Direction de la recherche, des études,
de l'évaluation et des statistiques
DREES

SERIE
SOURCES ET METHODES

DOCUMENT DE TRAVAIL

Analyse critique du développement d'indicateurs composites :
le cas de l'infarctus du myocarde après la phase aiguë

COMPAQH - INSERM

n° 19 – avril 2011

MINISTERE DU TRAVAIL, DE L'EMPLOI ET DE LA SANTE
MINISTERE DU BUDGET, DES COMPTES PUBLICS, DE LA FONCTION PUBLIQUE ET DE LA REFORME DE L'ÉTAT
MINISTERE DES SOLIDARITES ET DE LA COHESION SOCIALE

Cette publication n’engage que ses auteurs

Sommaire

Avant propos

À l'occasion de cette recherche, un groupe de travail a été constitué, composé de :

Mélanie Couralet (Projet COMPAQH), Sophie Guérin (Projet COMPAQH – stage de fin d'études ISUP), Marc Le Vaillant (Ingénieur de recherche CNRS – CERMES), Philippe Loirat (Projet COMPAQH), Etienne Minvielle (Projet COMPAQH).

Les auteurs remercient les cardiologues ayant participé à l'étude : Dr Pierre Aubry, Dr Loic Belle, Dr Simon Cattan, Pr Yves Cottin, Dr Frédéric Fossati, Pr Martine Gilard, Dr Michel Hanssen, Dr Eric Perchicot, Dr Elisabeth Pouchelon, Pr Christian Spaulding, Dr Jean-François Thébaut, Pr Patrice Virot, Dr Christian Ziccarelli, ainsi que M Kraay de la Banque Mondiale pour ses conseils avisés.

Note de synthèse

Ce rapport a pour objet de faire état du travail concernant la construction d'un **indicateur composite[1] portant sur la prise en charge de l'infarctus du myocarde après la phase aiguë**. Il fait suite à une demande de la Haute autorité de santé (HAS) et de la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) sur la « possibilité de créer un indicateur composite à partir des six indicateurs généralisés par la HAS en 2008 ».

[1] Un indicateur composite est une combinaison mathématique de plusieurs indicateurs qui représentent différentes dimensions d'un même concept.

Contexte

L'état des lieux montre que **les indicateurs composites sont largement diffusés dans de nombreux domaines :** plus de 300 sont développés aujourd'hui par les organismes internationaux. Un exemple connu est le Human Development Index (HDI) développé par les Nations Unies, qui permet un classement des pays à partir de trois dimensions (l'espérance de vie à la naissance, le taux d'éducation et le PIB par habitant). En France, le rapport sur la mesure des performances économiques et du progrès social, dit « rapport Stiglitz » (2009), recommande également la mise au point d'une mesure synthétique relative à l'estimation de la qualité de vie.

Dans le domaine de la santé, l'Organisation mondiale de la santé (OMS) a créé en 2000 un indicateur composite sur l'évaluation de la performance des systèmes de santé, classant les 191 états membres. Plus spécifiquement aux établissements de santé (ES), aux États-Unis, le National Committee for Quality Assurance (NCQA) a introduit en 1991 le programme HEDIS qui fournit un classement agrégé basé sur une série d'indicateurs ; en Angleterre, le NHS (National Health Service) diffuse les « League Tables » depuis 2001 ; enfin des initiatives privées (« Healthgrades » aux États-Unis, « Dr Foster » en Angleterre, les « palmarès » journalistiques en France, notamment) proposent des classements agrégés. À notre connaissance, spécifiquement à l'Infarctus du Myocarde, une seule initiative, CMS (Centers for Medicare & Medicaid Services), a développé un indicateur composite dans un objectif de comparaison, dans le cadre de son projet de paiement à la performance aux États-Unis.

En parallèle, **différentes controverses scientifiques accompagnent ces initiatives, ou expriment les plus grandes réserves quant à leur développement.** Comme en témoigne ce rapport, différentes études montrent les risques d'injustice associés au classement de producteurs de soins sur une telle base, et dénoncent la faible transparence des critères de choix des méthodes d'agrégation retenues [1-4].

Ce contexte exprime donc une lecture d'ensemble paradoxale : une tendance à l'augmentation des initiatives, alors que des réserves sont exprimées sur le plan scientifique.

Dans ce contexte, **l'objectif de cette étude** est double : (1) évaluer la sensibilité de plusieurs méthodes d'agrégation sur le classement des ES ; (2) si la publication d'un score agrégé et du classement qui en découle est envisagée, proposer des critères d'aide au choix entre les méthodes d'agrégation, sachant qu'il n'existe pas de « gold standard » par rapport auquel ces méthodes pourraient être évaluées.

Méthodes

Notre étude porte sur les 56 ES qui ont fait l'objet de l'expérimentation HAS en 2007, avec une analyse confirmatoire sur les données de la généralisation 2008 (n=275).

Une sélection des six indicateurs généralisés pouvant entrer dans le processus d'agrégation a été réalisée. Cinq indicateurs (pour lesquels au moins 30 dossiers de patients ont été évalués, ce qui permet la comparaison) ont été retenus : Prescription d'aspirine ; Prescription de bêtabloquant ; Mesure de la fraction d'éjection du ventricule gauche ; Prescription de statine et Sensibilisation aux règles hygiéno-diététiques.

Sept méthodes d'agrégation ont été sélectionnées à partir d'une revue de la littérature et adaptées au cas présent. Chaque méthode s'appuie sur un mode de construction du score spécifique reposant sur des objectifs différents. Les quatre premières partent des résultats des cinq indicateurs :

- Dans la méthode Indicator Average, l'indicateur composite est la moyenne des cinq indicateurs.
- Pour la méthode Budget Allocation Process (BAP), 13 cardiologues proposés par le Conseil National Professionnel de Cardiologie ont réparti un « budget » de 100 points entre les indicateurs à agréger. La moyenne des points ainsi attribués permet de calculer les poids des indicateurs, et le score composite est le résultat de leur somme pondérée.
- La méthode Benefit Of the Doubt (BOD) est une méthode statistique de maximisation sous contraintes qui permet, à partir des résultats observés des indicateurs, de définir un système de poids pour chaque ES qui lui soit le plus favorable possible.
- La méthode Unobserved Component Model (UCM) repose sur l'hypothèse selon laquelle la qualité est une grandeur non mesurable directement mais dont les différentes dimensions sont contenues de façon sous-jacente dans les mesures des indicateurs initiaux.

Dans les trois méthodes suivantes, le score composite agrège les résultats des cinq indicateurs, mais en revenant au niveau du patient :

- L'approche **All-or-None** consiste à donner un score nul si une seule étape du processus de soin n'a pas été respectée. Le score composite est égal à la proportion de patients dont la prise en charge a été complète.
- Dans la méthode **Patient Average**, le résultat pour un patient est égal à la moyenne des indicateurs. Le score composite est obtenu en faisant la moyenne des résultats des patients.
- La méthode Denominator-Based Weight (DBW) fournit un indicateur composite égal au ratio de la somme des processus réalisés pour tous les patients d'un ES, divisé par la somme des processus qui auraient dû être réalisés pour tous les patients.

Lorsqu'il n'y a aucune donnée non applicable, ce qui est le cas pour les cinq indicateurs retenus, les méthodes Indicator Average, Patient Average et DBW donnent les mêmes résultats : seules cinq méthodes sont donc comparées.

Pour chaque méthode d'agrégation, un score composite par ES est calculé, puis l'ES est classé dans l'une des trois catégories (« + », « = » et « − »), déterminée par la position de l'intervalle de confiance de son score composite par rapport à la moyenne.

Résultats

La méthode BOD fournit les scores les plus élevés ; à l'inverse, la méthode All-or-None produit les scores composites les plus faibles. Par ailleurs, **les résultats montrent une sensibilité du classement en fonction de la méthode d'agrégation.** D'une part, la concordance (coefficient kappa) est forte entre les trois méthodes BAP, BOD et Indicator Average ; et faible entre la méthode UCM et les autres. D'autre part, les ES changent de classe en fonction des méthodes pour les données de l'expérimentation (80 % dans le cadre de l'expérimentation (n=56), et 71 % dans le cadre des données de la généralisation (n=275)). Si l'on excepte la méthode UCM dont plus de 80 % des ES sont classés « = », 59 % (respectivement 48 % sur les données de la généralisation) des ES changent de classe en fonction des méthodes.

Discussion

L'étude sur l'agrégation permet de mettre en évidence le fait que les méthodes ne sont pas équivalentes entre elles puisqu'elles ont une influence sur le classement inter-établissements. Ce résultat confirme les données de la littérature, en le démontrant dans le cas d'un processus homogène, la prise en charge de l'infarctus du myocarde, et en tenant compte de l'incertitude.

Étant donné ce résultat, l'absence d'argument statistique en faveur d'une méthode, et l'absence d'équivalence entre ces mêmes méthodes, **deux orientations peuvent être prises**.

La première est de considérer que les réserves quant à l'emploi d'une méthode d'agrégation sont suffisantes pour ne pas développer de scores composites. Cette approche présente l'avantage de prendre en considération les limites statistiques qui ont été exprimées dans la littérature ces dernières années. La diffusion des résultats des indicateurs individuels et de la comparaison inter-établissement aux professionnels resterait nécessaire, car utile à leur démarche d'amélioration de la qualité, mais ces informations seraient sans doute peu lisibles pour le public.

Si l'on considère que la demande du public doit être satisfaite, on peut considérer qu'un score composite est nécessaire, ce d'autant que des scores composites sont utilisés dans différentes initiatives et d'autres domaines (économique avec la Banque mondiale, ou rapport « Stiglitz » sur la performance économique et le progrès social). Dans ce cas, la présentation des résultats doit s'accompagner **d'une information** sur les critères du choix de la méthode utilisée, ainsi que sur les limites inhérentes aux résultats.

Une minimisation du risque d'injustice de classer à tort un établissement « − » peut être envisagée, à travers différentes règles, par exemple, remonter tous les établissements dans la

catégorie supérieure, « = » ou « + » lorsqu'une autre méthode les classe mieux (dans notre étude, cela signifierait que 45 établissements seraient remontés, et seuls 3 resteraient « − » quelle que soit la méthode utilisée). Il pourrait aussi être envisagé de ne considérer que les résultats stables toute méthode confondue (ce qui concerne dans notre application 11 ES, 1 « + », 7 « = » et 3 « − »). L'explicitation des logiques sous-jacentes aux méthodes peut également constituer une aide à la décision, comme l'ont suggéré certains auteurs :

- Les méthodes, Indicator Average, Patient Average et DBW, en attribuant des poids égaux, ont l'avantage de la simplicité.
- La méthode All-or-None récompense l'excellence et met en valeur la coordination des soins.

Bien que le calcul soit différent, ces quatre premières méthodes donnent la même importance aux indicateurs, c'est-à-dire aux recommandations qui leurs sont sous-jacentes.

- La méthode BAP confère une légitimité professionnelle à l'indicateur composite.

Elle hiérarchise les recommandations par l'intermédiaire de l'avis d'experts.

- La méthode BOD valorise les établissements en allouant des poids élevés pour les indicateurs pour lesquels ils sont les plus performants.

Elle valorise le respect des recommandations (les résultats) en attribuant les plus grands poids aux indicateurs où l'ES est le plus performant.

- Enfin, la méthode UCM est une méthode qui se distingue de toutes les autres : la qualité est vue comme une variable latente commune à l'ensemble des indicateurs, et extraite des données au moyen d'un modèle mathématique.

Elle s'affranchit donc des considérations précédentes sur les recommandations.

Conclusion

En conclusion, nous considérons que le développement d'un score composite pour présenter les résultats des indicateurs de prise en charge de l'infarctus du myocarde après la phase aiguë est, malgré les réserves, un « mal nécessaire ». Des initiatives sont engagées en santé et dans d'autres domaines, et il est fort à parier que cette tendance, déjà perceptible, va s'accroître, la demande sociale étant plutôt en faveur de ces scores. Le rôle du scientifique dans ce cas de figure, en tant qu'aide à la décision des pouvoirs publics, nous semble d'accompagner prioritairement ce développement en exprimant explicitement les réserves, et en maintenant une « veille » de recherche sur les évolutions du débat.

Autrement dit, **si une méthode devait être retenue**, **la transparence** sur les critères du choix de la méthode utilisée, ainsi que les limites inhérentes aux résultats, semble nécessaire. **L'adhésion des professionnels** nous paraît également être un élément incontournable au développement de tels scores. Enfin, **la poursuite à titre expérimental de travaux de recherche** sur l'agrégation d'autres indicateurs cliniques, notamment dans le cadre des filières de soins (ou dans l'esprit de ce que les anglo-saxons nomment les « care bundles »), et à l'échelle de l'établissement, mériterait d'être poursuivie.

Références

1. **Jacobs R *et al.*** How robust are hospital ranks based on composite performance measures ? *Med Care*, 2005; 43(12):1177-84.
2. **Reeves D *et al.*** Combining multiple indicators of clinical quality : an evaluation of different analytic approaches. *Medical Care*, 2007 ; 45(6):489-96.
3. **O'Brien SM *et al.*** Exploring the behavior of hospital composite performance measures : an example from coronary artery bypass surgery. *Circulation*, 2007; 116(25):2969-75.
4. **Shwartz M *et al.*** Estimating a composite measure of hospital quality from the Hospital Compare database : differences when using a Bayesian hierarchical latent variable model versus denominator-based weights. *Med Care*, 2008 ; 46(8):778-85.

Introduction

Ce rapport a pour objectif de faire état du travail concernant la construction d'un score composite[2] portant sur la prise en charge de l'infarctus du myocarde après la phase aiguë. Ce travail fait suite à une demande de la Haute autorité de santé (HAS) et de la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) dans un courrier du 6 août 2009, sur la « possibilité de créer un indicateur composite à partir des six indicateurs généralisés par la HAS en 2008 ou à défaut un rapport présentant la justification de la non pertinence scientifique de la construction d'un tel indicateur ». Cette demande est partie du constat qu'en termes de diffusion des résultats des établissements de santé sur la prise en charge de l'infarctus du myocarde après la phase aiguë sur le site du ministère Platines (DREES), un score agrégé serait plus compréhensible et permettrait une meilleure transparence vis à vis du grand public.

Il n'existe pas de « gold standard » en matière de méthode d'agrégation, et de fait, la littérature fait état de plusieurs méthodes utilisées dans divers programmes internationaux ou études d'évaluation de la qualité. La revue de la littérature montre que leur application peut produire des scores composites et des résultats en termes de classements différents [1-5].

Le travail statistique se définit en trois temps :

- une présentation de sept méthodes parmi les plus couramment utilisées, adaptées à notre cas ;
- un descriptif des scores agrégés obtenus par ces méthodes ;
- une analyse des différences observées sur le classement des établissements de santé dans l'application de ces sept méthodes.

Dans une perspective d'aide à la décision de la HAS et la DREES dans le recours éventuel à une méthode, les critères et arguments qui peuvent orienter le choix sont également discutés.

[2] Pour des commodités d'expression, nous utilisons indifféremment les termes de scores et d'indicateurs composites dans ce rapport.

Contexte

Infarctus du myocarde : Éléments de contexte

Chaque année en France, il est estimé qu'environ 100 000 personnes sont atteintes d'infarctus du myocarde (IDM) [6]. Parmi les patients pris en charge, 7 % décèdent dans le premier mois et au total 13 % décèdent au cours de la première année. Cette mortalité a été réduite de moitié en dix ans grâce à une amélioration globale de la prise en charge.

Lors d'un infarctus du myocarde, après le traitement de la phase aiguë, un bilan est réalisé pour mettre en route un traitement adapté aux facteurs de risque cardiovasculaires et à la fonction cardiaque.

Ce bilan comprend notamment la recherche et la prise en charge du tabagisme et du diabète – facteurs majeurs de risque cardiovasculaire qui, non traités, augmentent le risque de récidive et la mortalité – ainsi que la réalisation d'examens pour évaluer la fonction cardiaque.

Le traitement est d'abord médicamenteux dit « BASI » associant 4 types de médicaments : bêtabloquant (B), antiagrégant plaquettaire (A), statine (S) et inhibiteur de l'enzyme de conversion (I). Il est complété par une réadaptation cardiaque (rééducation à l'effort) qui doit être envisagée dans les suites immédiates de l'infarctus du myocarde.

Le traitement, débuté à l'hôpital, a pour but de soulager le patient, de prolonger sa vie, et d'éviter l'aggravation ou la récidive de l'IDM. Bien suivi, il contribue également à éviter les autres complications de la maladie cardio-vasculaire, tels que l'accident vasculaire cérébral, l'artériopathie oblitérante des membres inférieurs, et l'insuffisance cardiaque [6].

De nombreuses recommandations de pratiques cliniques existent pour assurer la qualité de la prise en charge de l'IDM. Plus spécifiquement, des recommandations validées par les sociétés nationales ont été élaborées pour la prescription des traitements médicamenteux, pour la prise en charge sur le plan lipidique et l'éducation en vue de l'arrêt du tabac [7].

Généralisation d'indicateurs

Depuis 2006, la HAS, en coopération avec le ministère de la Santé (Direction de l´hospitalisation et de l´organisation des soins-DHOS, DREES), est engagée dans la généralisation d'indicateurs de qualité et la comparaison inter-établissements. Ce processus s'inscrit dans le prolongement du tableau de bord des infections nosocomiales réalisé par le ministère de la santé (DHOS, Direction Générale de la Santé-DGS) et la mise en ligne en janvier 2007 par le ministère de la santé du site Platines (DREES), plate-forme d'informations sur les établissements de santé.

En 2008, onze indicateurs de qualité ont fait l'objet d'un recueil généralisé dans les établissements de santé MCO (Médecine Chirurgie Obstétrique). Ces indicateurs concernent trois procédures de qualité : la tenue du dossier patient, qui se décline en 4 indicateurs ; la tenue du dossier anesthésique, qui recouvre un indicateur ; et enfin le respect des bonnes pratiques de prise en charge hospitalière de l'infarctus du myocarde après la phase aiguë, qui s'apprécie à travers six indicateurs.

Indicateurs généralisés sur la qualité de prise en charge de l'IDM après la phase aiguë

Les six indicateurs généralisés par la HAS et pris en compte dans cette étude sont les suivants (détails en annexe I) :

- Prescription d'aspirine et de clopidogrel à l'issue du séjour du patient ;
- Prescription de bêtabloquant à l'issue du séjour du patient ;
- Mesure de la Fraction d'éjection du ventricule gauche (FEVG) (Niveau 1) ; et pour les patients dont la FEVG est ≤ 40%, prescription d'Inhibiteur de l'enzyme de conversion (IEC) à l'issue du séjour du patient (Niveau 2) ;
- Prescription de statine à l'issue du séjour du patient (Niveau 1) ; et pour les patients qui ont fait l'objet d'une prescription de statine, prescription d'un bilan lipidique à distance (Niveau 2) ;
- Sensibilisation aux règles hygiéno-diététiques durant et à l'issue du séjour du patient ;
- Délivrance de conseils pour l'arrêt du tabac durant ou à l'issue du séjour du patient.

Ces indicateurs ont été conçus par le projet COMPAQH (COordination pour la mesure de la performance et l'amélioration de la qualité hospitalière – INSERM) entre 2004 et 2006, en collaboration avec la Société française de cardiologie [8] : les indicateurs ont été testés sur leur faisabilité, leur reproductibilité, leur validité interne et leur pouvoir discriminant. Ils sont également utilisés dans les initiatives internationales qui évaluent la qualité de la prise en charge de l'IDM à la sortie [9-12]. Ils sont fondés sur des recommandations de bonnes pratiques cliniques.

Le score de chaque indicateur est binaire : il vaut 1 si le processus de soin apparaît dans le dossier médical, 0 sinon. Le score d'un indicateur au niveau de l'établissement de santé est calculé comme le pourcentage de patients pour lesquels le processus a été satisfait ; il est accompagné d'un intervalle de confiance à 95 %.

C'est sur ces six indicateurs généralisés qu'a été menée l'étude sur l'agrégation.

Les indicateurs composites

Définition

Un « **indicateur composite** » est une combinaison mathématique (ou « **agrégation** ») de plusieurs indicateurs qui représentent différentes dimensions d'un même concept [13]. D'autres terminologies existent : il est également possible de parler d'index, de score ou d'indice agrégé.

La définition d'un indicateur composite donnée par la Commission européenne (Joint Research Center - JRC) [14] est la suivante :

« *Les indicateurs composites sont basés sur des sous-indicateurs qui n'ont pas d'unité de mesure (significativement) commune.* » Pour construire un indicateur composite, plusieurs **méthodes d'agrégation** existent.

Éléments historiques

L'utilisation des indicateurs composites a toujours été source de controverses [15, 16]. Cependant, ces réserves n'ont pas été jugées suffisantes pour empêcher leur développement [17]. Ils permettent une comparaison simple des unités, et sont donc facilement compréhensibles par l'opinion publique, et/ou le régulateur/payeur. Néanmoins, étant donné qu'il n'existe pas de méthode unique, claire et évidente pour les construire, les résultats sont sensibles à la méthodologie.

