation explicite, formulée à l'endroit des utilisateurs de GPC, de procéder à l'évaluation critique de la méthodologie ainsi que du contenu des recommandations avant de les incorporer dans leur pratique (5, 6).

Plusieurs outils ou instruments (les deux termes sont employés indifféremment) d'évaluation de la qualité des GPC ont été développés au fil des années mais il s'avère qu'ils se concentrent davantage sur les aspects méthodologiques qui ont mené à l'élaboration du GPC et trop peu sinon pas du tout sur le contenu des recommandations, ce qui, pourtant, apparaît essentiel afin

de déterminer la mesure dans laquelle ces recommandations sont effectivement fiables et valides (3) et ultimement de décider si elles doivent être implantées ou non. Ceci est d'autant plus vrai qu'un haut niveau de qualité méthodologique n'est pas garant de la qualité du contenu des recommandations d'un GPC (5, 7).

Le problème de l'évaluation du contenu scientifique des GPC demeure entier. Idéalement, un GPC devrait recommander aux cliniciens ce qu'il convient de faire dans des situations précises et une évaluation positive de la qualité d'un GPC devrait conforter les cliniciens sur le bien-fondé et la validité de ces recommandations. Sans évaluation fiable du contenu d'un GPC, les cliniciens peuvent demeurer dubitatifs face à la décision à prendre de suivre ou non les recommandations émises et se trouver dans une sérieuse impasse lorsque, par exemple, des GPC émettent des recommandations contradictoires tel que rapporté encore récemment (8). Dans ces cas, les GPC occasionnent précisément le contraire de l'effet recherché, puisqu'ils suscitent des interrogations alors qu'ils sont plutôt sensés y répondre.

1.2 Question de recherche

Le besoin est manifeste de la part des intervenants visés par un GPC et particulièrement des cliniciens accaparés par leur pratique de disposer d'un outil fiable ou d'une façon de faire permettant de procéder à l'évaluation de la qualité des GPC incluant la validité des recommandations qui en sont issues. La question de recherche qui émerge de ce qui qui a été avancé jusqu'ici est donc : quels sont les instruments qui permettent d'évaluer le plus adéquatement possible la qualité d'un GPC, incluant la validité du contenu scientifique des recommandations?

CHAPITRE 2 : ÉTATS DES CONNAISSANCES

Dans un premier temps, il convient d'exposer certaines généralités sur les GPC. Cet exercice s'avère pertinent dans la mesure où il permet non seulement de définir ce dont il est question mais également de cerner l'ampleur avec laquelle les GPC se sont développés à travers les années et surtout l'importance qu'ils peuvent représenter pour un clinicien. L'enjeu ultime pour les cliniciens étant de pouvoir éventuellement s'assurer de la validité d'un GPC d'une façon suffisamment satisfaisante pour avoir confiance dans le contenu des recommandations, eux dont la tâche première est de dispenser des soins et services et non de synthétiser de façon critique les données probantes en lien avec l'objet de leur pratique.

2.1 Généralités sur les guides de pratique

2.1.1 Définition

Les GPC peuvent être considérés comme des énoncés élaborés de façon systématique pour assister le clinicien et le patient dans leurs décisions relatives aux soins de santé appropriées lors de circonstances cliniques spécifiques (1) ou encore des énoncés incluant des recommandations destinées à optimiser le soin au patient reposant sur une revue systématique des données probantes et une évaluation des bénéfices et risques des différentes options de traitement (9).

Dans le contexte québécois, l'Institut national de santé publique du Québec définit simplement les guides de pratiques comme des « recommandations formulées de façon systématique pour soutenir les praticiens et les patients dans leurs décisions concernant les soins à prodiguer selon les symptômes cliniques » (10) (page3). La Fédération québécoise des centres de réadaptation en déficience intellectuelle et en trouble envahissant du développement propose, quant à elle, une définition davantage opérationnelle sous plusieurs aspects en présentant un GPC comme « un ouvrage décrivant la manière d'exercer une activité dans un contexte spécifique s'appuyant sur les valeurs et les principes reconnus. Il guide les choix visant l'amélioration des services. Il est généralement produit par des comités d'experts et devrait contenir des standards de pratiques basés sur des données probantes. Il doit être régulièrement révisé sur la base des nouvelles données disponibles » (Mercier 2010 cité par (7)).

Dans la pratique médicale, le recours aux GPC est plus que nécessaire étant donné l'augmentation fulgurante de ce que l'on peut désigner, au sens large, comme la littérature scientifique (11) dont la croissance annuelle contemporaine avoisine les 8-9 % (12) faisant en sorte que les cliniciens ne peuvent plus demeurer au fait des bases en rapide expansion des connaissances reliées à la santé (9).

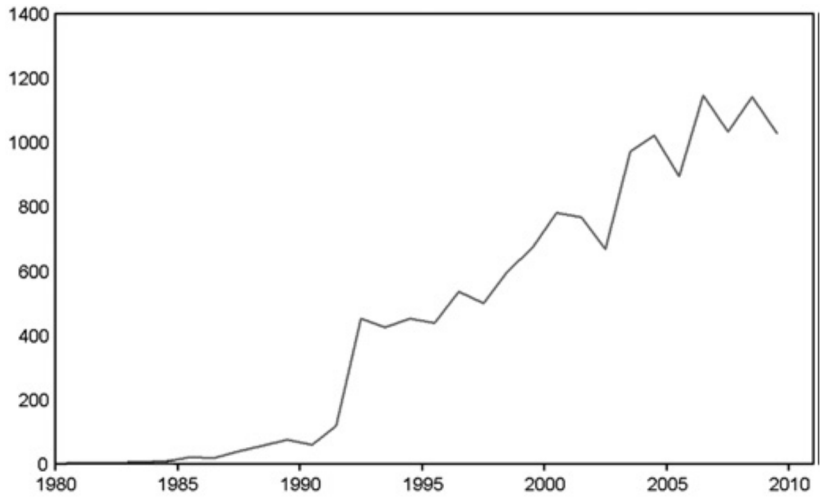
En termes pratiques, cela signifie que, pour un clinicien de carrière dont la durée de pratique totale s'échelonnera sur une période de 35-40 ans, l'ensemble du corps de connaissances doublera de 4 à 5 reprises. Comme ce clinicien consacrera la grande majorité de son temps à œuvrer auprès des patients et non pas uniquement mettre à jour ses connaissances, il devient essentiel pour lui ou elle de pouvoir s'en remettre à une 'autorité compétente' qui pourra effectivement remplir cette tâche. En outre, le fait que la qualité des données publiées doive elle-même être remise en question ajoute à cette problématique. En fin de compte, les cliniciens sont de plus en plus bombardés avec un volume énorme d'évidences et, qui plus est, de valeur incertaine ce qui rend l'obtention de données probantes synthétisées et évaluées d'une façon critique fondamentale pour la pratique clinique (9).

Le principal bénéfice des GPC est d'améliorer la qualité des décisions cliniques ainsi que la qualité des soins reçus par les patients (13) en réduisant l'écart entre les résultats de la recherche et les modalités de pratique actuellement en cours (14). Effectivement, en favorisant le recours à des pratiques affichant les meilleurs aspects touchant la sécurité, l'efficacité potentielle ou expérimentale (*efficacy*), l'efficacité réelle ou pragmatique (*effectiveness*), les impacts éthiques, psychologiques, organisationnels et économiques démontrés aux dépens de celles aux profils inférieurs, les GPC offrent le potentiel d'améliorer la prise en charge des patients, réduire la mortalité, optimiser leur qualité de vie, le tout possiblement à meilleurs coûts. De plus, les GPC offrent une clarification sur les pratiques à adopter ou délaisser simplifiant du même coup la prise de décision lors de situations cliniques particulières.

2.1.2 Évolution des guides de pratiques

Depuis leur apparition dans la littérature dans les années 80, le nombre de GPC s'est accru de façon importante, particulièrement au début des années 90 tel que le montre la figure 1.

Figure 1 Nombre de guides de pratiques identifiés avec PubMed, selon l'année de publication.



Tiré d'Alonso-Coello *et al.* 2010 (14).

Suivant la même tendance, le nombre de publications traitant de divers aspects reliés aux GPC notamment la façon de les élaborer, leur contenu, leur portée et la façon de les diffuser et de les implanter s'est accru également, tout comme une certaine inquiétude en ce qui a trait aux variations des recommandations ainsi que de la qualité des GPC (14). Qui plus est, l'évolution rapide du corps des connaissances touche aussi la validité du contenu des recommandations dont la durée est estimée à environ 3 ans, signifiant ainsi qu'une mise à jour des GPC devrait systématiquement être entreprise à cet intervalle (15).

2.1.3 Développement et contenu d'un guide de pratique

La pertinence de consacrer une section du présent travail à au processus de développement d'un GPC tient au fait que chacune des étapes constitue autant de portes d'entrée potentielles à une source d'invalidité. Hypothétiquement, le parfait instrument d'évaluation de la qualité d'un GPC serait en mesure de rendre compte de l'intégrité de chacune des étapes.

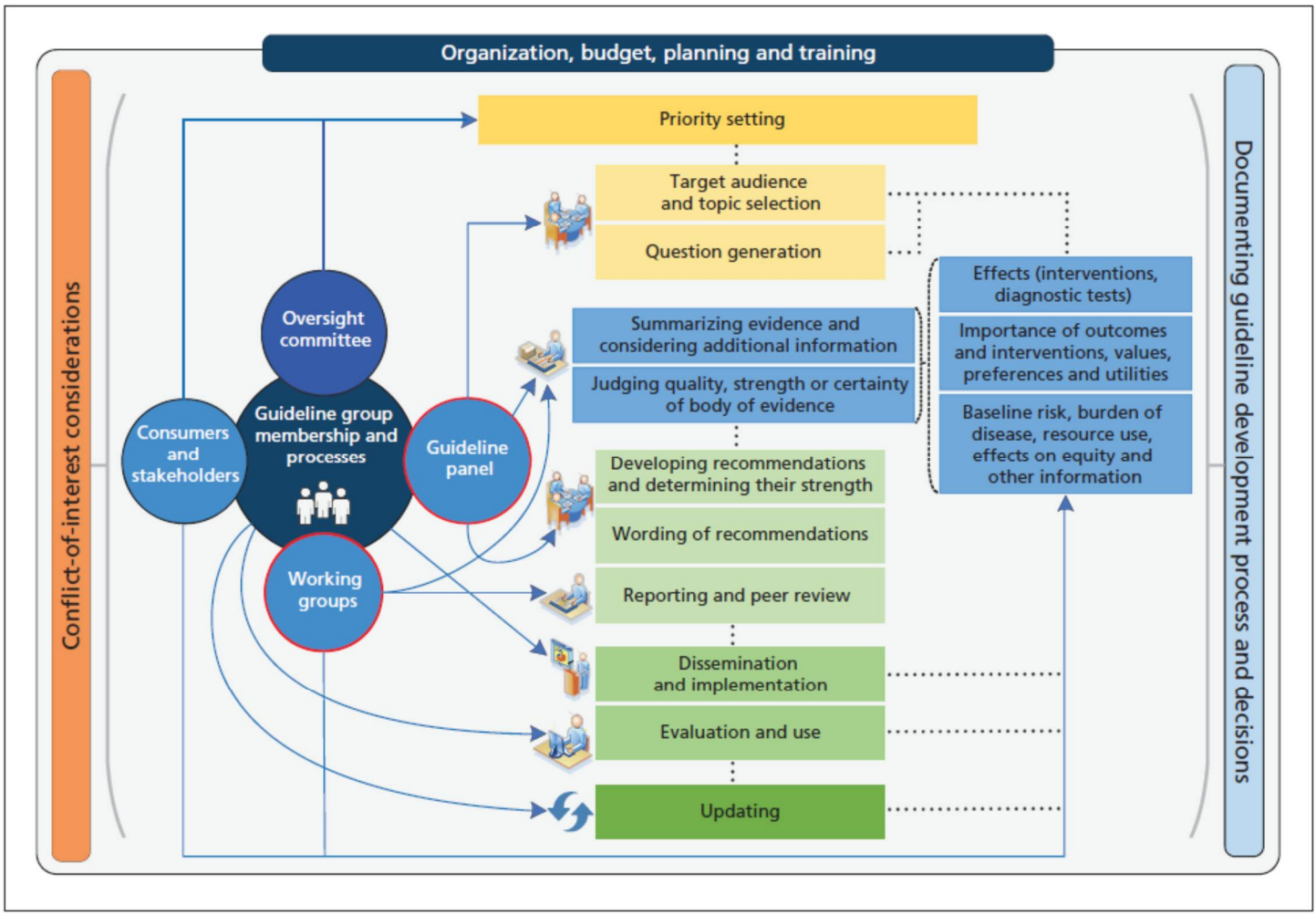
L'élaboration de GPC peut s'effectuer selon quatre approches principales, énumérées selon la rigueur de leur développement et dont chacune tente de répondre le plus adéquatement possible aux limitations de la précédente (7, 16) :

- Le consensus informel; soit un groupe d'experts déterminant d'un commun accord les pratiques à favoriser en utilisant une démarche non définie;
- Le consensus formel; soit une approche similaire, à la différence que le processus est formalisé et que le consensus se construit alors systématiquement selon des règles précises et définies explicitement, favorisant ainsi une plus grande transparence et la considération équitable de l'ensemble des opinions exprimées, ce qui limite les biais potentiels d'un consensus informel;
- L'approche centrée sur les données probantes; cette approche implique le recours à une méthode systématique de collection des études et données pertinentes, ainsi que d'émettre des recommandations en lien direct avec la revue effectuée. Une certaine rigueur scientifique entre alors en jeu et les opinions des experts reposent sur leur interprétation des évidences contenues dans la littérature pouvant elles-mêmes faire l'objet d'une gradation;
- L'approche explicite; cette approche comprend la précédente à laquelle s'ajoute un effort de circonscrire et d'expliciter les impacts attendus (bénéfice potentiel, dommages possibles et coût des options) et comprend notamment les opinions et préférences des patients.

De façon très claire, l'approche explicite est de loin la plus rigoureuse et la plus exigeante en termes de ressources nécessaires pour son élaboration.

Plusieurs plans de développement plus ou moins détaillés ont été proposés dans la littérature. L'illustration schématique de Schünemann *et al.* (17) reproduite à la figure 2, démontre bien l'ampleur de la tâche que représente l'élaboration, en bonne et due forme, d'un guide de pratique.

Figure 2 Vue d'ensemble du processus de développement d'un guide de pratique.



Tiré de Schunemann *et al.*,2014 (17).

Les étapes illustrées de même que la participation des différents membres du groupe de développement sont inter reliées et ne s'effectuent pas nécessairement de façon séquentielle. Le comité de direction ainsi que les groupes de soutien (par exemple, méthodologiste, économiste de la santé, équipe de la revue systématique et le secrétariat pour le soutien administratif) travaillent en collaboration, en tenant compte des enjeux exprimés par les parties prenantes (17).

Le groupe SIGN (Scottish Intercollegiate Guidelines Network) a développé au cours des 20 dernières années, 144 GPC en date de septembre 2015 et proposent une série de 50 étapes s'échelonnant sur une durée d'une trentaine de mois afin de réaliser adéquatement un GPC[1]. Sans être aussi exhaustif, le tableau suivant suggéré par Lortie *et al.* (7) résume les étapes principales :

Tableau 1 Étapes principales de la réalisation d'un guide de pratique.

1. Sélection du sujet.
2. Délimitation du but et du public cible anticipé (scope).
3. Adaptation du guide.
4. Formation du groupe de développement.
5. Participation des patients.
6. Élaboration des questions.
7. Revue systématique.
8. Critères d'inclusion et d'exclusion des études et des données.
9. Évaluation des données probantes ou des preuves.
10. Développement des recommandations.
11. Élaboration d'une stratégie d'implantation.
12. Consultation pour évaluer la version préliminaire du guide de pratique.
13. Rédaction de la version abrégée.
14. Planification de l'évaluation, révision et mise à jour.

Tiré de Lortie *et al.* 2012 (7).

[1] Disponible au http://www.sign.ac.uk/pdf/50steps.pdf.

Il a été estimé que le développement d'un guide prend d'un à trois ans et qu'il en coûte de 100 000 $ à 800 000 $ (1, 9), dont une grande partie est reliée à l'étape de revue systématique.
L'ensemble de ces étapes sert manifestement le but de garantir la rigueur du processus afin d'offrir un produit fini de la plus grande validité possible et exempt de biais. Force est de constater que plusieurs auteurs de GPC omettent bon nombre des étapes énumérées précédemment.
Ainsi, dans le but de réduire le travail requis par la revue de littérature et ultimement les coûts qui y sont associés, on retrouve des GPC contenant des critères de recension des écrits très restrictifs (langue, années de publication, types de bases de données, nature des documents retenus…), néanmoins systématiques ce qui est un moindre mal. Toutefois, quelques GPC omettent un si grand nombre d'étapes que la rigueur de développement en est affectée ce qui n'est peut-être pas sans influencer la qualité des recommandations qui sont émises. L'hypothèse de «compromis» sur la revue de littérature qui pourraient avoir été faits par certains auteurs de différents GPC sur un même sujet, dans le but d'en réduire les coûts, a d'ailleurs été invoquée afin d'expliquer les recommandations discordantes issues de ceux-ci (18).

