[Dealing with missing, abnormal and incoherent data in E3N cohort study]

Stéphanie Garcia-Acosta, Françoise Clavel-Chapelon

Revue d'épidémiologie et de santé publique 1999;47(6):515-23

Gestion des données manquantes, aberrantes et incohérentes dans l'étude de cohorte E3N

Dealing with missing, abnormal and incoherent data in E3N cohort Study

S. GARCIA-ACOSTA[1,2], F. CLAVEL-CHAPELON[1]

[1]*INSERM U521, Institut Gustave-Roussy, rue Camille Desmoulins, 94805 Villejuif, France*

[2]*Laboratoires Rhône-Poulenc Rorer, 15, rue de la Vanne, 92545 Montrouge cedex, France*

Background: The E3N Study, "Etude Epidémiologique auprès de femmes de la Mutuelle Générale de l'Education Nationale", is a cohort study, aiming at studying cancer risk factors on 100,000 women. Even if the incidence of problematic (missing, incoherent, etc.) data is low, any multivariate analysis which would be based only on complete subjects would rely on a too small sample, which would not necessarily be representative of the studied population. Results could thus be biased. **Methods:** Our dealing with problematic data includes: 1) the identification of problematic data: locating these data, looking for their source and differentiating their process of existence, 2) the definition of the methodology, and 3) the implementation of the methods: deductive method, cold-deck, and multiple imputation for Missing At Random data. **Results:** We looked at the number of individuals on which an analysis on 19 variables could be undertaken. The management of missing data made exploitable one fourth of the cohort, i.e.74.6% of individuals instead of 50.5%. Moreover, for 89.0% of subjects, one variable at most (out of the 19 studied) has missing datum. **Conclusions:** The main difficulty does not stand so much in the choice and implementation of methods to deal with problematic data than in the identification of their process of existence. Most of what was gained was due to the simplest methods: cold-deck and deductive method.

Cohort ; missing data ; multiple imputation

Position du problème : E3N, Etude Epidémiologique auprès de femmes de la Mutuelle Générale de l'Education Nationale, est une étude de cohorte (100 000 femmes) prospective, ayant pour objectif l'étude des facteurs de risque de cancer chez la femme. Même si l'incidence des données problématiques (manquantes, incohérentes, etc.) est faible, une analyse statistique multivariée, se fondant uniquement sur les sujets à données complètes, ne porterait que sur un nombre insuffisant d'individus, et non nécessairement représentatif de la population étudiée, d'où des résultats potentiellement biaisés. **Méthodes :** La gestion des données problématiques mise en œuvre inclut : 1) l'identification des données problématiques : repérage, recherche de la source de l'existence de ces données et différenciation selon leur processus d'existence, 2) la définition de la méthodologie, et 3) l'application des méthodes retenues : méthode déductive, cold-deck, et imputation multiple pour les données Missing At Random. **Résultats :** Le bilan présenté ici a été effectué en termes d'individus exploitables pour une analyse particulière portant sur 19 variables. Ce travail de gestion des données problématiques a permis de gagner un quart de la cohorte, en passant de 50,5% d'individus exploitables à 74,6%. Et 89,0% des individus ne présentent alors qu'une variable à donnée manquante sur les 19 variables pré-sélectionnées. **Conclusions :** La difficulté fondamentale ne réside pas tant dans le choix et l'application des méthodes de gestion des données problématiques que dans la qualification de leur processus d'existence. L'essentiel du gain en termes de sujets exploitables est attribuable aux méthodes les plus simples, à savoir le « cold-deck » et la méthode déductive.

Cohorte ; données manquantes ; imputation multiple

Position du problème

L'étude E3N, Etude Epidémiologique auprès de femmes de la Mutuelle Générale de l'Education Nationale (MGEN), est une étude prospective ayant pour objectif l'étude des facteurs de risque de cancer chez la femme. Elle s'appuie sur une cohorte de quelques 100 000 femmes, âgées de 40 à 65 ans à l'inclusion en 1990 et adhérentes à la MGEN. Les participantes répondent régulièrement à des auto-questionnaires sur leur mode de vie et leur état de santé. Différents thèmes sont abordés suivant le questionnaire : grossesses, traitements hormonaux, alimentation…. Les questionnaires sont relativement longs et complexes, se voulant complets et précis sur le thème abordé, sans négliger les questions d'ordre général.