Dans les années 1970, les tous premiers indices composites concernaient le risque d'insolvabilité des pays, ainsi que d'autres problèmes économiques. Au cours du temps, les sujets couverts se sont diversifiés, en même temps que le nombre d'institutions qui élaborent de tels indices a augmenté. En 2005, Bandura écrit : « *La quantité d'indices composites créés est montée en flèche, tout particulièrement ces 15 dernières années. Environ 80 % d'entre eux ont été construits dans la période 1991-2005* » [18].

L'une des principales utilisations des indices composites est l'évaluation des performances et des progrès sociotechniques au cours du temps. Ils servent à identifier et à suivre les actions nécessaires à l'amélioration des performances [15]. Ils sont également largement exploités pour créer des classements, et ainsi évaluer les performances relatives et leur évolution.

La performance des pays est désormais évaluée par ces indicateurs composites dans de nombreux domaines incluant, entre autres, l'économie, l'éducation, la santé, la corruption, les droits de l'homme, l'environnement, la recherche et l'innovation, entre autres. Des échelles de qualité de vie anglo-saxonnes ont été développées depuis de nombreuses années : SIP (Sickness Impact Profile) [19], NHP (Nottingham Health Profile) [20], MOS SF-36 (Medical Outcomes Study Short 36-item Health Survey) [21]. Les raisonnements entre scores composites et échelles de mesure [22] sont proches mais distincts, la littérature entre les deux domaines se recoupant rarement. Dans le cas des échelles de qualité de vie, ils diffèrent par la nature des données utilisées et par leur rapport à la dimension globale étudiée.

Les rapports présentant des indicateurs composites sont nombreux : plus de 300 sont développés aujourd'hui par les organismes internationaux [23]. Dans de nombreux domaines, un indicateur composite a été créé pour évaluer les pays et les comparer. Par exemple, les Nations Unies classent les pays avec le Technology Achievement Index et le Human Development Index [24] sur la base des travaux de Sen [25] ; la Banque mondiale les classe avec un score agrégé d'indicateurs de gouvernance [26].

Dans le domaine de la santé

De même, des initiatives internationales développent et diffusent des indicateurs composites dans le domaine de la santé. Leur utilisation est aussi controversée [1-3, 15, 27-29]. La principale critique provient du fait que des choix plus ou moins explicites sont faits à différents niveaux : sur la sélection des indicateurs et le système de pondération choisi, notamment. Ces choix peuvent entraîner des différences dans le classement final, ce qui serait source d'injustice. De plus, le score agrégé peut cacher de mauvais résultats sur un ou plusieurs indicateurs/dimensions, ce qui peut être préjudiciable dans un emploi d'aide à

l'amélioration. Enfin, l'incertitude autour du score composite et donc du classement qui en résulte est trop peu souvent appréciée.

Le classement de l'OMS

En 2000, l'OMS est la première organisation à publier un classement de 191 pays selon la performance de leur système de santé dans le *Rapport sur la santé dans le monde* [30].

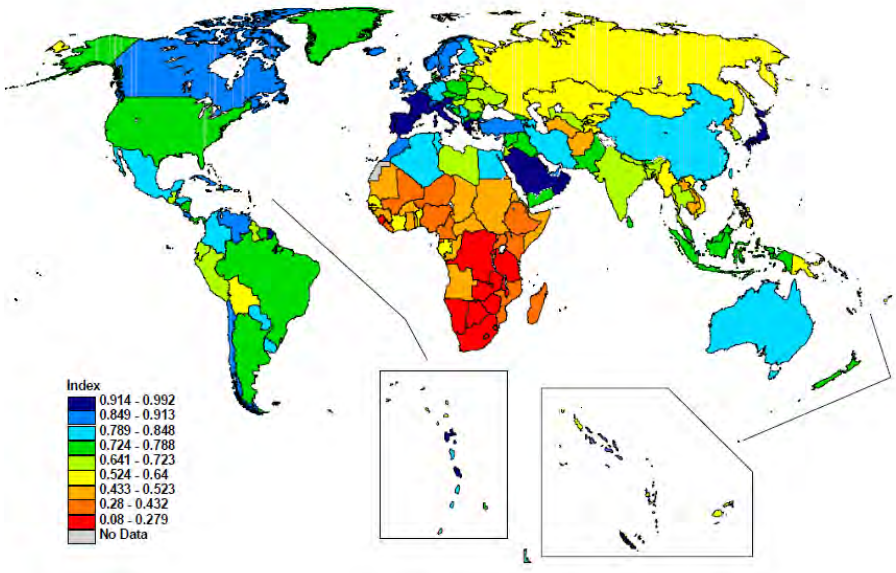
Cinq indicateurs individuels sont utilisés pour construire le score de performance des systèmes de santé :

- Le niveau de santé générale (level of Health) ;
- La distribution de la santé dans la population (Health Inequality) ;
- Le degré général de la réactivité du système de santé (Responsiveness) ;
- La distribution de cette réactivité (Responsiveness Inequality) ;
- La répartition/équité de la contribution financière (Fairness of financial contribution).

Les scores indicateurs sont tout d'abord ramenés entre 0 et 100, puis pondérés sur la base d'une enquête de 1006 informateurs issus de 125 pays.

L'indicateur composite prend des valeurs comprises entre 0 (Sierra Leone) et 0,99 (France). Chaque mesure est accompagnée d'un intervalle de confiance. Les résultats sont présentés sous la forme d'un tableau, mais également sous forme de carte [31].

Figure 1 - Répartition de la performance des systèmes de santé dans 191 états membres de l'OMS en 1997



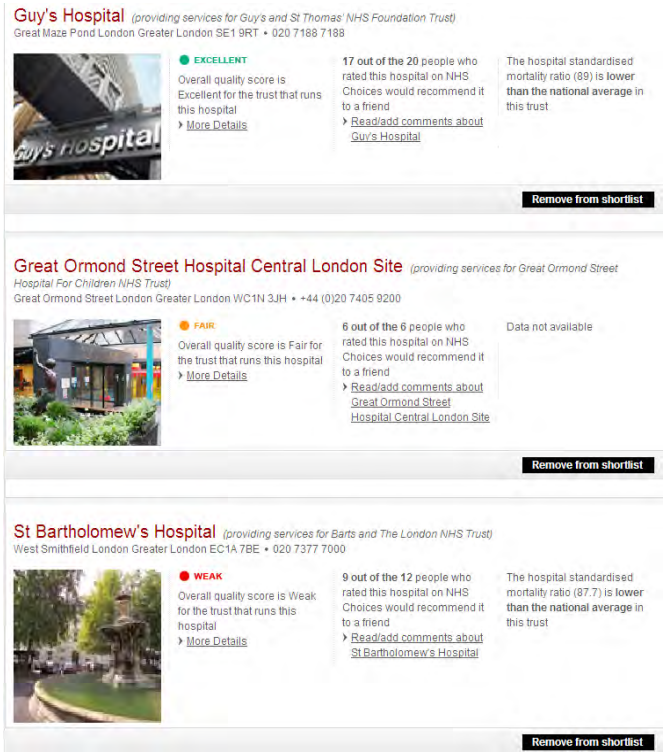
Sources : OMS, 2000.

Le classement de l'OMS a déclenché un grand nombre de commentaires et de débats, en particulier sur la méthodologie. Des critiques ont été émises sur la façon de mesurer la performance, le traitement des données manquantes ou la pauvre qualité des données, la manière de déterminer les pondérations des indicateurs, notamment [27, 29].

Des classements hospitaliers

À l'étranger, parmi les structures publiques, les « League tables » classent les établissements de santé selon leur performance globale : en Angleterre, la Care Quality Commission (anciennement Healthcare Commission) a établi un classement en quatre catégories : « Excellent », « Good », « Fair » et « Weak » sur la base d'une série d'indicateurs, dont le NHS diffuse les résultats [32]. Trois dimensions sont définies, dans lesquelles se déclinent les indicateurs : les « core standards » (standards applicables à tous les ES qui constituent un seuil minimal d'objectif à atteindre), « existing commitments » (engagements existants de l'ES) et « national priorities » (objectifs prioritaires définis nationalement et applicables à l'ES). Par exemple, pour les établissements MCO, le score composite est calculé sur la base de 47 indicateurs.

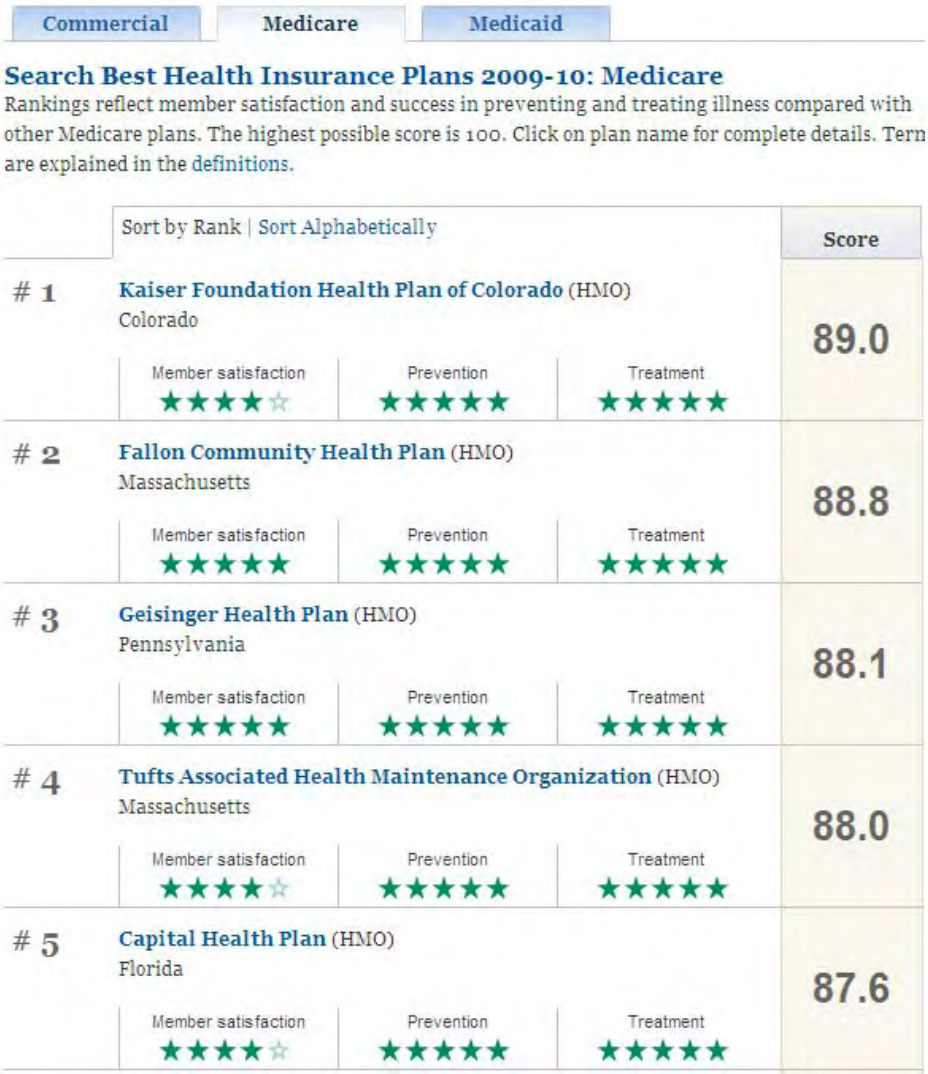
Figure 2 - Exemple de classement hospitalier par le NHS



Sources : NHS choices 2010

Aux États-Unis, le NCQA développe des comparaisons entre réseaux de soin à partir d'indicateurs de qualité portant sur la satisfaction des patients, la prévention et le traitement des pathologies [33].

Figure 3 - Exemple de classement hospitalier par le NCQA



Sources : U.S. News 2009-10 rankings.

Aux États-Unis toujours, le Centers for Medicare & Medicaid Services (CMS) et la Joint Commission on the Accreditation of Healthcare Organizations (JCAHO) proposent des classements basés sur de nombreux indicateurs diffusés par l'intermédiaire des sites « hospitalcompare » et « quality check » respectivement [34, 35].

Parallèlement à ces démarches, d'autres classements, guidés par des finalités commerciales, existent. Citons par exemple « Healthgrades » aux USA [36] et « Dr Foster » en Angleterre [37].

En France

Le 16 mars 2004, le ministre de la santé a annoncé la conception d'un tableau de bord des infections nosocomiales pour les établissements de santé [38]. L'objectif de ce tableau de bord était d'encourager tous les établissements de santé à mesurer leurs actions dans le

domaine de la lutte contre les infections. Les établissements les plus impliqués sont ainsi valorisés et les autres incités à progresser.

Le tableau de bord est une façon de présenter un certain nombre d'informations simples et sélectives. Il permet un suivi dans le temps et des comparaisons entre les établissements, facteurs d'amélioration de la qualité. Il répond à une demande légitime d'information et de transparence de la part des usagers. La publication du premier indicateur Indice Composite d'Action de Lutte contre les Infections Nosocomiales (ICALIN) 2004 a marqué le début de cette démarche. Le titre même comporte le terme « composite », mais en l'occurrence, il s'agit d'items qui couvrent un même thème dans le cadre d'une échelle de mesure à trois dimensions, et dont la corrélation inter-items a été testée par dimension [39].

Les tableaux de bord suivants se sont enrichi de nouveaux indicateurs : ICSHA (indicateur de consommation de solutions hydro-alcooliques), SURVISO (surveillance des infections du site opératoire) et ICATB (indice composite de bon usage des antibiotiques). Afin d'améliorer la lecture de ce tableau de bord des infections nosocomiales, le ministère chargé de la Santé a développé un score agrégé, élaboré à partir des résultats de chacun des indicateurs. Les usagers ont ainsi à leur disposition un affichage simplifié des quatre indicateurs sous forme d'une classe de A à F et d'une note sur 100 par catégorie d'établissements. La technique de pondération est « à dire d'experts » : des experts ont classé les quatre indicateurs selon leur importance relative puis le poids relatif de chaque indicateur dans le score agrégé a été fixé ainsi [40] :

- ICALIN = 40 %
- ICSHA = 30 %
- SURVISO = 20 %
- ICATB = 10 %.

Par ailleurs, et depuis quelques années, les médias publient régulièrement des classements des établissements de santé, basés sur des scores composites [41-44]. Ces « palmarès » sont créés selon des méthodologies souvent floues, mais leur succès prouve qu'ils répondent à une vraie demande du public.

En France, le rapport de la Commission sur la mesure des performances économiques et du progrès social, dit rapport « Stiglitz » (2009), recommande la mise au point d'une mesure synthétique relative à l'estimation de la qualité de vie [45].

Indicateurs et indicateurs composites sur l'infarctus du myocarde à l'international

Seize sites d'organisations internationales ont été visités.

Des indicateurs relatifs à l'infarctus du myocarde sont développés à l'étranger, avec l'Australian Council on Healthcare Standards (ACHS) [46], le BQS en Allemagne [47] et le National Committee for Quality Assurance (NCQA) aux États-Unis [48]. Concernant plus particulièrement la prise en charge, des initiatives telles que le Canadian Cardiovascular Outcome Research Team (CCORT) [9], la JCAHO [12], l'Agency for Healthcare Research and Quality (AHRQ) [11] et le CMS [10] ont construit des indicateurs.

Un score composite pour représenter la prise en charge de l'IDM a été construit par la JCAHO [12], l'AHRQ [11] et CMS [49]. L'AHRQ et la JCAHO le présentent dans leurs rapports annuels respectifs au niveau national, et par Etat pour l'AHRQ. Dans les deux cas, il est constitué des indicateurs : prescription d'aspirine dans les premières 24h et à la sortie, prescription de bêtabloquant dans les premières 24h et à la sortie, prescription d'IEC à la sortie et conseils pour l'arrêt du tabac pendant l'hospitalisation. L'indicateur composite constitue pour ces deux initiatives un outil de suivi de la performance de la prise en charge de l'IDM au niveau national et par Etat et non un outil de comparaison inter-établissements.

De son côté, le CMS, par l'intermédiaire du site « hospitalcompare », propose un classement des établissements sur cette pathologie par indicateur. En parallèle, il développe depuis 2003 le projet « Hospital Quality Incentive Demonstration » dont l'objectif est de déterminer si l'incitation financière est un levier à l'amélioration de la qualité des soins. La participation au projet est volontaire et concerne 230 établissements de santé. À cet égard, cinq indicateurs composites relatifs à cinq pathologies (dont l'IDM) ont été créés afin de comparer les hôpitaux. Le classement permet à la fois de rémunérer et de mettre en valeur publiquement les meilleurs hôpitaux.

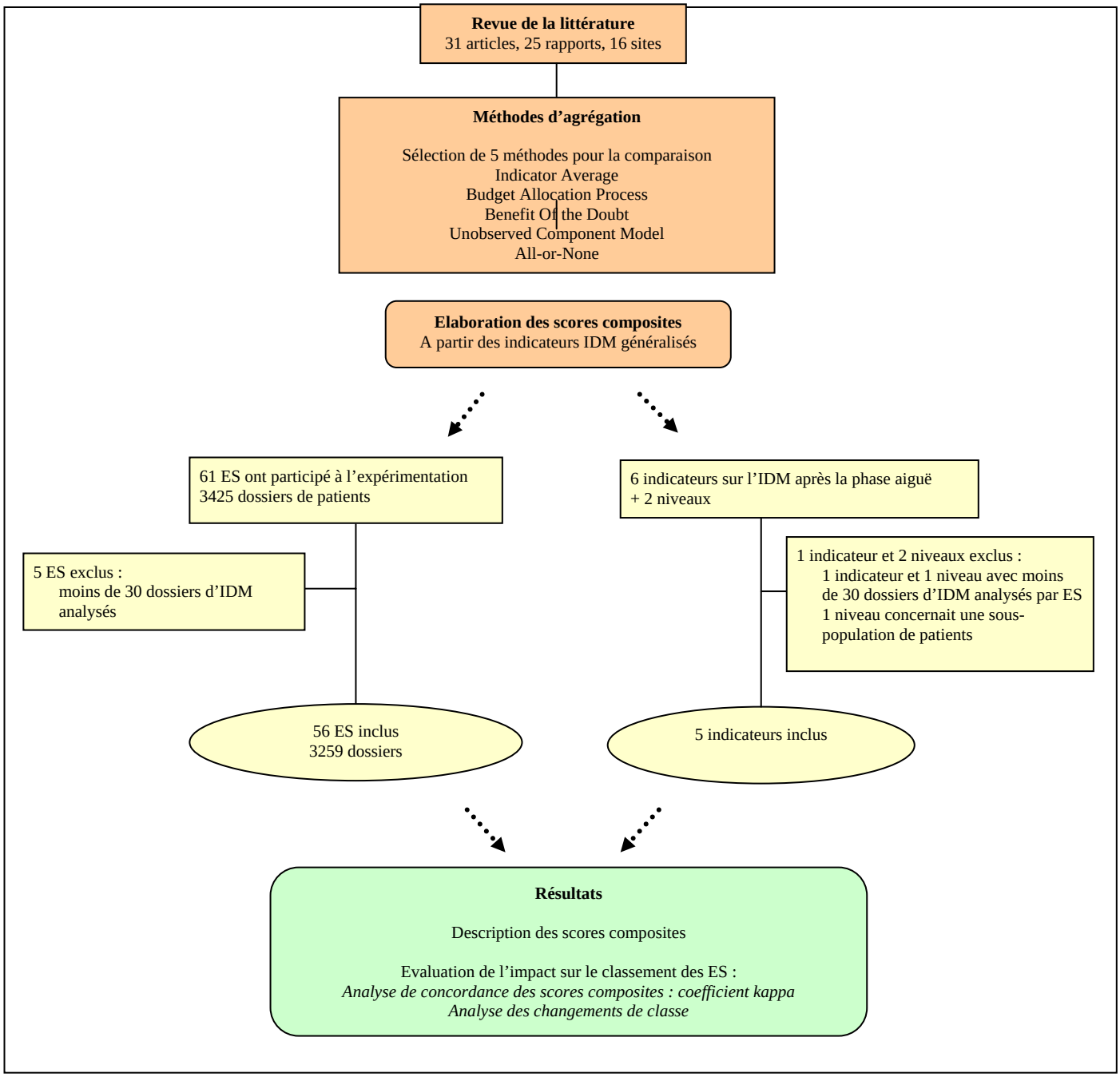
Pour ces trois organisations, la méthode consiste à agréger des indicateurs individuels en donnant à chacun un poids égal.

Des études se sont servies de scores composites d'indicateurs de pratique clinique, dont l'infarctus du myocarde, afin d'illustrer leur propos mais sans discuter la méthodologie de construction [50-55]. Certains auteurs ont même souligné l'intérêt d'avoir des scores composites dans le cadre de l'IDM [55].