Le concept de «validité interne» a été invoqué comme élément clé d'un bon GPC (19) et réfère à la minimisation des biais potentiels pouvant éventuellement affecter les recommandations. Cela comprend, entre autres, la façon dont les données probantes sont repérées, sélectionnées, interprétées et utilisées comme fondements aux recommandations.
Les sections suivantes, ponctuées d'exemples, traitent ces aspects, qui interviennent lors de la réalisation même d'un GPC et dont l'intégrité devrait idéalement être mesurée par un bon outil d'évaluation.

2.1.4 Étendue de la revue de littérature

L'un des problèmes rencontrés avec des GPC dont la littérature citée n'a pas été revue de façon systématique est certainement le choix des articles sur lesquels se fondent la réflexion du GPC et les recommandations qui en découlent. Le danger de la présence de biais est grand et le plus évident est que les articles retenus l'aient été parce qu'ils soutiennent les propos des auteurs,

qu'ils aient été choisis à la pièce ou comme le dit l'expression anglophone qu'ils aient fait l'objet de «*cherry picking*» et conséquemment qu'ils ne représentent qu'une partie seulement de la réalité de l'état des connaissances.

Sans pour autant y avoir un biais de sélection clair, certains GPC portant sur un même sujet reposent sur des éléments de littérature dont la disparité n'est pas sans soulever un certain questionnement.

Ce genre de situation a précisément été relatée dans une revue sur les GPC portant sur l'hypertension (20). Les auteurs ont notamment mis en évidence qu'à travers les GPC, des différences majeures quant à la nature des recommandations reliées à la gestion pharmacologique de l'hypertension existaient, et ce, même lorsque les auteurs de GPC ont prétendu qu'ils ont lié la gradation de leurs recommandations au niveau des évidences. Cette variation était vraisemblablement reliée soit à la stratégie de recherche des auteurs, au processus de sélection de l'évidence scientifique ou encore à la façon dont les recommandations ont été formulées.

Dans une revue de 15 GPC sur le diabète de type II provenant de 13 pays différents, Burger *et al.* rapportent des résultats pour le moins révélateurs à cet effet. Sur le grand total de 1033 articles cités (excluant les revues), 185 (18%) se retrouvent dans deux GPC ou plus et seulement 10 (moins d'un pour cent) se retrouvent dans six GPC (21). Une concordance géographique significative s'est révélée entre la localisation des auteurs et celles des articles qu'ils ont cités. Ce biais ne semble toutefois pas avoir d'influence puisque les auteurs de la revue notent un consensus sur les recommandations des modalités de traitement préconisées (21).

Dans une revue de 21 GPC portant sur le recours à la glace pour le traitement des entorses de la cheville, les auteurs ont identifié huit études pertinentes à ce sujet. Alors qu'un seul GPC citait six de ces études, six GPC citaient quatre études ou plus et neuf GPC n'en citaient aucune (18). Les auteurs suggèrent comme explication potentielle que les auteurs de GPC aient pu réaliser un compromis sur la revue de littérature pour sauver du temps ou des ressources.

L'outil d'évaluation AGREE sera présenté ultérieurement mais il convient de mentionner ici qu'une limite importante de cet instrument a été relevée par les auteurs de la revue en question

dans la mesure où ils ont noté qu'un certain nombre de GPC ont été évalués positivement par l'instrument au niveau de la rigueur de développement malgré que la sélection des études fût inadéquate. Qui plus est, cette déficience avait été notée pour la première version de l'instrument AGREE et n'a pas été corrigée dans la seconde version, l'AGREEII, version qui demeure toujours en vigueur.

2.1.5 Interprétation de la littérature comme fondement aux recommandations

L'appréciation ou le jugement quant à la justesse d'une recommandation est, de toute évidence, une tâche ardue. Outre les questions que soulèvent l'envergure ou l'étendue de la revue de littérature si tant est qu'une ait effectivement été réalisée, demeure le point sans doute le plus crucial, soit d'évaluer si l'interprétation des études considérées est adéquate ou non. Ainsi, il est évident qu'une interprétation erronée de résultats d'études dont la forme affiche une grande rigueur scientifique aura des conséquences tout à fait pernicieuses surtout si le GPC dans laquelle elle se retrouve est, du reste, bien conçu et positivement évalué par un instrument sensé bien le faire.

Malheureusement, il est navrant de constater que des situations révélant des interprétations de données inexactes, voire aberrantes, sont relativement courantes dans la littérature. Qu'il s'agisse de l'articulation, en toute bonne foi « d'une vision différente », d'un simple manque de rigueur ou encore de malhonnêteté intellectuelle, ceci constitue un problème sérieux qui afflige la littérature scientifique en général, et du même coup les GPC. Il ne semble pas y avoir d'autre solution apparente que de réviser l'interprétation en question directement en comparant avec les données primaires et d'évaluer la justesse de l'interprétation.

Si tant est que l'interprétation de la littérature scientifique ou des données probantes soit juste, les liens avec les recommandations formulées dans un GPC doivent être explicites. L'IOM spécifie même que les évidences doivent avoir préséance sur le jugement expert (2). Il est également précisé que dans le cas où les évidences empiriques auraient d'importantes

limitations et que les experts parviendraient à des conclusions allant à l'encontre des dites évidences, le fondement des raisons supportant cet écart devrait être minutieusement exposé (2) Conscients que les experts doivent régulièrement se prononcer malgré l'absence de données probantes, les auteurs expliquent que le raisonnement scientifique général et les principes normatifs supportant les jugements devraient également être décrits.

Ces dernières recommandations revêtent une importance exerçant une réelle influence sur la qualité du contenu des GPC. Ainsi, le facteur déterminant la valeur d'une recommandation est bel et bien l'exposition explicite des fondements ayant mené à cette recommandation, sans égard nécessairement à la valeur scientifique intrinsèque des fondements en questions. Ainsi, un raisonnement juste, basé sur une interprétation appropriée de données de plus faible rigueur scientifique, permettra d'émettre des recommandations de plus grande valeur que dans la situation inverse, soit un raisonnement erroné, basé sur une interprétation inappropriée de données de plus grande rigueur scientifique. Ultimement, une opinion judicieuse pourrait donc s'avérer plus utile qu'une interprétation inadéquate de méta-analyses. L'évaluation des GPC sera traitée explicitement plus loin dans ce mémoire.

Dans les sections précédentes, nous avons essentiellement exposé comment devrait être conçu un guide de pratique de qualité. Maintenant dans la section suivante, seront présentés des exemples illustrant un manque de rigueur s'étant immiscé dans l'une ou l'autre des étapes de fabrication du GPC. La pertinence d'exposer de tels exemple tient du fait qu'un instrument d'évaluation devrait permettre de repérer ces situations dans lesquelles les GPC n'ont pas été élaborés dans les règles de l'art et ainsi mettre en garde l'utilisateur de GPC de cette situation.

2.2 Situations illustrant le manque de rigueur dans le développement et le produit de lignes directrices

D'un point de vue pragmatique, les GPC devraient essentiellement correspondre à des outils commodes, à jour, logiques, et pratiques pour le clinicien occupé par la pratique quotidienne, le

but étant l'optimisation les résultats obtenus par le patient par l'utilisation raisonnable et judicieuse de moyens' (22).

Or, rares sont les GPC en mesure d'afficher toutes ces caractéristiques. Même si l'objet même souhaité d'un GPC est de faire consensus, il n'en demeure pas moins que bon nombre de GPC font l'objet de critiques. Les récriminations portent principalement sur le manque de rigueur des GPC, tant au niveau du processus d'élaboration, de l'interprétation des données que dans le contenu des recommandations auquel s'ajoute la gestion des conflits d'intérêts.

Une étude a démontré que les GPC développés par un organisme international d'envergure tel que l'Organisation Mondiale de la Santé (OMS), pourtant perçu comme une autorité hautement crédible disséminant des recommandations sur les soins de santé solidement fondées (environ 200 guides sont publiés annuellement (23)), étaient principalement basées sur les opinions d'experts et avaient rarement recours à une revue systématique des données probantes (24). Bien que depuis 2007, les méthodes de développement des GPC de l'OMS se soient généralement améliorées étant plus systématiques et transparentes, notamment au niveau de la rigueur du développement, l'uniformité n'était pas encore acquise en 2013, faisant en sorte que le problème de qualité est encore d'actualité (25, 26).

Dans une revue de littérature récente portant sur des sous-spécialités de médecine interventionnelle de cardiologie, gastroentérologie, néphrologie et pneumologie, Feuerstein *et al*. ont recensé 149 guides de pratique dont 64 % ne contenaient pas de gradation du niveau d'évidences considérées. Sur 3425 recommandations étudiées, seulement 11 % étaient supportées par des évidences (données probantes) de niveau A soit celles issues de méta analyses ou d'études randomisées sans limitations importantes. Puisqu'il s'agit bel et bien ici d'une moyenne, il est pertinent de mentionner que quelques GPC ne contenaient aucune preuve de ce niveau. Ainsi, 89 % des recommandations reposaient sur des niveaux de preuve inférieurs, B et C, dont la méthodologie la plus robuste correspond à une étude randomisée comportant d'importantes limitations telles que des erreurs méthodologiques, jusqu'à la méthodologie la plus faible soit l'émission d'une simple opinion (27).

Le phénomène similaire est observé dans le cas des GPC de l'American Academy of Pediatrics alors qu'une étude de 2013 (28) rapporte que 23 % seulement des 394 recommandations contenues dans les 28 GPC considérés reposaient sur des études dites expérimentales (étude

clinique randomisée, étude clinique quasi-expérimentale), 46% sur des études observationnelles et 31% sur des opinions d'experts ou encore sans référence.
Dans d'autres champs thérapeutiques tels que les maladies infectieuses, les auteurs d'une revue sur les GPC touchant l'utilisation d'antibiotiques dans le traitement des infections pulmonaires ou des voies urinaires soulèvent d'importantes faiblesses dans l'ensemble des GPC sur le sujet. Ainsi, selon les auteurs, la pertinence des données des études primaires n'est pas discutée et, qui plus est, la relation entre ces données et les recommandations émises ne sont que rarement décrites. D'importantes réserves sont émises sur les stratégies de recherches, les considérations épidémiologiques et écologiques (caractéristiques propres aux habitants d'une ville ou d'un pays), les méthodes utilisées pour formuler les recommandations ainsi que sur la gestion de conflits d'intérêts (29).

Il s'avère donc que les auteurs de bon nombre de GPC ne procèdent pas à des revues systématiques de la littérature, introduisant du même coup des biais dans la sélection des données probantes considérées dont nul ne peut mesurer l'étendue ou l'impact. De plus, la qualité parfois aléatoire des processus suivis pour formuler les recommandations handicape sévèrement la qualité globale de plusieurs GPC.
Dans le cas des GPC circonscrivant des situations cliniques et contextes d'applications similaires, comme les auteurs ont accès sensiblement aux mêmes données primaires c.-à-d. méta-analyses, études randomisées, etc. on pourrait s'attendre à ce qu'ils en fassent une interprétation, peut-être différente, mais somme toute congruente et qu'ultimement, les recommandations émises aillent dans le même sens. Dans la pratique toutefois, la littérature foisonne de GPC dont les recommandations, à une certain époque ou encore aujourd'hui, manquent de constance, sont parfois discordantes ou sinon carrément contradictoires. Les quelques exemples suivants en témoignent.

Dans une récente étude sur la rénographie diurétique comparant les GPC de la Society of Nuclear Medicine (SNM) et de l'European Association of Nuclear Medicine (EANM), les interprétations de la qualité du drainage divergent à un point tel qu'elles influencent directement le diagnostic (30). Ainsi, certains cas représentatifs d'une obstruction selon les critères de la

SNM, sont jugés normaux par l'EANM. Les auteurs en concluent que cette situation est inacceptable et qu'un accord entre les deux groupes est requis de façon urgente (30).
Dans le traitement des maladies inflammatoires de l'intestin, l'American College of Gastroenterology (ACG), l'American Gastroenterological Association (AGA), la Canadian Association of Gastroenterology (CAG) ainsi que la European Crohn's and Colitis Organisation (ECCO) sont en faveur du recours au médicament infliximab pour maintenir une rémission de la colite ulcéreuse alors que la British Society of Gastroenterology (BSG) prend précisément la position contraire en recommandant de ne pas utiliser l'infliximab pour des cas équivalents (31).

Dans les GPC issus d'un champ thérapeutique relativement simple comme l'utilisation de glace dans le traitement des entorses aiguës de la cheville, la durée totale d'application recommandée varie de 45 minutes à 14 heures (18).

Dans le cas de l'analyse de l'urine chez des patients pédiatriques asymptomatiques, l'American Academy of Family Physicians and US Preventive Services Task Force le désapprouve (*recommend against*) alors que l'American Academy of Pediatrics, malgré la reconnaissance que les études observationnelles et randomisées supportent de ne pas avoir recours au test de dépistage, le recommande néanmoins ce qui, de l'avis de Liberati *et al.*, représente '*simply an unqualified consensus statement without any reference to the level of evidence supporting it'* (32).

Les situations exposées précédemment sont autant d'exemples dans lesquels la consultation de la littérature dans le but d'orienter la pratique clinique s'avère infructueuse au point tel de placer le clinicien dans une certaine impasse dont la seule issue est de prendre lui-même position.

La question des conflits d'intérêts déborde le cadre du présent travail, mais il demeure pertinent de l'aborder brièvement étant donné les répercussions importantes qu'elle peut entraîner. Même avec un plus grand recours aux méthodes améliorant l'identification et la classification des évidences supportant explicitement les recommandations, le manque d'attention au conflit d'intérêts individuel et organisationnel peut s'avérer la plus grande menace à la création de GPC

digne de confiance (33). Ceci s'avère d'autant plus justifié considérant les données rapportées par Feuerstein *et al.*, à l'effet que 40 % des recommandations émises par les 149 GPC considérées dans son étude étaient de niveau C soit l'émission d'opinion (27), ce qui n'est pas sans jeter un doute sur la valeur des recommandations émises.

Ce point de vue semble tout à fait fondé dans la réalité (34). Ainsi, une situation assez unique s'est présentée relativement au traitement de la thrombocytopénie idiopathique dans laquelle deux guides de pratiques ont été successivement publiés dans un court laps de temps, soit 15 mois, dans la revue Blood (35, 36). Les deux groupes d'auteurs, principalement des hématologistes, se distinguant notamment par leurs liens avec l'industrie pharmaceutique ont donc eu accès à des pools de données probantes quasi identiques. Si les recommandations quant au diagnostic sont relativement similaires, d'importantes différences sont constatées dans les recommandations ayant trait au traitement de la condition, les médecins n'ayant pas de liens avec l'industrie privilégiant une approche plus conservatrice (34).

2.3 Évaluation des guides de pratique

Les quelques exemples cités précédemment illustrent qu'il existe de façon manifeste une importante variabilité dans la façon dont les GPC ont été développés au fil des ans, qui n'est pas sans possiblement influencer la qualité du produit fini, pour tant que l'on considère qu'un GPC de qualité ne puisse être généré que par un processus rigoureux et détaillé. En fait, les désaccords entre GPC peuvent survenir pour des raisons valables ou non (37) et ne constituent pas la preuve que les guides de pratiques sont de piètre qualité (38). Parmi les raisons valables, on peut penser à tout ce qui a trait au contexte d'application des GPC. D'aucuns remettraient en question la pertinence de considérer des caractéristiques propres à un contexte d'utilisation, dans la mesure où ils peuvent grandement varier d'une juridiction à l'autre, et de façon suffisamment importante pour influencer le choix des meilleures approches à adopter. Contexte socio-économique (39), incidence et prévalence d'une condition, comme l'obésité par exemple, nombre et disponibilité de dispositifs technologiquement avancés, comme un appareil de résonnance magnétique ou un robot chirurgical, sont des exemples qui peuvent influencer la façon dont des GPC recommandent de prendre en charge des patients. Ainsi, cette situation,

loin d'être problématique, est plutôt souhaitable, puisque chacun des GPC remplit son mandat d'éclairer le clinicien quant aux décisions à prendre pour traiter de façon optimale les patients dont il a la charge puisqu'il peut alors identifier et suivre le GPC qui correspond à sa propre réalité.