Du fait de leur longueur (multiplicité des items), de leur complexité et du caractère auto-administré, les réponses aux questionnaires comportent inévitablement des données manquantes, aberrantes ou incohérentes ; il existe aussi des données non reconnues à la lecture optique. Le nombre de ces données « problématiques » est tel qu'une analyse statistique qui se fonderait uniquement sur les sujets à données complètes ne porterait que sur des individus en nombre restreint, et surtout non nécessairement représentatif de la population étudiée, d'où des résultats qui risqueraient d'être biaisés.

Une réflexion sur les données problématiques a été réalisé en 1997 sur 19 variables du premier questionnaire, en vue d'une étude particulière. Nous nous proposons de présenter ici la démarche qui a été suivie.

Méthodes

Les méthodes retenues devaient permettre d'une part le repérage et la classification des données problématiques présentes dans les 19 variables explicatives pré-sélectionnées pour l'étude, et d'autre part leur traitement.

Typologies théorique et pratique des données manquantes et problématiques

La littérature statistique distingue trois types de données manquantes selon leur processus d'existence [1]. Cette distinction est de première importance puisque la validité des méthodes de gestion des données manquantes en dépend.

L'hypothèse la plus forte que l'on puisse faire sur les données manquantes est qu'elles sont *Missing Completely At Random* (M.C.A.R.). Les données manquantes d'une variable explicative X sont M.C.A.R. si la probabilité qu'une observation de X soit manquante est indépendante de sa valeur et de celles des autres variables recueillies, en particulier des autres variables explicatives pré-sélectionnées.

Une hypothèse moins contraignante est que les données manquantes sont seulement *Missing At Random* (M.A.R.). Dans ce cas, la probabilité qu'une observation de X soit manquante est fonction des valeurs prises par certaines covariables de X, i.e. par certaines autres variables recueillies. Mais elle ne doit pas être fonction de la vraie valeur de X. Ainsi, M.A.R. équivaut à « M.C.A.R. conditionnellement aux autres variables de la base de données ».

Enfin, certaines données manquantes sont *Missing Not At Random* (M.N.A.R). Le processus d'existence d'une donnée M.N.A.R., i.e. sa probabilité d'être manquante, dépend (au moins) de sa vraie valeur (et éventuellement des valeurs prises par les autres variables recueillies).

La première étape consiste donc à qualifier le processus d'existence des données problématiques de chaque variable.

En premier lieu, nous avons cherché les sources et types des données problématiques rencontrées dans l'étude E3N ; nous en avons distingué quatre. Les réponses par nature impossibles (une croix au lieu d'un chiffre) et les données absurdes sur le plan médical ou physiologique (un âge aux premières règles égal à 95 ans) ont été regroupées sous le terme de « *données aberrantes* ». Certaines données étaient manifestement *« incohérentes »* par rapport aux autres renseignements fournis par l'enquêtée (un âge aux premières règles supérieur à l'âge à la ménopause). Pour une non-réponse de l'enquêtée ou un défaut dans la saisie (des pages entières n'ont pas été lues par le lecteur optique alors qu'elles auraient dû l'être), nous avons choisi l'expression « *données manquantes car non rapportées* » ; les données problématiques relatives à ces 2 sources ont été regroupées, même si potentiellement elles n'obéissent pas au même type de processus d'existence, parce qu'elles ne pouvaient pas être distinguées dans la base de données (sauf à reprendre l'ensemble des 100000 questionnaires, variable par variable). Enfin, lorsqu'il y avait incompatibilité des réponses (à la fois oui et non en réponse à la question « avez-vous déjà été enceinte ? ») ou pour toute réponse illisible par le lecteur optique ou les opératrices de saisie, les données ont été dites « *manquantes car non reconnues* ». Nous avons choisi, dans l'étude E3N, de gérer toutes ces données problématiques comme données manquantes, mais chaque variable a été « flaguée » de manière à conserver l'information détaillée ci-avant. Ceci a permis une première approche du processus d'existence de certaines données manquantes : en effet, les données problématiques aberrantes, incohérentes et non reconnues peuvent raisonnablement être gérées comme des données M.C.A.R. dans la mesure où la femme a répondu (elle ne s'est pas dérobée devant la question).