Méthodes

Les principes méthodologiques sont présentés d'une manière synthétique dans la figure 4 ci-dessous.

Figure 4 - Schéma de la méthode



Revue bibliographique sur les méthodes d'agrégation

Stratégie de recherche

Nous avons réalisé une revue systématique de la littérature via Pubmed, avec les mots-clés suivants :

- *("index"[Title] OR "score"[Title] OR "indicator"[Title]) AND ("composite"[Title]) OR "synthetic"[Title])*
- *"Quality Indicators, Health Care"[Mesh] AND "composite"[Title/Abstract]*

La recherche a été effectuée sans limitation de date mais en la restreignant aux publications en français et en anglais. Les deux recherches ont respectivement mis en lumière 158 et 60 articles, dont 118 et 58 publiés depuis 2000. Sur tous ces articles passés en revue, 15 références pertinentes ont finalement été sélectionnées sur la base du titre et du résumé.

Une recherche complémentaire, notamment sur les principes des méthodes d'agrégation, a été effectuée : 23 références supplémentaires (16 articles et 7 rapports) ont été trouvées.

En parallèle, de nombreux sites internet ont été visités de façon approfondie de manière à identifier les rapports publiés sur le sujet. Au total, 35 sites ont été visités et 18 rapports référencés.

Principales sources méthodologiques

Bien que la littérature concernant les indicateurs composites soit abondante, les sources traitant de la méthodologie de construction des indicateurs composites datent essentiellement des années 2000.

En particulier, l'OCDE (Organisation de coopération et de développement économiques) et la Commission Européenne (JRC) ont beaucoup exploité le sujet, avec notamment la rédaction commune d'un guide méthodologique sur la construction d'indicateurs composites : *Handbook on Constructing Composite Indicators, Methodology and User Guide* [15].

En Angleterre, le centre d'économie de la santé de l'université de York a aussi travaillé sur ce thème, en particulier sur les indicateurs composites en santé, leur construction et leur implication dans le classement des hôpitaux [1, 27, 28, 56, 57].

Les questionnements méthodologiques relatifs à l'agrégation des indicateurs sont d'ailleurs toujours d'actualité : l'European Science Foundation organise notamment une école d'été en 2010 dont l'un des thèmes est « Methodological aspects of the use of composite scores in cross-cultural research ».

La construction d'un indicateur composite : Principes généraux

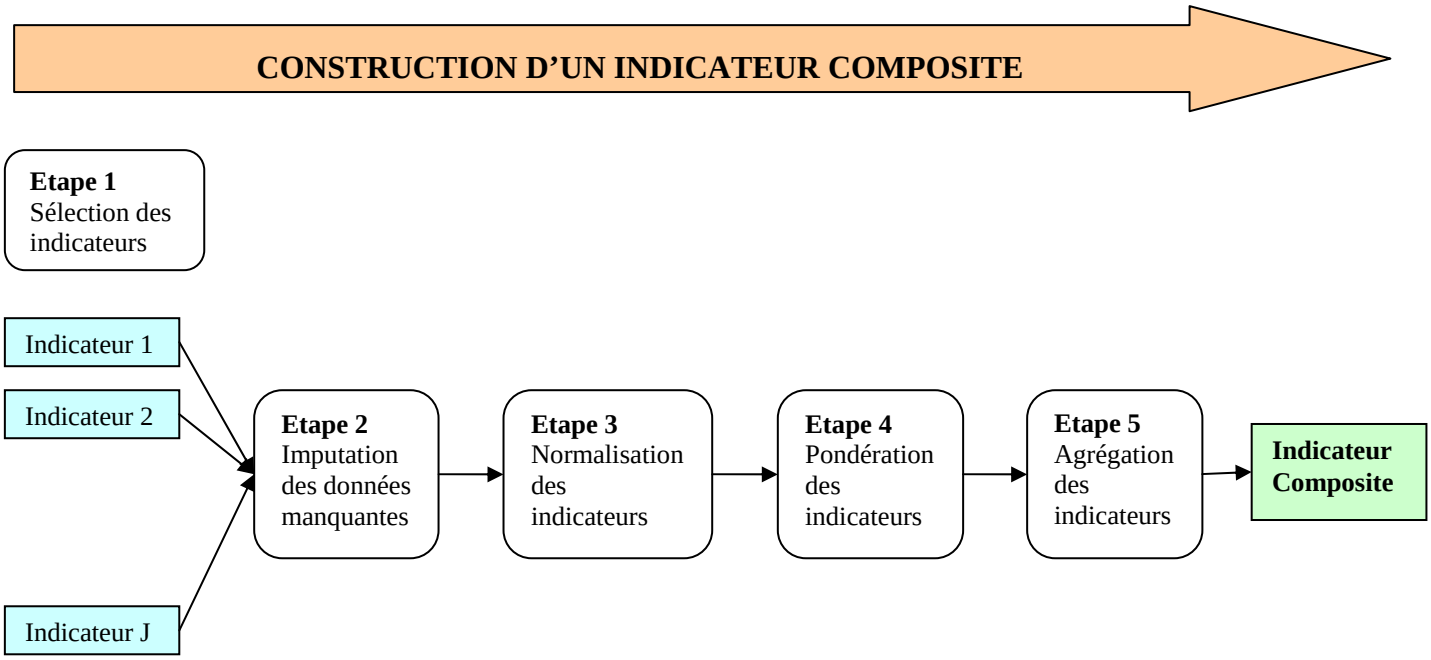
La construction d'un indicateur composite ne répond pas à une méthodologie unique mais dépend des données disponibles et des objectifs que le promoteur poursuit. Aucun modèle n'est a priori meilleur qu'un autre. Par ailleurs, il n'existe pas de « gold standard » par rapport auquel une méthode pourrait être considérée comme la plus appropriée.

La méthodologie de construction d'un indicateur composite comporte plusieurs étapes [15] :

1. la sélection des indicateurs pertinents ;
2. le traitement des données manquantes ;
3. la normalisation des indicateurs si les unités de mesure sont très différentes ;
4. la pondération des indicateurs ;
5. l'agrégation des indicateurs.

Les indicateurs composites vont dépendre essentiellement des poids rattachés aux indicateurs : les résultats seront très sensibles au choix du schéma de pondération.

Figure 5 - Etapes de construction d'un indicateur composite



En raison des spécificités de notre étude certaines étapes ont pu être omises dans notre travail de construction d'un indicateur composite relatif à la qualité de la prise en charge de l'IDM après la phase aiguë, ainsi :

- Les indicateurs choisis sont ceux qui ont été généralisés par la HAS en 2008.
- La mise en œuvre d'une procédure d'imputation des données manquantes n'a pas été nécessaire. En effet, le recueil des données concernant chaque patient est réalisé en une seule étape qu'il est impossible de valider si tous les indicateurs n'ont pas été

remplis. Aucune donnée « pour un patient » ne peut donc être manquante, mais il peut en revanche y avoir des données « non applicables ». Par exemple, il est inutile de donner des conseils sur l'arrêt du tabac à un non fumeur, et le patient correspondant n'aura donc pas de valeur pour l'indicateur « Délivrance de conseils pour l'arrêt du tabac ».

- Tous les indicateurs calculés au niveau de l'établissement sont compris entre 0 et 1. Il n'est donc a priori pas nécessaire de les normaliser afin de les ramener à la même échelle.

L'essentiel de notre travail a donc porté sur l'étape de la pondération des indicateurs, et plusieurs méthodes seront présentées dans le paragraphe suivant.

L'étape finale d'agrégation des indicateurs conduisant au calcul du score composite a été réalisée par la mise en œuvre d'une méthode d'agrégation linéaire : l'indicateur composite sera donc pour chaque méthode de pondération le résultat de la somme pondérée des indicateurs. Il existe d'autres méthodes d'agrégation (notamment géométrique), mais la plus simple et la plus utilisée est la méthode linéaire.

Il faut noter que le terme « agrégation » est utilisé soit pour caractériser l'ensemble des étapes de la construction d'un indicateur composite, soit pour désigner la dernière étape du processus qui consiste à réunir en un seul score les indicateurs pondérés. Dans ce rapport, il sera employé pour indiquer l'ensemble des étapes menant à l'obtention d'un indicateur composite, c'est-à-dire la pondération et l'agrégation linéaire des indicateurs pondérés.

Population d'analyse et sélection des indicateurs

Ce travail est exploratoire et consiste, avec le jeu de données à disposition, à comprendre l'impact de l'utilisation des méthodes. À cet égard, pour des raisons pédagogiques et de lisibilité des résultats, l'étude a été conduite à partir de l'échantillon des établissements ayant participé à l'expérimentation des indicateurs en 2007 (61 ES). Cependant, en complément, une analyse confirmatoire sur les données de la généralisation 2008 (649 ES) a été réalisée.

Certains indicateurs ou niveaux d'indicateurs et certains ES ont été exclus du processus d'agrégation, et ce pour deux raisons :

- Selon les consignes de la HAS, un indicateur ne peut être utilisé dans une comparaison inter-établissements de santé que s'il a été évalué dans un minimum de 30 dossiers patients. Si l'établissement a moins de 30 dossiers, l'indicateur est calculé malgré tout mais son score n'est pas comparé à ceux des autres.
  Ainsi, afin de ne retenir que les indicateurs dont les résultats permettent la comparaison des établissements, ceux qui présentent un effectif de moins de trente dossiers par établissement pour tous les établissements ont été exclus. Cela concerne les indicateurs « Prescription d'inhibiteur de l'enzyme de conversion » (niveau 2 de l'indicateur 3) et « Délivrance de conseils pour l'arrêt du tabac » (indicateur 6). En conséquence également, les établissements qui ont évalué moins de 30 dossiers ou sont non répondants ont été exclus : 5 ES de l'échantillon « expérimentation » et 374 de l'échantillon « généralisation » sont concernés (plus de la moitié des ES concernés par la généralisation prennent en charge moins de 30 IDM par an). Le processus

d'agrégation concerne donc 56 ES ayant participé à l'expérimentation et 275 ES de la généralisation. Chaque ES a évalué entre 30 et 60 dossiers, ce qui représente une base de données de 3259 dossiers pour l'expérimentation et 14966 pour la généralisation.

- Par ailleurs, il a été décidé de ne pas retenir les niveaux 2, liés par construction aux niveaux 1, afin d'assurer une représentation équivalente de chacun des indicateurs et d'éviter toute redondance.

La construction d'un indicateur composite de la prise en charge de l'infarctus du myocarde après la phase aiguë est donc réalisée sur cinq indicateurs pour lesquels aucune donnée n'est non applicable : Prescription d'aspirine, Prescription de bêtabloquant, Mesure de la FEVG, Prescription de statine et Sensibilisation aux règles hygiéno-diététiques.

Méthodes d'agrégation sélectionnées

Suite à la revue de la littérature, dix méthodes d'agrégation ont été identifiées. Seules les plus adaptées au jeu de données sur l'infarctus du myocarde après la phase aiguë sont présentées ici. Cela concerne sept méthodes : « Indicator Average », « Benefit Of the Doubt », « Unobserved Component Model », « Budget Allocation Process », « All-or-None », « Patient Average » et « Denominator-Based Weight ».

Les quatre premières sont des méthodes de pondération des indicateurs au niveau de l'établissement, et la somme pondérée des indicateurs fournit l'indicateur composite. Les trois suivantes utilisent l'information au niveau du patient pour calculer le score global de l'établissement.

Nous faisons ici une description synthétique, un développement plus complet étant proposé en annexe.

Indicator Average

La manière la plus intuitive d'agréger des indicateurs consiste à réaliser la moyenne de leurs scores. Cela sous-entend que le même poids est attribué à chaque indicateur.

De nombreux exemples existent où la même importance est accordée à tous les indicateurs, comme le « Environmental Sustainability Index – ESI » (par le World Economic Forum et le Joint Research Centre of the European Commission [58]), le « Human Development Index – HDI », ou le « Technology Achievement Index – TAI » (par les Nations Unies [24]).

Tableau 1 - Exemple de calcul du score « Indicator Average »

	Indic 1	Indic 2	Indic 3	Indic 4	Indic 5	Score Indicator Average
ES 1	0,77	0,83	0,74	0,92	0,86	0,83
ES 2	0,98	0,97	1,00	0,98	0,93	0,97
ES 3	0,90	0,63	0,63	0,85	0,17	0,64
...						

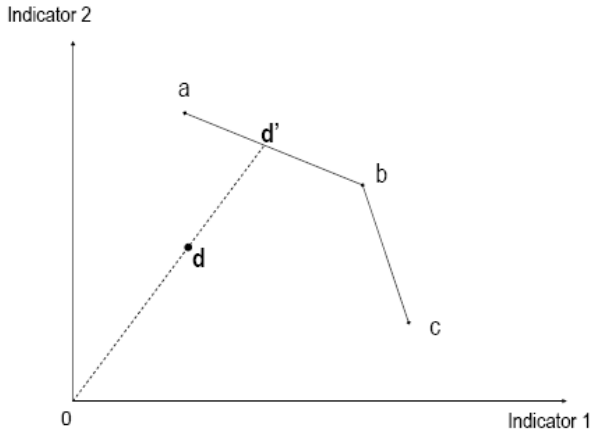
Benefit Of the Doubt (BOD)

L'approche BOD est une application de la méthode Data Envelopment Analysis (DEA) spécifique à la construction d'indicateurs composites [15, 59, 60]. L'approche DEA permet de construire une « frontière d'efficacité » convexe qui « enveloppe les données », et qui sert de référence (« benchmark ») pour mesurer la performance relative des établissements. La frontière d'efficacité est constituée des « meilleurs » établissements [61].

La méthode BOD est très utilisée dans de nombreux domaines, par exemple pour fournir un système de pondération alternatif pour le HDI [62, 63] et pour le TAI [59]. Dans le domaine de la santé, des travaux de recherche ont été menés afin d'étudier la possibilité d'utiliser cette méthode pour mesurer l'efficacité des centres de santé publiques au Ghana [64] ou la performance des hôpitaux Zambiens [65].

Elle peut être illustrée par un exemple simple [15], dans lequel quatre établissements (**a**, **b**, **c**, **d**) sont évalués par deux indicateurs. Ainsi, en dimension 2, il est possible de représenter graphiquement le score de chaque établissement :

Figure 6 - Exemple d'une frontière de performance (méthode BOD)



Sources : OECD, 2008.

Les établissements **a**, **b** et **c** sont les plus performants, un score de 1 leur est accordé. Ils constituent la frontière. La performance de l'établissement **d** est mesurée par : Od/Od', et est par définition inférieure à 1.

Le jeu de poids de chaque établissement dépend donc de sa position par rapport à la frontière, tandis que le benchmark correspond au point idéal (**d'** est le benchmark de **d** dans l'exemple). La particularité de la méthode est que le jeu de poids des indicateurs du score composite est déterminé par l'examen des données observées, et est particulier à chaque établissement.

Plus généralement, le score composite d'un établissement e est le ratio de la performance actuelle sur la performance benchmark :

$$IC_e = \frac{actual\ overall\ \ performance}{benchmark\ overall\ \ performance}$$

Le jeu de poids permettant de définir le score de chaque établissement est obtenu en résolvant un problème de maximisation sous contraintes, qui sont, d'une part la non-négativité des poids $w_{e,j}$ et d'autre part le fait que tous les scores composites doivent être inférieurs à 1. Le score composite est la somme des indicateurs pondérés :

$$IC_e = \max_{w_{e,j},\, j=1,\ldots,J} \left\{ \sum_{j=1}^{J} w_{e,j} y_{e,j} \right\}$$

avec

$$\begin{cases} 0 \le \sum_{j=1}^{J} w_{e,j} y_{e,j} \le 1 \; ; \; e = 1,\ldots,E & \text{(c1)} \\ w_{e,j} \ge 0 \; ; \; j = 1,\ldots,J & \text{(c2)} \end{cases}$$

Une démonstration complète du raisonnement se trouve en annexe II.

La maximisation a l'effet suivant : pour chaque établissement de santé, les poids les plus importants sont accordés aux indicateurs pour lesquels cet établissement accomplit la meilleure performance. L'idée centrale est qu'une bonne performance relative d'un établissement sur un indicateur indique que cet établissement considère la dimension comme importante. En d'autres termes, la méthode BOD accorde le « bénéfice du doute » aux établissements.

La performance d'un établissement est relative : elle est comparée à celle des établissements qui ont la « meilleure pratique ». Ces établissements benchmark peuvent être identifiés et pris pour modèles par les autres établissements dans leur objectif d'amélioration.

Par ailleurs, il est parfois utile de rajouter des contraintes afin que le problème d'optimisation ait une solution [60]. Deux contraintes supplémentaires sont rajoutées dans cette étude : chaque poids ne doit pas être inférieur à 10 % du total, ni supérieur à 85 % du total.

Tableau 2 - Exemple de calcul du score « Benefit Of the Doubt »

	Indic 1	Indic 2	Indic 3	Indic 4	Indic 5	Maximisation		pondérée	Score BOD
ES 1	0,77	0,83	0,74	0,92	0,86	→ sous contraintes	Poids spécifiques aux ES	→ des indicateurs	0,90
ES 2	0,98	0,97	1,00	0,98	0,93				1
ES 3	0,90	0,63	0,63	0,85	0,17				0,79
…									

Unobserved Component Model (UCM)

La méthode UCM est utilisée par la Banque Mondiale pour évaluer la gouvernance de 212 pays dans le monde à partir de 6 dimensions de la gouvernance [26].

La construction d'un indicateur composite de qualité par la méthode UCM repose sur le principe selon lequel la qualité est une grandeur hypothétique non mesurable directement mais présente de façon sous-jacente dans les mesures des indicateurs initiaux. Ce point de départ range le modèle UCM dans la famille des modèles à erreur sur les variables qui repose dans ce cas précis sur deux hypothèses :

- chaque variable observée peut être écrite sous la forme de la somme d'une variable latente et d'un terme d'erreur. Ceci revient à considérer que chaque indicateur mesure un aspect particulier d'un concept sous-jacent, identifié ici à la qualité globale de la prise en charge de l'IDM ;
- la liaison entre chaque indicateur et l'indicateur composite est linéaire.

Le modèle UCM se présente sous la forme d'un ensemble d'équations structurelles dans lesquelles interviennent des variables observées (les indicateurs), une variable latente (l'indicateur composite) et un ensemble de coefficients fixes à estimer. L'expression générale du modèle est :

$$y_{j,e} = \alpha_j + \beta_j \times \left(g_e + \varepsilon_{j,e} \right)$$

Où : $y_{j,e}$ est la mesure de l'indicateur j pour l'établissement de santé e ; g_e est une variable latente – estimée par l'indicateur composite – que l'on cherche à mesurer ; α_j et β_j sont des paramètres à estimer et $\varepsilon_{j,e}$ un résidu de moyenne nulle et d'écart-type σ_j (à estimer) permettant de capter deux sources d'incertitude : a) l'incertitude liée à la mesure de chaque indicateur, b) l'incertitude liée au caractère imparfait de la liaison entre chaque indicateur et le phénomène à mesurer.

L'estimation de la quantité g_e est donnée par la moyenne de la distribution conditionnelle des scores initiaux standardisés par les coefficients α_j et β_j :

$$IC_e = \mathrm{E}\left[g_e / y_{1,e}, .., y_{J,e} \right] = \sum_{j=1}^{J} w_j * \tilde{y}_{j,e}$$

Avec les poids :

$$\forall j = 1, ..., J, \; w_j = \frac{1/\sigma_j^2}{1 + \sum_{j=1}^{J} 1/\sigma_j^2}$$

Et les indicateurs standardisés :

$$\forall j \in \{1, ..., J\}, \forall e \in \{1, ..., E\}, \; \tilde{y}_{j,e} = \frac{y_{j,e} - \alpha_j}{\beta_j}$$

La variance de cette distribution conditionnelle :

$$\mathrm{V}\left[g_e / y_{e,1}, ..., y_{e,J(e)} \right] = \left(1 + \sum_{j=1}^{J(e)} \sigma_j^{-2} \right)^{-1},$$

peut être utilisée comme mesure de la précision de l'indicateur composite.

Plus de précisions sur cette méthode se trouvent en annexe III.

Tableau 3 - Exemple de calcul du score « Unobserved Component Model »

	Indic 1	Indic 2	Indic 3	Indic 4	Indic 5				Score UCM
ES 1	0,77	0,83	0,74	0,92	0,86	**Estimation**	$\tilde{y}_j$	**. pondérée**	-0,31
ES 2	0,98	0,97	1,00	0,98	0,93				1,42
ES 3	0,90	0,63	0,63	0,85	0,17	des paramètres	w_j	des indicateurs	-1,68
…									

Budget Allocation Process (BAP)

La méthode BAP a été utilisée par la Commission européenne pour construire l'E-Business Readiness Index (2003), un indicateur composite dont le but est d'évaluer l'équipement ou l'usage des TIC (Technologies de l'Information et de la Communication) de 27 pays européens [66].

