Régression logistique : intérêt dans l'analyse de données relatives aux pratiques médicales

The Use of Logistic Regression in the Analysis of Data Concerning Good Medical Practice

Aminot I [1], Damon MN [2]

Résumé
La régression logistique est un des modèles d'analyse multivariée explicatif couramment utilisé en épidémiologie. Son emploi, rendu aisé par l'utilisation de logiciels statistiques, permet le contrôle des biais de confusion. La mesure d'association calculée dans ce modèle est l'*odds-ratio* (ou rapport de cotes en français), qui quantifie la force de l'association entre la survenue d'un événement, représentée par une variable dichotomique, et les facteurs susceptibles de l'influencer, représentés par des variables explicatives. Le choix des variables explicatives intégrées au modèle repose sur une connaissance préalable du phénomène étudié afin de ne pas omettre de facteurs de confusion déjà identifiés. Les auteurs exposent les principes fondamentaux de la régression logistique et les principales étapes de sa réalisation. A l'aide de deux exemples (qualité du suivi des malades diabétiques, mortalité hospitalière après infarctus du myocarde), la démonstration est faite de l'intérêt d'utiliser cet outil statistique dans les études du service médical de l'assurance maladie, notamment celles évaluant les pratiques professionnelles.
Rev Med Ass Maladie 2002;33,2:137-143

Mots clés : analyse multivariée, régression logistique, évaluation, pratiques professionnelles.

Summary
Logistic regression is one of the commonly used models of explicative multivariate analysis utilized in epidemiology. Its use, which has become easier with modern statistical software, allows researchers to control confusion bias. It measures the *odds-ratio,* a quantification of the association probability between a given occurrence, represented by a dichotomic variable, and factors susceptible to influence it, represented by explicative variables. The choice of explicative variables integrated into the model is based on previous information on the study subject and is aimed at avoiding the confusion factors which have already been identified. The authors explain the fundamental principles of logistic regression and the steps involved in its application. By using two examples (the quality of the follow up care given to diabetics and in-hospital mortality after acute myocardial infarction), they demonstrate the value this statistical tool can have in studies performed by the medical service of the national health care fund, particularly in studies designed to evaluate professional practice.
Rev Med Ass Maladie 2002;33,2:137-143

Key words: multivariate analysis, logistic regression, evaluation, professional practice.

[1] Médecin-conseil, Direction régionale du service médical de Normandie (CNAMTS)
[2] Médecin-conseil, chef de service, Échelon local du service médical d'Elbeuf (CNAMTS)
Adresse pour correspondance : Dr Isabelle Aminot, Direction régionale du service médical de Normandie (CNAMTS), Avenue du Grand Cours, F-76108 Rouen cedex, e-mail : isabelle.aminot@ersm-normandie.cnamts.fr

I. INTRODUCTION

Les épidémiologistes, lors de l'exploitation de données médicales, doivent souvent décrire un événement ou un phénomène lui-même influencé par la survenue d'autres événements ou phénomènes appelés facteurs d'exposition. Une difficulté à laquelle sont souvent confrontés ces chercheurs est l'erreur systématique dans la mesure d'association entre deux évènements due à la présence de facteurs de confusion [1].

Au contraire des autres biais (biais d'information, biais de sélection), les biais de confusion sont contrôlables lors de l'analyse statistique. La méthode d'ajustement de Mantel-Haenszel permet de prendre en compte un nombre limité de facteurs de confusion et n'est applicable aisément que lorsque les variables sont qualitatives à deux classes. Par ailleurs, elle nécessite de s'assurer au préalable que le facteur de confusion supposé n'est pas un facteur d'interaction [2]. Cet ensemble de conditions limite considérablement le champ des analyses et par là même la possibilité de conclure de façon exacte pour un grand nombre d'études.

L'arrivée des logiciels de statistiques a permis l'emploi d'analyses multivariées explicatives capables de contrôler l'ensemble des biais de confusion, quel que soit le type de variables et d'obtenir une quantification de l'association entre l'événement étudié et chacun des facteurs l'influençant, tout en tenant compte de l'effet simultané des autres facteurs. La régression logistique est un des modèles d'analyses multivariées couramment employé en épidémiologie. L'objectif de cet article est de rappeler ses principes fondamentaux et d'en illustrer l'intérêt à travers deux exemples issus d'études du service médical de l'Assurance maladie.