Dans le cas où deux GPC (ou plus) traitant d'une même problématique dans un même contexte émettent des recommandations n'abondant pas dans le même sens, la problématique de l'évaluation est d'autant plus primordiale puisque le clinicien doit faire un choix quant au GPC à privilégier. L'un doit forcément être «meilleur» que l'autre ou, du moins, davantage susceptible de générer une façon de traiter les patients qui s'avérera supérieure. Théoriquement, les critères-clés d'un GPC devraient être identifiés et l'outil d'évaluation devrait essentiellement permettre d'apprécier dans quelle mesure un GPC rencontre effectivement ces critères.
Dans le document original de 1990, l'IOM cite 8 caractéristiques ou attributs qu'il serait souhaitable de retrouver dans un GPC bien conçu et parmi ceux-ci, la validité est considérée l'attribut plus critique quoiqu'il se révèle être également le plus difficile à évaluer. La validité est définie de la façon suivante : «*Practice guidelines are valid if, when followed, they lead to the health and cost outcomes projected for them, other things being equal. A prospective assessment of validity will consider the projected health outcomes and costs of alternative courses of action, the relationship between the evidence and recommendations, the substance and quality of the scientific and clinical evidence cited, and the means used to evaluate the evidence*» (2)(page 10). Cet énoncé soutient donc que la valeur d'un GPC repose d'abord sur les effets mesurables de son application sur les variables d'intérêt et conséquemment qu'elle ne peut, ultimement, être déterminée qu'en rétrospective comme cela a déjà été fait (40, 41). Toutefois, cette option n'est pas satisfaisante dans l'immédiat et d'ici à ce que cela puisse être effectué, une évaluation de la validité d'un GPC effectuée de façon prospective doit comprendre la relation entre les recommandations et les données probantes, la nature même des données scientifiques et cliniques ainsi que les moyens d'évaluer celles-ci.

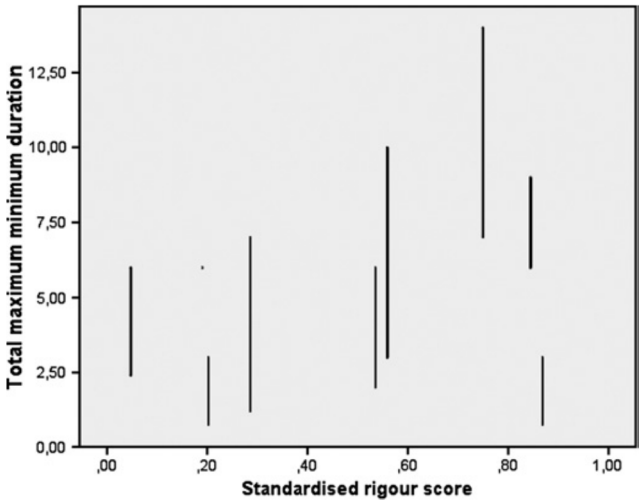
Problématique de l'évaluation du contenu scientifique

Ceci nous amène à explorer la problématique de l'évaluation du contenant *vs* du contenu des GPC. Il est permis de penser que les deux aillent généralement de pair, c'est-à-dire qu'un GPC bien 'conçu' dont la qualité a été évaluée favorablement, contiennent des recommandations valides issues d'une interprétation juste de la littérature scientifique. Toutefois, comme l'indiquent les auteurs de l'instrument AGREE, ce n'est pas le cas dans la pratique. «*A well reported guideline may contain flawed recommendations and, conversely, an unsystematically constructed one may provide sound evidence*» (Moher 1998, cité par (19)),

L'étude de Watine *et al.* 2006, sur les GPC portant sur la biologie médicale, illustre tout à fait cette réalité (5). Les auteurs ont identifié 11 GPC concernant les tests de laboratoire reliés au traitement du cancer du poumon non à petites cellules (*non-small cell lung cancer*) publiés au cours des 10 années précédentes. L'évaluation du contenant ou du processus d'élaboration du GPC a été obtenue en appliquant l'instrument AGREE, tandis que l'évaluation du contenu a été réalisée en comparant les recommandations à une revue systématique des données probantes préalablement effectuée par les mêmes auteurs. Les résultats révèlent une absence de relation entre les deux mesures. Ainsi, un GPC bien noté avec l'AGREE peut renfermer des recommandations de faible validité et vice-versa. Comme l'indiquent les auteurs, ce constat est inquiétant dans la mesure où le clinicien confronté à des recommandations de lignes directrices contradictoires n'a aucun moyen facile pour identifier quelle ligne directrice devrait être suivie (5) générant ainsi la situation précisément contraire à celle étant recherchée.

De façon similaire, Van de Velde *et al.* ont démontré, dans des GPC portant sur le recours à la glace pour le traitement des entorses de la cheville, une absence de relation entre les notes relatives à la rigueur de développement, telle que déterminée par l'AGREE, et la nature de la recommandation, en l'occurrence la durée totale d'application (figure 3) (18). On y voit clairement que la recommandation de la durée d'application varie pour des GPC ayant des notes de rigueur de développement similaires et inversement, soit que des durées d'application similaires sont recommandées par des GPC ayant des notes de rigueur de développement fort différentes.

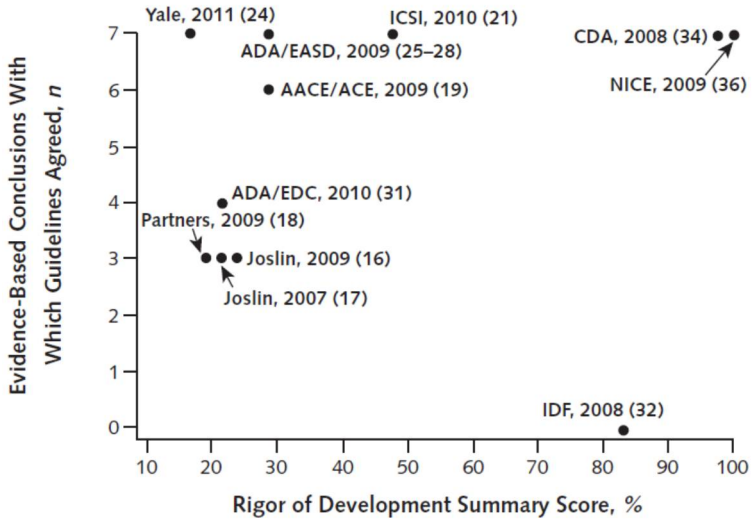
Figure 3 Valeurs maximales et minimales exprimées en heures de la durée globale d'application de glace dans le traitement recommandé pour les entorses à la cheville, en fonction de la note normalisée de rigueur de développement.



Tiré de Van de Velde *et al.* 2011 (18).

On retrouve également un phénomène semblable dans l'étude de Benett *et al.* portant sur les GPC en trait avec la médication orale pour le traitement du diabète de type II (42). Dans ce travail, les recommandations de chacun des GPC étaient comparées aux conclusions d'une revue systématique des données probantes préalablement réalisée. Puisque cette revue comportait 7 conclusions dites «basées sur les évidences», les auteurs ont comparé dans quelle mesure chacune des 7 conclusions de la revue systématique faisait l'objet de recommandations dans chacun des GPC donnant ainsi une note de 0 à 7. Le graphique suivant illustre la relation entre le nombre de conclusions concordantes avec la revue systématique en fonction de la rigueur de développement pour chacun des GPC.

Figure 4 Relation entre la rigueur de développement et le nombre de conclusions fondées sur les données probantes émises par chacun des GPC.



Tiré de Bennett *et al.* 2012 (42).

Bien qu'à l'évidence, la même conclusion que dans l'exemple précédent s'applique, c'est-à-dire que le nombre de recommandations «valides» varie pour des GPC ayant des notes de rigueur de développement similaires et inversement, soit que des nombres de recommandations «valides» similaires sont émis par des GPC ayant des notes de rigueur de développement fort différentes. Cependant, et ce de façon tout aussi déconcertante qu'incompréhensible, les auteurs concluent néanmoins que les GPC ayant reçu les meilleures notes de qualité c.-à-d. à l'évaluation de la rigueur de développement, contiennent davantage de recommandations en accord avec les conclusions basées sur les données probantes (42). Ceci expose manifestement la problématique de l'interprétation des données dont il sera fait mention plus loin dans ce mémoire.

D'importantes différences ont également été notées dans une revue de GPC sur le traitement de l'apnée du sommeil, non seulement en ce qui concerne le niveau des études sur lesquelles reposent les recommandations, mais également au niveau du contenu même des recommandations et de la nature des articles retenus, et ce malgré des notes quasi parfaites de tous les domaines AGREE (37).

De toute évidence, la rigueur de développement mesurée par l'instrument AGREE ne renseigne en rien sur la qualité et la validité des recommandations émises. Cette précision au niveau de la portée de l'instrument AGREE est ouvertement reconnue par les auteurs de l'instrument qui mentionnent également poursuivre des activités de recherche liées au concept de validité clinique et de pertinence (19)

Force est de constater qu'une certaine incompréhension puisse également exister au sein de la communauté scientifique, et qu'elle puisse même se révéler de façon explicite, par exemple lorsque Winther *et al.* affirment de façon erronée que l'instrument AGREE procure une évaluation de la validité prédite d'un GPC (43), ou encore une récente publication collective de l'Institut national d'excellence en santé et en services sociaux (INESSS) stipulant, toujours à tort, que la grille AGREE II permet d'évaluer la probabilité que l'observation des recommandations favorisera l'obtention des résultats escomptés (44). Ainsi, il semble que cette conception erronée quant aux prétentions mêmes de l'instrument AGREE soit à la source d'une certaine insatisfaction exprimée à son égard par certains individus. En d'autres mots, des auteurs entretenant la conviction qu'AGREE permet l'évaluation complète et totale de la qualité d'un GPC (processus et contenu) se trouvent déçus de constater que ce n'est précisément pas le cas.

Dans une étude comparant les GPC portant sur le traitement du syndrome de l'apnée du sommeil, dans laquelle des recommandations contradictoires ont notamment été relevées, Aarts *et al.* soutiennent que: «*The reported guidelines in our study met the AGREE quality criteria… Nevertheless, differences in conclusions, levels of evidence, and cited references were revealed. Therefore, assessing guidelines using these quality criteria alone may result in misleading comfort. The AGREE criteria proved to be not very helpful for assessing quality of the resulting guidelines because too many different items were added and calculated in a difficult-to-interpret summary domain score, where signal disappears in the rather large noise*»(37)

Conclusion de l'état des connaissances

L'élaboration des GPC est une entreprise complexe dont les nombreuses étapes sont autant de portes d'entrée dans lesquelles des sources d'invalidité peuvent s'immiscer et ultimement corrompre l'intégrité du produit fini et ainsi compromettre l'atteinte du bénéfice anticipé ayant motivé la conception même de la démarche. Pour qu'un GPC puisse réellement être bénéfique à la pratique clinique, il doit non seulement contenir des recommandations valides mais il faut également que cette validité puisse être étayée par un instrument d'évaluation adéquat dans la réalisation de cette tâche.

CHAPITRE 3: ARTICLE

Scientific content appraisal of clinical practice guidelines: adding to the current AGREEII instrument

Full title: **Scientific content appraisal of clinical practice guidelines: adding to the current AGREEII instrument**

Running header (shortened title): Scientific content appraisal of clinical practice guidelines

Forename(s) and surnames of authors: François Désy PhD, Nicolay Ferrari PhD, and Luigi Lepanto MD, MSc, FRCPC

Authors's affiliations:
École de santé publique,
Département de gestion, d'évaluation et de politique de santé
Université de Montréal
Québec, Canada

Corresponding author:
Luigi Lepanto
Unité d'évaluation des technologies et des modes d'intervention en santé
CHUM - Pavillon S
850, rue St-Denis, porte S05-504
Montréal (Québec) H2X 0A9
Telephone : 514 890-8000 ext. 36132
Email : luigi.lepanto.chum@ssss.gouv.qc.ca

Declaration of conflicts of interest: The authors have no conflicts of interest to declare

Abstract

Objective: Clinical practice guidelines (CPGs) are statements that contain recommendations for optimizing patient care based on a thorough review of evidence that should be explicitly and consistently linked to the content of the recommendations. Many CPGs display a lack of rigor in their development and are of suboptimal quality. Therefore, the challenge is to determine the quality of a CPG, not only regarding the development process, but also the value of the content of the recommendations.

The objective of this paperwork is to examine the tools available to evaluate the quality of CPGs, in particular the validity of the content of the recommendations, and potentially to propose ways to improve in this respect.

Methods: Two systematic reviews were conducted in order to identify 1) existing systematic reviews of appraisal tools and 2) additional evaluation tools developed more recently.

Results: The first review identified 3 documents presenting 47 evaluation instruments and the second led to the identification of 5 additional tools.

The AGREEII tool is considered the gold standard. A reflection on situations in which appropriate scientific content may go unnoticed led to suggest that the following aspects should be specifically assessed for each of the key recommendations in a CPG:

• Whether there is an explicit link between the recommendation and the supporting evidence;

• Whether the evidence used is still up to date;

• Whether the evidence used still applies to current patients.

Conclusion: Our suggested additions to the AGREEII tool are likely to enhance the ability to gain insight into the value of the scientific content of CPG recommendations.

Keywords: clinical practice guidelines, appraisal, evaluation, AGREE

Résumé

Objectif : Les guides de pratique clinique (GPC) sont des énoncés comprenant des recommandations pour optimiser les soins aux patients en fonction d'un examen approfondi des données probantes dont les liens avec le contenu des recommandations devraient être explicites et constants. De nombreux GPC affichent un manque de rigueur dans leur développement et leur qualité est sous-optimale. Par conséquent, le défi consiste à déterminer la qualité d'un GPC, non seulement le processus de développement, mais aussi la valeur du contenu des recommandations.

L'objectif est de caractériser les outils disponibles pour évaluer la qualité des GPC et notamment la validité du contenu des recommandations et potentiellement proposer une amélioration à cet égard.

Méthodes : Deux revues systématiques ont été menées afin d'identifier 1) les revues systématiques existantes et 2) les outils d'évaluation supplémentaires développés plus récemment.

Résultats: La première revue a identifié 3 documents présentant 47 instruments d'évaluation et la seconde revue a permis de trouver 5 outils supplémentaires.

L'outil AGREEII est considéré comme l'étalon d'or et une réflexion sur des situations dans lesquelles un contenu scientifique inapproprié pourrait passer inaperçu a mené à la suggestion d'évaluer spécifiquement pour chacune des recommandations clés:

• Si le lien est explicite entre les recommandations et les données probantes sur lesquelles elles reposent;

• Si les données probantes utilisées demeurent à jour;

• Si les données probantes utilisées s'appliquent toujours aux patients actuels.

Conclusion : L'ajout suggéré à l'outil AGREEII est susceptible d'améliorer la capacité d'avoir un aperçu de la valeur du contenu scientifique des recommandations d'un GPC.

Mots-clés : Guide de pratique clinique, lignes directrices, évaluation, AGREE

Introduction

Clinical practice guidelines (CPGs) are defined in several ways in the literature (1). Certain aspects are invariably invoked, including the rigor of their development as well as their ability to inform the clinician as to the most appropriate treatment option to favor, based on an appropriate interpretation of the most recent evidence.

The work of Field and Lohr of the Institute of Medicine (IOM) in the early 90s (2, 3) established key benchmarks that have greatly influenced the development of CPGs. In fact, it has outlined what CPGs should contain and also how rigorously they should be developed and eventually evaluated, all aspects being based primarily on credibility and accountability.

In order to be considered credible, CPGs should explicitly and consistently expose the direct links between the content of the recommendations and the scientific and clinical evidence upon which these were developed (2). However, it is clear that across the broad spectrum of the scientific literature, this is not the case and that CPGs show substantial differences in their methodological quality, development process, content, and that several recommendations are based on weak or low quality evidence (4, 5). This situation is of concern, as CPGs may end up providing inappropriate recommendations and thus compromise achievement of the expected benefit. This reality is precisely what led to the explicit recommendation made to CPGs users to critically appraise the methodology as well as the content of the recommendations before incorporating them into practice (6, 7).

CPG quality is a multifactorial concept which encompasses all elements likely to enable a CPG to fulfill its expected benefit, i.e., providing sound guidance for clinical decision making. Several quality assessment tools have been developed over the years, but these in fact focus more on the methodological aspects that led to the development of the CPG and too little if not at all on the actual recommendations. The latter approach is however, essential in order to determine the extent to which the recommendations are indeed reliable and valid (3) and, ultimately, to decide whether they should be implemented or not. This is particularly true because a high level of methodological quality does not in itself guarantee the quality of the content of a CPG's recommendations (6, 8).

The challenge of evaluating the scientific content of CPGs remains. Ideally, a CPG should provide guidance to clinicians on what could and should be done in specific clinical situations. A positive assessment of CPG quality should give clinicians confidence about the rationale and strength of the recommendations. Without an accurate assessment of CPG content, clinicians may remain skeptical

about whether or not to follow recommendations and may find themselves at a serious impasse when the CPG's basis appears weak or, worst, when different CPGs make conflicting recommendations as was recently the case (9). In such situations, CPGs have precisely the opposite of the desired effect since they raise questions rather than providing answers.

Objective

There is a clear need for CPG stakeholders, and particularly clinicians occupied with their practice, to have an easy-to-use tool or process to accurately assess the quality of CPGs, including the validity of recommendations.

The overall objective of the present work is to characterize what is available to the scientific community to assess the quality of CPGs and particularly the validity of recommendations. A critical reflection was subsequently carried out on the themes addressed by the various tools and on potential situations in which inappropriate scientific content could otherwise go unnoticed. The objective of this reflection is to suggest refinements of the appraisal process which could improve the ability of various stakeholders to judge the overall validity of a CPG.

Methods

An initial systematic review was undertaken to identify previously published systematic reviews of CPG appraisal instruments. A second systematic review was then performed in order to identify any new appraisal tools published since the last systematic review retrieved.