Dans notre étude, il a été demandé à treize cardiologues[3] proposés par le Conseil National Professionnel de Cardiologie de répartir un « budget » de 100 points entre les cinq indicateurs, en se basant sur leur expérience et leur jugement de l'importance relative des indicateurs, importance qui peut être jugée « sur un plan clinique ou sur le sens donné au résultat ».

Les poids de chaque indicateur sont calculés en faisant la moyenne des budgets alloués.

All-or-None

All-or-None est une méthode qui permet de calculer un score composite d'abord au niveau du patient, puis au niveau de l'établissement [67].

Aux États-Unis, l'AHRQ utilise dans son rapport national un indicateur composite All-or-None pour évaluer la prise en charge des patients diabétiques [68]. CMS exploite également cette méthode pour calculer un indicateur composite (l'Appropriate Care Measure) qui est le résultat de l'agrégation de 10 indicateurs dont 5 sur la prise en charge de l'infarctus du myocarde, 2 sur celle de l'insuffisance cardiaque et 3 sur celle de la pneumonie. Il est utilisé dans un de ses programmes de recherche [69].

Le score d'un patient est 1 si, pour tous les indicateurs, le processus a été rempli ; il est de 0 si au moins une des conditions n'est pas remplie. Le score au niveau de l'établissement est égal à la moyenne de ces scores. Un exemple simple est le suivant :

Tableau 4 - Exemple de calcul du score « All-or-None »

	Indic 1	Indic 2	Indic 3	Indic 4	Indic 5	Score "All-or-None" patient
Patient 1	1	0	1	1	1	0
Patient 2	0	.	1	0	.	0
Patient 3	1	1	1	1	1	1
Patient 4	1	1	1	.	1	1
Patient 5	1	1	0	0	0	0
…						

[3] Dont la liste se trouve en annexe IV.

Le score All-or-None au niveau de l'établissement se calcule ainsi :

$$Score\_All-or-None = moyenne\left(scores\ des\ patients\right) = \frac{2}{5} = 0,4$$

Patient Average

La méthode Patient Average a été utilisée afin de réaliser une étude longitudinale, sur cinq ans, de la qualité des soins de l'asthme, du diabète et des maladies cardiaques, en Angleterre [70].

Le score d'un patient est le ratio du nombre de processus remplis (scores 1) sur le nombre de processus applicables (que le patient aurait dû recevoir), c'est-à-dire la moyenne de ses réponses. Le score de l'établissement est égal à la moyenne des scores des patients.

Tableau 5 - Exemple de calcul du score « Patient Average »

	Indic 1	Indic 2	Indic 3	Indic 4	Indic 5	Score "Patient Average" patient
Patient 1	1	0	1	1	1	0,8
Patient 2	0	.	1	0	.	0,33
Patient 3	1	1	1	1	1	1
Patient 4	1	1	1	.	1	1
Patient 5	1	1	0	0	0	0,4
…						

Le score Patient Average au niveau de l'établissement se calcule ainsi :

$$Score\_Patient\ Average = moyenne\left(scores\ des\ patients\right) = \frac{3,53}{5} = 0,71$$

Il faut remarquer que s'il n'y a pas de donnée manquante ou non applicable, c'est-à-dire si tous les patients sont évalués sur tous les indicateurs disponibles, alors la méthode Patient Average est équivalente à la méthode Indicator Average.

Denominator-Based Weight (DBW)

Il s'agit de la méthode utilisée par la JCAHO [12], l'AHRQ [11] et CMS [10] dans le contexte présenté au paragraphe 0.

Le score est calculé directement au niveau de l'établissement. Il correspond au ratio de la somme des processus remplis pour tous les patients de l'établissement sur la somme des processus applicables pour tous les patients de l'établissement.

Tableau 6 - Exemple de calcul du score « DBW »

	Indic 1	Indic 2	Indic 3	Indic 4	Indic 5	Numérateur DBW	Dénominateur DBW
Patient 1	1	0	1	1	1	4	5
Patient 2	0	.	1	0	.	1	3
Patient 3	1	1	1	1	1	5	5
Patient 4	1	1	1	.	1	4	4
Patient 5	1	1	0	0	0	2	5
…							
						16	22

Le score DBW se calcule ainsi :

$$Score\_DBW = \frac{somme\ des\ numérateurs}{somme\ des\ dénominateurs} = \frac{16}{22} = 0,73$$

Par ailleurs, de la même manière que pour Patient Average, si tous les patients sont évalués sur tous les indicateurs disponibles, c'est-à-dire s'il n'y a pas de donnée manquante ou non applicable, alors la méthode Denominator-Based Weight est équivalente à la méthode Indicator Average.

Au final, pour cette étude, cinq méthodes d'agrégation référencées et adaptées au jeu de données seront comparées : Indicator Average, Benefit Of the Doubt (BOD), Unobserved Component Model (UCM), Budget Allocation Process (BAP) et All-or-None.

Prise en compte de l'incertitude dans le score composite

Une fois la méthode d'agrégation choisie, il est indispensable de prendre en compte l'incertitude liée aux indicateurs (rappelons que ces derniers sont calculés à partir d'un échantillon de 30 à 60 dossiers patients) et en conséquence d'accompagner les scores composites d'intervalles de confiance. Il existe plusieurs méthodes pour les calculer, l'une d'entre elles est le Bootstrap [71-73]. Plus de détails sur la méthode se trouvent en annexe V.

Des intervalles de confiance à 95 % sont générés à l'aide de cette méthode autour des scores composites pour chaque établissement de santé.

Classement des établissements et impact des différents scores composites

Une fois que l'indicateur composite et son intervalle de confiance ont été calculés pour chaque établissement, les résultats peuvent être présentés sous la forme d'un tableau ou d'un graphique.

Mode de classement

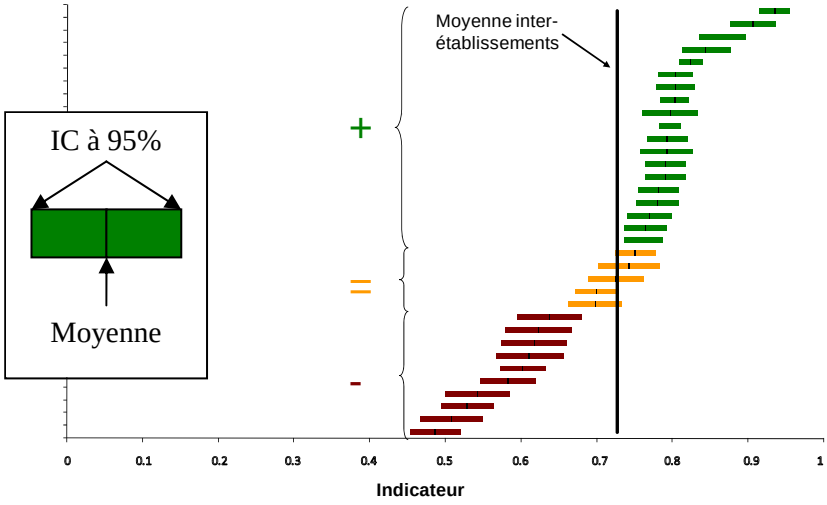
Lorsqu'il existe des fluctuations autour du résultat d'un indicateur dues à un échantillonnage (des dossiers patients dans notre cas), le rang des établissements comme méthode de classement n'est pas fiable. Ce constat reste vrai dans le cadre des indicateurs composites et est même renforcé puisque la littérature montre que même lorsque l'on dispose de données exhaustives, il existe une incertitude autour du score composite et du rang qui en résulte [1]. Une alternative au classement des établissements par leur rang est leur répartition dans des classes [74], qui est préférable dans ce contexte.

Les indicateurs de processus généralisés par la HAS (en particulier, les indicateurs portant sur la prise en charge de l'infarctus du myocarde après la phase aiguë) sont calculés à partir de 60 dossiers de patients tirés au sort dans chaque établissement. Les établissements ne sont pas comparés selon leurs rangs respectifs, mais selon leurs classes. La HAS propose un classement des établissements en trois classes à partir des intervalles de confiance à 95 % [8, 75] :

- Si la borne inférieure de l'intervalle de confiance à 95 % est supérieure à la moyenne inter-établissements, l'établissement est dans la classe +
- Si la borne supérieure de l'intervalle de confiance à 95 % est inférieure à la moyenne inter-établissements, l'établissement est dans la classe –
- Sinon, l'établissement est dans la classe =

Ces classes sont représentées de manière schématique dans la figure 7.

Figure 7 - Méthode de classement des établissements en trois classes



Les établissements peuvent être classés dans les trois classes (+, = et –) en fonction de leur score composite et de son intervalle de confiance de la même manière qu'ils le sont pour les indicateurs individuels.

Evaluation de la concordance entre les méthodes d'agrégation

Le degré d'accord du classement (en catégories) des établissements par les méthodes deux à deux peut être estimé par le coefficient kappa, introduit par Cohen en 1960 [76].

L'accord ou la « concordance » entre un ou plusieurs jugements résulte de la somme d'une composante « aléatoire » (hasard) et d'une composante d'accord « véritable » (réel). Le coefficient kappa propose de chiffrer l'intensité ou la qualité de l'accord réel. C'est un indice qui permet de « retirer » la portion de hasard ou de subjectivité de l'accord entre les techniques.

Il est donné par :

$$\kappa = \frac{\%\ \text{concordance observée} - \%\ \text{concordance due au hasard}}{1 - \%\ \text{concordance due au hasard}}$$

Le coefficient Kappa vaut 1 quand la concordance observée est totale ; et 0 quand la concordance observée est égale à la concordance due au hasard [77, 78].

Le degré d'accord est catégorisé par Landis et Koch [79] en six classes allant de l'accord « très mauvais » à l'accord « excellent » :

Tableau 7 - Caractérisation du kappa

Accord	Coefficient Kappa
Excellent	$\kappa \geq 0,81$
Bon	$0,61 \leq \kappa \leq 0,80$
Moyen	$0,41 \leq \kappa \leq 0,60$
Médiocre	$0,21 \leq \kappa \leq 0,40$
Mauvais	$0 \leq \kappa \leq 0,20$
Très Mauvais	$\kappa < 0$

Bien que cette catégorisation soit arbitraire, elle fournit malgré tout des repères utiles.

Dans cette étude, le coefficient « kappa pondéré » [80], introduit par Cohen en 1968, sera utilisé. Il présente l'avantage de considérer qu'un désaccord de deux classes entre deux méthodes est plus grave qu'un désaccord d'une classe. En d'autres termes, si une méthode classe un établissement dans la classe + et une autre dans la classe −, le coefficient kappa sera plus faible que si la seconde méthode classait l'établissement dans la classe =.
Plus d'informations sur le coefficient Kappa se trouvent en annexe VI.

Résultats

Après une analyse descriptive des indicateurs pris individuellement, les résultats des cinq méthodes d'agrégation : Indicator Average, Benefit Of the Doubt (BOD), Unobserved Component Model (UCM), Budget Allocation Process (BAP) et All-or-None, sont présentés.

Analyse descriptive

Le tableau suivant présente les résultats relatifs aux 5 indicateurs entrant dans le processus d'agrégation.

Tableau 8 - Statistiques descriptives des indicateurs

Indicateurs	N	Moyenne	Ecart-type	Min	Max
Prescription d'aspirine et de clopidogrel	56	0.91	0.12	0.2	1
Prescription de bêtabloquant	56	0.85	0.12	0.47	1
Mesure de la FEVG	56	0.88	0.12	0.45	1
Prescription de statine	56	0.89	0.07	0.73	1
Sensibilisation aux règles hygiéno-diététiques	56	0.33	0.27	0	0.83

Les dossiers des patients évalués mentionnaient qu'en moyenne, 91 % des patients ont reçu de l'aspirine, 85 % des bêtabloquants, 88 % ont fait l'objet d'une mesure de la Fraction d'Ejection Ventriculaire Gauche, 89 % d'une prescription de statine et enfin 33 % des patients ont été sensibilisés aux règles hygiéno-diététiques.

Tableau 9 - Corrélations entre les indicateurs

	Aspirine	Bêtabloquant	FEVG	statine	Règles hygiéno-diététiques
Aspirine	1	0.01	0.07	0.11	0.06
Bêtabloquant		1	0.29	0.18	-0.04
FEVG			1	0.29	0.28
Statine				1	0.25
Règles hygiéno-diététiques					1

Dans cette étude, les indicateurs ne présentent pas de corrélation significative.

Comparaison des méthodes d'agrégation

Les résultats de la comparaison des méthodes sont détaillés sur le jeu de données de l'expérimentation (56 ES), puis sont présentés sur le jeu de données de la généralisation (275 ES).

Pondérations

Le tableau suivant présente les poids donnés aux indicateurs par les méthodes de pondération BAP, Indicator Average, BOD et UCM. Il est à noter que comme la méthode All-or-None fournit un indicateur composite au niveau du patient puis au niveau de l'établissement (pas de poids calculé par indicateur), la comparaison des poids avec les autres méthodes ne peut être réalisée.

Tableau 10 - Poids des indicateurs

	Budget Allocation Process (BAP)	Indicator Average	Benefit Of the Doubt* (BOD)	Unobserved Component Model (UCM)**
Prescription d'Aspirine	0,24	0,2	0,38	0,01
Prescription de Betabloquant	0,21	0,2	0,19	0,08
Mesure de la FEVG	0,19	0,2	0,2	0,66
Prescription de Statine	0,22	0,2	0,13	0,15
Sensibilisation aux règles hygiéno-diététiques	0,13	0,2	0,11	0,1

* : les poids BOD sont spécifiques à chaque établissement. Les poids du tableau correspondent à la moyenne des pondérations BOD des 56 établissements.
** : pour ce tableau, les poids issus de la méthode UCM ont été ramenés à une somme de 1, pour assurer la comparabilité avec les autres méthodes.

Ici, les poids donnés par les experts (méthode BAP) sont proches des poids Indicator Average, sauf pour le poids accordé à l'indicateur « Sensibilisation aux règles hygiéno-diététiques » par la méthode BAP, qui est sensiblement plus faible que celui donné par la méthode Indicator Average. Par rapport à la méthode Indicator Average, l'approche BOD alloue un poids plus important à l'indicateur « Prescription d'Aspirine », au détriment des indicateurs « Prescription de Statine » et « Sensibilisation aux règles hygiéno-diététiques ». En ce qui concerne la méthode UCM, un poids très élevé est accordé à l'indicateur « Mesure de la FEVG ».

Scores composites et classements

L'annexe VII présente les scores et les classes des indicateurs individuels et composites pour chacun des 56 établissements de santé de l'échantillon. La répartition des ES dans les trois classes selon les méthodes d'agrégation est exposée dans le tableau 11.

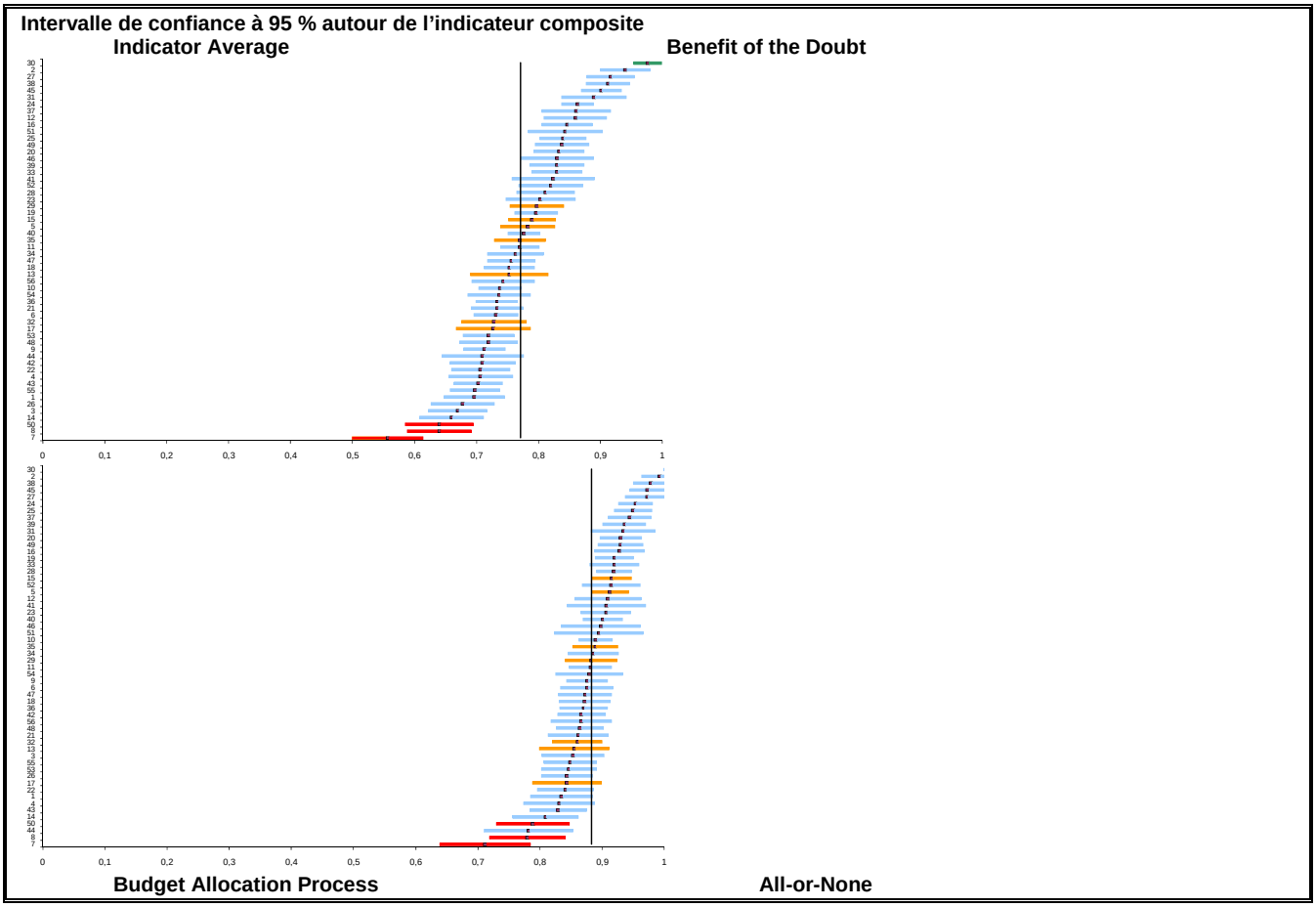
Tableau 11 - Répartition des ES dans les trois classes

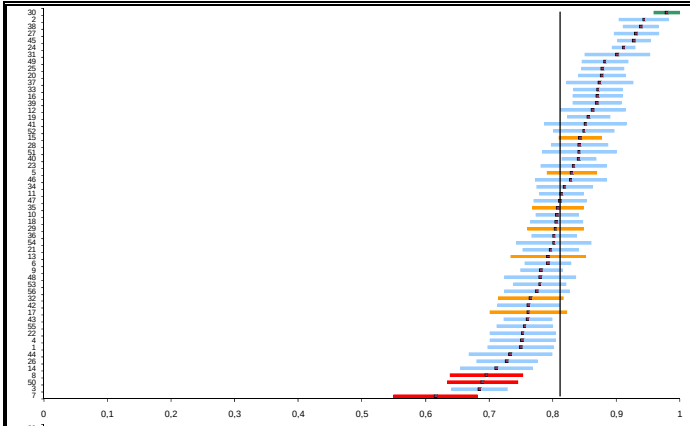
Classe	Indicator Average	BOD	BAP	All-or-None	UCM
+	29%	25%	29%	27%	2%
=	43%	59%	50%	27%	91%
–	29%	16%	21%	46%	7%

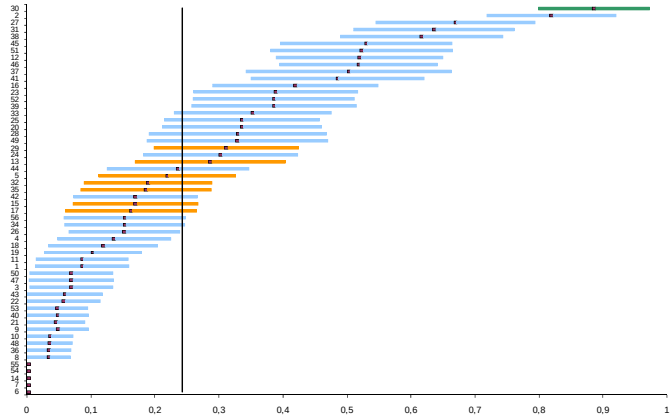
La méthode Indicator Average produit la répartition la plus égale dans les trois classes. À l'inverse, la méthode UCM classe 91 % des établissements dans la classe =. La pondération All-or-None classe 46 % des établissements dans la classe –.

Les scores composites accompagnés de leurs intervalles de confiance à 95 %, sont représentés graphiquement dans la figure 8.

