

Faisabilité de l'évaluation d'une observation médicale par combinaison de multiples critères : méthode NICTALOP

*Feasibility of a multi-criteria medical observation assessment:
the NICTALOP approach*

Laurent BONNEMAINS^{1, 2, 3}, Francois MARCON¹ et Marc BRAUN^{2, 3, 4}

¹ Cardiologie Infantile, Centre Hospitalier Universitaire de Nancy Brabois, Nancy, France

² Bureau de pédagogie, Faculté de médecine, Université de Lorraine, Nancy, France

³ IADI, INSERM U947, Nancy, France

⁴ Neuro-radiologie, Centre Hospitalier Universitaire de Nancy Central, Nancy, France

Manuscrit reçu le 9 septembre 2011 ; commentaires éditoriaux formulés aux auteurs le 28 mai 2012 ; accepté pour publication le 25 février 2013

Mots-clés

Rédaction médicale ;
observation médicale ;
évaluation ;
deuxième cycle
d'études médicales

Résumé – Contexte : La rédaction d'observations ou de comptes-rendus médicaux est un exercice difficile mais d'importance majeure pour garantir la qualité de la prise en charge des patients. La méthode d'évaluation des observations influe probablement sur la qualité de leur rédaction. **But :** Étudier la faisabilité de l'utilisation d'une méthode d'évaluation à multiples critères baptisée NICTALOP pour une évaluation des performances des étudiants de deuxième cycle des études médicales lors d'un exercice de rédaction d'observation médicale. **Méthodes :** 28 étudiants en quatrième, cinquième ou sixième année d'études médicales ont rédigé l'introduction des observations médicales correspondant à deux situations médicales (l'une simple et l'autre complexe), avant et après une formation spécifique de quelques heures. Ces observations ont été évaluées selon cinq critères qui définissent l'acronyme NICTALOP : Nombre d'Idées-clés, Choix des Termes, verAcité des concepts, LOngueur et Position des idées. Un score global a été calculé combinant ces critères. **Résultats :** A l'évaluation initiale, les étudiants en fin d'études médicales obtiennent de meilleurs scores pour l'histoire simple ($p = 0,02$) mais il n'y a pas de différence pour l'histoire complexe ($p = 0,4$). Les étudiants qui assistent à l'ensemble de la formation améliorent significativement leurs scores ($p = 0,01$). Cette progression fait disparaître les différences de niveaux concernant l'histoire simple ($p = 0,5$) et crée une différence concernant l'histoire complexe, mieux restituée par les étudiants expérimentés ($p = 0,05$). **Conclusion :** La méthode NICTALOP permet une évaluation fine de la qualité des observations médicales qui met en évidence les écarts entre étudiants liés à l'expérience ou aux enseignements reçus.

Keywords

Written medical communication; medical observations; evaluation; undergraduate medical student

Abstract – Context: Written inter-physician communication, i.e. the writing of observations and medical reports, is a difficult task that directly affects the quality of patient management. Yet it is overlooked in medical school programs. **Goal:** to study the feasibility of a multi-criteria evaluation approach called NICTALOP to teach post-graduate students and evaluate their ability in writing medical observations. **Methods:** 28 medical students in the 4th to the 6th year at the school of medical sciences wrote up an introduction on medical observations related to one simple and one complex case, both before and after a special training. The observations were evaluated using the five NICTALOP (French acronym) criteria included in this system: Number of Ideas, Choice of Terms, Accuracy of ideas, Length of sentence, Position of idea. A global score was determined using the various criteria. **Results:** At the initial evaluation, experienced students showed better scores for the simple case ($p = 0.02$) but there was no difference for the complex one ($p = 0.4$). Students who followed the entire training improved their score ($p = 0.01$). The progress eliminated any level differences for the simple case ($p = 0.5$) but created one for the complex case that experienced students had a better time summarizing ($p = 0.05$). **Conclusion:** The NICTALOP approach promotes a detailed evaluation of the quality of medical observations. It highlights the gap between students and reflects their various experience levels and training.

Introduction

De nombreuses études se sont intéressées à l'analyse de la communication médecin-malade^[1, 2] mais la communication entre médecins a été assez peu étudiée^[3]. Elle est pourtant au cœur du métier de médecin qui s'exerce de plus en plus en réseau. Tous les actes médicaux donnent normalement lieu à la rédaction d'un compte-rendu, d'une observation ou d'un courrier médical, dont le contenu et la forme devraient en principe permettre une compréhension complète de la situation décrite. Les courriers médicaux constituent des éléments indispensables à l'efficacité et à la continuité des soins. Pourtant, une grande variabilité des contenus des courriers médicaux a parfois été constatée^[4, 5]. Certaines sociétés savantes ont proposé des référentiels permettant d'évaluer et d'améliorer la qualité des courriers^[6] et certains organismes officiels ont tenté de préciser le contenu et les objectifs de ces courriers^[7].

La pratique de la rédaction médicale commence en fait dès l'externat avec la rédaction d'observations médicales. L'observation médicale est le document écrit par lequel l'externe rend compte de son examen clinique, de façon méthodique et systématique, selon un plan structuré abordant successivement : les données d'identification du patient, le motif d'hospitalisation, l'« histoire de la maladie » recueillie par