II. PRINCIPE DE LA RÉGRESSION LOGISTIQUE

En épidémiologie, plusieurs modèles d'analyse multivariée sont couramment utilisés : régression linéaire multiple, régression logistique, régression de Poisson, modèle de Cox, etc. Effectuer une régression, c'est tenter de réduire les données d'un phénomène complexe en une loi mathématique simplificatrice. La fonction logistique (qui a donné son nom au modèle) possède des caractéristiques mathématiques expliquant son emploi dans un modèle d'analyse de données épidémiologiques : elle varie de 0 à 1 comme la probabilité de survenue d'un événement ; sa représentation graphique, de forme sigmoïde, correspond assez fidèlement au modèle de relation entre la survenue d'une maladie et un facteur d'exposition ; enfin, elle permet le calcul aisé des *odds-ratios* [3] (ou rapports de cotes en français).

Le modèle de régression logistique permet d'estimer la force de l'association entre une variable qualitative à deux classes (dichotomique) appelée variable *dépendante* et des variables qui peuvent être qualitatives ou quantitatives appelées variables *explicatives* ou indépendantes. La variable dépendante est la survenue ou non de l'événement étudié – maladie, acte médical, … – et les variables explicatives sont des facteurs susceptibles d'influencer la survenue de l'événement (facteurs d'exposition ou facteurs de confusion).

La régression logistique peut être univariée mais son intérêt réside dans son utilisation multivariée puisqu'elle permet, alors, d'estimer la force de l'association entre la variable dépendante et chacune des variables explicatives, tout en tenant compte de l'effet simultané de l'ensemble des autres variables explicatives intégrées dans le modèle. L'association ainsi estimée est dite « ajustée » sur l'ensemble des autres facteurs.

Même si des adaptations permettent de l'appliquer à certains cas particuliers, le modèle de régression logistique requiert, en principe, certaines conditions : indépendance des différentes observations entre elles, normalité de la distribution des variables quantitatives intégrées dans le modèle, et linéarité de la relation entre chacune de ces variables quantitatives et la variable dépendante.

III. MÉTHODE

La réalisation pratique d'un modèle de régression logistique comporte plusieurs étapes :

1. La qualité d'une régression logistique repose, avant tout, sur le **choix des variables explicatives** que l'on est susceptible d'intégrer au modèle. Ce choix est fondé sur la pertinence clinique et sur la connaissance de facteurs de confusion avérés ou supposés. C'est pourquoi, une recherche bibliographique approfondie est, au préalable, obligatoire.
2. Il est nécessaire ensuite **d'étudier chacune de ces variables** : analyse de la distribution des variables qualitatives selon leurs différentes modalités et, s'il y a lieu, regroupement de ces dernières ; étude de

[1] Il y a facteur de confusion lorsqu'il y a une erreur systématique d'estimation de l'association entre la variable dépendante et une variable explicative, cette erreur étant due à la présence d'une autre variable.

[2] Il y a interaction entre deux variables explicatives si l'association de l'une des deux variables avec la variable dépendante varie selon la valeur de l'autre variable

[3] Ceux-ci mesurent l'association d'un facteur donné à la survenue d'un évènement quel que soit le schéma d'étude : enquêtes transversales, enquêtes cas-témoins et même enquêtes de cohorte. L'odds-ratio varie entre zéro et l'infini. En l'absence d'association, il tend vers 1, et à l'inverse, lorsque les variables sont fortement liées, il tend vers zéro ou vers l'infini.

l'existence d'une relation linéaire entre chacune des variables quantitatives explicatives et la variable dépendante. Si, pour une variable, cette condition n'est pas vérifiée, on procédera à la transformation de celle-ci en une variable ordinale en créant des classes dont le choix repose sur des critères cliniques et statistiques.