En second lieu, des tests statistiques ont permis de départager les hypothèses M.C.A.R. contre M.A.R., même si en pratique les données M.C.A.R. sont particulièrement rares : pour cela, nous avons étudié les associations statistiques entre statut observationnel d'une variable (réponse problématique ou non) et les valeurs prises par les covariables. En revanche, l'hypothèse M.A.R. *versus* M.N.A.R. ne peut pas être testée,

en l'absence de données supplémentaires [2] : sauf dans le cadre d'une simulation, les valeurs sous-jacentes sont inconnues et cela empêche naturellement de tester si la probabilité d'être manquantes dépend des vraies valeurs. Une recherche bibliographique et des hypothèses comportementales ont alors permis d'apporter de nouveaux éléments pour la classification des données manquantes.

Revue des méthodes de gestion des données manquantes

La littérature statistique fournit des critiques de chaque méthode proposée selon les caractéristiques de la distribution des données manquantes et de leur processus d'existence [3-7].

Parmi les méthodes réputées « simples », on distingue la méthode déductive et la méthode dite « cold-deck », les méthodes d'analyse des sujets complets ou disponibles, les méthodes de prédiction par régression et la méthode de l'indicatrice.

La *méthode déductive* se fonde essentiellement sur des règles déterministes (une femme qui a accouché au moins une fois a nécessairement été enceinte). Le *cold-deck* consiste à rechercher les informations manquantes dans des enquêtes comparables portant sur les mêmes individus : dans le cas d'une étude de cohorte telle que E3N, il peut s'agir de données issues d'un autre questionnaire (date de naissance, fait d'avoir déjà subi une interruption volontaire de grossesse à une date donnée …).

La *méthode des sujets complets* (complete-case analysis ou listwise deletion) consiste à ne faire porter l'analyse que sur les sujets sans donnée manquante, tandis que la méthode des données disponibles (available-case analysis) utilise le plus grand nombre de sujets complets pour l'estimation individuelle de chaque paramètre.

Les *méthodes des moindres carrés* imputent des valeurs aux données manquantes des variables explicatives, sur la base de régressions soit *déterministes* (les résidus du modèle sont arbitrairement posés à 0), soit *stochastiques* (on attribue aux résidus des valeurs vraisemblables, par exemple générées aléatoirement) expliquant la variable explicative par d'autres variables recueillies. On peut choisir par exemple de remplacer la donnée manquante soit par la moyenne observée de la variable concernée sur l'ensemble de la population (*méthode de la moyenne*), ce qui correspond à une régression avec pour variables explicatives seulement un terme constant et un terme d'erreur ; soit par la valeur observée pour un individu répondeur, le « donneur » (hot-deck), choisi en fonction de sa distance (en termes de profil, pour un certain nombre d'autres variables à déterminer) à l'individu receveur pour lequel on doit estimer la donnée manquante ; soit par le ratio $y_i=x_i*y_r/x_r$, où y_r et x_r sont respectivement la moyenne des y_i (variable à données manquantes) et celle des x_i (autre variable recueillie supposée proportionnelle à y_i) sur l'ensemble des répondants (*prédiction par le ratio*) ; soit enfin par l'estimation de la moyenne conditionnellement aux valeurs observées des autres variables recueillies (*prediction-imputation method*) , ce qui correspond à une régression avec pour variables explicatives un terme constant, des covariables (parmi les autres variables recueillies) et un terme d'erreur.

Enfin, la *méthode de l'indicatrice* (*indicator method*) associe chaque variable présentant des données manquantes à sa variable indicatrice de données manquantes dans le terme explicatif du modèle final de régression expliquant la variable « end-point ».

Les méthodes plus élaborées se classent en trois groupes : les méthodes du maximum de vraisemblance (ou pseudo-vraisemblance), les méthodes à imputations multiples, et les méthodes à équations d'estimation pondérées.

Parmi les *méthodes du maximum de vraisemblance ou pseudo-vraisemblance*, la plus connue utilise l'algorithme EM [6]. Il s'agit de modéliser la variable « end-point » en fonction des variables explicatives sur les individus sans donnée manquante (1ère étape M) : on obtient alors une première estimation des coefficients des variables explicatives. Ensuite, le modèle obtenu est appliqué aux individus à donnée manquante pour certaines variables explicatives pour estimer leur donnée manquante à partir de la connaissance de leur variable « end-point » et de l'estimation des paramètres du modèle (1ère étape E). Une 2ème étape M est conduite sur l'ensemble des individus, avec pour données des variables explicatives la réunion des données observées et des données estimées lors de la précédente étape. Une nouvelle estimation des données manquantes est réalisée à partir de cette nouvelle estimation des paramètres du modèle (2ème étape E), et ainsi de suite jusqu'à l'obtention de paramètres stables.