« l'interrogatoire », les données objectives recueillies lors de l'examen physique, la synthèse des résultats d'examens complémentaires éventuellement disponibles et enfin une conclusion comportant un résumé syndromique et la formulation argumentée d'hypothèses diagnostiques. Il paraît donc logique d'inscrire l'enseignement de la rédaction médicale au programme du deuxième cycle des études médicales (DCEM)^[3]. Cependant, en France, les étudiants en DCEM sont essentiellement confrontés pour l'évaluation de leurs écrits à la méthode des mots-clés qui est utilisée lors des épreuves classantes nationales (ECN). L'importance de cet examen est primordiale dans le parcours des étudiants puisqu'il conditionne le choix de la faculté pour l'internat et de la spécialité. De ce fait, on observe souvent une focalisation des étudiants sur le concept de mots-clés, qui induit des effets notables sur leurs productions écrites. On note souvent une faible hiérarchisation des idées puisqu'il est quasiment impossible de rendre compte de la structure d'un texte par la méthode des mots-clés. Le recours aux associations de qualificatifs partiellement redondants est fréquent et s'explique par la volonté de citer le « bon » mot-clé (par exemple : tachycardie ventriculaire droite infundibulaire au lieu de tachycardie infundibulaire). On pourrait citer également l'usage de certaines expressions stéréotypées, telles que « bi-antibiothérapie à large spectre adaptée à

l'antibiogramme ». La longueur de la réponse ou la présence d'éléments discutables, voire faux, ne sont quasiment jamais prises en compte lors de la correction des examens écrits, sauf si le comité de rédaction de la question l'a prévu. Nous pensons qu'une prise en compte de ces différents éléments serait utile pour l'évaluation d'une observation médicale car les données prises en compte orientent fortement les apprentissages. L'objectif de cette étude était d'évaluer la faisabilité d'une évaluation recourant à de multiples critères pour apprécier la qualité de la rédaction d'une observation médicale considérée en tant qu'exercice d'apprentissage pour des étudiants de DCEM et utilisable notamment dans le cadre d'une démarche formative.

Principes

Critères d'évaluation

Renaud^[8] retient trois critères principaux d'évaluation de la qualité d'une observation médicale : l'exhaustivité des renseignements, la précision des termes et la concision. Nous proposons de retenir deux critères additionnels : la justesse des concepts, afin de pénaliser les observations comportant des renseignements faux et la position des concepts, afin d'encourager nos étudiants à faire un effort de hiérarchisation. Au total, la méthode d'évaluation élaborée est basée sur cinq critères : 1) l'exhaustivité des concepts (les idées principales ont-elles été énoncées ?) ; 2) la précision des termes choisis (par exemple et par ordre croissant : « saignement dans la tête », hématome cébral, hématome sous-dural ou hématome sous-dural chronique) ; 3) la justesse des concepts (ce qui a été dit est-il vrai, probable, peu sûr, douteux ou faux ?) ; 4) la longueur de l'observation (présence de périphrases, de répétitions) ; 5) la position des concepts les uns par rapport aux autres (l'ordre est-il logique ?).

Ces cinq critères définissent l'acronyme NICTALOP(NI = Nombre d'idées, CT = Choix des Termes, A = vrAi/fAux, LO = LOngueur, P = Position). Les

trois premiers (NI, CT et A) correspondent à des critères essentiels, à rechercher en premier et les deux derniers (LO et P) à des critères mineurs, que l'étudiant ne doit rechercher qu'une fois les critères majeurs raisonnablement satisfaits.

Évaluation séparée de chaque critère

L'analyse de chaque observation médicale selon ces cinq critères permet de générer respectivement cinq scores élémentaires. Dans un premier temps, l'observation médicale doit être analysée au regard de groupes de mots ne véhiculant qu'une idée, appelés ci-après « concepts » (cette répartition peut être demandée à l'étudiant pour simplifier la correction) :

- 1) Le critère NI est apprécié à l'aide d'un score basé sur la méthode des mots-clés en acceptant tous les synonymes (notamment sans prendre en compte la précision des termes : pyélonéphrite et infection des reins sont équivalents pour ce critère). Le score global NI correspond à la somme des valeurs associées à chaque « idée »- clé.
- 2) Le critère CT est évalué selon le choix des termes utilisés pour désigner chaque concept dans l'observation : les termes choisis désignent-ils en médecine une situation unique et précise ? Une expression d'un registre courant et non professionnel (« l'enfant est tombé ») correspond à un score de 0. Toute amélioration dans la précision des termes donne un score un peu plus élevé. Par exemple, « chute avec lipothymie » correspond à un score intermédiaire. Une expression permettant une identification complète du concept pour un médecin (« malaise vagal », « hématome sous-dural chronique ») correspond à un score élevé. Dans notre expérience il n'est pas utile de différencier plus de 3 ou 4 niveaux pour le critère CT. Le score global CT est la moyenne des scores de chaque concept mentionné dans l'observation.
- 3) Le critère A est évalué selon la justesse de chaque concept mentionné dans l'observation. Un concept complètement faux (« crise d'épilepsie » pour un malaise vagal) correspond à un score négatif. Un concept qui ne peut être exclu (« trouble du

rythme paroxystique » pour un malaise d'allure vagal) correspond à un score nul et un concept vrai donne un score positif. Plus le concept est sûrement vrai ou faux, plus la valeur absolue du score augmente. Dans notre expérience, il n'est pas utile de différentier plus de trois niveaux de probabilité : possible (0), peu probable (1 point), très probable (2 points) et sûr (3 points). La note globale A correspond à la moyenne des scores de chaque concept.

- 4) Le quatrième critère « LO » (pour « L'ongeur ») est apprécié à l'aide d'un nombre négatif dont la valeur absolue correspond au nombre de mots (nom, verbe, adjetif ou adverbe) que l'on peut supprimer dans le groupe identifié pour exprimer la même idée. Par exemple, un motif d'hospitalisation commençant par « bilan de malaise » au lieu de « malaise » donne un score LO de -1. Le score global LO correspond à la somme des scores de chaque concept.
- 5) Le cinquième critère « P » (pour « Position ») est apprécié à l'aide d'un nombre négatif qualifiant le nombre d'erreurs de positionnement du concept, c'est-à-dire le nombre d'écart entre la position choisie dans la copie et les positions acceptables pour le concept. Ainsi, un concept placé en position 1 (principale hypothèse diagnostique selon l'étudiant par exemple) alors que les experts le place en position trois ou quatre est affecté de -2. Dans notre expérience, il n'est pas utile de différentier plus de trois ou quatre niveaux de hiérarchisation des concepts. Le score global P correspond à la somme des scores de chaque concept.