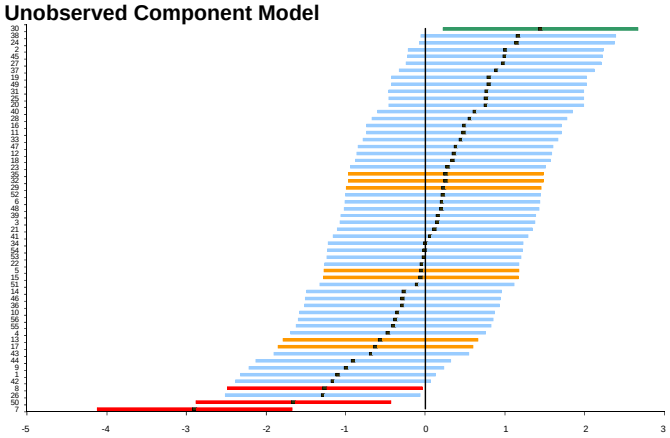
Figure 8 - Graphiques des scores composites avec intervalle de confiance à 95 % selon les 5 méthodes de pondération







Unobserved Component Model



Les ES anonymisés se trouvent en ordonnée, et les scores composites en abscisse
11 ES sont classés dans la même classe quelle que soit la méthode, dont :
1 ES reste dans la classe + (en vert)
7 ES restent dans la classe = (en orange)
3 ES restent dans la classe – (en rouge)

Les scores composites moyens calculés par les différentes méthodes d'agrégation sont donnés dans le tableau 12. Étant donné qu'il n'y a aucune donnée non applicable, les méthodes Patient Average, DBW et Indicator Average sont strictement équivalentes. Seule la dernière sera présentée. Les statistiques usuelles des scores composites sont les suivantes :

Tableau 12 - Scores composites moyens

	Budget Allocation Process (BAP)	Indicator Average	Benefit Of the Doubt (BOD)	All-or-None	UCM*
Moyenne des Scores composites	0,81	0,77	0,88	0,24	0
Ecart-type	0,07	0,08	0,06	0,22	0,79
Minimum	0,61	0,55	0,71	0	-2,92
Maximum	0,98	0,97	1,00	0,88	1,42
Médiane	0,81	0,76	0,88	0,17	0,13

* Les scores issus de la méthode UCM sont standardisés. Les résultats ne sont pas comparables aux autres.

Concordance et impact sur le classement selon les méthodes

Le premier résultat a trait à la méthode UCM. La concordance (tableau 19 de l'annexe VIII) entre cette méthode et les autres est comprise entre 0,14 (avec All-or-None) et 0,24 (avec BAP et avec BOD), et est donc caractérisée de « mauvaise » à « médiocre ». Par ailleurs, 36 % des établissements classés dans une classe par UCM sont classés dans une autre classe par BOD, et ce chiffre monte jusqu'à 64 % entre UCM et All-or-None (tableau 20 de l'annexe VIII).

Par ailleurs, en considérant les cinq méthodes, seuls 11 établissements sur 56 ne changent pas de classe quelle que soit la méthode de pondération choisie (en couleur dans le tableau 18, annexe VII). Ceci peut s'expliquer par le caractère sélectif de la méthode UCM : 91 % des ES sont classés par cette méthode dans la classe = (tableau 11).

Dans cet ensemble de mesures, la méthode UCM occupe une position particulière : les poids alloués aux indicateurs sont atypiques, sa concordance avec les autres méthodes est faible et elle est très sélective puisqu'elle classe une très grande partie des établissements dans la classe moyenne. Les résultats de la méthode UCM influent fortement sur la comparaison des classements inter-établissements.

Le caractère singulier des résultats de la méthode UCM justifie une approche comparative portant sur les résultats des autres méthodes.

Concordance et impact sur le classement sans UCM

Les graphiques des indicateurs composites accompagnés de leurs intervalles de confiance à 95 % pour les quatre méthodes de pondération Indicator Average, Benefit of the Doubt, Budget Allocation Process et All-or-None se trouvent en annexe IX.

La concordance entre les méthodes deux à deux est présentée dans le tableau 19 en annexe VIII. Les résultats varient de l'accord « excellent » entre les méthodes Indicator Average et BAP (kappa=0,84) à « médiocre » entre BOD et All-or-None (kappa=0,34).

Le tableau 20 montre les établissements qui changent de classe par méthodes (deux à deux) : changement d'une classe (classe +/= ou classe =/−) ou changement de deux classes (classe +/−). Deux établissements sont classés par All-or-None dans la classe − et par BAP dans la classe +. Les méthodes BAP et BOD sont celles pour lesquelles le plus petit nombre d'établissements (8) change de classe selon la méthode, ce qui confirme le résultat précédent de leur forte concordance. Par ailleurs, 54 % des établissements sont classés différemment selon que le score composite est calculé par la méthode BOD ou la méthode All-or-None.

Seuls 23 établissements sur 56 (41 %) sont classés dans la même classe quelle que soit la méthode (en couleur dans le tableau 18), dont 8 dans la classe +, 7 dans la classe = et 8 dans la classe −.

Par ailleurs, il est à noter que sur les 23 établissements classés au moins une fois +, 15 obtiennent un classement différent avec une autre méthode ; respectivement, sur les 28 établissements classés au moins une fois −, 20 obtiennent un classement différent. Il y a donc une sensibilité du classement à la méthode utilisée.

Résultats sur les données de la généralisation

Calcul du score composite sur les données de la généralisation

Les mêmes analyses ont été réalisées sur les données de la généralisation 2008. 275 établissements ont pu évaluer les indicateurs sur 30 dossiers ou plus, ce qui représente 14 966 dossiers analysés.

Les résultats sont similaires à ceux obtenus sur l'échantillon : 79 établissements (soit 29 %) restent stables sur les 5 méthodes, et 143 (52 %) sur 4 méthodes (sans UCM).

Tableau 13 - Comparaison des résultats de l'expérimentation et de la généralisation

	Nombre d'ES qui sont dans la même classe quelle que soit la méthode parmi les 5 testées	**Nombre d'ES qui sont dans la même classe quelle que soit la méthode parmi les 4 testées (sans UCM)**
Sur les données de la généralisation (275 ES)	79 (29 %)	143 (52 %)
Sur les données de l'expérimentation (56 ES)	11 (20 %)*	23 (41 %)*

* Test du khi-deux non significatif dans les 2 cas.

La répartition des établissements dans les 3 classes est équivalente même si les résultats sont généralement mieux dispersés sur les données de la généralisation.

Tableau 14 : Répartition des établissements dans les trois classes

Classe	**Indicator Average**	**BOD**	**BAP**	**All-or-None**	**UCM**
+	36%	39%	37%	29%	5%
=	41%	50%	47%	32%	83%
−	23%	11%	17%	40%	12%

UCM reste la méthode la moins discriminante mais néanmoins 17 % des établissements sont classés + ou −.

Les tableaux de concordance et de changement de classes sont présentés en annexe XIII.

Calcul du score composite incluant tous les indicateurs (et niveaux) disponibles

La généralisation porte actuellement sur six indicateurs de qualité dont deux à 2 niveaux. Leur agrégation en un seul indicateur composite peut être réalisée lorsqu'elle se fait à partir des résultats des patients. Trois méthodes sont donc utilisées : Patient Average, All-or-None et Denominator-Based Weight. Les résultats des trois méthodes sur ces six indicateurs (dont 2 à 2 niveaux) sont calculés à partir de l'échantillon des 56 ES de l'expérimentation.

Le tableau 21 de l'annexe X présente les scores composites et les classes des établissements en fonction des trois méthodes de pondération. Les scores composites accompagnés de leurs intervalles de confiance, sont représentés graphiquement en annexe XI. La distribution des établissements dans les trois classes se trouve dans le tableau 15 :

Tableau 15 : Répartition des établissements dans les trois classes

	All-or-None	Patient Average	Denominator-Based Weight
+	13%	32%	32%
=	61%	27%	29%
−	27%	41%	39%

Le tableau 22 de l'annexe XII présente les coefficients kappa pour évaluer la concordance entre les méthodes d'agrégation. L'accord est excellent entre les méthodes Patient Average et Denominator-Based Weight (0,98), et moyen entre ces deux méthodes et All-or-None (respectivement 0,50 et 0,52).

Le tableau 23 de l'annexe XII montre les établissements qui changent de classe par méthode (deux à deux) : changement d'une classe (classe +/= ou classe =/−) ou changement de deux classes (classe +/−). Seul un établissement n'est pas classé dans la même classe selon Patient Average et DBW. 41 % des établissements sont classés dans une classe différente selon DBW et All-or-None.

33 établissements sont classés dans la même classe quelle que soit la méthode (en couleur dans le tableau 21), dont 7 dans la classe +, 13 dans la classe = et 13 dans la classe −.

Cependant, un certain nombre de **limites** sont associées à la prise en compte des six indicateurs (dont 2 à 2 niveaux). Tout d'abord, trois méthodes seulement sont compatibles, ce qui élimine des méthodes présentant des avantages. De plus, cela confère à certaines dimensions de la prise en charge, où il y a deux niveaux, un poids plus important dans le score final. Enfin, certains établissements pourraient être désavantagés du fait de leur case-mix (exemple : prescription d'IEC).

Discussion

Synthèse des résultats et limites

Une synthèse des résultats montre les points suivants :

- L'**étude des pondérations** des indicateurs montre que les trois méthodes « Budget Allocation Process », « Indicator Average » et « Benefit Of the Doubt » ont des résultats similaires. Ces trois méthodes donnent toutes un poids moyen aux indicateurs « Prescription de bétabloquant » et « Mesure de la FEVG ». Il est intéressant de noter que les deux méthodes BOD et BAP allouent un poids moins important à l'indicateur « Sensibilisation aux règles hygiéno-diététiques », ce qui signifie que d'une part, les experts ont considéré cet indicateur comme moins important que les autres, et que d'autre part, les établissements ont globalement des résultats moins bons sur cet indicateur que sur les autres, c'est-à-dire, selon la logique de la méthode BOD, que les établissements considèrent également cet indicateur comme moins important que les autres. Un raisonnement similaire peut être appliqué à l'indicateur « Prescription d'aspirine » auquel un poids plus important est alloué par BOD et BAP (dans une moindre mesure), et qui est donc considéré par les professionnels et par les établissements comme l'indicateur le plus important. Cependant, la méthode BAP accorde à l'indicateur « Prescription de statine » un poids moyen, ce qui signifie que les experts considèrent cet indicateur comme aussi important que les autres. La méthode BOD, quant à elle, lui alloue un poids faible malgré les très bons scores des établissements sur cet indicateur, ceux-ci étant en général un peu meilleurs sur un ou deux autres indicateurs (« Prescription d'aspirine » et « Prescription de bétabloquant »). La méthode UCM est très divergente puisqu'elle donne un poids nul à l'indicateur « Prescription d'aspirine », contrairement à toutes les autres qui lui accordent le plus grand poids, et un poids très élevé à l'indicateur « Mesure de la FEVG ».

- La méthode UCM mise à part, la **concordance** la plus faible entre les méthodes deux à deux est observée pour les méthodes All-or-None et Benefit Of the Doubt, avec un coefficient kappa caractérisé comme « médiocre ». Ceci peut s'expliquer par la grande différence de logique entre ces deux méthodes. All-or-None produit un score composite élevé pour les établissements qui ont le plus de patients pour lesquels tous les indicateurs ont été remplis et recherche donc l'excellence, alors que la pondération BOD valorise les établissements et leur accorde le bénéfice du doute en allouant un poids plus important aux indicateurs pour lesquels ils sont plus performants. À l'inverse, les méthodes BAP et BOD d'une part, BAP et Indicator Average d'autre part présentent une excellente concordance, ce qui se comprend par l'observation des poids alloués aux indicateurs, voisins dans les deux cas. Les scores composites sont donc proches, et les classes des établissements également.

En utilisant 5 méthodes pour agréger 5 indicateurs relatifs à la prise en charge de l'infarctus du myocarde après la phase aiguë, les classements des établissements qui en résultent changent. Ce résultat est important car il souligne que le recours à une méthode d'agrégation peut être source d'injustice pour un établissement de santé dans le cadre d'un classement inter-établissements, dans la mesure où il peut se retrouver mieux classé avec d'autres méthodes.

Ces résultats doivent être considérés **en tenant compte de certaines limites** : les résultats et conclusions présentés sont spécifiques à ce jeu d'indicateurs, et seraient certainement

différents avec un autre jeu de données. De plus, 6 indicateurs dont 2 à 2 niveaux sont généralisés. Or pour la construction des scores composites, une sélection a été effectuée : 5 indicateurs sont conservés, et aucun des 2 niveaux ; cette sélection est justifiée dans le cadre de ce rapport. Enfin, il faut souligner qu'en ce qui concerne des indicateurs de type « conformité à de bonnes pratiques cliniques » comme ceux étudiés, le système de poids idéal pourrait être déterminé en fonction de l'impact de chaque indicateur sur la mortalité ou l'espérance de vie. Cependant, ces données ne sont pas disponibles.

Malgré ces limites, nous pouvons tirer une tendance générale. La littérature fait état de plusieurs études qui ont cherché à comparer des méthodes d'agrégation avec pour conclusions que le classement est sensible à la méthode utilisée [1-5]. Notre étude, basée sur un jeu d'indicateurs relatif à une même prise en charge, et en tenant compte de l'incertitude autour du score composite (classement en catégories), confirme ces résultats.

Recommandations

Etant donné ce résultat, l'absence d'argument statistique en faveur d'une méthode, et l'absence d'équivalence entre ces mêmes méthodes, **deux orientations peuvent être prises**.

La première est de considérer que les réserves quant à l'emploi d'une méthode d'agrégation sont suffisantes pour ne pas développer de scores composites. Cette approche présente l'avantage de prendre en considération les limites statistiques qui ont été exprimées dans la littérature ces dernières années.

La seconde est de considérer qu'un score composite est malgré tout nécessaire, mais qu'il doit être accompagné d'une présentation explicite des réserves précédentes. Cette approche présente l'intérêt de considérer un score composite dans ce cas précis alors même que dans d'autres cadres différentes initiatives l'ont déjà réalisé (économique avec la Banque mondiale), ou le préconisent (rapport « Stiglitz » sur la performance économique et le progrès social). Par ailleurs, le développement des raisonnements par filière et parcours de soins va conduire à re-questionner le besoin de score composite. A cet égard, certains auteurs recommandent de développer des scores composites dans le cas de la prise en charge générale de l'infarctus du myocarde [55, 81], notamment pour affirmer des évaluations plus intégrées de la qualité ou, « care bundles ».

Dans ce cas, plusieurs éléments devraient être pris en considération :

- Dans la plupart des programmes internationaux qui présentent des scores composites, les établissements sont classés selon leur rang respectif (leur score moyen). Cependant la littérature et notre étude montrent la nécessité de prendre en compte l'incertitude liée au résultat, quelle que soit la qualité et l'exhaustivité des données [1]. Il nous parait donc important de se borner à des résultats par classe.
- Devant ce constat, certains auteurs proposent d'utiliser les résultats des différentes méthodes [2]. Des conventions pourraient être établies telles que par exemple : considérer seulement les résultats stables toute méthode confondue ou affecter à l'établissement le classement qui lui est le plus favorable. Certains auteurs préconisent aussi la transparence sur le choix de la méthode qui a été fait [1]. Dans ce cas, une

explicitation des logiques sous-jacentes aux méthodes peut constituer une aide à la décision [2, 3].

Critères de choix des méthodes d'agrégation

- L'indicateur composite calculé par la méthode **Indicator Average,** moyenne arithmétique des scores obtenus pour chaque indicateur, est une mesure de la qualité moyenne des établissements de santé. Son utilisation est largement répandue dans de nombreux domaines, ce qui s'explique probablement par la simplicité de sa mise en œuvre et par sa transparence. Aisément compréhensible par le public, cette méthode repose cependant sur une hypothèse qui n'est pas anodine : tous les indicateurs ont la même importance dans le score composite. Dès lors, il faut s'assurer avec cette approche que cette hypothèse reflète bien la réalité (force des recommandations, redondance entre les indicateurs, égale fiabilité des données).

- Les méthodes **Patient Average** et **Denominator-Based Weight** sont proches de la méthode Indicator Average : elles considèrent que tous les indicateurs ont la même importance dans l'indicateur composite, mais en partant, dans le mode de calcul, des scores obtenus au niveau du patient.

- L'approche **Benefit Of the Doubt** se base sur les données observées pour attribuer un « jeu de poids » à chaque établissement : le benchmark est composé des meilleurs établissements, et non défini de manière théorique. Il s'agit donc d'une performance relative (sur distribution observée), par opposition à une performance absolue (référence à dires d'experts). Ce type d'approche peut représenter une cible plus réaliste pour les établissements. De plus, avec cette méthode, au moins un établissement recevra systématiquement un score maximum de 1. Elle valorise les établissements en leur allouant des poids spécifiques et élevés pour les indicateurs pour lesquels ils sont les plus performants. Avec tout autre système de pondération, le score composite de l'établissement serait plus bas que celui qu'il obtient avec la méthode BOD. L'idée centrale est qu'une bonne performance relative d'un établissement sur un indicateur indique que cet établissement considère cette dimension particulière comme importante, et que cet indicateur mérite donc un poids considérable. Comme son nom l'indique, la méthode accorde le bénéfice du doute aux établissements de santé. Elle nécessite cependant la plupart du temps la mise en place de contraintes sur les poids qu'il faut définir, afin de ne pas obtenir trop d'établissements avec un score maximum.

- La prise en compte de l'avis d'experts, dans la méthode **Budget Allocation Process**, confère une légitimité professionnelle à l'indicateur composite. Ses autres avantages sont la simplicité et la transparence. Dans cette étude, elle donne des poids proches de ceux obtenus par la méthode Indicator Average (sauf pour l'indicateur « Sensibilisation aux règles hygiéno-diététiques », dont le poids est plus faible), mais ceci pourrait être différent avec un autre jeu d'indicateurs.

- Comme la méthode Indicator Average, la méthode **All-or-None** confère la même importance à chaque indicateur individuel, mais le calcul est différent. La méthode All-or-None estime que si une seule étape du processus n'est pas respectée, c'est toute la prise en charge du patient qui est considérée comme un échec : le score composite représente le pourcentage de patients pour lesquels toutes les étapes du processus de soin ont été

respectées. La méthode récompense l'excellence et met en valeur la coordination des soins. Etant donné que cette approche est drastique, il est indispensable que les indicateurs utilisés soient fiables et bien conçus. Nolan & Berwick [67] considèrent que la méthode All-or-None reflète l'intérêt et les désirs des patients : ceux-ci préfèrent, a priori, recevoir un soin complet plutôt que partiel. Cependant, les scores composites qui en résultent sont très faibles comparés aux autres méthodes (moyenne = 0.24), ce qui pourrait, dans le cadre d'une diffusion publique des résultats, être interprété négativement.

- Les scores composites obtenus par la méthode **Unobserved Component Model** se basent aussi sur les données observées pour calculer les poids. Les scores composites sont standardisés, et perdent ainsi leur lisibilité, ce qui peut rendre difficile leur compréhension. Contrairement aux autres méthodes, UCM part du principe que la qualité est une grandeur hypothétique non directement mesurable mais contenue dans chacun des indicateurs mesurés. Cette méthode est actuellement utilisée par la Banque Mondiale pour évaluer et comparer la gouvernance des pays pour laquelle ils obtiennent des résultats discriminants. Cependant, les indicateurs qu'ils utilisent sont fortement corrélés, ce qui n'est pas le cas des indicateurs relatifs à la prise en charge de l'infarctus du myocarde après la phase aiguë ; cela pourrait expliquer la largeur des intervalles de confiance et donc pourquoi, avec notre méthode de classement, une majorité des établissements (plus de 80 %) sont classés « = ».

En résumé, les méthodes Indicator Average, Patient Average, DBW et All-or-None ont comme point commun de donner la même importance à tous les indicateurs, c'est-à-dire aux recommandations sous-jacentes à ces indicateurs. Alors que la méthode BAP hiérarchise les recommandations par l'intermédiaire de l'avis d'experts. La méthode BOD quant à elle valorise le respect des recommandations (les résultats), en attribuant le plus grand poids là où l'établissement est le plus performant. Enfin, la méthode UCM s'affranchit de ces considérations en calculant les poids des indicateurs à l'aide d'un modèle mathématique.

Conclusion

L'étude sur l'agrégation a permis tout d'abord de mettre en évidence le fait que les méthodes ne sont pas équivalentes puisqu'elles ont une influence sur le classement des établissements de santé. Ces résultats confirment les connaissances sur le sujet, en soulignant les réserves scientifiques relatives à l'emploi d'un score agrégé. Elles démontrent que même dans le cas d'un concept homogène, « la prise en charge de l'infarctus du myocarde », et en tenant compte de l'incertitude, la sensibilité du classement inter-établissements à la méthode utilisée existe.