For the first systematic review, the MEDLINE (through PubMed) and EMBASE (through OVID) databases were searched from January 1st 1990 to December 31st 2015. The search strategy consisted of a combination of entire or truncated forms of the following words: guideline, appraisal, evaluation, scientific, tool, instrument, and systematic.

Articles had to be English or French language systematic reviews on CPG appraisal tools or on the quality evaluation of CPGs. Two authors (FD and NF) selected the documents. All screened documents were either excluded or retained for consideration based on abstract reading. Eligibility was determined after full-text reading. The quality of included documents was evaluated with the AMSTAR instrument (10) and disagreements between reviewers were resolved by consensus.

For the second systematic review, the MEDLINE (through PubMed), EMBASE (through OVID) and EBM Reviews databases were searched from January 1st 2011, to coincide with the end of the literature search of the most recent systematic review (11), and up to December 28th 2015. The search strategy consisted of a combination of entire or truncated forms of the following words: guideline, appraisal, quality, evaluation, scientific, analysis, comparison, attributes, evidence, grading, AGREE, and review. A manual search based on the bibliographies of the retrieved documents was also performed.

The same two authors (FD and NF) selected the documents as previously described. A scientific watch with the same search strategy in PubMed was carried out until December 1st 2017.

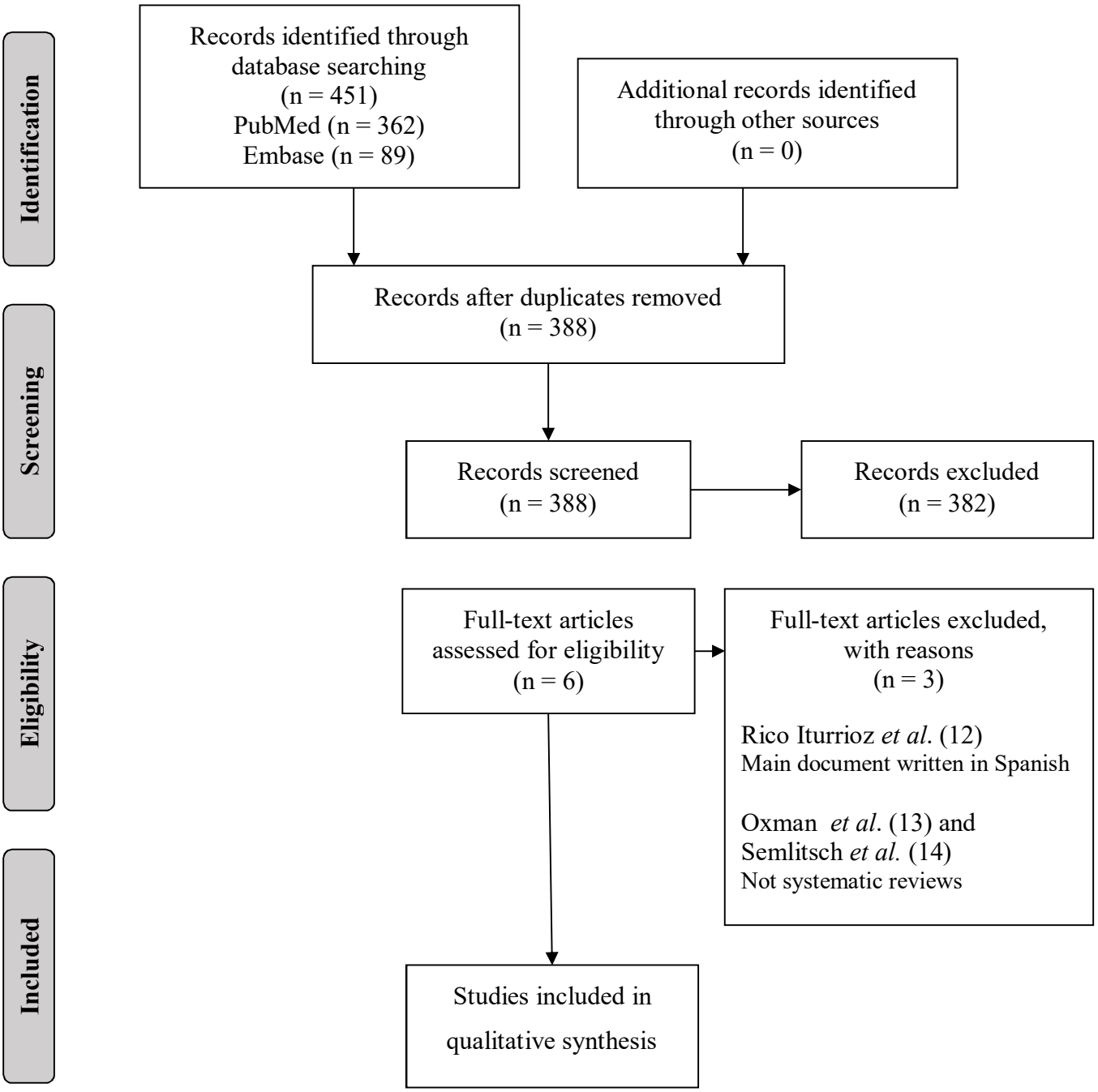
The term 'review' was considered in the broadest way possible in order to identify articles about reviewing, appraising or making some sort of judgement about CPGs. As a first step, all articles performing some kind of CPG assessment were kept. As a second step, we determined for each article if the approach used was original or if it was the strict application of an already known appraisal instrument. If the article used a new appraisal instrument that was generic enough to be applied to different clinical settings, it was retained. Articles were not kept if they addressed the evaluation of CPG implementation or laboratory analyses. Disagreements between reviewers were resolved by consensus. Since there is no appraisal tool for a CPG evaluation instrument, the quality of the retrieved documents could not be measured.

Results

Systematic review of systematic reviews

The first search yielded 451 documents from which 3 were selected. Figure 1 shows the PRISMA flowchart of the selected articles.

Figure 1 PRISMA flowchart of selected articles for the systematic review of systematic reviews on CPG appraisal instruments.



From the 6 documents assessed for eligibility, three were excluded. One was excluded because the abstract was in English but the main document was in Spanish (12) and two because the methodology used did not correspond that of a systematic review (13, 14). The three documents retained are in table 1.

Table 1 Systematic reviews of CPG appraisal instruments

Author, Year, Reference	Period of the search strategy	Number of tools identified	Classification of items and domains assessed	Recommended instrument (s)
Siering, 2013 (11)	1995-2011	40	34 items under 13 domains	AGREEII, 2009 (15) DELBI, 2008 (16) ADAPTE, 2010 (17)
Vlayen, 2005 (18)	1966-2003	24	50 items under 10 domains	Cluzeau, 1999 (19)
Graham, 2000 (20)	1966-1999	15	44 items under 10 domains	Cluzeau, 1999 (19) Shaneyfelt, 1999 (21)

The three systematic reviews identified were evaluated with the AMSTAR appraisal tool which does not provide a way to convert the total score into a qualitative category (i.e. poor, good…)(10). Given that the appraisal results were 7 'yes' out of a possibility of 9 for the most recent review and 5 'yes' for each of the remaining two others, all retrieved reviews were deemed to be of satisfactory quality.

The time period covered by the reviews extends over 45 years but the oldest instrument examined was published in 1992. This instrument was created by the IOM and its influence has been such that Graham and collaborators reported that 8 instruments out of the 15 in their review were directly based on it (20).

As a whole, the 3 systematic reviews gathered 47 CPG appraisal tools (suppl.1) after the removal of those published in a language other than English. From 1992 to 2012, the number of appraisal tools grew at a steady rate of approximately 5 new instruments every two years.

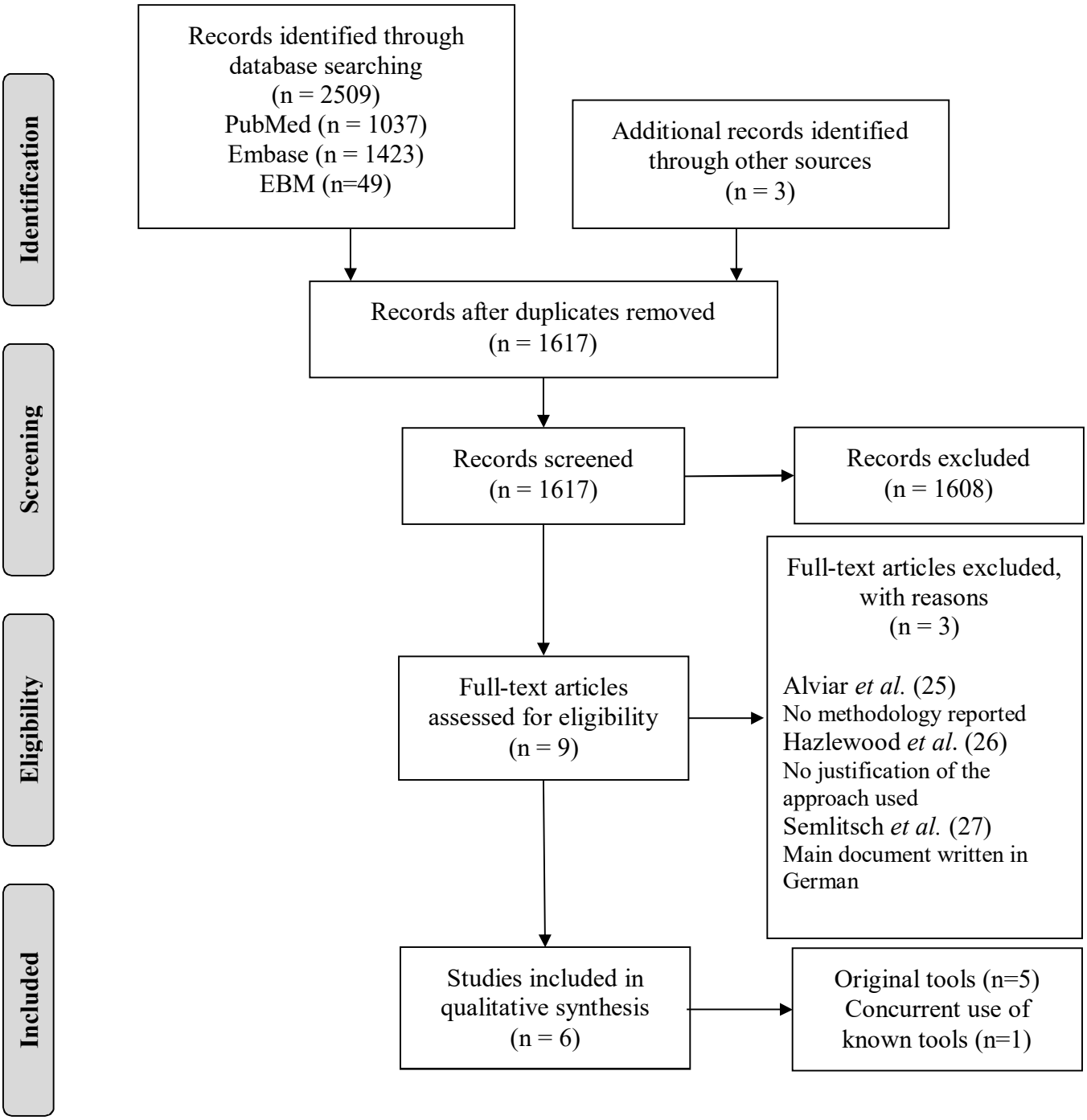
As far as recommended tools are concerned, the 3 systematic reviews point to the work of Cluzeau and collaborators that served as a basis for the AGREE tool development, for which a second version has been published in 2009 (15).

Systematic review of appraisal tools

The purpose of this second review was to identify any new appraisal instrument that has been developed since the literature search of the most recently published systematic review (11).

The second search strategy yielded 2509 documents to which 3 documents from others sources were added: one from the manual search (22) and two from the scientific watch (23, 24). At the end of the selection process, 6 articles were retained. Figure 2 shows the PRISMA flowchart of the selected articles.

Figure 2 PRISMA flowchart of selected articles for the systematic reviews of additional original CPGs appraisal instruments or approach.



Of the 9 documents assessed for eligibility, three were excluded. One was excluded because no methodology was reported (25). A second document was excluded because it did not provide justification of the approach used (26). Finally, the third one was excluded because only the abstract was in English and the document itself was in German (27). However, this new appraisal tool was published in English several months later and the latter document was identified through the scientific watch (24).

The 5 new tools retained for the current systematic review are shown with the 5 instruments recommended by the previous systematic review in table 2.

Table 2 CPG appraisal instruments identified in the current systematic review or suggested by previous systematic reviews.

Current systematic review				
Author, year, tool, reference	**Number of questions/domains**	**Domain specific appraisal**	**Global evaluation**	**Validated instrument**
Shaughnessy, 2017 GTRUST, (23)	8/3	No	No	No
Siebenhofer, 2016 MiChe, (24)	8/4	No	Yes	Yes
Coroneos, 2014 (22)	8/3	Yes	No	No
Grimmer, 2014 iCAHE, (7)	14/6	Yes	No	No
Brouwers, 2012 GRS, (28)	5+2/5	Yes	Yes	No

Previous systematic reviews				
Author, year, tool, reference	**Number of questions/domains**	**Domain specific appraisal**	**Global evaluation**	**Validated instrument**
ADAPTE, 2010 (17)	43/3	No	No	Yes
AGREEII, 2009 (15)	23/6	Yes	Yes	Yes
Kopp, 2008 DELBI (16)	34/8	Yes	No	No
Shaneyfelt, 1999 (21)	25/3	No	No	Yes
Cluzeau, 1999 (19)	37/3	Yes	Yes	Yes

The domains referred to by the authors encompass different areas or themes under which individual appraised items can be conceptually grouped. These are presented in Table 3.

Table 3 Domains assessed by specific CPG appraisal instruments

Author, year, reference	Domains assessed
Shaughnessy, 2017 (23)	Relevance threats Evidence threats Interpretation threats
Siebenhofer, 2016 (24)	Quality of guideline creation Quality of reporting Quality of presentation Quality of evidence synthesis
Coroneos, 2014 (22)	Are the recommendations valid? What recommendations are made? Will the results help me in caring for my patients?
Grimmer, 2014 (14)	Scope and purpose Stakeholder involvement Rigour of development Currency Availability Summary
Brouwers, 2012 (28)	Process of development Presentation style Completeness of reporting Clinical validity Overall quality
ADAPTE, 2010 (17)	Search and selection of evidence Scientific validity of guidelines Acceptability/applicability
AGREEII, 2009 (15)	Scope and purpose Stakeholder involvement Rigor of development Clarity of presentation Applicability Editorial independence
Kopp, 2008 (10)	Scope and purpose Stakeholder involvement Rigor of development Clarity of presentation Applicability Editorial independence Applicability to the German healthcare system
Shaneyfelt, 1999 (21)	Format and development Identification and summary of evidence Formulation of recommendations
Cluzeau, 1999 (19)	Rigor of development

One last approach (26) was retained because of the originality of undertaking a two-step evaluation process by first evaluating the quality of the evidence using the GRADE tool (29) and then evaluating the quality of the CPG using the AGREEII (9).

Discussion

The aim of the present work was to characterize what is available to the scientific community to assess the quality of CPGs and particularly the validity of recommendations, and to make suggestions for refining the appraisal process. The current work identified 5 new instruments which add to the other 47 already reported by previous systematic reviews (11, 18, 20).

This proliferation of CPG appraisal tools over the years clearly reflects the dissatisfaction that the scientific community has had since the beginning of CPGs regarding the available means to properly evaluate them. Each newly developed tool was arguably an attempt at producing a finished product that would somehow perform better than anything else available at the time.

AGREEII as the gold standard

Based on the documents consulted, there is an agreement that the AGREEII tool (9), although not perfect, is considered the gold standard of CPG appraisal instruments. The AGREEII (9) is at the end of a refinement assembly line which started with the very first published appraisal instrument by the Institute of Medecine in 1992 (2). Physicians were not the intended users of this comprehensive provisional instrument that consisted of 142 questions addressing 46 items under 7 main domains. From this foundation, Cluzeau and collaborators (19) produced a more easier-to-use tool in 1999 which was later used as the basis for the development of the AGREE tool in 2003 (30) and the AGREEII in 2009 (15). Cluzeau's document and both AGREE tools have all been recommended by authors of the systematic reviews (11, 18, 20) as it was felt they were complete and comprehensively validated, an opinion that still holds today. It is true that from the sole standpoint that an ideal CPG evaluation tool should allow validation of the presence or absence of major items required to be found in a sound CPG, the AGREEII balances being complete without being excessive (Table 3). Moreover, the AGREEII offers the advantage of providing a domain-specific appraisal along with a global CPG assessment, and is one of the few instruments having gone through a form of validation (Table 2).