Combinaison des critères d'évaluation

Une fois les cinq critères évalués séparément, il est souhaitable de combiner les cinq évaluations obtenues (les cinq scores élémentaires) pour obtenir un score global. L'identification de critères majeurs et mineurs nous a incités à proposer une combinaison qui pondère l'évaluation fondée sur les critères majeurs par une petite correction provenant des critères mineurs. Ce type de correction est fréquent dans le monde scolaire ou universitaire (points négatifs pour fautes d'orthographes par exemple) et de

nombreux autres tests utilisent des bonus/malus qui viennent pondérer le score obtenu (par exemple, en fonction de la vitesse de réponse dans le *Wechsler Intelligence Scale for Children -WISC*^[9]). Puisque l'objectif principal de l'évaluation est d'encourager les étudiants à satisfaire les trois critères majeurs simultanément (c'est-à-dire : NI et CT et A), il est logique de proposer que le score global soit le produit des scores élémentaires correspondant aux critères majeurs ($NI \times CT \times A$). En effet, la multiplication est l'opération arithmétique équivalente au « et » logique. L'addition qui est plus usitée dans le contexte des grilles de corrections correspond à un « ou » logique. C'est pourquoi il est possible dans un barème de type additif d'obtenir un score moyen en ayant zéro à un critère donné. Il n'est sans doute pas souhaitable de donner la moyenne à une observation contenant autant de concepts vrais que de concept faux (score élémentaire A = 0), quand bien même la plupart des mots-clés y seraient présent et les termes utilisés précis. L'usage du produit des trois scores élémentaires obtenus pour chaque critère majeur (NI, CT et A) se justifie également par le constat d'une potentialisation de ces différents éléments. Par exemple, un concept faux est d'autant plus dangereux qu'il est pertinent dans le contexte et exprimé dans un langage médical. Avec cette formule, un texte « théoriquement parfait » correspond à un score élevé, un texte vide est affecté d'un score nul et un texte contenant beaucoup de concepts faux correspond à un score négatif.

Pour prendre en compte l'ensemble scores élémentaires dans un seul score final, nous proposons donc la formule $S = (NI \times CT \times A) \times 100 + LO + 4 \times P$, après normalisation des scores NI, CT et A (pour les ramener à une valeur entre 0 et 100 %). Les erreurs de placement ont été pondérées arbitrairement avec un coefficient quatre par rapport aux erreurs de longueur.

Méthodes

Population

Vingt-huit externes présents en stage de cardiologie infantile au Centre hospitalier universitaire de Nancy

Tableau I. Principales caractéristiques de la population d'étudiants.

Effectif, <i>n</i>	28
Répartition selon le genre	
masculins, <i>n</i> (%)	16 (57 %)
féminins, <i>n</i> (%)	12 (43 %)
Répartition en année d'étude :	
DCEM2 : 4 ^e année, <i>n</i> (%)	10 (36 %)
DCEM3 : 5 ^e année, <i>n</i> (%)	8 (28 %)
DCEM4 : 6 ^e année, <i>n</i> (%)	10 (36 %)
ÉTUDIANT redoublant, <i>n</i> (%)	2 (7 %)
Heures d'enseignement reçues, moyenne (écart-type)	2,5 (0,9)

DCEM : Deuxième cycle des études médicales.

au cours de l'année universitaire 2010-2011 ont accepté de participer à l'étude. Ils ont tous été évalués une première fois le jour de leur entrée en stage, avant tout enseignement. Ceux qui étaient présents la dernière semaine de leur stage ont été évalués une deuxième fois. Ces évaluations se sont déroulées par écrit avec anonymisation des copies. Les principales caractéristiques de ces étudiants sont présentées dans le tableau I.

Dispositif pédagogique

Dans le cadre de cette étude concernant des étudiants hospitaliers dont la durée de présence en stage est d'environ un mois, seule l'introduction de l'observation médicale a été abordée. Les externes ont assisté en fonction de leur disponibilité (vacances, absentéisme, repos de garde) à des séances d'enseignement hebdomadaires en petits groupes (effectif maximal = 6) où ils devaient à tour de rôle présenter oralement un patient du service avec accès possible à leur propres notes et le lendemain d'une visite complète où chaque cas était discuté. La discussion des cas au cours de ces visites était libre et la nature précise des informations discutées n'a pas été précisée dans le protocole de l'étude. De même, la participation des étudiants à ces visites (ou leur absence) n'a pas été recueillie. Au cours d'une même séance

d'enseignement, chaque étudiant était à tour de rôle évaluateur et évalué. En effet, chaque étudiant était incité à évaluer les présentations des autres membres du groupe et les avis étaient confrontés à l'arbitrage de l'enseignant (LB). Les présentations étaient évaluées selon les cinq critères déjà cités, qui ont été explicités dès la première séance.

Formulaires d'évaluation

Les évaluations initiales et finales se sont déroulées par écrit et toutes les séances ont été dirigées selon le même principe par le même enseignant (LB). Avant le début de la séance, chaque étudiant avait reçu l'introduction de deux observations médicales avec des zones à compléter correspondant au motif d'hospitalisation et au contexte de l'hospitalisation. Le modèle du formulaire utilisé est présenté sur la figure 1. Pour l'introduction de la première observation, correspondant à une situation médicale simple, l'imprimé proposait seulement deux espaces blancs à compléter correspondant au motif d'hospitalisation et au contexte. Pour l'introduction de la deuxième observation, correspondant à une situation médicale plus complexe, l'imprimé proposait aussi un espace pour le motif d'hospitalisation mais le contexte était hiérarchisé en deux niveaux avec une liste de un à cinq éléments principaux et des éléments secondaires. Sur chaque questionnaire, l'étudiant devait en outre s'identifier (avec anonymisation), noter le nombre de séances d'enseignement auxquelles il avait participé (entre zéro et quatre) et son année d'étude (quatrième à sixième année des études médicales).