3. On procède ensuite à l'analyse des liaisons entre chacune des variables explicatives et la variable dépendante : on réalise une **analyse univariée** ; les *odds-ratios* calculés sont bruts. Deux catégories de variables explicatives pourront être intégrées dans un **modèle de départ** : celles pour lesquelles l'association avec la variable dépendante est suffisamment forte sans toutefois être trop stricte afin de ne pas omettre d'éventuels facteurs de confusion (p-value inférieure ou égale à 0,20, et non pas 0,05, seuil habituellement retenu) et celles qui ont un intérêt clinique avéré en dehors de tout critère d'association (elles sont rares : ce sont des variables dites « forcées »).

4. **Plusieurs stratégies sont possibles** pour parvenir à un modèle final qui devra porter le maximum d'informations tout en ayant un nombre limité de variables afin de faciliter l'interprétation : les plus employées sont les procédures dites « pas à pas descendantes ou pas à pas ascendantes ». La déclinaison des modèles permettra de rechercher les phénomènes d'interaction ou de confusion qu'il faudra prendre en compte lors de l'interprétation. Certaines variables seront impérativement conservées dans le modèle : la variable explicative d'intérêt principal et les facteurs de confusion.

5. En fin d'analyse, plusieurs modèles finaux peuvent s'avérer satisfaisants sur un plan statistique. Parmi ceux-ci, on retiendra le modèle le plus adéquat avec le phénomène constaté : des tests d'adéquation permettent de guider le statisticien.

IV. EXEMPLES

Deux exemples sont présentés ci-après : ils portent sur des données analysées par le service médical de l'Assurance maladie. Les modèles de régression logistique ont été réalisés à l'aide du logiciel *Statistical Package for Social Science*™ (SPSS version 9.0).

A. Facteurs associés à un meilleur suivi chez les patients diabétiques

Le premier exemple porte sur des données concernant les patients diabétiques affiliés au régime général de l'Assurance maladie d'Aquitaine, traités exclusivement par hypoglycémiants oraux. Ces données sont issues d'une enquête nationale [1, 2].

Il avait été mis en évidence que les patients ayant eu recours au moins une fois à un endocrinologue libéral ou à un spécialiste de médecine interne ainsi que les patients exonérés du ticket modérateur au titre de l'une des trente affections de longue durée (ALD_{30}) [4] bénéficiaient d'un meilleur suivi de leur diabète.

Les examens cliniques et biologiques recommandés par l'Agence nationale d'accréditation et d'évaluation en santé étaient plus fréquemment pratiqués chez ces patients et notamment le dosage de l'HbA_1c (59,3 % de dosages de l'HbA_1c pour ceux ayant eu recours au spécialiste *versus* 33,7 % et 39,3 % de dosages de l' HbA_1c pour ceux exonérés du ticket modérateur *versus* 23,3 %). La question de l'interférence d'autres facteurs tels que l'âge, l'existence de complications cardio-vasculaires ou la fréquence des consultations chez le généraliste se pose alors. L'emploi d'un modèle de régression logistique nous permet de confirmer ou d'infirmer le constat initial et de le quantifier en prenant en compte les éventuels facteurs de confusion et en ajustant sur chacune des variables présentes.

La réalisation d'un dosage d'HbA_1c dans les six mois précédents est considérée comme l'indicateur d'un suivi de bonne qualité. Cet indicateur est la variable dépendante. Les variables explicatives initialement retenues sont les suivantes : « âge », « sexe », « traitement à visée cardiovasculaire associé », « traitement hypolipémiant associé », « existence d'une exonération du ticket modérateur et modalité de cette exonération », « consultation spécialisée en endocrinologie ou médecine interne » et « nombre de consultations par un généraliste dans l'année précédente ». La variable quantitative « âge », vérifiant l'hypothèse de linéarité, est maintenue dans sa forme initiale. Par contre, la variable « nombre de consultations par un généraliste dans l'année précédente », ne vérifiant pas cette condition, doit être transformée en une variable ordinale. Après étude de sa distribution en fonction de la variable dépendante, trois classes sont retenues : 0 à 2 consultations, 3 à 8 consultations et plus de 8 consultations annuelles.

Parmi ces variables susceptibles d'être intégrées, le sexe est exclu du modèle de départ car n'étant pas associé à la variable dépendante : p = 0,999 (Tableau I). Par ailleurs, l'existence de facteurs d'interaction ou de confusion n'est pas mise en évidence.