The need for speed

Trying to improve on this level of satisfaction, several groups have devoted a fair amount of effort developing assessment tools that would require less time to use than the AGREEII. Compared to the several hours needed to complete the first appraisal instrument (2), it takes 12 to 20 minutes to use the AGREEII depending on the expertise level of the appraiser (7). Recently developed rapid assessment tools improved further in this area with reported average required times of 13 minutes (based on 12

appraisers) for the MiChe (24) and an impressive 3-7 minutes (based on 3 appraisers) for iCAHE, (7). Confirming once again the gold standard status of the AGREEII, authors of rapid assessment instruments report the comparison of the results obtained by their newly developed tool against those of the AGREEII and have interpreted a lack of difference as confirmation of good performance (7, 23, 24, 28). This implies that the gain in assessment speed has not been earned at the expense of the ability to discern 'good' CPGs from 'bad' ones.

Validity appraisal

In spite of the fact that greater rapidity represents a major advantage in terms of ease of use, the fact remains that one, if not the main objective of appraising a CPG is to establish its validity. The appraised level of validity is likely to be commensurate with the level of confidence stakeholders will have when it will be time to implement recommendations embedded in the CPG. The challenge is that validity is the most difficult dimension to properly assess (2). In the AGREEII, the validity aspect is mainly captured within the 'rigor of development' domain, although editorial independence (or the lack thereof) can also have an impact on the validity of the whole process, to a lesser extent. Thus, potential threats to validity in terms of the building process undertaken and final format such as unbalanced representation of stakeholders, questionable literature search strategy, ambiguous wording, etc. would be noted and graded accordingly.

This leads to the imperative assessment of clinical content, since it represents the foundation on which recommendations are built. Even if a CPG is concisely written, by authors free of conflicts of interest, taking into account the different perspectives of the right stakeholders, this does not guarantee that the evidence collected was appropriately interpreted and incorporated. This could create a situation in which there would be a certain discrepancy between the content of recommendations and the actual scientific evidence, despite the CPG being positively assessed. Field and Lohr issued this warning some 25 years ago (2) and it remains relevant today. In fact, quality appraisal of the clinical content of guidelines is quite challenging as it is time-consuming, requires highly qualified personnel and may need additional information not available in the guidelines themselves (11). Such a limitation has also been acknowledged by the AGREE's authors, as they stated that their instrument was designed to assess the process of guideline development and how well this process is reported and thus the clinical content of the guideline nor the quality of evidence that underpins the recommendations (15). This observation applies to other appraisal tools as well, representing the 'common deficit' evoked by Vlayen in 2005 (18). In fact, there seems to be no evaluation tool enabling the structured and comprehensive assessment of the content of guideline recommendations with special regard to their reliability and validity (4).

Improvement proposal

In an attempt to improve CPG clinical content appraisal, we reflected on this current limitation of the AGREEII and tried to come up with a variety of scenarios where some level of invalidity could sneak in, and how these hypothetical situations could be flagged. These reflections led us to propose three recommendations.

The quality of different recommendations is not equal within one guideline

Firstly, from a conceptual point of view, it should be recognized that the quality of a practice guide is not uniform across the different elements of which it is composed.
In its present form, the items in the AGREE tool do not allow for this, as they are only assessed at the level of the whole guideline (4). As an example, item 12 - *there is an explicit link between the recommendations and the supporting evidence*, forces, in a way, the evaluator to make a judgment representing the average level of this item across all the recommendations. Now, it is clear that this link can be more or less explicit depending on the recommendation and that it would be highly desirable, when using an appraisal tool, to be able to distinguish each situation in order to adequately inform the interested stakeholder.

The *explicit link* concept could also be looked at as two distinct components coming into play: the accuracy of the evidence interpretation (i.e. if the recommendation content is truly in line with what the evidence says), and the level of strength of the evidence (i.e. if the recommendation adequately reflect the level of evidence in terms of its certainty).

In the end, a gain in the fine aspect of the assessment of item number 12 could positively impact the overall value of the CPG assessment process.

The first recommendation is to:

- Specifically assess item number 12- *there is an explicit link between the recommendations and the supporting evidence* for each of the main recommendations of a CPG.

Keeping up with innovation

Depending on the therapeutic field, medical innovations (drugs, devices, or intervention modes) can sometimes introduce drastic paradigm changes that reduce the time window for evidence to stay current and adequately describe contemporary reality. In high technology fields such as interventional cardiology, for example, this could be of significant importance. Ideally, a positive evaluation of a CPG would reassure various stakeholders that not only was evidence interpreted correctly and adequately used for a recommendation, but also that the considered evidence was indeed relevant to the current paradigm and was not contaminated by data that are no longer relevant.

The second recommendation is to:

- Specifically assess whether the evidence used for each of the key recommendations is still up to date.

Along the same lines, the use of innovations create growing subgroups of patients possibly with slightly different characteristics, especially if a new disruptive technology begins to be widely diffused (diabetics treated with and insulin pump instead of injections for example). In such instances, there would be, for a certain time at least, a mixed pool of evidence combining data from patients treated ‘the old way’ and ‘the new way’. Again, a positive evaluation of a CPG should reassure different stakeholders that the evidence used to make recommendations adequately applies to current patients.

The third recommendation is to:

- Specifically assess whether the evidence used for each of the key recommendations still applies to current patients.

The current work has some strengths and weaknesses. The systematic review search strategy and scientific watch were very broad and allowed retrieval of numerous relevant papers on CPG appraisal. It seems unlikely that an appraisal tool or procedure has been missed. The recommendations included in the present work were humbly made on the basis of our reflection and discussions of situations in which a CPG could receive a positive evaluation despite making inadequate recommendations for practice. The proposed improvements have not been validated nor submitted for external expert review.

Conclusion

With over 50 different CPG appraisal tools available, the choice of one over the other rests on the specific need of the appraiser in terms of the nature of the items to be assessed. The AGREEII tool is the end product result of a multiple iteration process and has earned the status of being considered the gold standard of CPG appraisal. Recently developed appraisal tools allow rapid assessments of CPGs without apparent compromise on the capacity to distinguish between CPGs of different overall quality when compared to using the AGREEII.
In terms of validity, the assessment of the clinical content of CPGs remains a challenge. In an attempt to identify situations in which inappropriate scientific content could otherwise go unnoticed, suggestions to refine current AGREEII items have been made:

- Specifically evaluate item number 12 - *There is an explicit link between the recommendations and the supporting evidence* for each of the key recommendations;
- Specifically assess whether the evidence used for each of the key recommendations is still up to date.
- Specifically assess whether the evidence used for each of the key recommendations still applies to current patients.

Further research will be needed to clarify how overall appraisal of the clinical content of a guideline can be included in appraisal tools with a reasonable use of resources (11). We consider the suggested additions to the AGREEII tool as a proposal that deserves to be looked into further as means to succeed in this endeavour.

1. Graham R. MM, Miller Wolman D., Greenfield S., Steinberg E.,. Clinical Practice Guidelines We Can Trust. US: National Academies Press (US); 2011. Available from: http://www.ncbi.nlm.nih.gov/books/NBK209539.

2. Institute of Medicine Committee on Clinical Practice Guidelines. In: Field MJ, Lohr KN, editors. Guidelines for Clinical Practice: From Development to Use. Washington (DC): National Academies Press (US) by the National Academy of Sciences.; 1992.

3. Institute of Medicine Committee to Advise the Public Health Service on Clinical Practice Guidelines. In: Field MJ, Lohr KN, editors. Clinical Practice Guidelines Directions for a New Program. Washington (DC): National Academies Press (US) by the National Academy of Sciences.; 1990.

4. Eikermann M, Holzmann N, Siering U, Ruther A. Tools for assessing the content of guidelines are needed to enable their effective use--a systematic comparison. BMC research notes. 2014;7:853.

5. Abdelsattar ZM, Reames BN, Regenbogen SE, Hendren S, Wong SL. Critical evaluation of the scientific content in clinical practice guidelines. Cancer. 2015;121(5):783-9.

6. Watine J, Friedberg B, Nagy E, Onody R, Oosterhuis W, Bunting PS, et al. Conflict between guideline methodologic quality and recommendation validity: a potential problem for practitioners. Clinical chemistry. 2006;52(1):65-72.

7. Grimmer K, Dizon JM, Milanese S, King E, Beaton K, Thorpe O, et al. Efficient clinical evaluation of guideline quality: development and testing of a new tool. BMC medical research methodology. 2014;14:63.

8. Lortie M, IRSST (Québec), Canadian Electronic Library (Firm). Bilan des connaissances sur les guides de pratique en santé enseignements clés et transférabilité pour la santé et la sécurité au travail. Montéal, Qué.: Institut de recherche Robert-Sauvé en santé et en sécurité du travail; 2012. Available from: http://myaccess.library.utoronto.ca/login?url=http://site.ebrary.com/lib/utoronto/Top?id=10575156.

9. Watine J, Wils J, Augereau C. Evaluation of the methodological quality of two contradictory guidelines recently published by the Haute autorite de sante. Annales de biologie clinique. 2017;75(1):101-6.

10. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10.

11. Siering U, Eikermann M, Hausner E, Hoffmann-Esser W, Neugebauer EA. Appraisal tools for clinical practice guidelines: a systematic review. PloS one. 2013;8(12):e82915.

12. Rico Iturrioz R, Gutierrez-Ibarluzea I, Asua Batarrita J, Navarro Puerto MA, Reyes Dominguez A, Marin Leon I, et al. [Assessment of clinical practice guidelines evaluation. Scales and criteria]. Revista espanola de salud publica. 2004;78(4):457-67.

13. Oxman AD, Schunemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 16. Evaluation. Health research policy and systems. 2006;4:28.

14. Semlitsch T, Blank WA, Kopp IB, Siering U, Siebenhofer A. Evaluating Guidelines: A Review of Key Quality Criteria. Deutsches Arzteblatt international. 2015;112(27-28):471-8.

15. The AGREE Next Steps Research Consortium. Grille AGREE II 2009 [cited 2015 29 septembre].

16. Kopp I, Thole, H., Selbmann, H.K., Ollenschlaeger,G. German Instrument for Methodological Guideline Appraisal Deutsches Instrument zur methodischen Leitlinien-Bewertung (DELBI). 2008.

17. ADAPTE Collaboration. Guideline adaption: a resource toolkit; version 2.0. http://www.g-i-n.net/document-store/adapteresource-. 2010. Available from: http://www.g-i-n.net/document-store/adapteresource-.

18. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua. 2005;17(3):235-42.

19. Cluzeau FA, Littlejohns P, Grimshaw JM, Feder G, Moran SE. Development and application of a generic methodology to assess the quality of clinical guidelines. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua. 1999;11(1):21-8.

20. Graham ID, Calder LA, Hebert PC, Carter AO, Tetroe JM. A comparison of clinical practice guideline appraisal instruments. International journal of technology assessment in health care. 2000;16(4):1024-38.

21. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. Jama. 1999;281(20):1900-5.

22. Coroneos CJ, Voineskos SH, Cornacchi SD, Goldsmith CH, Ignacy TA, Thoma A. Users' guide to the surgical literature: how to evaluate clinical practice guidelines. Canadian journal of surgery Journal canadien de chirurgie. 2014;57(4):280-6.

23. Shaughnessy AF, Vaswani A, Andrews BK, Erlich DR, D'Amico F, Lexchin J, et al. Developing a Clinician Friendly Tool to Identify Useful Clinical Practice Guidelines: G-TRUST. Annals of family medicine. 2017;15(5):413-8.

24. Siebenhofer A, Semlitsch T, Herborn T, Siering U, Kopp I, Hartig J. Validation and reliability of a guideline appraisal mini-checklist for daily practice use. BMC Med Res Methodol. 2016;16:39.

25. Alviar CL, Bangalore S, Messerli FH. Optimal blood pressure targets in 2014 - Does the guideline recommendation match the evidence base? Hipertension y Riesgo Vascular. 2015;32(2):71-82.

26. Hazlewood GS, Akhavan P, Schieir O, Marshall D, Tomlinson G, Bykerk V, et al. Adding a "GRADE" to the quality appraisal of rheumatoid arthritis guidelines identifies limitations beyond AGREE-II. Journal of clinical epidemiology. 2014;67(11):1274-85.

27. Semlitsch T, Jeitler K, Kopp IB, Siebenhofer A. [Development of a workable mini checklist to assess guideline quality]. Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen. 2014;108(5-6):299-312.

28. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. The Global Rating Scale complements the AGREE II in advancing the quality of practice guidelines. Journal of Clinical Epidemiology. 2012;65(5):526-34.

29. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.

30. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Quality & safety in health care. 2003;12(1):18-23.

31. Hayward RS, Wilson MC, Tunis SR, Bass EB, Rubin HR, Haynes RB. More informative abstracts of articles describing clinical practice guidelines. Annals of internal medicine. 1993;118(9):731-7.

32. Selker HP. Criteria for adoption in practice of medical practice guidelines. The American journal of cardiology. 1993;71(4):339-41.

33. Hayward RS, Wilson MC, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? The Evidence-Based Medicine Working Group. Jama. 1995;274(7):570-4.

34. Mendelson EB. The development and meaning of appropriateness guidelines. Radiologic clinics of North America. 1995;33(6):1081-4.

35. Woolf SH. Practice guidelines: what the family physician should know. American family physician. 1995;51(6):1455-63.

36. (SIGN). SIGN. Clinical Guidelines: Criteria for Appraisal for National Use.: Edinburgh: Royal College of Physicians. 1995.

37. Mottur-Pilson C. Internists' evaluation of guidelines: the IMCARE Practice Guidelines Network. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua. 1995;7(1):31-7.

38. Liddle J, Williamson, M., Irwig, L. Method for Evaluating Research Guidelines Evidence (MERGE). In: Department NSW, Health o, editors. Sydney,1996.

39. Savoie I, Green, C., Bassett, K., et al. Critical appraisal criteria for clinical practice guidelines. Vancouver: BC Office of technology Assessment,; 1996 (cited by Vlayen et al. 2005).

40. Ward JE, Grieco V. Why we need guidelines for guidelines: a study of the quality of clinical practice guidelines in Australia. The Medical journal of Australia. 1996;165(10):574-6.

41. Calder L, Hébert, P., Carter, A., Gaham, I. Review of published recommendations and guidelines for the transfusion of allogenieic red blood cell and plasma. Can Med Assoc J 1997;156(11 Suppl):S1-S8.

42. Cook DJ, Ellrodt AG, Calvin J, Levy MM. How to use practice guidelines in the intensive care unit: Diagnosis and management of unstable angina. Critical care medicine. 1998;26(3):599-606.

43. Veale B, Weller D, Silagy C. Clinical practice guidelines and Australian general practice. Contemporary issues. Australian family physician. 1999;28(7):744-9.

44. Fields SD. Clinical practice guidelines. Finding and appraising useful, relevant recommendations for geriatric care. Geriatrics. 2000;55(1):59-63.

45. Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet (London, England). 2000;355(9198):103-6.

46. Marshall JK. A critical approach to clinical practice guidelines. Canadian journal of gastroenterology = Journal canadien de gastroenterologie. 2000;14(6):505-9.

47. Sanders GD, Nease RF, Jr., Owens DK. Design and pilot evaluation of a system to develop computer-based site-specific practice guidelines from decision models. Medical decision making : an international journal of the Society for Medical Decision Making. 2000;20(2):145-59.

48. Savoie I, Kazanjian A, Bassett K. Do clinical practice guidelines reflect research evidence? Journal of health services research & policy. 2000;5(2):76-82.

49. Criteria for evaluating treatment guidelines. The American psychologist. 2002;57(12):1052-9.

50. Foy R, MacLennan G, Grimshaw J, Penney G, Campbell M, Grol R. Attributes of clinical recommendations that influence change in practice following audit and feedback. J Clin Epidemiol. 2002;55(7):717-22.

51. Fretheim A, Williams JW, Jr., Oxman AD, Herrin J. The relation between methods and recommendations in clinical practice guidelines for hypertension and hyperlipidemia. The Journal of family practice. 2002;51(11):963-8.

52. Guyatt G. HR, Richardson WS., Green L., Wilson MC., Sinclair J., Cook D., Glasziou P., Detsky A., Bass E. Moving from evidence to action. In: Guyatt G. DR, editor. User's guides to the medical literature Chicago: AMA Press; 2002. p. 175-99.

53. Hart RG, Bailey RD. An assessment of guidelines for prevention of ischemic stroke. Neurology. 2002;59(7):977-82.

54. Reed GM, McLaughlin CJ, Newman R. American Psychological Association policy in context. The development and evaluation of guidelines for professional practice. The American psychologist. 2002;57(12):1041-7.

55. Baxter N, Palda, VA. Guidelines for colorectal surgery. Semin Colon Rectal Surg. 2003;14:19-25.

56. Hasenfeld R, Shekelle PG. Is the methodological quality of guidelines declining in the US? Comparison of the quality of US Agency for Health Care Policy and Research (AHCPR) guidelines with those published subsequently. Quality & safety in health care. 2003;12(6):428-34.

57. Hutchinson A, McIntosh A, Anderson J, Gilbert C, Field R. Developing primary care review criteria from evidence-based guidelines: coronary heart disease as a model. The British journal of general practice : the journal of the Royal College of General Practitioners. 2003;53(494):690-6.

58. Kulig M, Schulte E, Willich S. Comparing methodological quality and consistency of international guidelines for the management of patients with chronic heart failure. European journal of heart failure. 2003;5(3):327-35.