Présentation des histoires médicales pour l'évaluation

Deux situations médicales distinctes étaient alors présentées aux candidats par oral. La première situation « simple » (histoire S) correspondait à une jeune fille de quatorze ans hospitalisée pour un malaise d'allure vagal à l'arrêt de l'effort au cours d'un match sportif au collège. L'autre situation « complexe » (histoire C) correspondait à un enfant de huit mois hospitalisé pour difficultés alimentaires dans un contexte d'hématome sous-dural chronique apparu au décours

Fig. 1. Formulaire fourni aux étudiants.

d'une chirurgie cardiaque et accompagné de troubles de la déglutition avec retard de développement psychomoteur majeur. Les situations étaient à chaque fois présentées selon le même principe, bien expliqué aux participants avant le début de la séance. Les étudiants pouvaient interroger leur patient (ou sa mère pour l'histoire C) aussi longtemps qu'ils le souhaitaient dans la limite de 45 minutes maximum. Les étudiants étaient encouragés à poser une question à tour de rôle. L'enseignant (LB) jouait le rôle du patient (ou de sa mère) et répondait le plus honnêtement possible, en gardant à l'esprit la volonté de faire émerger par l'interrogatoire la totalité des concepts. L'ordre dans lequel ces concepts ont été mentionnés a dépendu bien sûr du déroulement de l'interrogatoire. Les étudiants pouvaient aussi avoir accès à tout moment au dossier médical du patient et interroger son médecin référent. L'objectif de cette méthode de présentation était de garantir que toutes les équipes avaient accès aux mêmes données tout en laissant les étudiants accéder eux même activement à ces données. Après cette première période de simulation d'interrogatoire, chaque étudiant devait remplir les blancs de l'imprimé qui était à sa disposition.

Construction de la grille de correction

Afin de permettre une correction la plus objective possible en tenant compte des situations médicales exposées aux étudiants, plusieurs médecins du service, dits « médecins experts », ont accepté de participer à l'exercice et ont rempli eux aussi un formulaire pour servir de référence. Chaque concept présent dans au moins un des formulaires de référence (remplis par des médecins experts) a été retenu comme « concept attendu » et il lui a été affecté une position attendue (motif, contexte, contexte secondaire) ainsi qu'une valeur d'importance (appelé « poids du concept ») correspondant au nombre de copies d'expert le mentionnant.

Évaluation des observations

La progression des scores S suit une loi géométrique d'ordre trois puisque le score total S est basé sur le

produit de trois scores élémentaires. Le score correspondant à une observation moyenne est $S = (0,5)^3 \times 100 = 12,5$ et une observation correspondant au premier quartile aura $S = (0,75)^3 \times 100 = 42$. Pour simplifier l'analyse, les observations ont donc été réparties en trois catégories : la mention « A » a été attribuée aux très bonnes observations avec score $S \geq 42$ (correspondant au 1^{er} quartile des scores possibles) ; la mention « B » a été attribuée aux bonnes observations dont le score $S \geq 12$ (correspondant à la moyenne) ; la mention « C » a été attribuée aux autres copies, jugées insuffisantes, avec un score $S < 12$.

Évaluation des étudiants et de l'enseignement

Les scores S ont été comparés selon l'expérience des étudiants et la complexité de l'histoire par un test de comparaison de moyennes U de Mann et Whitney. Les étudiants ont été répartis en deux groupes selon leur ancienneté, la limite étant le milieu de la cinquième année d'étude médicale (DCEM3). La progression des étudiants a été évaluée en comparant deux à deux les scores par tests de comparaisons appariées T de Wilcoxon. Le recours à des tests non paramétriques a été dicté par le faible effectif des sous-groupes comparés. La valeur $p = 0,05$ a été retenue comme seuil de significativité. Les tests statistiques ont été effectués avec le logiciel R 2.10.1 (*R Foundation for Statistical Computing*, Vienne, Autriche)^[10].

Résultats

Vingt-huit étudiants ont été soumis à l'évaluation initiale mais seulement 23 à l'évaluation finale. Les cinq perdus de vus (18 % de l'effectif) étaient absents pendant la dernière semaine de leur stage. La répartition selon l'année d'étude et le nombre d'heures d'enseignement effectivement reçues est donnée dans le tableau I.

Le tableau II présente la liste des « concepts » des deux histoires parmi les copies de référence et dans les copies d'étudiants. Le tableau III fournit quelques exemples de correction d'observations issues de l'histoire S.

Tableau II. Liste des concepts utilisés pour décrire la situation médicale S. À gauche ont été mentionnés les concepts issus des questionnaires soumis aux experts avec le poids et la localisation de chaque concept. À droite apparaissent les autres concepts proposés par les étudiants (concepts supplémentaires).

Concepts de référence,	(Poids, Position)	Concepts énoncés par les étudiants
Malaise	3 M	Vomissement
D'allure vagale	3 M	Chute
A l'arrêt de l'effort	2 M	Perte de connaissance
Sans antécédent	3 C	Trauma crânien Epilepsie Fibrillation auriculaire paroxystique Céphalée Vers 9h30 Match de volley-ball Enfant sportif Enfant non diabétique Enfant non épileptique Pas d'antécédents familiaux Pas de traitement Pas de tabac
Difficultés alimentaires	3 M	Vomissements
Alimentation parentérale	2 C	RGO
Attente de jejunostomie	3 C	Hypotonie
Atrésie Tricuspidale Ia	3 C	Insuffisance cardiaque
Blalock	3 c	Retard staturo-pondéral
2 reprises chirurgicales	2 c	Traitemen
Plastie artère pulmonaire gauche	1 c	(Hématome) non évacué
Plicature diaphragmatique	1 c	(Retard) depuis 2 mois
Réanimation compliquée	1 c	(Opération) à la naissance
Hématome sous dural chronique	3 C	Absence de tenue de tête, assise
Découvert en post-chirurgie	3 c	Fraction d'Ejection = 50 %
Retard psychomoteur	3 c	Sans pathologie génétique
Cécité corticale	1 c	Pèse 5 kg
Trouble déglutition	2 cC	Pas de dyspnée, Pas de fièvre Né à terme, grossesse normale, ... Rétrécissement aortique Cyanogène Maltraitance Asthénie

Légende : M = concept situé dans le motif d'hospitalisation.