Ajustés sur les autres variables, c'est-à-dire toute chose étant égale par ailleurs, les deux facteurs les plus fortement associés à la qualité du suivi sont le recours à un endocrinologue ou à un interniste (OR ajusté = 2,71) et l'existence d'une exonération du ticket modérateur, surtout au titre de l'ALD_{30} (OR ajusté = 2,17). Un nombre de consultations par un médecin généraliste compris entre trois et huit l'année précédente est également lié, mais

[4] Articles L. 322-3.3° et D. 322-1 du Code de la sécurité sociale

Tableau I
Analyse multivariée des facteurs liés à l'indicateur de qualité de suivi des patients diabétiques traités par hypoglycémiants oraux (Région Aquitaine - avril 1998 à mars 1999)

Facteurs associés	Effectif total	HbA_1c effectuée		Analyse univariée			Analyse multivariée		
		Effectif	%	OR[a] brut	IC[b] 95 %	p-value	OR[a] ajusté	IC[b] 95 %	p-value
Age[c]						**$< 10^{-4}$**			**$< 10^{-4}$**
	33 961	12 051	35,5	**0,99**	[0,988-0,992]		**0,99**	[0,989-0,993]	
Sexe						**ns**			
masculin	17 621	6 253	35,5	1					
féminin	16 340	5 798	35,5	0,99	[0,96-1,05]				
Traitement cardio-vasculaire						**$< 10^{-4}$**			**10^{-3}**
non	8 406	3 142	37,4	1			1		
oui	25 555	8 909	34,9	**0,90**	[0,85-0,94]		**0,90**	[0,86-0,96]	
Traitement hypolipémiant						**0,036**			0,025
non	21 158	7 418	35,1	1			1		
oui	12 803	4 633	36,2	**1,05**	[1,01-1,09]		**1,06**	[1,01-1,11]	
Exonération du ticket modérateur						**$< 10^{-4}$**			**$< 10^{-4}$**
sans	7 527	1 757	23,3	1			1		
ALD_{30}	25 138	9 884	39,3	**2,13**	[2,01-2,26]		**2,17**	[2,04-2,31]	
Autre ETM	1 296	410	31,6	**1,52**	[1,34-1,73]		**1,62**	[1,42-1,84]	
Consultation par spécialiste						**$< 10^{-4}$**			**$< 10^{-4}$**
non	31 650	10 681	33,7	1			1		
oui	2 311	1 370	59,3	**2,86**	[2,62-3,11]		**2,71**	[2,48-2,96]	
Nombre de consultations par un généraliste						**$< 10^{-4}$**			**$< 10^{-4}$**
0-2	1 459	419	28,7	1			1		
3-8	10 309	3 733	36,2	**1,41**	[1,25-1,59]		**1,63**	[1,44-1,85]	
9 et plus	22 193	7 899	35,6	**1,37**	[1,22-1,54]		**1,50**	[1,32-1,69]	

[a] *Odds-ratio : les odds-ratios statistiquement différents de 1 sont notés en gras.*
[b] *Intervalle de confiance à 95 %.*
[c] *L'âge étant une variable quantitative, l'odds-ratio correspond à une augmentation de 1 an. Pour obtenir l'odds ratio correspondant à une augmentation d'âge de 5 ans, il suffit d'élever ce dernier à la puissance cinq. Ainsi, l'*odds-ratio *ajusté était de 0,95 et son intervalle de confiance à 95 % de : 0,94-0,96.*

moins fortement, à une meilleure qualité de suivi (OR ajusté = 1,63). Un effet seuil affecte cependant cette variable puisque la multiplication des consultations au-delà de huit ne semble pas apporter plus de gain. L'existence d'un traitement hypolipémiant est très faiblement associée à un meilleur suivi. Par contre, on constate une diminution de la qualité du suivi lorsque l'âge des patients augmente (OR ajusté = 0,95 pour une augmentation de cinq ans) et lorsque les patients bénéficient d'un traitement à visée cardiovasculaire (OR ajusté = 0,90).