59. Organization WH. Guidelines for WHO guidelines. 2003.

60. Shiffman RN, Shekelle P, Overhage JM, Slutsky J, Grimshaw J, Deshpande AM. Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization. Annals of internal medicine. 2003;139(6):493-8.

61. Nonino F, Liberati A. Essential requirements for practice guidelines at national and local levels. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology. 2004;25(1):2-7.

62. Hindley C, Hinsliff SW, Thomson AM. Developing a tool to appraise fetal monitoring guidelines for women at low obstetric risk. Journal of advanced nursing. 2005;52(3):307-14.

63. Sanderlin BW, AbdulRahim N. Evidence-based medicine, part 6. An introduction to critical appraisal of clinical practice guidelines. The Journal of the American Osteopathic Association. 2007;107(8):321-4.

64. Chou R. Using Evidence in Pain Practice: Part I: Assessing Quality of Systematic Reviews and Clinical Practice Guidelines. Pain Medicine. 2008;9(5):518-30.

65. Using Evidence in Pain Practice: Part II: Interpreting and Applying Systematic Reviews and Clinical Practice Guidelines. Pain Medicine. 2008;9(5):531-41.

66. Hargrove P, Griffer M, Lund B. Procedures for using clinical practice guidelines. Language, speech, and hearing services in schools. 2008;39(3):289-302.

67. Pentheroudakis G, Stahel R, Hansen H, Pavlidis N. Heterogeneity in cancer guidelines: should we eradicate or tolerate? Annals of oncology : official journal of the European Society for Medical Oncology. 2008;19(12):2067-78.

68. Chong CA, Chen IJ, Naglie G, Krahn MD. How well do guidelines incorporate evidence on patient preferences? Journal of general internal medicine. 2009;24(8):977-82.

69. SELECT: evaluation and implementation of clinical practice guidelines: a guidance document from the American Professional Wound Care Association. Advances in skin & wound care. 2010;23(4):161-8.

70. Linskey ME. Defining excellence in evidence-based medicine clinical practice guidelines. Clinical neurosurgery. 2010;57:28-37.

71. Warriner RA, 3rd, Carter MJ. The current state of evidence-based protocols in wound care. Plastic and reconstructive surgery. 2011;127 Suppl 1:144s-53s.

Supplement 1

List of identified CPG appraisal instrument

Systematic review				Appraisal instrument identified				
Graham 2000 (20)	Vlayen 2005 (18)	Siering 2013 (11)	Current	Author, year, reference	Number of questions	Domain specific appraisal	Global appraisal	Validation
X	X			Lohr and Field, 1992 (2)	142	Yes	No	No
X	X			Hayward et al. 1993 (31)	10	No	No	No
X	X			Selker, 1993 (32)	10	Yes	No	No
X	X	X		Hayward et al. 1995 (33)	13	No	No	No
X	X			Mendelson 1995 (34)	8	-	-	No
X	X	X		Wolf 1995 (35)	10	No	No	No
X	X			SIGN 1995 (36)	50	-	-	No
	X	X		Mottur-Pilson 1995 (37)	18	No	Yes	No
X	X	X		Liddle 1996 (38)	14	No	No	Yes
X	X			Savoie 1996 (39)	33	-	-	No
X	X	X		Ward and Grieco, 1996 (40)	18	No	No	No
X	X	X		Calder et al 1997(41)	26	No	No	No
		X		Cook et al 1998 (42)	9	No	No	No
		X		Veale et al 1999 (43)	7	No	No	No
X	X	X		Shaneyfelt, 1999 (21)	25	No	No	Yes
X	X			Cluzeau 1999 (19)	37	Yes	Yes	Yes
		X		Fields 2000 (44)	8	No	No	No
	X	X		Grilli et al 2000 (45)	3	No	No	Yes
	X	X		Marshall 2000 (46)	9	No	No	No
	X	X		Sanders et al 2000 (47)	15	Yes	Yes	No
		X		Savoie et al 2000 (48)	51	No	No	No
		X		APA 2002 (49)	47	No	No	No
		X		Foy et al 2002 (50)	13	No	No	No
		X		Freitheim et al 2002 (51)	8	No	No	No
		X		Guyatt et al 2002 (52)	4	No	No	No

		X		Hart and Bailey 2002 (53)	9	No	No	No
	X			Reed et al 2002 (54)	-	-	-	No
	X			AGREE 2003 (30)	23	Yes	No	Yes
		X		Baxter and Palda 2003 (55)	12	No	No	No
		X		Hasenfeld and Shekelle 2003 (56)	30	No	No	No
	X			Hutchinson et al 2003 (57)	5	Yes	No	No
		X		Kulig 2003 (58)	13	Yes	Yes	Yes
		X		WHO 2003 (59)	25	No	No	No
	X	X		Shiffman 2003 (60)	18	No	No	No
		X		Nonino and Liberati 2004 (61)	6	No	No	No
		X		Hindley 2005 (62)	18	No	Yes	Yes
		X		Sanderlin and AbdulRahim 2007 (63)	5	No	No	No
		X		Chou 2008 (64, 65)	26	No	No	No
		X		Hargrove et al 2008 (66)	18	No	No	Yes
		X		Pentheroudakis et al 2008 (67)	24	No	No	No
		X		Chong et al 2009 (68)	11	No	No	Yes
		X		AGREEII 2009 (15)	23	Yes	Yes	Yes
		X		APWCA 2010 (69)	11	No	No	No
		X		Linskey 2010 (70)	9	No	No	No
		X		ADAPTE 2010 (17)	43	No	No	Yes
		X		Kashyap et al 2011 cited by (11)	30	No	No	Yes
		X		Warriner and Carter 2011 (71)	11	No	No	No
			X	Brouwers 2012 (28)	5+2	Yes	Yes	No
			X	Grimmer 2014 (7)	14	Yes	No	No
			X	Coroneos 2014 (22)	8	Yes	No	No
			X	Siebenhofer 2016 (24)	8	No	Yes	Yes
			X	Shaughnessy 2017 (23)	8	No	No	No

CHAPITRE 4: DISCUSSION

4.1 Discussion des deux revues systématiques

Le premier objectif du présent projet était de caractériser ce qui est à la disposition de la communauté scientifique pour évaluer la qualité des GPC et particulièrement la validité du contenu des recommandations.

La démarche a permis d'identifier 52 instruments ayant été développés au cours des 25 dernières années. D'ailleurs, le taux de production de nouveaux instruments s'est remarquablement maintenu constant à près de 2 instruments par année pendant cette période. Cette multiplication d'outils de mesure traduit bien l'insatisfaction que la communauté scientifique en général a ressentie sur les moyens dont elle disposait pour juger adéquatement des GPC. Chacun des auteurs a vu, dans la publication de l'instrument fraîchement élaboré, un produit fini qui leur convenait davantage que tout ce qui était disponible auparavant.

Bien que chacun de ces instruments partage le même but d'évaluer la qualité de GPC, ils présentent étonnamment des caractéristiques fort différentes, tant dans les domaines évalués que dans le nombre de question requis pour le faire tel qu'il est possible de le constater au tableau en annexe de l'article. Ainsi, à une extrémité du continuum, se trouve le tout premier instrument, conçu par l'IOM en 1992, et possiblement le plus exhaustif avec 142 items à considérer (1) (page 346) alors que Guyatt *et al.* (45) proposent seulement 4 questions pour juger de la validité des recommandations de GPC à partir d'un scénario clinique donné.

Statut de précurseur de l'instrument de l'IOM

Lors des travaux exécutés en 1990 par l'IOM sur le développement, l'implantation et l'évaluation des GPC, les membres du comité avaient noté qu'il n'existait alors rien qui permette de juger de la qualité, la fidélité et la validité du contenu d'un GPC. C'est pourquoi l'une des tâches d'un second comité avait été de concevoir et développer un tel instrument (1) (page 346). Le but de l'outil proposé était de fournir une méthode explicite pour examiner la justesse (*soundness*) de telles directives de pratiques cliniques et d'encourager leur développement systématique. Dans ce contexte, l'évaluation de la justesse visait tant le processus du

développement du guide de pratique clinique que du résultat final qui en découle. Qui plus est, les auteurs avaient mentionné spécifiquement que l'intention était d'éviter des situations dans lesquelles des GPC n'étant pas rigoureusement compatibles avec la preuve scientifique, seraient néanmoins évalués comme bons sur la considération unique de critères relatifs aux procédures suivies dans son développement.
Au sujet de l'importance relative des attributs, les auteurs du guide avaient exposé la prépondérance de la validité qui faisait l'objet de 22 questions illustrant manifestement la préoccupation du comité d'évaluer la justesse des directives plutôt qu'uniquement l'acceptabilité du processus de développement.

Qui plus est, l'esprit dans lequel l'IOM considérait l'évaluation d'un GPC est qu'il s'agissait d'une entreprise d'envergure à peine plus restreinte que l'élaboration du GPC lui-même. En fait, cet instrument, provisoire puisqu'il était appelé à être amélioré subséquemment, était destiné non pas à des médecins, cliniciens, patients ou autres non professionnels, mais bien à un ou des groupes d'individus alliant des expertises au niveau de l'expérience clinique du traitement de patients atteints de la condition visée par le GPC, de l'expérience de recherche sur les conditions et technologies en lien avec le GPC et de compétences méthodologiques relatives à l'élaboration de GPC. C'est ce groupe d'évaluateurs qui était pressenti pour rapporter, sous forme de résumé, les résultats de l'application de l'instrument proposé.
Il est manifeste que la communauté scientifique était d'abord à la recherche d'un instrument beaucoup plus convivial. En guise de rétroaction, les premiers utilisateurs ont majoritairement émis le commentaire que l'utilisation de l'outil était modérément à très difficile, l'un d'eux ayant même été qualifié d'instrument de torture intellectuelle (1) (page 349). Plusieurs auteurs ont alors entrepris de proposer des versions «améliorées» de l'outil de l'IOM, si bien que Graham *et al.* rapportent que 8 instruments sur les 15 recensés lors de la première revue en 2000 (46) sont directement basés sur l'instrument de l'IOM. Ce souci de convivialité s'est d'ailleurs maintenu dans l'élaboration subséquente des autres instruments et est possiblement devenu le principal attribut des outils les plus récents.

Un aspect intéressant observé dans les revues systématique retenues est l'identification d'un certain nombre de thèmes, sous lesquels peuvent se regrouper les éléments spécifiquement

évalués par chacun des outils. Ceci s'avère profitable dans la mesure où l'utilité d'un instrument d'évaluation est commensurable à sa capacité de bien identifier la présence ou l'absence des attributs qu'il est souhaitable de retrouver dans un GPC.

Par conséquent, une sélection parmi les autres instruments repérés peut s'effectuer sur la base de la désirabilité de retrouver dans un instrument donné, certaines caractéristiques telles que :

- la capacité d'évaluer un guide de pratique selon les thèmes présentés dans les lignes précédentes;
- la capacité d'émettre un jugement global sur le GPC évalué;
- l'instrument ait fait l'objet d'un processus de validation.

AGREE, l'étalon d'or

L'instrument AGREE *l'Appraisal of Guidelines for Research & Evaluation* (AGREE) (3) rassemble ces caractéristiques et, à la consultation de l'ensemble des documents recueillis dans le présent travail, il apparaît comme l'outil de référence ou l'étalon d'or des instruments d'évaluation de GPC.

L'instrument AGREE est le fruit d'une collaboration internationale dont le travail visait à répondre à «un besoin urgent de critères internationalement reconnus qui soient valables, fiables et utiles pour des buts d'évaluation divers dans des pays différents» (19). Il est le produit final d'un processus itératif de développement qui origine du tout premier instrument (1) (page 346) à partir duquel Cluzeau *et al.* (47) avaient produit en 1999 un instrument plus convivial qui a lui-même servi de base au développement du premier instrument AGREE en 2003 (19) et de sa plus récente version, l'AGREEII, en 2010 AGREEII (48, 49). Le document de Cluzeau et les deux outils AGREE ont tous été recommandés par les auteurs des revues systématiques (46, 50, 51) car ils ont été jugés complets et validés de manière exhaustive, opinion qui reste valable aujourd'hui. Il est vrai que du seul point de vue qu'un outil d'évaluation CPG idéal devrait permettre la validation de la présence ou de l'absence d'éléments majeurs que l'on doit trouver dans un bon GPC, l'AGREEII semble atteindre cet équilibre sans être excessif (Tableau 3 de l'article). De plus, l'AGREEII offre l'avantage de fournir une évaluation spécifique à un domaine ainsi qu'une évaluation globale du GPC et est l'un des rares instruments ayant fait l'objet d'une

forme de validation (Tableau 2 de l'article). L'instrument a été, à ce jour, traduit en 33 langues et cité dans plus de 600 articles[2]

Le besoin de rapidité

Pour tenter d'améliorer le niveau de satisfaction obtenu par le recours à l'AGREEII, plusieurs groupes ont consacré une part équitable de leurs efforts à l'élaboration d'outils d'évaluation qui nécessiteraient moins de temps d'utilisation. Par rapport aux quelques heures nécessaires pour compléter le premier instrument d'évaluation (1) (page 346), il faut compter entre 12 et 20 minutes pour utiliser l'AGREEII en fonction du niveau d'expertise de l'évaluateur (6). Les outils d'évaluation rapide récemment développés se sont améliorés dans cette zone avec des temps moyens requis de 13 min pour le MiChe (52) et un impressionnant 3-7 min pour l'iCAHE, mais la dernière mesure était basée sur seulement 3 évaluateurs (6). Confirmant une fois de plus le statut d'étalon d'or de l'AGREEII, les auteurs d'instruments d'évaluation rapide rapportent la comparaison des résultats obtenus par leur nouvel outil avec celui produit par l'AGREEII et interprètent un manque de différence comme une confirmation de bonne performance (6, 52-54). Cela signifie que le gain de rapidité d'évaluation n'a pas été obtenu au détriment de la capacité de discerner les «bons» CPG des «mauvais».

Évaluation de la validité

En dépit du fait qu'une plus grande rapidité d'utilisation représente un avantage majeur en termes de facilité d'utilisation, il n'en reste pas moins que l'un des principaux objectifs de l'évaluation d'un GPC est d'établir son niveau de validité. Le niveau de validité évalué est susceptible d'être proportionnel au niveau de confiance que les parties prenantes auront quand il sera temps de mettre en œuvre les recommandations intégrées dans le GPC. Le défi est que la validité est la dimension la plus difficile à évaluer correctement (1) (p 375). Dans l'AGREEII, l'aspect de validité est principalement capté dans la rigueur du domaine du développement, bien que l'indépendance éditoriale (ou l'absence) puisse également avoir un impact sur la validité de

[2] Information disponible sur le site https://www.agreetrust.org/

l'ensemble du processus. Par conséquent, les menaces potentielles à la validité en termes de processus telles que: représentation déséquilibrée des parties prenantes, stratégie de recherche documentaire douteuse, formulation ambiguë, etc. seraient notées et classées en conséquence.

Cela conduit à l'évaluation impérative du contenu clinique puisqu'il représente le fondement premier à partir duquel les recommandations sont élaborées. Même si un GPC est rédigé de manière concise, par des auteurs sans conflit d'intérêts, en tenant compte des différentes perspectives des parties prenantes, cela ne garantit pas nécessairement que les données probantes recueillies aient été correctement interprétées et incorporées dans les différentes recommandations. Cela pourrait créer une situation dans laquelle il existerait un certain degré de discordance entre le contenu de la recommandation et les preuves scientifiques réelles malgré une évaluation positive du GPC. Field et Lohr avaient publié cet avertissement il y a 25 ans (1) (p 209) et qui demeure tout à fait pertinent même aujourd'hui. En fait, l'évaluation de la qualité du contenu clinique des lignes directrices s'avère ardue car elle prend du temps, demande un personnel hautement qualifié et peut nécessiter des informations supplémentaires non disponibles dans les GPC eux-mêmes (51). Ces limites ont également été reconnues par les auteurs de l'AGREE, car ils spécifiaient clairement d'emblée que l'instrument avait été conçu pour évaluer le processus de développement du GPC ainsi que la mesure dans laquelle il avait été rapporté et que l'outil n'évaluait pas le contenu clinique du GPC ni la qualité des données probantes soutenant les recommandations (19).

4.2 Proposition d'amélioration

Le second objectif du présent projet était de proposer une nouvelle façon de faire, susceptible d'améliorer la capacité des différentes parties prenantes de juger de la validité des recommandations d'une GPC. Ainsi, une réflexion sur divers scénarios dans lesquels un certain niveau d'invalidité pourrait passer inaperçu par l'application de l'outil AGREEII dans sa forme actuelle a mené à l'élaboration de trois recommandations.

La qualité des recommandations n'est pas uniforme dans un GPC

D'abord, d'un point de vue conceptuel, il faut savoir que la qualité d'un GPC n'est pas uniforme à travers les différents éléments qui le composent.

Dans sa forme actuelle, les éléments de l'outil AGREE ne permettent pas de mettre en évidence ce fait car ils ne sont évalués qu'au niveau de l'ensemble du GPC (3).

Le tableau suivant résume les différents aspects évalués et met en évidence si la teneur de l'évaluation peut varier au sein d'un même GPC, notamment au gré des diverses recommandations.