C = concept situé dans le contexte en position principale.

c = concept situé dans le contexte en position secondaire.

Tableau III. Exemple de correction d'une copie de l'histoire S avec les scores de chaque concept, les notes globales, le score et la mention obtenus.

	Concepts	NI	CT	A	LO	P
Motif	Malaise	3	3	3		
	et chute sur la tête	0	1	0	-1	
	ayant entraîné une perte de connaissance	0	2	1	-1	
	Vers 9h30	0	3	3	-1	-2
Contexte	En jouant au volley	0	2	3	-2	-2
	Sportive	0	3	3	-1	
	Non diabétique	0	3	3	-1	
	Ne fumant pas	0	1	3	-1	
	Notes globales (et maxima possibles)	3/11	18/24	19/24	-8	-4
	Notes globales normalisées	27 %	75 %	79 %		
	Score	100*NI*CT*A = 16			-8	-16
		S = -8				
	Mention	C				

* Les scores NI et CT ont été attribués selon une échelle de 0 à 3. Les scores A ont été élaborés selon une échelle de -3 à +3.

Tableau IV. Scores élémentaires moyens NI, CT, A, LO et P et scores S moyens lors de l'évaluation initiale selon l'expérience de l'étudiant et la difficulté de l'histoire.

	Histoire Simple	Histoire Complexé	p
Début de DCEM	NI = 77 % CT = 39 % A = 71 % LO = 1,1 P = 1,7	NI = 55 % CT = 55 % A = 77 % LO = 2,8 P = 5,3	0,8
p	0,02	0,4	
Fin de DCEM	NI = 86 % CT = 73 % A = 75 % LO = 0,8 P = 1,0	NI = 63% CT = 65% A = 60% LO = 0,6 P = 1,8	0,07

DCEM : Deuxième cycle des études médicales.

La séparation entre début et fin de DCEM a été fixée à la moitié de la cinquième année.

Les scores NI, CT et A ont été normalisés et exprimées entre 0 et 100.

Le score des formulaires de référence (remplis par des médecins experts) était de 63 (+/-12) (mention A). Le score NICTALOP moyen lors de l'évaluation initiale en fonction de l'ancienneté de l'étudiant et de la complexité de l'histoire médicale est rapporté sur le tableau IV et illustré sur la figure 2. Lors de l'évaluation initiale, les étudiants en début d'étude obtiennent des scores plus faibles que ceux en fin de formation pour l'histoire S ($p = 0,02$) mais cette

différence n'est pas retrouvée pour l'histoire C ($p = 0,4$). Lors de l'évaluation finale, les étudiants qui ont assisté à trois ou quatre cours obtiennent des scores en progression significative ($p = 0,01$) mais pas ceux qui n'ont participé qu'à un ou deux cours ($p = 0,8$). Lors de cette évaluation finale, il n'y a plus de différence de scores entre les étudiants de début et de fin de formation pour l'histoire S ($p = 0,5$) mais il existe une différence en faveur des étudiants expérimentés

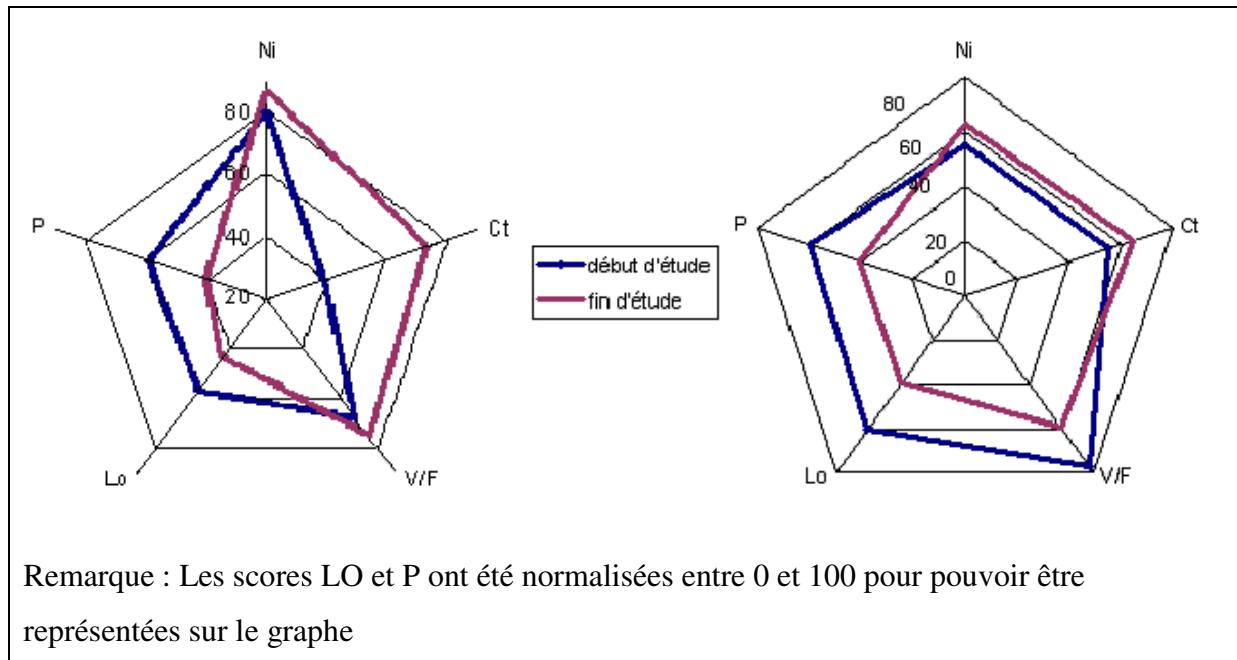


Fig. 2. Scores élémentaires composant le score NICTALOP, lors de l'évaluation initiale, en fonction de l'expérience des étudiants et de la complexité de l'histoire (simple à gauche et complexe à droite).

pour l'histoire C ($p = 0,05$). L'évaluation des observations médicales d'étudiants et leur évolution après enseignement est représentée dans les graphes sur la figure 3.