Le principal intérêt de la régression logistique est donc de fournir des *odds-ratios* qui sont ajustés sur l'ensemble des variables retenues dans le modèle. Ceux-ci nous ont permis de constater que le recours à des spécialistes et l'exonération du ticket modérateur pour ALD_{30} étaient effectivement deux facteurs favorisant un bon suivi du patient diabétique et ce quel que soit l'âge des patients, l'existence ou non d'un traitement à visée cardio-vasculaire, l'existence ou non d'un traitement hypolipémiant et la fréquence des consultations par un géné-raliste. Dans cet exemple, il existe peu de différence entre les *odds-ratios* bruts et ajustés, ce qui nous permet d'affirmer qu'aucun des facteurs explicatifs étudiés ne représente un facteur de confusion vis-à-vis des autres facteurs.

Même si, théoriquement, la méthode traditionnelle de Mantel-Haenszel reste applicable, elle devient vite fastidieuse quand le nombre de variables augmente puisque l'analyse stratifiée qu'elle implique porte sur un nombre de strates qui est le produit du nombre de modalités de chacune des variables.

De plus, les logiciels statistiques tels que SPSS ne permettent pas d'appliquer la méthode de Mantel-Haenszel au-delà de deux variables explicatives à deux modalités chacune. Par ailleurs, cette méthode nous aurait contraints à utiliser l'âge sous forme de classes d'âge et non pas sous forme d'une variable quantitative continue.

B. Facteurs associés à la mortalité hospitalière chez des patients en phase aiguë d'infarctus du myocarde

Le deuxième exemple concerne une étude dont le champ était l'ensemble des patients hospitalisés

pour un infarctus aigu du myocarde dans les établissements de santé aquitains [3]. Un des objectifs de cette étude était de déterminer les facteurs associés à la mortalité hospitalière et *in fine* de savoir si, en tenant compte des différentes caractéristiques des patients, facteurs de confusion potentiels, le secteur sanitaire hospitalier de prise en charge du patient était associé à cet événement (en clair, si les taux de décès étaient statistiquement plus élevés dans certains secteurs).

La mortalité hospitalière est la variable dépendante. Les variables explicatives qualitatives intégrées dans le modèle de départ sont : l'âge, le sexe, les antécédents de cardiopathie, la localisation de l'infarctus, le délai de prise en charge, la procédure de revascularisation et le secteur sanitaire hospitalier. L'âge est une variable quantitative continue. La relation de cette variable avec la mortalité hospitalière n'étant pas linéaire, cette variable a été scindée en classes dont les bornes ont été fixées selon des critères statistiques renforcés par des arguments cliniques retrouvés dans la littérature médicale. Les analyses univariées permettent de constater que chacune de ces variables est associée à la mortalité : *p-value < 0,2* (Tableau II).

L'analyse multivariée permet de prendre en compte les variables « sexe », « âge » et « localisation de l'infarctus » liées à la mortalité hospitalière, dans l'étude de l'association entre la mortalité hospitalière et chacune des deux variables suivantes : « procédure de revascularisation » et « secteur sanitaire hospitalier ». Si l'on compare les *odds-ratios* ajustés aux *odds-ratios* bruts calculés lors des analyses univariées, on constate une nette diminution du risque de mortalité lié au sexe féminin – OR brut = 2,53 ; OR ajusté = 1,39 –, cette association restant

Tableau II
Analyse multivariée des facteurs liés à la mortalité hospitalière des patients pris en charge en établissements de santé pour infarctus du myocarde (Région Aquitaine – Année 1999)