Tableau 2 Caractère uniforme ou variables au sein d'un GPC des énoncés constituant l'instrument d'évaluation AGREE II.

Domaine	Énoncés	S'applique de façon uniforme à l'ensemble du GPC	Peut varier d'une recommandation à l'autre
Domaine 1. Champ et objectifs	1. Le ou les objectifs de la RPC sont décrits explicitement.	X	
	2. La ou les questions de santé couvertes par la RPC sont décrites explicitement.	X	
	3. La population (patients, public, etc.) à laquelle la RPC doit s'appliquer est décrite explicitement.	X	
Domaine 2. Participation des groupes concernés	4. Le groupe ayant élaboré la RPC inclut des représentants de tous les groupes professionnels concernés.	X	
	5. Les opinions et les préférences de la population cible (patients, public, etc.) ont été identifiées.		X
	6. Les utilisateurs cibles de la RPC sont clairement définis.	X	
Domaine 3. Rigueur d'élaboration	7. Des méthodes systématiques ont été utilisées pour rechercher les preuves scientifiques.	X	
	8. Les critères de sélection des preuves sont clairement décrits.	X	
	9. Les forces et les limites des preuves scientifiques sont clairement définies.		X

	10. Les méthodes utilisées pour formuler les recommandations sont clairement décrites.	X	
	11. Les bénéfices, les effets secondaires et les risques en termes de santé ont été pris en considération dans la formulation des recommandations.		X
	12. Il y a un lien explicite entre les recommandations et les preuves scientifiques sur lesquelles elles reposent.		X
	13. La RPC a été revue par des experts externes avant sa publication.	X	
	14. Une procédure d'actualisation de la RPC est décrite.	X	
Domaine 4. Clarté et présentation	15. Les recommandations sont précises et sans ambiguïté.		X
	16. Les différentes options de prise en charge de l'état ou du problème de santé sont clairement présentées.	X	
	17. Les recommandations clés sont facilement identifiables.	X	
Domaine 5. Applicabilité	18. La RPC décrit les éléments facilitant son application et les obstacles.		X
	19. La RPC offre des conseils et/ou des outils sur les façons de mettre les recommandations en pratique.		X
	20. Les répercussions potentielles de l'application des recommandations sur les ressources ont été examinées.		X
	21. La RPC propose des critères de suivi et/ou de vérification.18. La RPC offre des conseils et/ou des outils sur les façons de mettre les recommandations en pratique.		X
Domaine 6. Indépendance éditoriale	22. Le point de vue des organismes de financement n'ont pas influencé le contenu de la RPC.		X
	23. Les intérêts divergents des membres du groupe ayant élaboré la RPC ont été pris en charge et documentés.	X	

Cette façon de faire consistant à évaluer chacune des recommandations de manière indépendante a déjà été invoquée (1) (page 392) (55, 56) et permet d'éviter de contraindre l'évaluateur à émettre un jugement qui s'applique à l'ensemble des recommandations contenues, émettant du même coup le postulat implicite que la validité scientifique est rigoureusement la même d'une recommandation à l'autre alors que le sens commun indique le contraire.

À titre d'exemple, le point 12 - *Il y a un lien explicite entre les recommandations et les preuves scientifiques sur lesquelles elles reposent*, oblige, en quelque sorte, l'évaluateur à émettre un jugement représentant le niveau moyen de ce point parmi toutes les recommandations. Or, il est évident que ce lien peut être plus ou moins explicite selon chacune des recommandations et qu'il serait hautement souhaitable, de pouvoir distinguer chacune de ces situations afin d'informer adéquatement les parties prenantes pour qui le résultat de l'évaluation est destiné.

Le concept de lien explicite pourrait également être considéré comme deux composantes distinctes entrant en jeu:

- l'exactitude de l'interprétation des données probantes, c.-à-d. si le contenu de la recommandation correspond vraiment à l'information véhiculée par les données probantes
- la congruence des niveaux de force des données probantes à ceux des recommandations.

La première recommandation consiste à:

- Évaluer spécifiquement le point numéro 12 - *Il y a un lien explicite entre les recommandations et les preuves scientifiques sur lesquelles elles reposent* pour **chacune** des principales recommandations.

De cette façon, un gain dans l'aspect de la fine précision ou granularité de l'évaluation de l'énoncé 12 pourrait avoir une incidence positive sur la valeur globale du processus d'évaluation du GPC particulièrement dans une situation hypothétique, mais fort probable, dans laquelle certaines recommandations, ou ne serait-ce qu'une seule, s'avère ne pas être adéquatement soutenue par les données probantes recueillies comparativement aux autres recommandations du même GPC.

GPC de domaines innovants

Selon différents domaines thérapeutiques, les innovations médicales: médicaments, dispositifs, ou modes d'intervention peuvent parfois introduire des changements de paradigmes drastiques qui réduisent la fenêtre temporelle au cours de laquelle les preuves restent à jour et décrivent adéquatement la réalité contemporaine. Cela pourrait revêtir une importance particulièrement significative dans les domaines de haute technologie.

Idéalement, une évaluation positive d'un GPC rassurerait les différentes parties prenantes que non seulement les données probantes ont été correctement interprétées et utilisées adéquatement dans la formulation des recommandations, mais aussi que les preuves considérées demeurent pertinentes au paradigme actuel et ne sont pas contaminées par des données dorénavant inadaptées.

Une telle situation est récemment survenue en neurologie ou la publication d'études randomisées d'importance en 2015 a provoqué un important changement dans la prise en charge des patients victime d'un accident vasculaire cérébral ainsi que la mise à jour des GPC canadiens (57) et américains (58) alors en vigueur.

La seconde recommandation consiste à:

- Évaluer spécifiquement si les données probantes sur lesquelles se fondent les principales recommandations sont encore actuelles.

Dans la même ligne de pensée, l'utilisation de ces innovations crée des sous-groupes croissants de patients avec des caractéristiques éventuellement légèrement différentes, en particulier si une nouvelle technologie perturbatrice (*disruptive*) commence à être largement utilisée, par exemple les patients victime d'AVC de l'exemple précédent maintenant traités à l'aide de dispositifs endovasculaires permettant le retrait mécanique du caillot sanguin et non uniquement par l'administration d'un agent pharmacologique ou encore des diabétiques maintenant traités avec une pompe à insuline et non plus au moyens de multiples injections quotidiennes. Dans de tels cas, il y aurait, au moins pendant un certain temps, un ensemble de données combinant à la fois

celles provenant de patients traités de l' «ancienne» façon et celles de patients traités de la «nouvelle» façon. Ainsi, il se pourrait fort bien que les recommandations destinées à optimiser le suivi de ces patients soient différentes selon le mode d'intervention (l'ancien ou le nouveau) dont le patient a bénéficié. Encore une fois, une évaluation positive d'un GPC devrait rassurer les différentes parties prenantes que les données probantes utilisées pour élaborer les recommandations s'appliquent adéquatement aux patients actuels.

La troisième recommandation consiste donc à :

- Évaluer spécifiquement si les données probantes sur lesquelles se fondent les principales recommandations s'appliquent toujours aux patients actuels.

4.3 Limites de la présente démarche

Le travail actuel a certes quelques forces et faiblesses. La stratégie de recherche de revue systématique et la veille scientifique étaient très larges et ont permis de récupérer de nombreux articles pertinents sur l'évaluation des GPC. Il semble peu probable qu'un outil ou une procédure d'évaluation ait été manqué. En revanche, la démarche n'a pas bénéficié de l'expertise d'un spécialiste de l'information.

Qui plus est, les articles rédigés dans une autre langue que le français ou l'anglais n'ont pas été considérés, de même que la littérature grise. Conséquemment, il est possible que des éléments pertinents à la présente problématique qui pouvaient s'y trouver aient été ignorés.

Les recommandations incluses dans le présent travail ont été faites humblement sur la base d'une réflexion sur des situations dans lesquelles un GPC pouvait recevoir une évaluation positive malgré des recommandations inadéquates. Les améliorations proposées n'ont pas été validées ou soumises à une expertise externe. En outre, il plaît à l'esprit que ces recommandations puissent permettre une évaluation plus juste des GPC mais cette hypothèse demeure à être vérifiée empiriquement.

5 CONCLUSION

Le développement de GPC peut être vu comme une longue chaine de montage dans laquelle s'imbriquent des processus techniques (revue systématique des données probantes pertinentes), des processus de jugement (interprétation de la revue systématique et inférence des recommandations) et des processus interpersonnels (établissement de consensus). La validité des recommandations qui en découlent peut être influencée négativement si l'un ou l'autre de ces processus est biaisé (9). Avec tant de biais potentiels pouvant mettre en péril la validité des recommandations, les communautés de pratique ont nécessairement besoin d'outils d'évaluation de la qualité des GPC incluant la validité des recommandations.

En termes de validité, l'évaluation du contenu clinique du CPG demeure un réel défi. Avec plus de 50 outils d'évaluation différents, le choix de l'un par rapport à l'autre repose évidemment sur le besoin spécifique de l'évaluateur en termes de nature des items à évaluer. Toutefois, la communauté scientifique reconnaît en quelque sorte à l'instrument AGREEII, résultat final d'un processus d'améliorations itératives multiples, le statut d'étalon-or de l'évaluation de GPC.

Il est bien convenu que d'autres recherches devront clarifier comment l'évaluation globale du contenu clinique d'une ligne directrice peut être incluse dans les outils d'évaluation des lignes directrices avec une utilisation raisonnable des ressources (10). Malgré le fait qu'AGREEII n'ait pas été spécifiquement conçu pour évaluer la validité des recommandations, une réflexion sur les situations dans lesquelles un contenu scientifique inapproprié pourrait passer inaperçu, a permis de potentiellement affiner l'outil actuel en ce sens.

6 BIBLIOGRAPHIE

1. Institute of Medicine Committee on Clinical Practice G. In: Field MJ, Lohr KN, editors. Guidelines for Clinical Practice: From Development to Use. Washington (DC): National Academies Press (US) by the National Academy of Sciences.; 1992.

2. Institute of Medicine Committee to Advise the Public Health Service on Clinical Practice Guidelines. In: Field MJ, Lohr KN, editors. Clinical Practice Guidelines Directions for a New Program. Washington (DC): National Academies Press (US) by the National Academy of Sciences.; 1990.

3. Eikermann M, Holzmann N, Siering U, Ruther A. Tools for assessing the content of guidelines are needed to enable their effective use--a systematic comparison. BMC research notes. 2014;7:853.

4. Abdelsattar ZM, Reames BN, Regenbogen SE, Hendren S, Wong SL. Critical evaluation of the scientific content in clinical practice guidelines. Cancer. 2015;121(5):783-9.

5. Watine J, Friedberg B, Nagy E, Onody R, Oosterhuis W, Bunting PS, et al. Conflict between guideline methodologic quality and recommendation validity: a potential problem for practitioners. Clinical chemistry. 2006;52(1):65-72.

6. Grimmer K, Dizon JM, Milanese S, King E, Beaton K, Thorpe O, et al. Efficient clinical evaluation of guideline quality: development and testing of a new tool. BMC medical research methodology. 2014;14:63.

7. Lortie M, IRSST (Québec), Canadian Electronic Library (Firm). Bilan des connaissances sur les guides de pratique en santé enseignements clés et transférabilité pour la santé et la sécurité au travail. Montéal, Qué.: Institut de recherche Robert-Sauvé en santé et en sécurité du travail; 2012. Available from: http://myaccess.library.utoronto.ca/login?url=http://site.ebrary.com/lib/utoronto/Top?id=10575156.

8. Watine J, Wils J, Augereau C. Evaluation of the methodological quality of two contradictory guidelines recently published by the Haute autorite de sante. Annales de biologie clinique. 2017;75(1):101-6.

9. Graham R. MM, Miller Wolman D., Greenfield S., Steinberg E.,. Clinical Practice Guidelines We Can Trust. US: National Academies Press (US); 2011. Available from: http://www.ncbi.nlm.nih.gov/books/NBK209539.

10. Koutsavlis AT. Dissémination des guides de pratiques chez les médecins. Bibliothèque naitionale du Québec: Institut national de santé publique du Québec; 2001.

11. Burls A. AGREE II—improving the quality of clinical care. The Lancet. 2010;376(9747):1128-9.

12. Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology. 2015.

13. Cecamore C, Savino A, Salvatore R, Cafarotti A, Pelliccia P, Mohn A, et al. Clinical practice guidelines: what they are, why we need them and how they should be developed through rigorous evaluation. European journal of pediatrics. 2011;170(7):831-6.

14. Alonso-Coello P, Irfan A, Sola I, Gich I, Delgado-Noguera M, Rigau D, et al. The quality of clinical practice guidelines over the last two decades: a systematic review of guideline appraisal studies. Quality & safety in health care. 2010;19(6):e58.

15. Martinez Garcia L, Sanabria AJ, Garcia Alvarez E, Trujillo-Martin MM, Etxeandia-Ikobaltzeta I, Kotzeva A, et al. The validity of recommendations from clinical guidelines: A survival analysis. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2014;186(16):1211-9.

16. Semlitsch T, Blank WA, Kopp IB, Siering U, Siebenhofer A. Evaluating Guidelines: A Review of Key Quality Criteria. Deutsches Arzteblatt international. 2015;112(27-28):471-8.

17. Schunemann HJ, Wiercioch W, Etxeandia I, Falavigna M, Santesso N, Mustafa R, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. Cmaj. 2014;186(3):E123-42.

18. Van de Velde S, Heselmans A, Donceel P, Vandekerckhove P, Ramaekers D, Aertgeerts B. Rigour of development does not AGREE with recommendations in practice guidelines on the use of ice for acute ankle sprains. BMJ quality & safety. 2011;20(9):747-55.

19. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Quality & safety in health care. 2003;12(1):18-23.

20. Al-Ansary LA, Tricco AC, Adi Y, Bawazeer G, Perrier L, Al-Ghonaim M, et al. A systematic review of recent clinical practice guidelines on the diagnosis, assessment and management of hypertension. PloS one. 2013;8(1):e53744.

21. Burgers JS, Bailey JV, Klazinga NS, Van Der Bij AK, Grol R, Feder G. Inside guidelines: comparative analysis of recommendations and evidence in diabetes guidelines from 13 countries. Diabetes care. 2002;25(11):1933-9.

22. Zagouri F, Peroukidis S, Tzannis K, Kouloulias V, Bamias A. Current clinical practice guidelines on chemotherapy and radiotherapy for the treatment of non-metastatic muscle-invasive urothelial cancer: A systematic review and critical evaluation by the Hellenic Genito-Urinary Cancer Group (HGUCG). Critical reviews in oncology/hematology. 2015;93(1):36-49.

23. Hill S, Pang T. Leading by example: a culture change at WHO. The Lancet. 2007;369(9576):1842-4.

24. Oxman AD, Lavis JN, Fretheim A. Use of evidence in WHO recommendations. The Lancet. 2007;369(9576):1883-9.

25. Sinclair D, Isba R, Kredo T, Zani B, Smith H, Garner P. World Health Organization guideline development: an evaluation. PloS one. 2013;8(5):e63715.

26. Burda BU, Chambers AR, Johnson JC. Appraisal of guidelines developed by the World Health Organization. Public health. 2014;128(5):444-74.

27. Feuerstein JD, Akbari M, Gifford AE, Hurley CM, Leffler DA, Sheth SG, et al. Systematic analysis underlying the quality of the scientific evidence and conflicts of interest in interventional medicine subspecialty guidelines. Mayo Clinic proceedings. 2014;89(1):16-24.

28. Isaac A, Saginur M, Hartling L, Robinson JL. Quality of reporting and evidence in American Academy of Pediatrics guidelines. Pediatrics. 2013;131(4):732-8.

29. Henig O, Yahav D, Leibovici L, Paul M. Guidelines for the treatment of pneumonia and urinary tract infections: evaluation of methodological quality using the Appraisal of Guidelines, Research and Evaluation II instrument. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2013;19(12):1106-14.

30. De Man KE, Troch ME, Dobbeleir AA, Duong HP, Goethals IM. Comparison of the EANM and SNM guidelines on diuretic renography in children. Nuclear Medicine Communications. 2015;36(5):486-8.

31. Feuerstein JD, Gifford AE, Akbari M, Goldman J, Leffler DA, Sheth SG, et al. Systematic analysis underlying the quality of the scientific evidence and conflicts of interest in gastroenterology practice guidelines. The American journal of gastroenterology. 2013;108(11):1686-93.

32. Liberati A, Buzzetti R, Grilli R, Magrini N, Minozzi S. Which guidelines can we trust?: Assessing strength of evidence behind recommendations for clinical practice. The Western journal of medicine. 2001;174(4):262-5.

33. Talwalkar JA. Improving the Transparency and Trustworthiness of Subspecialty-Based Clinical Practice Guidelines. Mayo Clinic proceedings.89(1):5-7.

34. George JN, Vesely SK, Woolf SH. Conflicts of interest and clinical recommendations: comparison of two concurrent clinical practice guidelines for primary immune thrombocytopenia developed by different methods. American journal of medical quality : the official journal of the American College of Medical Quality. 2014;29(1):53-60.