Discussion

Synthèse des résultats

La rédaction médicale est l'une des capacités au cœur du métier de médecin et conditionne d'une certaine façon la qualité de la prise en charge des patients en garantissant notamment la continuité des soins. On peut aussi postuler, dans une certaine mesure, qu'un compte-rendu médical mal rédigé soit le fruit d'une difficulté à concevoir clairement la situation médicale qu'il relate. Comme le mentionnait Boileau : « Ce qui se conçoit bien s'énonce clairement. Et les mots pour le dire arrivent aisément ». Pourtant l'enseignement universitaire de la rédaction d'une observation médicale est un exercice souvent

non systématisé. L'étudiant en médecine dans ce domaine est souvent auto-didacte. Sijstermans et al.^[3] ont fait ce constat en Hollande et proposé d'y remédier avec un enseignement assisté par ordinateur. L'amélioration des compétences des étudiants après l'enseignement y était mesurée par autoévaluation. La méthode d'évaluation de cette compétence est cruciale car elle conditionne la réflexion des étudiants et oriente leurs apprentissages. Les étudiants sont très habitués à une évaluation basée sur le repérage de mots-clés dans leurs réponses écrites. Cette méthode a l'intérêt d'une certaine simplicité d'utilisation mais ne motive pas les étudiants à prendre en compte d'autres aspects importants de la rédaction médicale que sont par exemple la précision, la concision ou la hiérarchisation des données.

Notre étude avait comme objectif d'évaluer la faisabilité d'une autre méthode d'évaluation, basée sur plusieurs critères. Le recours à une échelle multiple est fréquent en sciences du langage, en psychologie clinique ou en neuropsychiatrie, lorsqu'il s'agit d'évaluer des notions complexes comme l'intelligence (exemple du test WISC-R^[10]), la pensée

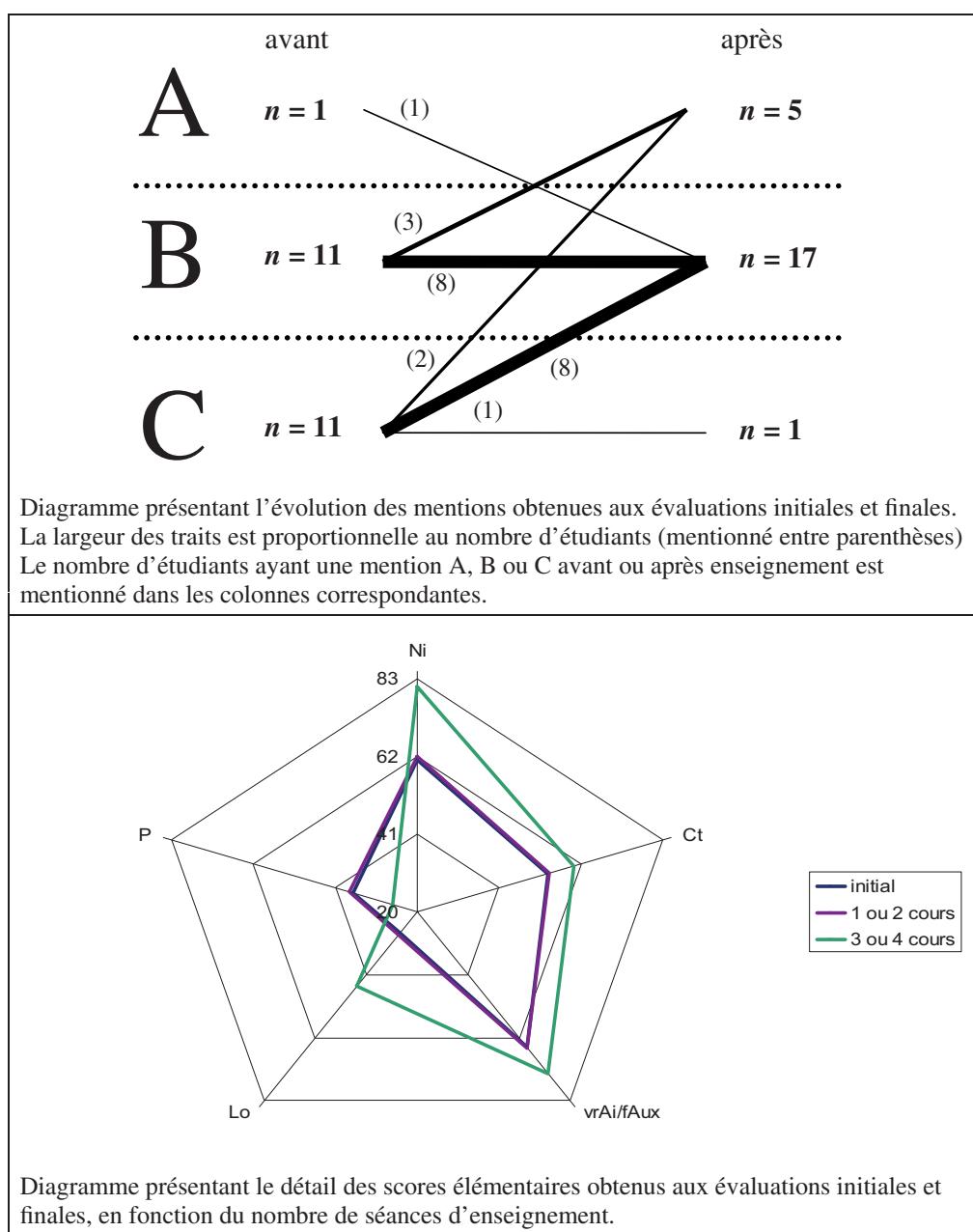


Fig. 3. Évolution des mentions NICTALOP après enseignement.