Deux options ont été discutées sur la base de ces résultats. L'une consiste à ne pas considérer de score composite, l'autre vise à développer ce score moyennant différentes précautions : expliciter les critères de choix entre les différentes méthodes (de ce fait une identification des logiques sous-jacentes à chaque méthode a été réalisée) et affirmer les réserves relatives à l'emploi de ces méthodes, ou établir des conventions à partir des résultats de toutes les méthodes afin de minimiser le risque de pénaliser injustement un ES dans son classement,. C'est cette dernière option que nous privilégions à des fins d'aide à la décision publique en tenant compte du contexte de demande sociale accrue.

Par ailleurs, la communication de la méthode et l'expression du résultat constituent un sujet à part entière. Pour certaines méthodes, le score est moins facilement compréhensible (UCM) et pour d'autres il peut apparaître peu encourageant (All-or-None), alors que ces méthodes présentent des avantages certains et permettent un classement des établissements. Ces questions sont relatives aux cibles visées dans l'emploi de ces scores composites. Par exemple, la compréhension de la méthode n'est peut être pas pour le « grand public » un critère essentiel, alors qu'il l'est probablement plus pour les professionnels.

De plus, si la décision de développer un score composite s'avérait retenue, l'adhésion des professionnels apparaît comme un élément incontournable, surtout lorsque les thèmes abordés reposent sur des recommandations de bonne pratique clinique (par exemple qualité de la prise en charge de l'accident vasculaire cérébral). Il apparaît également plus judicieux d'intégrer préalablement à la construction des indicateurs de qualité l'objectif de définir un indicateur composite scientifiquement et médicalement pertinent permettant la diffusion publique d'un score agrégé.

Enfin, il nous semble opportun d'envisager des perspectives de recherche, afin de porter un regard critique sur les initiatives engagées et d'étudier les éventuelles avancées dans le domaine. Cinq opérations peuvent être envisagées : i) l'agrégation d'autres indicateurs cliniques, notamment dans le cadre des filières de soins (ou dans l'esprit de ce que les anglo-saxons nomment les « care bundles ») ; ii) le lien entre chacun de ces scores composites et un indicateur de résultat tel que le taux de mortalité post-infarctus [82] ; iii) une meilleure connaissance de la méthode UCM ; iv) une analyse des changements de classes au cours du temps avec deux années consécutives de données sur la méthode retenue ; v) une agrégation plus générale au niveau de l'établissement.

Glossaire

ACHS : Australian Council on Healthcare Standards
AHRQ : Agency for Healthcare Research and Quality
BAP : Budget Allocation Process
BOD : Benefit of the Doubt
BQS : Bundesgeschäftsstelle Qualitätssicherung
CCORT : Canadian Cardiovascular Outcomes Research Team
CERMES : CEntre de Recherche MEdecine, Sciences, Santé et Société
CMS : Centers for Medicare & Medicaid Services
CNRS : Centre National de la Recherche Scientifique
COMPAQH : COordination pour la Mesure de la Performance et l´Amélioration de la Qualité Hospitalière
DBW : Denominator-Based Weights
DEA : Data Envelopment Analysis
DGS : Direction Générale de la Santé
DHOS : Direction de l´Hospitalisation et de l´Organisation des Soins
DREES : Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques
ES : Etablissement de Santé
ESI : Environmental Sustainability Index
FEVG : Fraction d'Éjection du Ventricule Gauche
HAS : Haute Autorité de Santé
HDI : Human Development Index
HEDIS : Healthcare Effectiveness Data and Information Set
IDM : Infarctus Du Myocarde
ICALIN : Indice Composite des Activités de Lutte contre les Infections Nosocomiales
ICATB : Indice Composite de bon usage des AnTiBiotiques
ICSHA : Indicateur de Consommation de Solutions Hydro-Alcooliques
IEC : Inhibiteur de l'Enzyme de Conversion de l'angiotensine
INSERM : Institut National de la Santé Et de la Recherche Médicale
ISUP : Institut de Statistiques de l'Université Paris 6
JCAHO : Joint Commission on the Accreditation of Healthcare Organizations
JRC : Joint Research Center
MCO : Médecine-Chirurgie-Obstétrique
NCQA : National Committee for Quality Assurance
NHS : National Health Service
OCDE : Organisation de Coopération et de Développement Economiques
OMS : Organisation Mondiale de la Santé
PIB : Produit Intérieur Brut
PSPH : établissement Privé participant au Service Public Hospitalier
SFC : Société Française de Cardiologie
SIP : Sickness Impact Profile
SURVISO : SURVveillance des Infections du Site Opératoire
TAI : Technology Achievement Index
UCM : Unobserved Component Model

Bibliographie

1. **Jacobs R *et al.*** How robust are hospital ranks based on composite performance measures? *Med Care*, 2005 ; 43(12):1177-84.
2. **Reeves D *et al.*** Combining multiple indicators of clinical quality: : an evaluation of different analytic approaches. *Med Care*, 2007 ; 45(6):489-96.
3. **O'Brien SM *et al.*** Exploring the behavior of hospital composite performance measures : an example from coronary artery bypass surgery. *Circulation*, 2007 ; 116(25):2969-75.
4. **Shwartz M *et al.*** Estimating a composite measure of hospital quality from the Hospital Compare database : differences when using a Bayesian hierarchical latent variable model versus denominator-based weights. *Med Care*, 2008; 46(8):778-85.
5. **Hermans E *et al.*** Combining road safety information in a performance index. *Accid Anal Prev*, 2008 ; 40(4):1337-44.
6. **HAS - Haute Autorité de Santé**. Prise en charge de l'infarctus du myocarde. [cited July 2009]; Available from: http://www.has-sante.fr/portail/jcms/c_736856/ensemble-ameliorons-la-prise-en-charge-de-linfarctus-du-myocarde.
7. **F. Delahaye *et al.*** *Recommandations de la société française de cardiologie concernant la prise en charge de l'infarctus du myocarde après la phase aiguë*. 2001.
8. **Corriol C *et al.*** [The COMPAQH project: : researches on quality indicators in hospitals]. *Rev Epidemiol Sante Publique*, 2008; ; 56 Suppl 3: :S179-88.
9. **CCORT - Canadian Cardiovascular Outcome Research Team**. CCORT Quality Indicators. [cited July 2009] ; Available from: http://www.ccort.ca/Research/QualityIndicators/CCORTCCSAMICHFQualityIndicators/tabid/67/Default.aspx.
10. **CMS - Centers for Medicare and Medicaid Services**. Hospital Measures. [cited July 2009] ; Available from: http://www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp.
11. **AHRQ - Agency for Healthcare Research and Quality**. Quality Measures AMI. [cited July 2009] ; Available from: http://www.qualitymeasures.ahrq.gov/search/searchresults.aspx?Type=3&txtSearch=acute+myocardial+infarction&num=20.
12. **JCAHO - Joint Commission on Accreditation of HealthCare Organizations**. Current Specification Manual for National Hospital Quality Measures. [cited July 2009] ; Available from: http://www.jointcommission.org/PerformanceMeasurement/PerformanceMeasurement/Current+NHQM+Manual.htm.
13. **Nardo M *et al.*** *Tools for Composite Indicators Building*. Ispra, Italy, European Commission, 2005.
14. **Saisana M & Tarantola S**. *State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development*. Ispra, Italy, European Commission, 2002.
15. **Nardo M *et al.*** *Handbook On Constructing Composite Indicators: : Methodology And User Guide*. Ispra, Italy, Organisation de Coopération et de Développement Economiques, Commission Européenne (JRC), 2008.
16. **Freudenberg M**. *Composite Indicators of Country Performance: : A Critical Assessment*. OECD Science, Technology and Industry Working Papers, OECD Publishing, 2003.

17. **Gadrey J & Jany-Catrice F**. Les nouveaux indicateurs de richesse. La Découverte. 2005.
18. **Bandura R**. *Measuring Country Performance and State Behavior : A Survey of Composite Indices*. New York, United Nations Development Programme, 2005.
19. **Bergner M *et al.*** The Sickness Impact Profile : development and final revision of a health status measure. *Med Care*, 1981 ; 19(8):787-805.
20. **Bucquet D & Condon S**. *Adaptation en français du Nottingham health profile et caractéristiques opératoires de la version française*. Villejuif, FRANCE, INSERM U164, 1988.
21. **Leplègue A *et al.*** Le questionnaire MOS SF-36. ESTEM. 1999.
22. **Streiner D & Norman G**. Health Measurement Scales, A practical guide to their development and use. Oxford Medical Publications. 1995.
23. **JRC**. An information server on composite indicators (methodologies and case studies). [cited January 2009] ; Available from: http://composite-indicators.jrc.ec.europa.eu/.
24. **United Nations**. Human Development Report. [cited August 2009] ; Available from: : www.undp.org.
25. **Fukuda-Parr S *et al.*** Readings in Human Development. Oxford University Press. Oxford ; 2003.
26. **Kaufmann D *et al.*** *Governance Matters VIII : Aggregate and Individual Governance Indicators for 1996-2008*. The World Bank, 2009.
27. **Smith P**. Developing Composite Indicators for Assessing Health System Efficiency. In: : OECD, editor. *Measuring Up : Improving Health System Performance in OECD Countries*. Paris : OECD ; 2002. p. 251-75.
28. **Jacobs R *et al.*** *Are composite measures a robust reflection of performance in the public sector?* York, United Kingdom, Centre of Health Economics, 2006.
29. *Report of the Scientific Peer Review Group on Health Systems Performance Assessment*. 2002.
30. **Organisation Mondiale de la Santé**. *Rapport sur la santé dans le monde 2000 : pour un système de santé plus performant*. 2000.
31. **Evans DB *et al.*** The Comparative Efficiency of National Health systems in producing Health : An Analysis of 191 Countries. [GPE Discussion Paper Series : No. 29], 2000.
32. **Healthcare Commission**. *The annual Health check 2008/09 assessing and rating the NHS*. National Health Service, Commission for Healthcare Audit and Inspection 2008, 2008.
33. **NCQA - National Committee for Quality Assurance**. Comparing and assessing health-plans. [cited January 2010] ; Available from : www.ncqa.org.
34. **CMS - Centers for Medicare and Medicaid Services**. Hospitalcompare. [cited January 2010] ; Available from: www.hospitalcompare.hhs.gov.
35. **JCAHO - Joint Commission on Accreditation of HealthCare Organizations**. Quality check. [cited January 2010] ; Available from: www.qualitycheck.org.
36. **Healthgrades**. Healthgrades's web site. [cited January 2010] ; Available from: www.healthgrades.com.
37. **DrFoster**. Dr Foster's web site. [cited Available from: www.drfosterhealth.co.uk.
38. **Ministère de la Santé**. Tableau de bord des Infections Nosocomiales. [cited January 2010] ; Available from: http://www.icalin.sante.gouv.fr/.
39. **Projet COMPAQH**. *Indicateurs du tableau de bord des infections nosocomiales*. Paris, France, Institut National de la Santé et de la Recherche Médicale, 2005.

40. **Ministère de la Santé de la Jeunesse et des Sports**. *Modalités de calcul et de classement du score agrégé du tableau de bord des infections nosocomiales*. 2008.
41. **Le Point**. Hôpitaux - Le palmarès 2009. *Le Point*, 2009:188-226.
42. **Falga P**. Palmarès des hôpitaux : la méthodologie de l'Express. L'EXPRESS. 2009. 01/21/2009.
43. **Zemouri A**. Le classement référence : Cliniques, le palmarès 2009. Le Figaro. 2009. 06/20/2009: p.48-68.
44. **Qafli M & Presles P**. Hôpitaux cliniques, le classement national 2010. Le Nouvel Observateur. 2009. 11/26/2009: p.75-100.
45. **Stiglitz J *et al.*** *Rapport de la Commission sur la mesure des performances économiques et du progrès social*. 2009.
46. **ACHS - Australian Council on Healthcare Standards**. ACHS indicators. [cited November 2009] ; Available from: http://www.achs.org.au.
47. **BQS**. Federal office for quality assurance. [cited January 2010] ; Available from: http://www.bqs-outcome.de/.
48. **NCQA - National Committee for Quality Assurance**. HEDIS & Quality Measurement. [cited January 2010]; ; Available from: http://www.ncqa.org/tabid/59/Default.aspx.
49. **CMS - Centers for Medicare and Medicaid Services**. Composite Indicator. [cited July 2009] ; Available from: http://www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp#TopOfPage.
50. **Landon BE *et al.*** Quality of care for the treatment of acute medical conditions in US hospitals. *Arch Intern Med*, 2006 ; 166(22):2511-7.
51. **Werner RM & Bradlow ET**. Relationship between Medicare's hospital compare performance measures and mortality rates. *JAMA*, 2006 ; 296(22):2694-702.
52. **Jha AK *et al.*** Care in U.S. hospitals--the Hospital Quality Alliance program. *N Engl J Med*, 2005 ; 353(3):265-74.
53. **Lindenauer PK *et al.*** Public reporting and pay for performance in hospital quality improvement. *N Engl J Med*, 2007 ; 356(5):486-96.
54. **Glickman SW *et al.*** Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA*, 2007 ; 297(21):2373-80.
55. **Weston CF**. Performance indicators in acute myocardial infarction : a proposal for the future assessment of good quality care. *Heart*, 2008 ; 94(11):1397-401.
56. **Jacobs R *et al.*** *Measuring performance : An examination of composite performance indicators*. 2004.
57. **Jacobs R *et al.*** Composite performance measures in the public sector. 2007.
58. **World Economic Forum**. Environmental Sustainability Index. [cited August 2009] ; Available from: http://sedac.ciesin.columbia.edu/es/esi/.
59. **Cherchye L *et al.*** *Creating Composite Indicators with DEA and Robustness Analysis: the case of the Technology Achievment Index*. Center for Economic Studies, European University College, European Commission, 2006.
60. **Cherchye L *et al.*** *'Benefit of the doubt' composite indicators*. Belgium, Center for Economic Studies, European University College, 2006.
61. **Jacobs R**. Alternative methods to examine hospital efficiency: data envelopment analysis and stochastic frontier analysis. *Health Care Manag Sci*, 2001; 4(2):103-15.
62. **Mahlberg B & Obersteiner M**. *Remeasuring the HDI by Data Envelopment Analysis*. Laxenburg, Austria, International Institute for Applied Systems Analysis, 2001.

63. **Despotis DK**. Measuring human development via data envelopment analysis: : the case of Asia and the Pacific. *Omega*, 2004.
64. **Akazili J *et al.*** Using data envelopment analysis to measure the extent of technical efficiency of public health centres in Ghana. *BMC Int Health Hum Rights*, 2008 ; 8:11.
65. **Masiye F**. Investigating health system performance: an application of data envelopment analysis to Zambian hospitals. *BMC Health Serv Res*, 2007 ; 7:58.
66. **W. Castaings & S. Tarantola**. *The 2007 European e-Business Readiness Index*. Joint Research Centre of the European Commission, 2008.
67. **Nolan T & Berwick DM**. All-or-none measurement raises the bar on performance. *JAMA*, 2006; ; 295(10):1168-70.
68. **AHRQ - Agency for Healthcare Research and Quality**. National Healthcare Quality Report. [cited November 2009]; Available from: http://www.ahrq.gov/qual/nhqr08/Chap1.htm.
69. **CMS - Centers for Medicare and Medicaid Services**. All-or-None Composite Score. [cited December 2009]; ; Available from: http://www.cms.hhs.gov/apps/QMIS/measure_details.asp?id=292.
70. **Campbell SM *et al.*** Improvements in quality of clinical care in English general practice 1998-2003: longitudinal observational study. *BMJ*, 2005; 331(7525):1121.
71. **Palm R**. Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres. *Biotechnol Agron Soc Environ*, 2002; 6(3):143–53.
72. **Efron B & Tibshirani R**. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1986; ; 1(1):54-75.
73. **Efron B *et al.*** Le Bootstrap et ses Applications. Centre International de Statistique et d'Informatique Appliquées CISIA ed.; 1995.
74. **Bird S *et al.*** Performance indicators: : good, bad, and ugly. *J R Statist Soc A*, 2005; ; 168(1):1-27.
75. **Hospital Report Research Collaborative**. Hospital e-Scorecard Report 2008: Acute Care. [cited July 2009]; Available from: http://www.hospitalreport.ca/downloads/2008/AC/2008_AC_cuo_techreport.pdf.
76. **Cohen J**. AÀ coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960; 20:37-46.
77. **Falissard B**. Mesurer la subjectivité en santé. MASSON. 2001.
78. **Bergeri I *et al.*** Pour tout savoir ou presque sur le coefficient kappa... *Médecine tropicale*, 2002; 62(6):634-6.
79. **Landis JR & Koch GG**. The measurement of observer agreement for categorical data. *Biometrics*, 1977; 33(1):159-74.
80. **Cohen J**. Weighted kappa: : nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*, 1968; 70(4):213-20.
81. **Peterson ED *et al.*** Association between hospital process performance and outcomes among patients with acute coronary syndromes. *JAMA*, 2006; 295(16):1912-20.
82. **Glickman SW *et al.*** Alternative Pay-for-Performance Scoring Methods: Implications for Quality Improvement and Patient Outcomes. *Med Care*, 2009; 47(10):1062-8.
83. **Charnes A *et al.*** Measuring the efficiency of decision making units. *European Journal of Operational Research*, 1978; 2:429-44.
84. **Centre for European Labour Market Studies**. *Benchmarking European Labour Market Performance with Efficiency Frontier Techniques*. Göteborg, Centre for European Labour Market Studies, 2000.

85. **Cherchye L *et al.*** *Legitimately diverse, yet comparable: On synthesising social inclusion performance in the EU*. Public economics, Center for Economic Studies, 2003.
86. **Nayar P & Ozcan Y**. Data Envelopment Analysis Comparison of Hospital Efficiency and Quality. *Journal of Medical Systems*, 2008; 32(3):193-9.
87. **European Commission**. *The EU Economy: 2004 Review*. Bruxelles, European Economy, 2004.

Annexes

Annexe I : Tableau des indicateurs

Tableau 16 - Description des indicateurs

Description	Numérateur Nombre de séjours pour lesquels le patient…	Dénominateur
IND1 (Aspirine/Clopidogrel)		
Prescription d'aspirine et de clopidogrel	À fait l'objet d'une prescription d'aspirine et de clopidogrel à l'issue du séjour (ordonnance de sortie), en l'absence de contre-indication à l'aspirine et au clopidogrel ; OU À fait l'objet d'une prescription d'aspirine (ordonnance de sortie) et présente une contre-indication au clopidogrel ; OU À fait l'objet d'une prescription de clopidogrel (ordonnance de sortie) et présente une contre-indication à l'aspirine ; OU Présente une contre-indication à l'aspirine et au clopidogrel	Nombre total de séjours inclus d'IDM (Infarctus du Myocarde).
IND2 (Betabloquant)		
Prescription d'un bêtabloquant	À fait l'objet d'une prescription de bêta-bloquants à l'issue du séjour (ordonnance de sortie) en l'absence de contre-indication ; OU Présente au moins une contre-indication relative aux bêta-bloquants ET dont le dossier contient une trace d'une discussion bénéfice/risque justifiant la prescription ou non de bêta-bloquants ; OU Présente au moins une contre-indication absolue aux bêta-bloquants	Nombre total de séjours inclus d'IDM.
IND3n1 (FEVG1)		
Mesure de la Fraction d'Ejection du Ventricule Gauche (FEVG)	Nombre de séjours de patients ayant fait l'objet d'une mesure de la FEVG	Nombre total de séjours inclus d'IDM.
IND3n2 (FEVG2)		
Prescription de l'inhibiteur de l'enzyme de conversion de l'angiotensine (IEC)	Nombre de séjours de patients dont la FEVG est ≤ 40 % ayant fait l'objet d'une prescription d'IEC à sa sortie (ordonnance de sortie), en l'absence de contre-indication	Nombre total de séjours inclus d'IDM avec une FEVG ≤ 40 %.
IND4n1 (Statine1)		
Prescription de statine	Nombre de séjours pour lesquels le patient a fait l'objet d'une prescription de statine à l'issue du séjour (ordonnance de sortie), sauf si : le patient est déjà sous un autre hypolipémiant avant l'IDM ; OU présente un taux de LDL-cholestérol inférieur à 1g/l ; OU présente une contre-indication aux statine	Nombre total de séjours inclus d'IDM.