35. Provan D, Stasi R, Newland AC, Blanchette VS, Bolton-Maggs P, Bussel JB, et al. International consensus report on the investigation and management of primary immune thrombocytopenia. Blood. 2010;115(2):168-86.

36. Neunert C, Lim W, Crowther M, Cohen A, Solberg L, Jr., Crowther MA. The American Society of Hematology 2011 evidence-based practice guideline for immune thrombocytopenia. Blood. 2011;117(16):4190-207.

37. Aarts MC, van der Heijden GJ, Rovers MM, Grolman W. Remarkable differences between three evidence-based guidelines on management of obstructive sleep apnea-hypopnea syndrome. The Laryngoscope. 2013;123(1):283-91.

38. Sun M, Zhang M, Shen J, Yan J, Zhou B. Critical Appraisal of international guidelines for the management of diabetic neuropathy: Is there global agreement in the internet era? International Journal of Endocrinology. 2015;2015(519032).

39. Gandhi S, Verma S, Ethier JL, Simmons C, Burnett H, Alibhai SM. A systematic review and quality appraisal of international guidelines for early breast cancer systemic therapy: Are recommendations sensitive to different global resources? Breast (Edinburgh, Scotland). 2015;24(4):309-17.

40. Desai SH, Jeong K, Kattan JD, Lieberman R, Wisniewski S, Green TD. Anaphylaxis management before and after implementation of guidelines in the pediatric emergency department. The Journal of Allergy and Clinical Immunology: In Practice. 2015;3(4):604-6.e2.

41. Deasy C, Bray JE, Smith K, Wolfe R, Harriss LR, Bernard SA, et al. Cardiac arrest outcomes before and after the 2005 resuscitation guidelines implementation: Evidence of improvement? Resuscitation. 2011;82(8):984-8.

42. Bennett WL, Odelola OA, Wilson LM, Bolen S, Selvaraj S, Robinson KA, et al. Evaluation of guideline recommendations on oral medications for type 2 diabetes mellitus: a systematic review. Annals of internal medicine. 2012;156(1 Pt 1):27-36.

43. Winther LP, Mitchell AU, Moller AM. Inconsistencies in clinical guidelines for obstetric anaesthesia for Caesarean section: a comparison of the Danish, English, American, and German guidelines with regard to developmental quality and guideline content. Acta anaesthesiologica Scandinavica. 2013;57(2):141-9.

44. Beauchamp S, Drapeau M, Dionne C, Duplantie JP. Cadre d'élaboration des guides de pratique dans le secteur des services sociaux. . In: (INESSS) Indeeseess, editor.: Bibliothèque et Archives nationales du Québec; 2015. p. 88.

45. Guyatt G. HR, Richardson WS., Green L., Wilson MC., Sinclair J., Cook D., Glasziou P., Detsky A., Bass E. Moving from evidence to action. In: Guyatt G. DR, editor. User's guides to the medical literature Chicago: AMA Press; 2002. p. 175-99.

46. Graham ID, Calder LA, Hebert PC, Carter AO, Tetroe JM. A comparison of clinical practice guideline appraisal instruments. International journal of technology assessment in health care. 2000;16(4):1024-38.

47. Cluzeau FA, Littlejohns P, Grimshaw JM, Feder G, Moran SE. Development and application of a generic methodology to assess the quality of clinical guidelines. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua. 1999;11(1):21-8.

48. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. Development of the AGREE II, part 1: performance, usefulness and areas for improvement. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2010;182(10):1045-52.

49. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. Development of the AGREE II, part 2: assessment of validity of items and tools to support application. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2010;182(10):E472-8.

50. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua. 2005;17(3):235-42.

51. Siering U, Eikermann M, Hausner E, Hoffmann-Esser W, Neugebauer EA. Appraisal tools for clinical practice guidelines: a systematic review. PloS one. 2013;8(12):e82915.

52. Siebenhofer A, Semlitsch T, Herborn T, Siering U, Kopp I, Hartig J. Validation and reliability of a guideline appraisal mini-checklist for daily practice use. BMC Med Res Methodol. 2016;16:39.

53. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. The Global Rating Scale complements the AGREE II in advancing the quality of practice guidelines. Journal of Clinical Epidemiology. 2012;65(5):526-34.

54. Shaughnessy AF, Vaswani A, Andrews BK, Erlich DR, D'Amico F, Lexchin J, et al. Developing a Clinician Friendly Tool to Identify Useful Clinical Practice Guidelines: G-TRUST. Annals of family medicine. 2017;15(5):413-8.

55. Hargrove P, Griffer M, Lund B. Procedures for using clinical practice guidelines. Language, speech, and hearing services in schools. 2008;39(3):289-302.

56. Calder L, Hébert, P, Carter, A., Gaham, I., . Review of published recommendations and guidelines for the transfusion of allogenieic red blood cell and plasma. Can Med Assoc J 1997;156((11 Suppl)):S1-S8.

57. Casaubon LK, Boulanger JM, Glasser E, Blacquiere D, Boucher S, Brown K, et al. Canadian Stroke Best Practice Recommendations: Acute Inpatient Stroke Care Guidelines, Update 2015. International journal of stroke : official journal of the International Stroke Society. 2016;11(2):239-52.

58. Powers WJ, Derdeyn CP, Biller J, Coffey CS, Hoh BL, Jauch EC, et al. 2015 American Heart Association/American Stroke Association Focused Update of the 2013 Guidelines for the Early Management of Patients With Acute Ischemic Stroke Regarding Endovascular Treatment: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke; a journal of cerebral circulation. 2015;46(10):3020-35.

Annexe 1 Évaluation des revues systématiques selon la grille AMSTAR

	AMSTAR – A measurement tool to assess the methodological quality of systematic reviews.		
Question	Graham ID, et al. A comparison of clinical practice guideline appraisal instruments. International journal of technology assessment in health care 2000;16(4):1024-38	Vlayen J, et al. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care. 2005 Jun;17(3):235-42.	Siering U, et al. Appraisal tools for clinical practice guidelines: a systematic review. PloS one 2013;**8**(12) e82915
1. Was an 'a priori' design provided? **The research question and inclusion criteria should be established before the conduct of the review.** **Note: Need to refer to a protocol, ethics approval, or predetermined/a priori published research objectives to score a "yes."**	oui. To identify and compare clinical practice guideline appraisal instruments.	Oui. A systematic review of the literature was carried out to identify and compare existing critical appraisal tools for clinical practice guidelines.	Oui. The aims of this systematic review were to identify and compare existing guideline appraisal tools to see if the landscape of tools had changed. We included articles with the following characteristics: • Publication described the most recent version of an appraisal tool for clinical guidelines • Availability of a full-text document (e.g., journal article or internet file). Articles were excluded that only described the content of guidelines, the guideline development process or the application of an appraisal tool already identified in another publication.
2. Was there duplicate study selection and data extraction? **There should be at least two independent data extractors and a consensus procedure for disagreements should be in place.** **Note: 2 people do study selection, 2 people do data extraction, consensus process or one person checks the other's work.**	Non.	Non. One investigator (J.V.) assessed the selected papers and retrieved all the different tools that appraised the quality of clinical practice guidelines.	Oui. Two reviewers (US, WHE) independently screened titles and abstracts of the retrieved citations to identify potentially eligible primary and secondary publications. The full texts were obtained and independently evaluated by the same two reviewers. Disagreements were resolved by consensus.
3. Was a comprehensive literature search performed? **At least two electronic sources should be searched. The report must include years and** **databases used (e.g., Central, EMBASE, and MEDLINE). Key words and/or MESH terms must be stated and where feasible the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study, and by reviewing the references in the studies found.**	Non. Only MEDLINE database was systematically searched. Manual searches of the retrieved articles' bibliographies were conducted, and articles from personal collections (PCH, AOC) were included. We also attempted to contact the developers of the identified appraisal instruments to determine whether they were aware of any other instruments that might have been missed.	Oui. A literature search of the English and non-English literature indexed in the Ovid–Medline database, Embase database, and Cinahl database was conducted.	Oui. Relevant primary and secondary publications (systematic and narrative reviews) in MEDLINE, EMBASE, the Cochrane Database of Systematic Reviews (Cochrane Reviews), the Database of Abstracts of Reviews of Effects (Other Reviews), the Health Technology Assessment Database (Technology Assessments), the NHS Economic Evaluation Database, and the Cochrane Methodology Register.

Note: If at least 2 sources + one supplementary strategy used, select "oui" (Cochrane register/Central counts as 2 sources; a grey literature search counts as supplementary).			
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion? **The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.** **Note: If review indicates that there was a search for "grey literature" or "unpublished literature," indicate "yes." SIGLE database, dissertations, conference proceedings, and trial registries are all considered grey for this purpose. If searching a source that contains** **both grey and non-grey, must specify that they were searching for grey/unpublished lit.**	Oui. Manual searches of the retrieved articles' bibliographies were conducted, and articles from personal collections (PCH, AOC) were included. We also attempted to contact the developers of the identified appraisal instruments to determine whether they were aware of any other instruments that might have been missed.	Oui. A manual search of the references of relevant articles was conducted. We also contacted the developers of the instruments identified to determine whether they were aware of any other instruments to include in our review. All articles that described the evaluation of clinical practice guidelines or the development of a guideline appraisal tool were included. No restriction was placed on abstracts, conference proceedings, or language.	Oui. The systematic search was limited to publications in German and English published after 1994. The search in all databases was performed in May 2011. In addition, we scrutinized the reference lists of the relevant primary and secondary publications retrieved in the above search to identify further publications. We included articles with the following characteristics: • Publication described the most recent version of an appraisal tool for clinical guidelines • Availability of a full-text document (e.g., journal article or internet file). Articles were excluded that only described the content of guidelines, the guideline development process or the application of an appraisal tool already identified in another publication.
5. Was a list of studies (included and excluded) provided? **A list of included and excluded studies should be provided.** **Note: Acceptable if the excluded studies are referenced. If there is an electronic link to the list but the link is dead, select "no."**	Oui. A total of 15 possible practice guideline appraisal instruments were identified by the search process. Two were identified by one of the instrument developers we contacted (23;24). Two (2;30) were eventually excluded because they were designed as guiding principles for guideline development and implementation rather than instruments for evaluating guidelines. A description of the 13 instruments is presented in Table 2.	Non. only a list of included studies is presented	Oui. Excluded publications are listed in online File S2. Supporting information 2 – Excluded studies (ordered by reasons for exclusion)
6. Were the characteristics of the included studies provided? **In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions and outcomes. The ranges of characteristics in all the** **studies analyzed e.g., age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.** **Note: Acceptable if not in table format as long as they are described as above.**	Oui. Table 3 presents the comparison of each instrument against the items generated during the first stage of the content analysis.	Oui. Table 1 provides an overview of the characteristics of the 24 tools and Table 3 dimensions covered by the critical appraisal tools	Oui. Table 2 shows the main formal characteristics of the 40 appraisal tools considered. Figures 2 and 3 compare the quality dimensions and items covered by the appraisal tools analysed.
7. Was the scientific quality of the included studies assessed and documented? **'A priori' methods of assessment should be provided (e.g., for effectiveness studies if the** **author(s) chose to include only randomized, double-blind, placebo controlled studies, or allocation concealment as**	Oui. Since an assessment tool for guideline appraisal tool does not exist, the most relevant characteristic that can be formally evaluated and reported is whether each guideline appraisal tool has been validated or not and the current review reports that for each individual guidelines.	Oui. Since an assessment tool for guideline appraisal tool does not exist, the most relevant characteristic that can be formally evaluated and reported is whether each guideline appraisal tool has been validated or not and the current review reports that for each individual guidelines.	Oui. Since an assessment tool for guideline appraisal tool does not exist, the most relevant characteristic that can be formally evaluated and reported is whether each guideline appraisal tool has been validated or not and the current review reports that for each individual guidelines.

inclusion criteria); for other types of studies alternative items will be relevant. Note: Can include use of a quality scoring tool or checklist, e.g., Jadad scale, risk of bias, sensitivity analysis, etc., or a description of quality items, with some kind of result for EACH study ("low" or "high" is fine, as long as it is clear which studies scored "low" and which scored "high"; a summary score/range for all studies is not acceptable).			
8. Was the scientific quality of the included studies used appropriately in formulating conclusions? The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review, and explicitly stated in formulating recommendations. Note: Might say something such as "the results should be interpreted with caution due to poor quality of included studies." Cannot score "yes" for this question if scored "no" for question 7.	Non. La qualité des outils n'est pas considérée, seulement si l'outil a été validé ou non	Non. La qualité des outils n'est pas considérée, seulement si l'outil a été validé ou non	Non. La qualité des outils n'est pas considérée, et non plus si l'outil a été validé ou non
9. Were the methods used to combine the findings of studies appropriate? For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e., Chi-squared test for homogeneity, I2). If heterogeneity exists a random effects model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e., is it sensible to combine?). Note: Indicate "yes" if they mention or describe heterogeneity, i.e., if they explain that they cannot pool because of heterogeneity/variability between interventions.	Non applicable	Non applicable	Non applicable
10. Was the likelihood of publication bias assessed? An assessment of publication bias should include a combination of graphical aids (e.g., funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test,Hedges-Olken). Note: If no test values or funnel plot included, score "no". Score "yes" if mentions that publication bias could not be assessed because there were fewer than 10 included studies.	Non applicable	Non applicable	Non applicable
11. Was the conflict of interest included? Potential sources of support should be clearly acknowledged in both the systematic review and the included studies.	Non rien n'est rapporté tant pour les auteurs de la revue ou des auteurs des études primaires .	Non rien n'est rapporté tant pour les auteurs de la revue ou des auteurs des études primaires .	No. Les auteurs de la revue ont rapport ne pas avoir reçu de financement de sources externs mais rien n,est mentionné pour les auteurs des études primaires incluses.

Note: To get a "yes," must indicate source of funding or support for the systematic review AND for each of the included studies.			

Annexe 2 Stratégie de recherche documentaire

Search strategy

Bases de données

MEDLINE (PubMed)

Search	Add to builder	Query	Items found
#6	Add	Search **((((#1) AND #2) AND #3) AND #4) AND #5**	362
#5	Add	Search **("1990"[Date - Publication] : "2015"[Date - Publication])**	16780947
#4	Add	Search **systematic*[Title/Abstract]**	342440
#3	Add	Search **(tool*[Title/Abstract]) OR instrument*[Title/Abstract]**	779855
#2	Add	Search **((apprais*[Title/Abstract]) OR evaluat*[Title/Abstract]) OR scientific[Title/Abstract]**	3028764
#1	Add	Search **guideline*[Title]**	63820

6	**5 and 1990:2015.(sa_year).**	**89**
5	1 and 2 and 3 and 4	115
4	"systematic*".ab,ti.	2864
3	(tool* or instrument*).ab,ti.	1686953
2	(apprais*or evaluat* or scientific).ab,ti.	1018782
1	guideline*.ti.	84259

Search strategy

Bases de données

MEDLINE (PubMed)

Search	Add to builder	Query	Items found
#5	Add	Search **(#1 AND #2 AND #3)** Filters: **Publication date from 2011/05/01 to 2015/12/28**	1037
#4	Add	Search **(#1 AND #2 AND #3)**	2293
#3	Add	Search **(((((((((quality[Title]) OR analys*[Title]) OR compara*[Title]) OR valid*[Title]) OR attributes[Title]) OR recommend*[Title]) OR review[Title]) OR evidence-based[Title]) OR grad*[Title]) OR agree[Title]**	1401212
#2	Add	Search **((apprais*[Title/Abstract]) OR evaluat*[Title/Abstract]) OR scientific[Title/Abstract]**	2659374
#1	Add	Search **guideline*[Title]**	56072

Embase (OvidSP) Embase 1974 to 2015 December 28

5	**limit 4 to yr="2011 - 2015"**	**1423**
4	1 and 2 and 3	2864
3	(quality or analys* or compara* or valid* or attributes or recommend* or review or evidence-based or grad* or agree).ti.	1686953
2	(apprais* or evaluat* or scientific).ti,ab.	3511724
1	guideline*.ti.	72456

EBM Reviews - Cochrane Database of Systematic Reviews 2005 to December 23, 2015, Database Field Guide EBM Reviews - ACP Journal Club 1991 to December 2015, Database Field Guide EBM Reviews - Database of Abstracts of Reviews of Effects 2nd Quarter 2015, Database Field Guide EBM Reviews - Cochrane Central Register of Controlled Trials

November 2015, Database Field Guide EBM Reviews - Cochrane Methodology Register 3rd Quarter 2012, Database Field Guide EBM Reviews - Health Technology Assessment 4th Quarter 2015, Database Field Guide EBM Reviews - NHS Economic Evaluation Database 2nd Quarter 2015

	# ▼	Searches	Results
☐	**5**	**limit 4 to yr="2011 - 2015" [Limit not valid in DARE; records were retained]** Submit	**49**
☐	4	1 and 2 and 3 Submit	156
☐	3	(quality or analys* or compara* or valid* or attributes or recommend* or review or evidence-based or grad* or agree).ti. Submit	107386
☐	2	(apprais* or evaluat* or scientific).ti,ab. Submit	216444
	1	guideline*.ti. Submit	2447