Facteurs associés	Effectif total	Décès		Analyse univariée			Analyse multivariée		
		Effectif	%	OR[a] brut	IC[b] 95 %	p-value	OR[a] ajusté	IC[b] 95 %	p-value
Sexe						**< 10^{-4}**			**0,0081**
masculin	1 808	201	11,1	1			1		
féminin	807	194	24,0	**2,53**	[2,03-3,14]		**1,39**	[1,09-1,77]	
Classes d'âge						**< 10^{-4}**			**< 10^{-4}**
< 65 ans	893	39	4,4	1			1		
65 - 75 ans	754	83	11,0	**2,71**	[1,83-4,02]		**2,43**	[1,63-3,64]	
> 75 ans	968	273	28,2	**8,50**	[6,06-12,21]		**6,44**	[4,40-9,43]	
Antécédent de cardiopathie						**< 10^{-4}**			**0,0486**
aucun	1 756	225	12,8	1			1		
angor	489	108	22,1	**1,93**	[1,49-2,49]		**1,40**	[1,06-1,84]	
infarctus du myocarde	370	62	6,8	**1,37**	[1,01-1,86]		1,21	[0,88-1,68]	
Localisation de l'infarctus						**0,0025**			**0,0073**
inférieure	1 078	133	12,3	1			1		
antérieure	1 288	225	17,5	**1,50**	[1,19-1,90]		**1,30**	[1,02-1,67]	
indéterminée	249	37	14,9	1,24	[0,84-1,84]		0,75	[0,50-1,13]	
Délai d'admission						**< 10^{-4}**			
≤ 6 heures	1 147	133	11,6	1					
plus de 6 heures	1 031	159	15,4	**1,39**	[1,08-1,78]				
indéterminé	356	82	23,0	**2,28**	[1,68-3,10]				
au cours de l'hospitalisation	81	21	25,9	**2,67**	[1,57-4,53]				
Procédure de revascularisation						**< 10^{-4}**			**0,0460**
sans	1 789	326	18,2	1			1		
thrombolyse pré-hospitalière	91	1	1,1	**0,05**	[0,01-0,36]		**0,12**	[0,02-0,88]	
thromb. hospit. ou angioplastie	735	68	9,3	**0,46**	[0,35-0,60]		0,80	[0,59-1,08]	
Secteur						**0,0261**			**0,0132**
Bordeaux-Arcachon-Blaye	893	133	14,9	1			1		
Libourne-Bergerac	238	52	21,8	**1,60**	[1,12-2,29]		1,46	[0,99-2,15]	
Périgueux-Sarlat	242	27	11,2	0,72	[0,46-1,12]		0,65	[0,41-1,02]	
Dax - Mont-de-Marsan	293	43	14,7	0,98	[0,68-1,43]		0,88	[0,68-1,43]	
Agen-Villeneuve	321	39	12,1	0,79	[0,54-1,16]		0,68	[0,60-1,32]	
Pau-Orthez	324	49	15,1	1,01	[0,71-1,45]		0,92	[0,63-1,35]	
Bayonne-Sud Landes	304	52	17,1	1,18	[0,83-1,68]		1,24	[0,85-1,80]	

[a] Odds-ratio *: les odds-ratios statistiquement différents de 1 sont notés en gras.*
[b] *Intervalle de confiance à 95 %.*

cependant statistiquement significative (p = 0,0081). Par contre, à âge, à sexe et à localisation de l'infarctus identiques, le risque de mortalité n'est pas plus élevé (statistiquement parlant) chez les patients ayant eu précédemment un infarctus du myocarde que chez des patients qui n'avaient aucun antécédent cardio-vasculaire. Le maintien conjoint des variables « procédure de revascularisation » et « délai de prise en charge » s'avère inutile. En effet, ces variables, très liées entre elles, sont redondantes dans le modèle. Seule la variable « procédure de revascularisation » considérée comme variable d'intérêt principal sera maintenue dans le modèle final. Par ailleurs, la thrombolyse pré-hospitalière se révèle comme étant la seule technique de revascularisation apportant un gain de survie : OR ajusté = 0,12 ; [0,01-0,36]. Enfin, en analyse univariée, les *odds-ratios* des différentes modalités de la variable « secteur sanitaire » varient de 0,72 à 1,60 en prenant comme référence le secteur où le centre hospitalo-universitaire est implanté (Bordeaux-Arcachon, Blaye ; OR = 1). Le secteur de Libourne-Bergerac apparaît comme ayant statistiquement plus de décès : OR brut = 1,60 ; [1,12-2,29].

En analyse multivariée permettant la prise en compte des caractéristiques de la population et de sa prise en charge, cette différence n'apparaît plus statistiquement significative : OR ajusté = 1,46 ; [0,998-2,15]. La variable doit, par contre, être maintenue dans le modèle, car il persiste entre les secteurs sanitaires des différences significatives en terme de mortalité.

Ainsi, la régression logistique multivariée a permis la comparaison des différents secteurs sanitaires en termes de mortalité hospitalière faisant suite à un infarctus aigu du myocarde, alors que les populations concernées n'étaient pas identiques tant pour ce qui concerne leurs caractéristiques générales que pour leur taux d'accès aux différentes procédures de revascularisation.