(exemple du test Sally-Anne^[11]) ou les capacités sociales (exemple du test EASE^[12]). La méthode NICTALOP utilise cinq critères différents dont trois majeurs et deux mineurs. Elle exige de ce fait un temps significatif de correction mais permet une analyse plus fine des performances d'un étudiant, permettant

d'orienter l'enseignant dans son choix pédagogique. Par exemple, les étudiants en début de formation dans notre étude avaient en moyenne un score A élevé mais un score CT plutôt bas. Cela correspond à une tendance du débutant à placer dans son introduction une quantité importante d'éléments véridiques mais

souvent peu importants et exprimés dans un langage peu médical. Parmi les concepts mentionnés dans le tableau II, l'heure de la chute, le sport pratiqué ou l'absence de tabagisme illustrent bien cette situation. Ce comportement semble d'autant plus manifeste que l'histoire rapportée est complexe. Les étudiants en fin de formation obtenaient par contre un score CT assez élevé et un score A moins bon, car ces étudiants formulaient plus facilement des hypothèses diagnostiques, qu'ils exprimaient souvent avec les termes médicaux appropriés mais qui n'étaient pas toujours justes. Ceci était d'autant plus vrai que l'histoire était complexe. Les examens universitaires sont souvent corrigés en France sur la base du repérage par le correcteur des mots-clés que dans la réponse de l'étudiant, ce qui correspond au premier critère de la méthode que nous proposons. De fait, dans une évaluation par mots-clés, le niveau sémantique est pris en compte de façon assez aléatoire par le correcteur qui doit faire l'effort de reconnaître un mot clé dans un synonyme ou une expression de sens proche. Enfin les éléments faux dans la réponse sont souvent ignorés s'ils ne correspondent pas à un mot-clé. La méthode NICTALOP permet de sanctionner l'excès de données médicales dans un compte-rendu ainsi que l'existence d'erreurs même en présence de concepts justes par ailleurs. La méthode d'évaluation à plusieurs critères encourage l'étudiant à choisir parmi ses hypothèses et à hiérarchiser ses idées. Cet effort de hiérarchisation et de sélection est très valorisé par l'existence de points négatifs en cas d'éléments faux et par l'utilisation d'une normalisation des notes CT et A par le nombre de concepts énoncés.

Le temps de correction est bien sûr un élément crucial dans le choix d'une méthode d'évaluation. Le temps universitaire et médical est précieux et certainement pas inépuisable. Dans notre expérience, l'usage de la méthode NICTALOP entraîne un doublème du temps de correction par rapport à une méthode plus simple par mots-clés. Cet allongement du temps de correction peut sans doute se compenser par une diminution du nombre de questions. Il s'agit de choisir un compromis entre la redondance des questions et la pertinence de la correction. Notre expérience montre qu'il est possible de demander aux étudiants (comme nous l'avons fait) de répartir eux

même leur texte en concepts-clés, ce qui ouvre la voie à une automatisation de la correction avec usage de l'outil informatique mais cela n'était pas l'objet de cette étude.

Limites de l'étude

Notre étude s'inscrit dans le cadre d'une expérience évaluative préliminaire et résolument pragmatique, visant à fournir aux étudiants une évaluation formative selon des critères faciles à exploiter. Le mode de construction de la méthode (choix des critères, élaboration du barème et de la pondération) s'est appuyé sur une argumentation didactique informelle et non sur une démarche méthodique et systématique. Des développements docimologiques pourraient éventuellement être envisagés à terme mais, en l'état, la méthodologie employée ne permet en aucune manière de vérifier qu'elle satisfait rigoureusement les critères métrologiques de validité et de fidélité.

Le choix de la formule $S = (NI \times CT \times A) \times 100 + LO + 4 \times P$ est discutable et notamment le choix du coefficient de pondération appliqué au score P est arbitraire. Ce choix a résulté d'un compromis entre simplicité et performance, avec comme objectif la prise en compte les erreurs de positionnement sans faire intervenir la subjectivité du correcteur. Toute évaluation basée sur l'analyse d'un texte comporte des éléments de subjectivité. Nous avons essayé de mettre en œuvre toutes les procédures envisageables pour limiter cette subjectivité.

Certains éléments sont cependant de nature à soutenir la viabilité de la méthode^[13]. Ainsi, les concepts clés ont été choisis par une méthode inspirée des tests de concordance de scripts avec utilisation d'experts ; la valeur attribuée et la position attendue de chaque concept clé dérivent de cette même méthode. De même, l'évaluation de la véracité ou de la précision des expressions utilisées par les étudiants a été faite après recueil de l'ensemble des expressions et la règle définie a été utilisée pour toutes les copies. Par ailleurs, davantage d'aspects de la rédaction médicale sont pris en compte que dans la méthode qui se limite au repérage des mots-clés. Ces différents éléments peuvent être mis au crédit d'une certaine validité de