Après avoir modélisé la probabilité de donnée manquante pour la variable explicative X en fonction des autres variables recueillies (*propensity score*) et après l'avoir appliquée à chaque individu (chaque individu est donc caractérisé par la probabilité, calculée *a posteriori* d'après ce modèle, qu'il avait de présenter une donnée manquante pour la variable X), l'*imputation multiple* consiste à générer plusieurs (n) copies de la base originelle. Dans chaque copie, la population est divisée en intervalles inter-quantiles de la probabilité calculée de donnée manquante, et la sous-distribution de la variable X observée chez les

répondeurs d'un tel intervalle est appliquée aux non-répondeurs du même intervalle. Ainsi, les données qui vont être imputées aux individus à données manquantes pour X et appartenant à cet intervalle sont générées stochastiquement à partir de la distribution empirique de X observée chez les individus sans donnée manquante de cet intervalle. Les analyses sont alors conduites sur chacune des bases de données imputées et les paramètres finaux du modèle de l'analyse sont calculés à partir des paramètres relatifs à chaque base de données [6] : le coefficient θ estimé d'une variable explicative est égal à la moyenne des coefficients θ_i (i décrivant l'ensemble des indices de chacune des bases de données imputées) estimés de cette variable dans chacune des analyses portant respectivement sur chacune des bases de données ($\theta = \sum_{i=1 \text{ à } n} \theta_i / n$), et la variance estimée V du coefficient θ est égale à la somme de la moyenne des variances estimées V_i relatives aux coefficients θ_i et à un terme reflétant l'incertitude quant au procédé d'imputation. $V = (\sum_{1 \le i \le n} V_i)/n + (1+1/n) * [\sum_{1 \le i \le n} (\theta_i-\theta)^2 / (n-1)]$

Les *méthodes à équations d'estimation pondérées* utilisent une modélisation du processus d'existence des données manquantes afin d'attribuer des poids aux covariables pour l'analyse par régression de la variable « end-point ».

Les propriétés de ces différentes méthodes en terme de biais des paramètres estimés et de leurs écarts-types sont résumées dans le *tableau 1*. Néanmoins, en deçà de 5% de données manquantes, quel qu'en soit le type de processus d'existence, les données peuvent être gérées comme si elles obéissaient à un processus M.C.A.R., donc par les méthodes les plus simples, sans risque de biais important.

Tableau 1. **Propriétés des différentes méthodes de gestion des données problématiques.**

Méthodes	Propriétés et caractéristiques	Utilisation
Déduction Cold-deck	Très simples et très fiables	A utiliser en priorité
Sujets complets Données disponibles	Biais si données non M.C.A.R.	Données M.C.A.R.
Prédiction par régression : - déterministe ou stochastique - moyenne - hot-deck - ratio - prédiction-imputation	Biais si données non M.C.A.R - stochastique préférable à déterministe - distorsion de la distribution et sous-estimation de la variance empirique - le mécanisme de réponse doit être globalement uniforme - la distribution doit être identique entre répondeurs et non-répondeurs - les écarts-types des coefficients finaux sont sous-estimés, mais moins qu'avec la moyenne ; l'application de formules de correction peut s'avérer complexe	Données M.C.A.R. Taux de non-réponses < 40%
Indicatrice	Biais même sous M.C.A.R. [1] et [2]	Jamais.
Maximisation de la (pseudo-) vraisemblance	Biais si données M.N.A.R.	Données M.C.A.R. ou M.A.R.
Imputations multiples	Biais si données M.N.A.R. Bonne estimation de la variance des coefficients estimés (pour un modèle multivarié normal).	Données M.C.A.R. ou M.A.R. Taux de non-réponses < 40%
Equations d'estimation pondérées	Les poids dérivent des processus d'existence des données problématiques : les coefficients de la modélisation sont très sensibles à la formulation de ceux-ci, qui sont malheureusement peu	A utiliser quand les processus d'existence sont connus avec certitude.

fiables.