IND4n2 (Statine2)		
Prescription pour un bilan lipidique	Nombre de séjours pour lesquels le patient a fait l'objet d'une prescription de statine à l'issue du séjour (ordonnance de sortie) ET d'une prescription pour la réalisation d'un bilan lipidique à distance (entre J30 et J90 après le début du traitement)	Nombre de séjours inclus d'IDM avec prescription de statine
IND5 (Règles Hygiéno-diététiques)		
Sensibilisation aux Règles Hygiéno-diététiques	Nombre de dossiers pour lesquels le patient a fait l'objet d'une sensibilisation aux règles hygiéno-diététiques (consultation diététicienne et/ou nutritionniste, et/ou participation à un atelier d'éducation thérapeutique, et/ou conseils consignés dans le dossier ou le courrier de sortie)	Nombre total de séjours inclus d'IDM.
IND6 (Tabac)		
Délivrance de conseils pour l'arrêt du tabac	Nombre de dossiers pour lesquels le patient a reçu des conseils pour l'arrêt du tabac, dont une trace est trouvée dans le dossier ou dans le courrier de sortie, ou par une consultation spécifique	Nombre total de séjours inclus d'IDM dont le patient est FUMEUR

Annexe II : Benefit of the doubt

La méthode DEA a été introduite par Charnes, Cooper et Rhodes en 1978 [83] et a depuis été surtout utilisée en économétrie, bien que ses applications puissent concerner tous les domaines. La question originelle était de mesurer l'efficacité relative d'unités (à l'origine, des industries), étant données des observations sur les quantités en entrée (input) et en sortie (output = quantité produite).

Au-delà des nombreuses contributions universitaires (par exemple : pour évaluer la politique européenne de lutte contre le chômage [84] ou la politique d'insertion sociale [85] ; ou pour classer les hôpitaux de l'état de Virginie, aux États-Unis [86]), la Commission Européenne elle-même a utilisé la technique de pondération DEA pour évaluer la performance des états membres par rapport aux objectifs de Lisbonne [87].

La méthode « Benefit of the Doubt » est l'application de la méthode DEA au champ des indicateurs composites.

Le plus souvent, l'indicateur composite est défini comme une somme pondérée des indicateurs :

$$IC_e = \sum_{j=1}^{J} w_{e,j}.y_{e,j}$$

où
IC_e = indice composite pour l'établissement e
$y_{e,j}$ = valeur de l'indicateur j pour l'établissement e, j = 1, …, J
w_j = poids associé à l'indicateur j
Dans le cas de l'approche BOD, l'indicateur composite est défini comme le ratio de sa performance actuelle sur une performance « benchmark » :

$$IC_e = \frac{actual\ overall\ \ performance}{benchmark\ overall\ \ performance} = \frac{\sum_{j=1}^{J} w_{e,j}.y_{e,j}}{\sum_{j=1}^{J} w_{e,j}.y_j^B}$$

Le benchmark est défini comme un établissement j qui, utilisant les poids $w_{e,j}$, obtient la somme pondérée maximale. Ce qui peut se traduire par :

$$IC_e = \frac{\sum_{j=1}^{J} w_{e,j}.y_{e,j}}{\max_{y_{g,i} \in \{établissements\}} \sum_{j=1}^{J} w_{e,j}.y_{g,j}}$$

La particularité de BOD est de déterminer un système de poids qui donne au composite la valeur la plus grande possible, sous certaines contraintes, ce système étant spécifique pour chaque établissement :

$$IC_e = \max_{w_{e,j}} \frac{\sum_{j=1}^{J} w_{e,j}.y_{e,j}}{\max_{y_{g,j} \in \{établissements\}} \sum_{j=1}^{J} w_{e,j}.y_{g,j}}$$

avec

$$\begin{cases} 0 \le IC_e \le 1 \; ; e = 1,...,E & \text{(c1)} \\ w_{e,j} \ge 0 \; ; j = 1,...,J & \text{(c2)} \end{cases}$$

Sachant que par construction la somme pondérée maximale vaut 1, l'équation se simplifie et devient :

$$IC_e = \max_{w_{e,j}} \sum_{j=1}^{J} w_{e,j}.y_{e,j}$$

sous les contraintes c1 et c2

Annexe III - Précisions sur la méthode UCM

Le modèle UCM appartient à la famille des modèles à erreur sur les variables et repose dans ce cas précis sur deux hypothèses :

1. Chaque variable observée peut être écrite sous la forme de la somme d'une variable latente et d'un terme d'erreur. Cette hypothèse revient à considérer que chaque indicateur mesure un aspect particulier d'un concept sous-jacent, identifié ici comme la qualité globale de la prise en charge de l'infarctus du myocarde. Le modèle est alors :

$$y_{j,e} = \alpha_j + \beta_j \times \left(g_e + \varepsilon_{j,e} \right)$$

Où : $y_{j,e}$ est la mesure de l'indicateur *j* pour l'établissement de santé *e* ; g_e une variable latente – estimée par l'indicateur composite – qu'on cherche à mesurer ; et $\varepsilon_{j,e}$ un résidu de moyenne nulle permettant de capter deux sources d'incertitude : a) l'incertitude liée à la mesure de chaque indicateur, b) l'incertitude liée au caractère imparfait de la liaison entre chaque indicateur et le phénomène à mesurer.

2. La deuxième hypothèse est que la liaison entre chaque indicateur et l'indicateur composite est linéaire. On a donc :

$$y_{j,e} = \alpha_j + \beta_j \times g_e + \varepsilon_{j,e}$$

Le modèle UCM se présente finalement sous la forme d'un ensemble d'équations structurelles dans lesquelles interviennent des variables observées (les indicateurs), une variable latente (l'indicateur composite) et un ensemble de coefficients fixes à estimer. L'expression générale du modèle est :

$$y_{j,e} = \alpha_j + \beta_j \times \left(g_e + \varepsilon_{j,e} \right)$$

Avec :

$$\mathrm{E}\left[\varepsilon_{j,e} \right] = 0$$

$$\mathrm{E}\left[\varepsilon_{j,e}^2 \right] = \sigma_j^2$$

$$\mathrm{Cov}\left[\varepsilon_{j,e}, \varepsilon_{j',e'} \right] = 0, \text{ pour } j \neq j' \text{ et } e \neq e'$$

Le résidu est donc supposé de moyenne nulle et de variance constante d'un établissement à l'autre au sein de chaque indicateur. D'un indicateur à l'autre les variances du résidu ε sont au contraire inégales. De plus ce terme d'erreur est supposé de covariance nulle entre les indicateurs ce qui signifie que chaque indicateur mesure un aspect particulier de la qualité indépendant des aspects mesurés par les autres.

Enfin, deux hypothèse permettent de simplifier l'estimation de la quantité g_e : a) l'indicateur composite IC_e est supposé être une variable aléatoire de variance 1, b) g_e et $\varepsilon_{j,e}$ sont supposés avoir une distribution gaussienne. Sous ces hypothèses l'indicateur composite (l'estimation de la quantité g_e) est donné par la moyenne de la distribution conditionnelle des scores initiaux standardisés par les coefficients α et β :

$$IC_e = \mathrm{E}\left[g_e / y_{1,e}, .., y_{J,e} \right] = \sum_{j=1}^{J} w_j * \tilde{y}_{j,e}$$

Avec les poids :

$$\forall j = 1,...,J,\ w_j = \frac{1/\sigma_j^2}{1 + \sum_{j=1}^{J} 1/\sigma_j^2}$$

Et les indicateurs standardisés :

$$\forall j \in \{1,...,J\}, \forall e \in \{1,...,E\},\ \tilde{y}_{j,e} = \frac{y_{j,e} - \alpha_j}{\beta_j}$$

La variance de cette distribution conditionnelle est donnée par :

$$\mathrm{V}\left[g_e / y_{e,1},...,y_{e,J(e)} \right] = \left(1 + \sum_{j=1}^{J(e)} \sigma_j^{-2} \right)^{-1}$$

et peut-être utilisée comme mesure de la précision de l'indicateur composite.

En résumé l'estimation de l'indicateur composite par la méthode UCM comporte 3 étapes :

- Étape 1 : estimation des coefficients fixes α_j, β_j, et σ_j^2 par la méthode du maximum de vraisemblance.
- Étape 2 : calcul des indicateurs standardisés et des poids w_j.
- Étape 3 : calcul des quantités IC_e et de la variance $\mathrm{V}\left[g_e / y_{e,1},...,y_{e,J(e)} \right]$.

Annexe IV - Liste des 13 cardiologues qui ont répondu à l'enquête pour la pondération Budget Allocation Process

- Dr Pierre Aubry
- Dr Loic Belle
- Dr Simon Cattan
- Pr Yves Cottin
- Dr Fréderic Fossati
- Pr Martine Gilard
- Dr Michel Hanssen
- Dr Eric Perchicot
- Dr Elisabeth Pouchelon
- Pr Christian Spaulding
- Dr Jean-François Thébaut
- Pr Patrice Virot
- Dr Christian Ziccarelli

Annexe V - Le Bootstrap

Rééchantillonnage

Considérons un échantillon composé de *K* observations $(x_1,...,x_K)$, prélevé de manière aléatoire et simple dans une population. Ces observations peuvent concerner un seul indicateur, ou, au contraire, être relatives à plusieurs indicateurs. Dans ce cas, les x_i représentent des vecteurs de dimension *J*, *J* étant le nombre d'indicateurs. Afin de ne pas alourdir les notations, nous ne distinguerons pas ces deux situations et, de manière plus condensée, nous désignerons l'échantillon initial par le symbole x, qu'il s'agisse d'un vecteur ou d'une matrice.

Le principe de la méthode bootstrap est de prélever, avec remise, une série d'échantillons aléatoires et simples de *K* observations dans l'échantillon initial, considéré comme une population. Ces échantillons successifs sont notés : $(x_1^*,...,x_B^*)$; *B* étant le nombre de rééchantillonnages effectués avec remise. Le but est d'estimer les caractéristiques du phénomène aléatoire qui a engendré ces données.

Pour bien comprendre cette étape de rééchantillonnage, considérons le cas simple où nous disposons d'un établissement de santé dans lequel un échantillon de 5 patients a été sélectionné, pour lesquels nous observons la réponse d'un patient à un seul indicateur, par exemple, le patient a-t-il bien eu une prescription d'aspirine après un infarctus aigu du myocarde ? Les observations forment un vecteur binaire.

Lors du rééchantillonnage, *B* échantillons de taille 5 sont formés. Ils sont composés des observations x_1, x_2, x_3, x_4 et x_5 tirées au hasard avec remise. Par exemple, il est possible d'avoir :

$x_1^* = (x_1, x_3, x_4, x_2, x_4)$

$x_2^* = (x_2, x_3, x_4, x_4, x_5)$

$\vdots$

$x_B^* = (x_1, x_1, x_3, x_5, x_5)$

Au final, *B* échantillons qui « ressemblent » à l'échantillon de départ sont obtenus.

Estimation de l'erreur-standard

Soit θ un paramètre de la population (la moyenne, la médiane…), et soit $\hat{\theta} = f(x_1,...,x_K)$ une estimation de ce paramètre, obtenue à partir des données de l'échantillon initial x.
Chaque échantillon obtenu par rééchantillonnage permet de calculer une répétition bootstrap (*bootstrap replication*) de l'estimation $\hat{\theta}$:

$\hat{\theta}_b^* = f(x_b^*),\ b = 1,...,B$

où la fonction *f* est la même que celle utilisée pour la définition de $\hat{\theta}$.

Disposant des *B* répétitions, la moyenne et l'écart-type des $\hat{\theta}_b^*$ peuvent être déterminés :

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

$$\hat{\sigma}_{\hat{\theta}^*} = \sqrt{\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2 \Big/ \left(B-1\right)}$$

L'écart-type $\hat{\sigma}_{\hat{\theta}^*}$ est une estimation de l'erreur-standard de l'estimateur du paramètre θ. Pour les situations où un estimateur de cette erreur-standard est disponible, et pour autant que les conditions d'application soient remplies, il est possible de montrer que lorsque *B* tend vers l'infini, l'écart-type des $\hat{\theta}_b^*$ tend vers le résultat analytique [71].

Efron et Tibshirani [73] proposent les règles empiriques suivantes pour le choix de *B* :

- un nombre réduit de répétitions (*B* = 25, par exemple) permet d'obtenir une première information et *B* = 50 est généralement suffisant pour avoir une bonne estimation de l'erreur-standard ;
- il est très rare que plus de 200 répétitions soient nécessaires pour estimer une erreur-standard.

Tableau 17 - Algorithme du bootstrap pour estimer l'erreur standard

1. Echantillon initial	**2.** Echantillons Bootstrap de taille *K*	**3.** *B* Réplications Bootstrap de $\hat{\theta}$	**4.** Estimation Bootstrap de l'erreur standard $\hat{\sigma}_{\hat{\theta}^*}$
	x_1^*	$\hat{\theta}_1^* = f\left(x_1^*\right)$	
x_1	x_2^*	$\hat{\theta}_2^* = f\left(x_2^*\right)$	
x_2	⋮	⋮	$\hat{\sigma}_{\hat{\theta}^*} = \sqrt{\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2 \Big/ \left(B-1\right)}$,
⋮	x_b^*	$\hat{\theta}_b^* = f\left(x_b^*\right)$	où $\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$
x_K	⋮	⋮	
	x_B^*	$\hat{\theta}_B^* = f\left(x_B^*\right)$	

Intervalles de Confiance

Méthode de l'erreur standard

Une première méthode consiste à définir l'intervalle de confiance par la méthode de l'erreur-standard (*standard boostrap confidence interval*) :

$$\left[\hat{\theta} - u_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}^*}; \hat{\theta} + u_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}^*}\right]$$

$u_{1-\alpha/2}$ étant le percentile $1-\alpha/2$ de la distribution normale centrée réduite et $1-\alpha$ étant le degré de confiance retenu.

Pour que cette approche soit satisfaisante, il faut que la distribution d'échantillonnage du paramètre étudié soit approximativement normale, que l'estimateur soit non biaisé, et que $\hat{\sigma}_{\hat{\theta}^*}$ soit une bonne estimation de l'erreur-standard de la distribution du paramètre.

La condition de normalité peut être vérifiée à partir de la distribution des $\hat{\theta}_b^*$. Le biais de l'estimateur peut être estimé, mais sa prise en compte risque d'augmenter la variance de l'estimateur. Enfin, la qualité de l'estimation de l'erreur-standard est liée au nombre de répétitions *B* considéré.

200 répétitions sont généralement suffisantes pour cette méthode.

Méthode des percentiles simples

Dans la méthode des percentiles simples (*simple percentile confidence interval*), les limites de confiance sont données par les percentiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique (c'est-à-dire des $\hat{\theta}_b^*$) :

$$\left[\hat{\theta}^*_{[\alpha/2]}; \hat{\theta}^*_{[1-\alpha/2]}\right]$$

Contrairement à la méthode de l'erreur-standard, la distribution d'échantillonnage du paramètre étudié ne doit pas être normale pour que la méthode des percentiles soit applicable. Par contre, le nombre de rééchantillonnages *B* doit être plus élevé que dans le cas de la méthode de l'erreur-standard, car il faut un plus grand nombre d'observations pour estimer un percentile avec une précision suffisante que pour estimer un écart-type. *B* sera par exemple de l'ordre de 1000.

Les tests de normalité sont effectués sur les 1000 estimations des indicateurs composites et ne sont pas rejetés. Les intervalles de confiance sont donc calculés avec la méthode de l'erreur standard pour toutes les méthodes.

Annexe VI - Le coefficient Kappa

Prenons l'exemple de l'accord entre 2 méthodes, en fonction du classement des établissements dans les classes +, = et –.

		Méthode 2			
		+	=	–	Total
Méthode 1	+	14 (a)	1 (b)	1 (c)	**16 (M)**
	=	2 (d)	10 (e)	5 (f)	**17 (N)**
	–	0 (g)	3 (h)	20 (i)	**23 (O)**
	Total	**16 (J)**	**14 (K)**	**26 (L)**	**56 (P)**

Le coefficient Kappa est défini comme suit :

$$\kappa = \frac{\%\ \text{concordance observée} - \%\ \text{concordance due au hasard}}{1 - \%\ \text{concordance due au hasard}}$$

La concordance observée est : (a + e + i)/P = (14 + 10 + 20)/56 = 79%

Cohen a défini le hasard comme correspondant au produit des marges :
[(J × M) + (K × N) + (L × O)] / P² = [(16 × 16) + (14 × 17) + (26 × 23)] / 56² = 35%

Dans cet exemple le Kappa vaut donc :
κ = (0.88 – 0.35) / (1-0.35) = 0.67
E
n pratique, le coefficient kappa peut servir à déterminer la fiabilité d'un instrument de mesure, l'accord inter-observateur ou inter-technique. De nombreux exemples existent où le coefficient kappa est utilisé pour évaluer le taux d'accord entre deux (ou plusieurs) observateurs, comme le diagnostic de sclérose (certain, probable, possible ou peu probable) par des neurologues [79].

Cependant, le coefficient kappa peut également être utilisé pour mesurer la concordance entre des techniques, ici entre les différentes méthodes d'agrégation, évaluées deux à deux.

Par ailleurs, un kappa pondéré existe [77, 80] : un poids plus important est accordé à l'accord entre deux méthodes lorsqu'elles ne diffèrent que d'une classe.

La concordance observée est alors :

$[(a + e + i) + w_1*(b+f+d+h) + w_2*(g+c)]/P = [(14 + 10 + 20) + w_1*(1+5+2+3) + w_2*(0+1)]/56$
(où w2<w1<1)

« Cette concordance pondérée traduit le fait que a, e et i correspondent à des situations de plus grande concordance que b, f, d et h, qui correspondent, à leur tour, à des situations de plus grande concordance que g et c » [77]. De la même manière, il est possible de calculer une concordance pondérée « due au hasard », et enfin un coefficient kappa pondéré.

En notant *i* les indices des lignes et *j* les indices des colonnes, les poids sont calculés ainsi :
$w_1 = \sum_i w_{i,j},\ w_2 = \sum_j w_{i,j},$ avec $w_{i,j} = 1 - (i-j)^2 / (3-1)^2$
Les coefficients kappa sont obtenus par la procédure Freq du logiciel ©SAS 9.1.

Annexe VII - Scores et classes des indicateurs individuels et des indicateurs composites (étude sur 5 indicateurs)

Tableau 18 - Scores et classes des indicateurs individuels et des indicateurs composites

	Indicateurs individuels										Indicateurs composites									
	Aspirine/ Clopidogrel		β-bloquant		FEVG		Statine		Règles hygiéno-diététiques		BAP		Indicator Average / Patient Average / DBW		BOD		UCM		All-or-None	
ES anonymisés	Score moyen	Classe	Score moyen	Classe	Score moyen	Classe	Score moyen	Classe	Score moyen	Classe	IC moyen	Classe	IC moyen	Classe	IC moyen	Classe	IC moyen	Classe	IC moyen	Classe
30	0,98	+	0,97	+	1,00	+	0,98	+	0,93	+	**0,98**	**+**	**0,97**	**+**	**1,00**	**+**	**1,42**	**+**	**0,88**	**+**
38	1,00	+	0,96	+	0,98	+	0,98	+	0,61	+	0,94	+	0,91	+	0,98	+	1,14	=	0,61	+
45	1,00	+	0,93	+	0,95	+	1,00	+	0,62	+	0,92	+	0,90	+	0,97	+	0,97	=	0,53	+
2	0,95	=	0,98	+	0,97	+	0,90	=	0,88	+	0,94	+	0,94	+	0,99	+	0,98	=	0,82	+
16	0,97	+	0,83	=	0,93	=	0,90	=	0,58	+	0,87	+	0,84	+	0,93	+	0,46	=	0,42	+
27	0,98	+	0,93	+	0,97	+	0,93	=	0,75	+	0,93	+	0,91	+	0,97	+	0,95	=	0,67	+
37	0,82	=	0,92	=	1,00	+	0,89	=	0,66	+	0,87	+	0,86	+	0,94	+	0,86	=	0,50	+
39	0,98	+	0,98	+	0,85	=	0,90	=	0,42	=	0,87	+	0,83	+	0,93	+	0,14	=	0,38	+
24	1,00	+	0,98	+	1,00	+	1,00	+	0,32	=	0,91	+	0,86	+	0,95	+	1,12	=	0,30	=
49	0,98	+	0,91	=	0,96	+	0,98	+	0,35	=	0,88	+	0,83	+	0,93	+	0,77	=	0,33	=
20	0,98	+	0,93	+	0,98	+	0,92	=	0,33	=	0,87	+	0,83	+	0,93	+	0,73	=	0,33	=
25	0,92	=	1,00	+	0,97	+	0,92	=	0,38	=	0,88	+	0,84	+	0,95	+	0,73	=	0,33	=
12	0,92	=	0,72	−	0,90	=	0,92	=	0,83	+	0,86	+	0,86	+	0,91	=	0,34	=	0,52	+
31	0,93	=	0,88	=	0,95	+	0,92	=	0,75	+	0,90	+	0,89	+	0,93	=	0,74	=	0,63	+
33	0,97	+	0,95	+	0,90	=	0,93	=	0,38	=	0,87	+	0,83	+	0,92	=	0,42	=	0,35	=
19	0,93	=	0,95	+	0,97	+	1,00	+	0,12	−	0,85	+	0,79	=	0,92	+	0,77	=	0,10	−
40	0,98	+	0,92	=	0,98	+	0,93	=	0,05	−	0,84	+	0,77	=	0,90	=	0,59	=	0,03	−
51	0,86	=	0,82	=	0,84	=	0,82	=	0,86	+	0,84	=	0,84	+	0,89	=	-0,13	=	0,52	+
28	0,90	=	0,88	=	1,00	+	0,84	=	0,43	=	0,84	=	0,81	=	0,92	+	0,53	=	0,33	=
23	0,96	+	0,81	=	0,98	+	0,79	=	0,46	=	0,83	=	0,80	=	0,90	=	0,26	=	0,39	+
41	0,95	=	0,90	=	0,86	=	0,86	=	0,53	+	0,85	=	0,82	=	0,90	=	0,03	=	0,48	+
46	0,77	−	0,83	=	0,74	−	0,92	=	0,86	+	0,83	=	0,83	=	0,90	=	-0,31	=	0,52	+
52	0,97	+	0,85	=	0,92	=	0,85	=	0,50	+	0,85	=	0,82	=	0,91	=	0,20	=	0,38	+
5	1,00	+	0,90	=	0,88	=	0,83	=	0,28	=	**0,83**	**=**	**0,78**	**=**	**0,91**	**=**	**-0,08**	**=**	**0,22**	**=**
13	0,92	=	0,90	=	0,80	=	0,80	=	0,33	=	**0,79**	**=**	**0,75**	**=**	**0,85**	**=**	**-0,59**	**=**	**0,28**	**=**
15	0,98	+	0,98	+	0,82	=	0,92	=	0,23	=	**0,84**	**=**	**0,79**	**=**	**0,91**	**=**	**-0,09**	**=**	**0,17**	**=**
17	0,92	=	0,72	=	0,84	=	0,78	=	0,36	=	**0,76**	**=**	**0,72**	**=**	**0,84**	**=**	**-0,65**	**=**	**0,16**	**=**
29	0,91	=	0,47	−	0,94	+	0,93	=	0,72	+	**0,80**	**=**	**0,79**	**=**	**0,88**	**=**	**0,20**	**=**	**0,31**	**=**
	Aspirine/ Clopidogrel		β-bloquant		FEVG		Statine		Règles hygiéno-diététiques		BAP		Indicator Average / Patient Average /		BOD		UCM		All-or-None	