contenu de la méthode NICTALOP. Par ailleurs, certains éléments plaident en faveur d'une certaine validité de construit de la méthode puisqu'elle met en évidence une différence entre les étudiants selon leur niveau de formation et les experts. Ainsi, les résultats des étudiants en fin de formation sont meilleurs à l'évaluation initiale que ceux des étudiants en début de formation, en tout cas pour l'histoire S. L'histoire médicale C correspondait à une situation médicale complexe difficile à comprendre et résumer pour les deux groupes en début de stage, ce qui explique l'absence de différence significative entre les deux groupes pour cette observation. Par contre, les évaluations de fin de stage montrent que les étudiants en début d'étude ont progressé et rattrapé le niveau de leurs collègues expérimentés pour l'histoire S, tandis que ces derniers parviennent à un meilleur niveau de compréhension des situations complexes qui explique l'apparition d'une différence de scores pour l'histoire C à la fin du stage entre les étudiants de début et de fin d'étude. De même, les étudiants qui participent à la formation entre les deux évaluations augmentent leur score au deuxième test. Cette augmentation n'est pas simplement due à un effet « répétition » (lorsque qu'un test est répété deux fois, une amélioration due à la disparition de l'effet surprise est possible), car ceux qui ont assisté à la moitié de la formation seulement n'améliorent pas leur performance. De même, les experts ont un score bien meilleur que les étudiants (S moyen = 63). La reproductibilité de la méthode n'a pas été directement évaluée mais, pour les cinq étudiants ayant assisté à moins de deux séances d'enseignement on constate une stabilité des scores et des scores globaux obtenus pour la même épreuve réalisée à un mois d'intervalle (Figure 3).

L'effectif de la population de notre étude est petit mais a été suffisant pour mettre en évidence des différences statistiquement significatives. La présence de cinq perdus de vue peut constituer un biais de sélection. La méthode n'a par ailleurs été testée que sur deux cas cliniques avec potentiellement un phénomène de dépendance aux cas (*case specificity*).

Enfin, la méthode a été utilisée à l'oral pour l'évaluation formative au cours des quatre séances d'enseignement et à l'écrit pour l'évaluation initiale et finale. Ce choix est certainement maladroit et un

exercice de rédaction à partir des cas cliniques présentés en groupe aurait dû compléter cet exercice.

Conclusion

Les compétences de rédaction d'un compte-rendu médical s'acquièrent dès le DCEM et peuvent être améliorées par un petit nombre de séances de travaux dirigés à partir de cas concrets issus du service qui accueille les étudiants. Pour évaluer ces compétences, nous proposons une nouvelle méthode d'évaluation pragmatique utilisable notamment dans le cadre d'une démarche formative, appelée NICTALOP et basée sur trois critères majeurs qui évaluent les idées formulées, le choix des termes et la véracité des concepts et deux critères mineurs qui évaluent la concision de la rédaction et la hiérarchisation des concepts.

Contributions

Laurent Bonnemains a conçu le protocole de recherche, effectué le recueil des données, participé à l'interprétation des résultats, à l'analyse statistique et rédigé la version initiale du manuscrit. Francois Marcon a participé à l'interprétation des résultats et à la lecture critique du manuscrit. Marc Braun a participé à la conception du protocole de recherche et à la lecture critique du manuscrit.

Approbation éthique

Le protocole d'étude n'a pas fait l'objet d'un examen devant un comité d'éthique.

Déclaration d'intérêts

Les auteurs déclarent qu'ils n'ont pas de conflit d'intérêt financier en lien avec l'objet de cet article. Ils déclarent tous une position hiérarchique à l'égard des étudiants dans le cadre de leurs responsabilités institutionnelles au sein de la faculté, relativement à l'organisation du stage ou dans l'enseignement.

Références

- Shapiro SM, Lancee WJ, Richards-Bentley CM. Evaluation of a communication skills program for first-year medical students at the University of Toronto. BMC Med Educ 2009;9:11.

2. Smith S; Hanson JL, Tewksbury LR, Christy C, Talib NJ, Harris MA, Beck GL, Wolf FM. Teaching patient communication skills to medical students: a review of randomized controlled trials. *Eval Health Prof* 2007;30:3-21.
3. Sijstermans R, Jaspers MW, Bloemendaal PM, Schoonderwaldt EM. Training inter-physician communication using the Dynamic Patient Simulator. *Int J Med Inform* 2007;76:336-43.
4. Dievart F. Le cardiologue et son courrier professionnel : Le fond et la forme. *Cardiologie Cardinale* 2010;26:67.
5. Engblom M, Alexanderson K, Rudebeck CE. Physicians' messages in problematic sickness certification: a narrative analysis of case reports. *BMC Fam Pract* 2011;12:18.
6. Union nationale de Formation continue et d'évaluation en médecine Cardio-Vasculaire. Lettre du cardiologue à son correspondant. 2010 [On-line] Disponible sur : <http://www.ufcv.org/index.php/fre/Recommendations-et-referentiels/Referentiels-UFCV>
7. Mercat FX. Les courriers entre médecins. CNOM. 1998 Jan. [On-line] Disponible sur : <http://www.conseil-national.medecin.fr/system/files/courriers.pdf>
8. Renaud MC. Guide pratique. Semiolegie & observation médicale. Paris : Editions ESTEM et EDLINE, 1996.
9. Weschler D. Échelle d'intelligence de Wechsler pour enfants, forme révisée WISC-R. (1^{ère} ed.). Paris : Centre de psychologie appliquée, 1981.
10. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2009.
11. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child's have a "theory of mind" ? *Cognition* 1985;21:37-46.
12. Comte-Gervais I, Giron A, Soares-Boucaud I, Poussin G. Evaluation de l'intelligence sociale chez l'enfant présentant des troubles spécifiques du langage oral. Présentation d'une échelle d'évaluation clinique. L'évolution psychiatrique 2008;73:353-66.
13. Amarenco G, Bayle B, Lagache D, Lapeyre E, Sheik Ismael S. Construction et validation des échelles de qualité de vie. *Ann Readaptation Med Phys* 2000;43:263-9.

Correspondance et offprints : Laurent Bonnemains. Cardiologie Infantile, CHU Nancy-Brabois, rue du Morvan, 54511 VANDOEUVRE-LES-NANCY, France.
Mailto : l.bonnemains@chu-nancy.fr