V. DISCUSSION

La régression logistique est un modèle d'analyse multivariée puissant qui permet d'analyser les relations entre la survenue d'un événement et chacun de ses facteurs associés tout en contrôlant les facteurs de confusion. Il permet de dépister aisément les interactions et leur prise en compte. Dans cet article, volontairement, le terme de « facteur associé » a été employé plutôt que celui de « facteur de risque ». En effet, le jugement de causalité requiert des conditions très précises auxquelles seules les études réalisées dans un objectif de recherche clinique peuvent satisfaire pleinement.

D'autres modèles d'analyses multivariées explicatives utilisent la fonction logistique et peuvent être considérés comme des « extensions » (au sens large du terme) du modèle de régression logistique : le modèle de régression logistique multinomiale qui permet l'emploi de variables dépendantes à plusieurs modalités, et le modèle de Cox qui permet d'étudier le délai de survenue d'un événement. La méthode d'analyse répond aux mêmes impératifs que ceux de la régression logistique.

L'ensemble de ces modèles appartient à la catégorie des analyses multivariées explicatives utilisées en épidémiologie qu'il faut bien distinguer des analyses multivariées descriptives (analyse en composantes principales, analyse factorielle des correspondances multiples, etc.) plus employées en Sciences humaines. Ces dernières ont, comme objectif, d'extraire l'information pertinente liée à la présence conjointe de nombreuses variables. Elles ne permettent pas de tenir compte des biais de confusion. Ce sont des modèles permettant une description des données. Ces deux catégories d'analyses, qualifiées toutes les deux de multivariées, diffèrent ainsi considérablement tant sur le plan de leur objectif que sur celui des principes statistiques sur lesquels elles reposent.

L'outil informatique et ses applications permettent de réaliser plus aisément ces analyses. Cependant, cette facilité d'emploi ne doit surtout pas faire oublier qu'une analyse de données n'aura de valeur qu'en réponse à un objectif pertinent. Les variables intégrées dans le modèle doivent être issues d'un raisonnement épidémiologique, clinique et les facteurs de confusion doivent avoir été recherchés. Chacune des différentes étapes de la méthode d'analyse doit avoir été scrupuleusement effectuée sous peine d'erreurs dans la conception et dans l'interprétation du modèle.

Parce qu'elles permettent de prendre en compte l'ensemble des biais de confusion connus et mesurés et l'ajustement sur d'autres facteurs associés, les analyses multivariées explicatives offrent un gain incontestable dans la compréhension des liens existants entre les différentes variables et donnent du sens à l'information délivrée. Il serait intéressant de développer leur utilisation dans le cadre des évaluations de pratiques médicales réalisées par le service médical de l'assurance maladie.

RÉFÉRENCES

1. Allemand H, Fender P. Un programme de santé publique pour une meilleure prise en charge des malades. Diabetes Metab 2000 ; 26 Suppl 6:7-9.

2. Weill A, Ricordeau P, Vallier N, Bourrel R, Fender P, Allemand H. Les modalités de suivi des diabétiques non insulino-traités en France métropolitaine durant l'année 1998. Diabetes Metab 2000 ; 26 Suppl 6: 39-48.

3. Fernandez L, Aminot I, Dupuy E, Degré A. Accès à une procédure de revascularisation chez les patients hospitalisés pour un infarctus aigu du myocarde en Aquitaine. Rev Med Ass Maladie 2001;32:219-26.

BIBLIOGRAPHIE

Beaucage C, Bonnier Viger Y, Aubin, M et al. Épidémiologie appliquée. Montréal : Gaëtan Morin ; 1996, p. 302-21.

Bouyer J, Hémon D, Cordier S et al. Épidémiologie. Principes et méthodes quantitatives. Paris : INSERM ; 1995, p. 227-90.

Dabis F, Drucker J, Moren A. Épidémiologie d'intervention. Paris : Arnette ; 1992, p.291-387.

Falissard B. Comprendre et utiliser les statistiques dans les sciences de la vie. Paris : Masson ; 1996, p.131-55.

Hosmer DW, Lemeshow S. Applied logistic regression. New-York : Wiley ; 1989.