													DBW							
ES anonymisés	**Score moyen**	**Classe**	**Score moyen**	**Classe**	**Score moyen**	**Classe**	**Score moyen**	**Classe**	**Score moyen**	**Classe**	**IC moyen**	**Classe**	**IC moyen**	**Classe**	**IC moyen**	**Classe**	**IC moyen**	**Classe**	**IC moyen**	**Classe**
32	0,80	−	0,68	−	0,97	+	0,92	=	0,27	=	**0,76**	**=**	**0,73**	**=**	**0,86**	**=**	**0,23**	**=**	**0,19**	**=**
35	0,97	+	0,65	−	0,95	+	0,92	=	0,35	=	**0,81**	**=**	**0,77**	**=**	**0,89**	**=**	**0,23**	**=**	**0,18**	**=**
6	0,84	=	0,96	+	0,93	=	0,89	=	0,02	−	0,79	=	0,73	=	0,87	=	0,18	=	0,00	−
10	1,00	+	0,95	+	0,79	=	0,91	=	0,02	−	0,80	=	0,73	=	0,89	=	-0,38	=	0,02	−
11	0,90	=	0,75	=	0,97	+	0,95	+	0,27	=	0,81	=	0,77	=	0,88	=	0,46	=	0,08	−
18	0,95	=	0,73	=	0,95	+	0,95	+	0,17	−	0,80	=	0,75	=	0,87	=	0,32	=	0,12	−
21	0,95	=	0,83	=	0,92	=	0,92	=	0,03	−	0,79	=	0,73	=	0,86	=	0,09	=	0,03	−
34	0,97	+	0,95	+	0,90	=	0,83	=	0,15	−	0,82	=	0,76	=	0,88	=	-0,03	=	0,15	−
36	0,97	+	0,92	=	0,78	=	0,97	+	0,02	−	0,80	=	0,73	=	0,87	=	-0,32	=	0,02	−
47	0,85	=	0,93	+	0,93	=	0,93	=	0,12	−	0,81	=	0,75	=	0,87	=	0,35	=	0,07	−
54	0,97	=	0,97	+	0,93	=	0,80	=	0,00	−	0,80	=	0,73	=	0,88	=	-0,03	=	0,00	−
56	0,95	=	0,55	−	0,82	=	0,95	+	0,43	=	0,77	=	0,74	=	0,86	=	-0,40	=	0,15	−
42	0,98	+	0,73	=	0,63	−	0,95	+	0,23	=	0,76	=	0,71	−	0,86	=	-1,19	=	0,17	=
9	0,95	=	0,98	+	0,63	−	0,95	+	0,03	−	0,78	=	0,71	−	0,87	=	-1,02	=	0,03	−
48	0,92	=	0,82	=	0,98	+	0,85	=	0,02	−	0,78	=	0,72	−	0,86	=	0,17	=	0,02	−
53	0,93	=	0,80	=	0,92	=	0,88	=	0,05	−	0,78	=	0,72	−	0,84	=	-0,05	=	0,03	−
3	0,20	−	0,98	+	0,88	=	0,97	+	0,30	=	0,68	−	0,67	−	0,85	=	0,13	=	0,07	−
26	0,97	+	0,63	−	0,67	−	0,90	=	0,21	−	0,73	−	0,67	−	0,84	=	-1,31	=	0,15	−
55	0,87	=	0,95	+	0,91	=	0,73	−	0,02	−	0,75	−	0,69	−	0,85	=	-0,42	=	0,00	−
44	0,82	=	0,70	−	0,78	=	0,78	=	0,45	=	0,73	−	0,71	−	0,78	−	-0,93	=	0,23	=
1	0,93	=	0,82	=	0,70	−	0,85	=	0,17	−	0,75	−	0,69	−	0,83	−	-1,12	=	0,08	−
4	0,82	=	0,78	=	0,80	=	0,92	=	0,20	−	0,75	−	0,70	−	0,83	−	-0,49	=	0,13	−
8	0,88	=	0,72	−	0,75	−	0,78	=	0,05	−	**0,69**	**−**	**0,64**	**−**	**0,78**	**−**	**-1,29**	**−**	**0,02**	**−**
14	0,68	−	0,88	=	0,93	=	0,78	=	0,00	−	0,71	−	0,66	−	0,81	−	-0,29	=	0,00	−
22	0,83	=	0,75	=	0,95	+	0,83	=	0,15	−	0,75	−	0,70	−	0,84	−	-0,07	=	0,05	−
43	0,92	=	0,88	=	0,80	=	0,82	=	0,08	−	0,76	−	0,70	−	0,83	−	-0,71	=	0,05	−
50	0,90	=	0,63	−	0,63	−	0,85	=	0,17	−	**0,69**	**−**	**0,64**	**−**	**0,79**	**−**	**-1,68**	**−**	**0,07**	**−**
7	0,83	=	0,75	=	0,45	−	0,73	−	0,00	−	**0,61**	**−**	**0,55**	**−**	**0,71**	**−**	**-2,92**	**−**	**0,00**	**−**
Moyenne inter-ES	**0,91**		**0,85**		**0,88**		**0,89**		**0,33**		**0,81**		**0,77**		**0,88**				**0,24**	
Etendue	0,2 − 1		0,47 − 1		0,45 − 1		0,73 − 1		0 − 0,93		0,61 − 0,98		0,55 − 0,97		0,71 − 1				0 − 0,88	

IC = Indicateur composite

Annexe VIII - Comparaison des méthodes de pondération (étude sur 5 indicateurs)

Tableau 19 - Coefficients kappa

	All-or-None	BAP	BOD	Indicator Average	UCM
All-or-None	1	0,44	0,34	0,54	0,14
BAP		1	0,80	0,84	0,24
BOD			1	0,69	0,24
Indicator Average				1	0,21
UCM					1

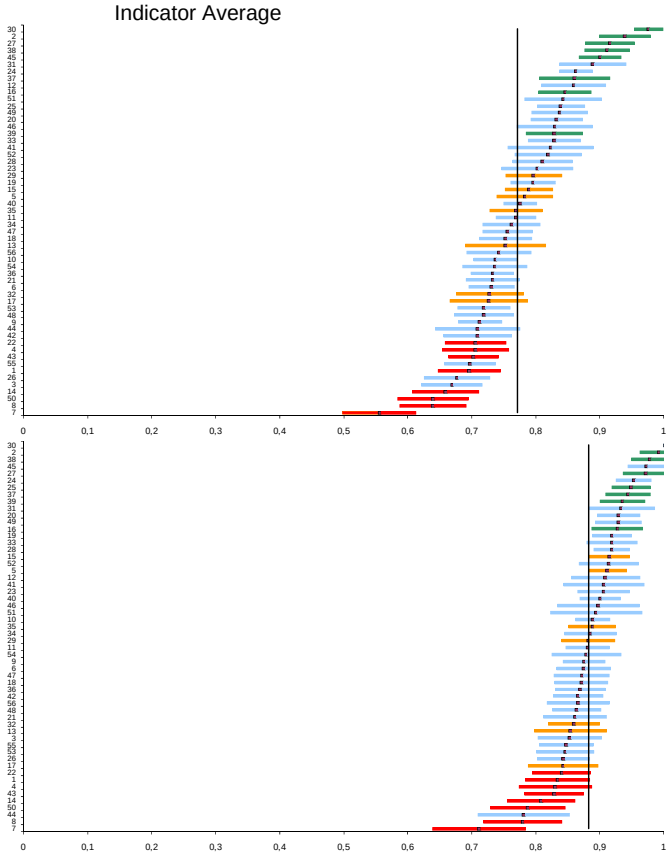
Tableau 20 - Changements de classe

		BOD	Indicator Average	UCM	BAP
All-or-None	1 classe	30 (54 %)	23 (41 %)	36 (64 %)	24 (43 %)
	2 classes	1 (2 %)	0	0	2 (4 %)
BOD	1 classe		13 (23 %)	20 (36 %)	8(14%)
	2 classes		0	0	0
Indicator Average	1 classe			27 (48 %)	7(13%)
	2 classes			0	0
UCM	1 classe				21 (38 %)
	2 classes				0

Annexe IX - Scores composites et intervalles de confiance à 95 % pour quatre méthodes d'agrégation

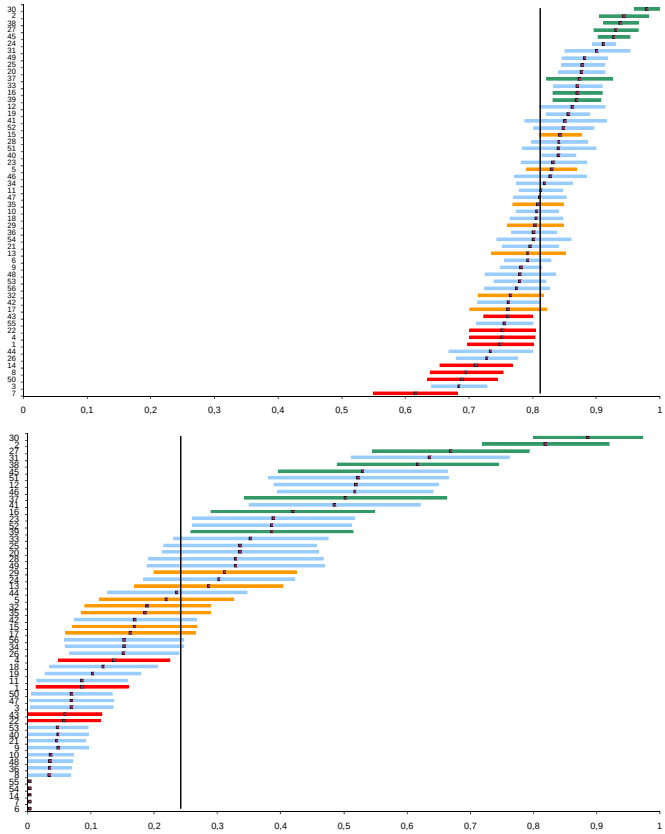
Indicator Average

Benefit Of the Doubt



Budget Allocation Process

All-or-None



Les ES anonymisés se trouvent en ordonnée, et les scores composites en abscisse

23 ES sont classés dans la même classe quelle que soit la méthode, dont :

- 8 ES restent dans la classe + (en vert)
- 7 ES restent dans la classe = (en orange)
- 8 ES restent dans la classe - (en rouge)

Annexe X - Scores et classes des indicateurs composites (étude sur 6 indicateurs dont 2 à 2 niveaux)

Tableau 21 - Scores composites et classes (sur 6 indicateurs dont 2 à 2 niveaux)

	All-or-None		Patient Average		Denominator-Based Weight	
OBS	Score	Classe	Score	Classe	Score	Classe
2	0,8	+	0,94	+	0,94	+
31	0,55	+	0,86	+	0,87	+
37	0,47	+	0,87	+	0,87	+
27	0,43	+	0,85	+	0,86	+
41	0,34	+	0,77	+	0,78	+
30	0,25	+	0,86	+	0,86	+
38	0,21	+	0,8	+	0,8	+
46	0,14	=	0,73	+	0,74	+
16	0,13	=	0,76	+	0,76	+
45	0,11	=	0,78	+	0,78	+
24	0,1	=	0,75	+	0,75	+
51	0,1	=	0,74	+	0,74	+
39	0,1	=	0,73	+	0,74	+
33	0,05	=	0,73	+	0,73	+
25	0	=	0,73	+	0,73	+
12	0	=	0,72	+	0,73	+
20	0	=	0,72	+	0,72	+
49	0	=	0,71	+	0,71	+
29	0,06	=	0,7	=	0,7	=
15	0,05	=	0,68	=	0,68	=
3	0,05	=	0,65	=	0,65	=
23	0,04	=	0,7	=	0,71	=
52	0,03	=	0,7	=	0,71	=
34	0,03	=	0,66	=	0,66	=
56	0,03	=	0,63	=	0,63	=
40	0,02	=	0,66	=	0,66	=
19	0,02	=	0,66	=	0,66	=
28	0	=	0,7	=	0,7	=
13	0	=	0,65	=	0,66	=
5	0	=	0,65	=	0,65	=
17	0	=	0,63	=	0,64	=
18	0,05	=	0,63	−	0,63	−
44	0,03	=	0,61	−	0,61	−
26	0,03	=	0,58	−	0,59	−
35	0,02	=	0,63	=	0,63	−
1	0,02	=	0,61	−	0,61	−
47	0,02	=	0,61	−	0,61	−
43	0,02	=	0,61	−	0,61	−
4	0,02	=	0,6	−	0,61	−
21	0	−	0,66	=	0,67	=
11	0	−	0,64	=	0,64	=
32	0	=	0,61	−	0,62	−
42	0	=	0,58	−	0,59	−
55	0	−	0,62	−	0,62	−

10	0	–	0,61	–	0,62	–
54	0	–	0,61	–	0,62	–
36	0	–	0,61	–	0,61	–

	All-or-None		Patient Average		Denominator-Based Weight	
IDM_1	Score	Classe	Score	Classe	Score	Classe
6	0	–	0,6	–	0,6	–
22	0	–	0,6	–	0,6	–
9	0	–	0,58	–	0,58	–
8	0	–	0,54	–	0,55	–
14	0	–	0,54	–	0,55	–
50	0	–	0,54	–	0,54	–
7	0	–	0,47	–	0,48	–
22	0	–	0,6	–	0,6	–
9	0	–	0,58	–	0,58	–

Annexe XI - Graphiques des scores composites avec intervalles de confiance à 95 % (sur 6 indicateurs dont 2 à 2 niveaux)

Figure 9 - All-or-None

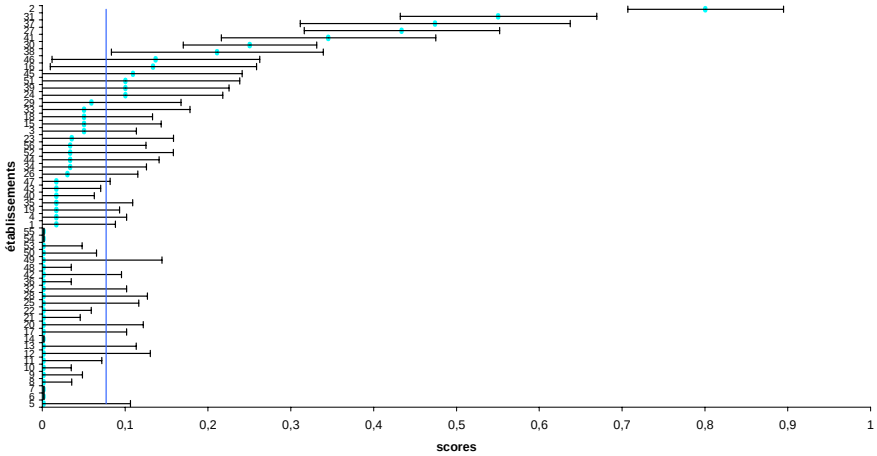


Figure 10 - DBW

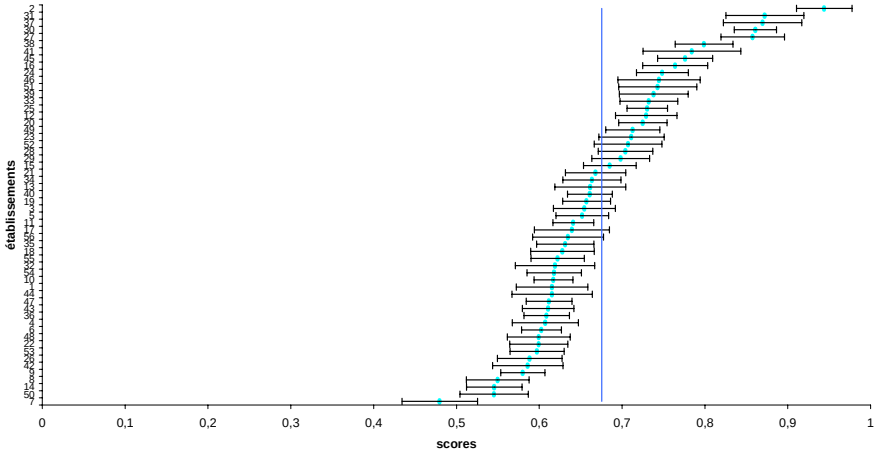
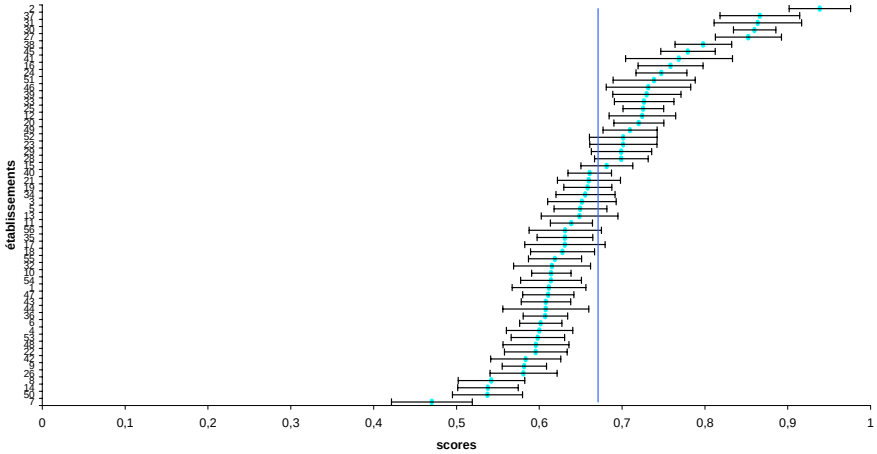


Figure 11 - Patient Average



Annexe XII - Comparaison des méthodes de pondération (étude sur 6 indicateurs dont 2 à 2 niveaux)

Tableau 22 - Coefficients kappa

	All-or-None	Patient Average	Denominator-Based Weight
All-or-None	1	0,50	0,52
Denominator-Based Weight		1	0,98
Patient Average		1	

Tableau 23 - Changements de classe

		Patient Average	Denominator-Based Weight
All-or-None	**1 classe**	22 (39 %)	23 (41 %)
	2 classes	0	0
Patient Average	**1 classe**		1 (2 %)
	2 classes		0

Annexe XIII - Comparaison des méthodes sur les données de la généralisation

Tableau 24 - Coefficients kappa

	All-or-None	BAP	BOD	Indicator Average	UCM
All-or-None	1	0.51	0,45	0.65	0.20
BAP		1	0,84	0.83	0.38
BOD			1	0,73	0,33
Indicator Average				1	0.31
UCM					1

Tableau 25 - Changements de classe

		BOD	Indicator Average	UCM	BAP
All-or-None	1 classe	116 (42 %)	83 (30 %)	160 (58 %)	107 (29 %)
	2 classes	8 (3 %)	1 (0.4 %)	0	6 (2 %)
BOD	1 classe		57 (21 %)	112 (41 %)	31 (11 %)
	2 classes		0	0	0
Indicator Average	1 classe			127 (46 %)	36 (13 %)
	2 classes			0	0
UCM	1 classe				107 (39 %)
	2 classes				0