Choix méthodologiques

Le premier travail a consisté à repérer les données problématiques. Afin de privilégier robustesse et simplicité, nous avons eu recours en première intention à la méthode déductive et à la méthode « cold-deck » sur l'ensemble des variables du questionnaire.

Les traitements ultérieurs ont été choisis en fonction du processus d'existence des données problématiques irrésolues et des caractéristiques des variables concernées. Pour les variables à données M.A.R., l'imputation multiple a été préférée à la maximisation de la pseudo-vraisemblance : ces deux méthodes ont en fait les mêmes propriétés mais l'imputation multiple a l'avantage d'intervenir en amont de la modélisation, de sorte que l'étape de modélisation s'en trouve simplifiée et que toutes les analyses peuvent être conduites sur les mêmes bases de données. Nous avons choisi de réaliser trois imputations car cela constitue généralement un bon compromis pour obtenir une bonne estimation de la variance des paramètres, sans trop multiplier les bases qui sont déjà de grande taille.

Nous avons ainsi réservé l'étape de la modélisation à la gestion des données problématiques qu'il n'aura pas été possible de traiter auparavant, dites données problématiques irrésolues. Cette étape consiste à réaliser les analyses sur chacune des 3 bases de données de manière distincte et à intégrer leurs résultats au calcul des paramètres estimés finaux et de leurs écarts-types. Cette étape n'a pas encore eu lieu car les données problématiques restantes de type potentiellement M.N.A.R. n'ont pas encore été traitées.

Résultats

Le bilan de chaque étape a pu être exprimé en termes d'individus exploitables pour l'étude envisagée, c'est-à-dire ceux pour lesquels les 19 variables classées prioritaires pour cette étude étaient disponibles. De la base initiale ont été écartés les questionnaires pour lesquels au moins une page entière était vide, ainsi que les individus ne respectant pas le protocole au regard de leur date de naissance. Nous avons ainsi travaillé sur 99 694 individus.

Dans la pratique, le repérage des données problématiques ainsi que leur gestion par la méthode « cold-deck » et la méthode déductive ont porté sur toutes les variables de la base de données et pas seulement sur les 19 variables pré-sélectionnées dont la liste est donnée au *tableau 2*. En effet, il était plus efficace de progresser parallèlement sur toutes les variables, dans la mesure où des variables non pré-sélectionnées intervenaient dans le repérage mais également dans la gestion des données problématiques des 19 variables pré-sélectionnées ; par exemple, si le nombre de grossesses (variable explicative non pré-sélectionnée) était nul, alors le nombre d'enfants nés vivants (variable explicative pré-sélectionnée) devait être nul, cette relation logique servant à la fois pour repérer une éventuelle incohérence et pour imputer la valeur 0 à la 2ème variable si elle était manquante alors que la 1ère était renseignée à 0. Parfois, les relations logiques impliquaient plus de 2 variables simultanément et il a fallu, en cas d'incohérence, décider quelles données privilégier.

Le *tableau 2* rend compte des pourcentages de données problématiques repérées et des non résolues par l'emploi de la méthode « cold-deck » et de la méthode déductive, ventilés selon la typologie que nous avons définie précédemment. Le repérage des données problématiques a permis de constater qu'avant tout traitement, seuls 50 358 individus étaient exploitables. Les premiers traitements seuls, à savoir les corrections logiques, déductions et « cold-deck », apportaient déjà 20 025 individus supplémentaires, ce qui portait à 70 383 le nombre d'individus pouvant être inclus dans l'analyse, soit 70,6 % des individus de la base.

Ce n'est qu'après l'application des méthodes « cold-deck » et déductive que l'on a considéré les données comme fiables, et que le processus d'existence des données manquantes des 19 variables pré-sélectionnées a été examiné.

Par souci de simplicité, nous n'avons pas cherché à qualifier le processus d'existence des données problématiques d'une variable par type (aberrantes, incohérentes, non reconnues, et non rapportées), mais globalement : c'est-à-dire que nous n'avons pas essayé de répondre à la question « Quel est le processus d'existence des données aberrantes (respectivement incohérentes, non reconnues et non rapportées) de la variable X ? » mais à la question « Globalement, quel est le processus d'existence des données problématiques de la variable X ? ». Ceci ne devait pas entraîner d'utilisation abusive des méthodes de

gestion dans la mesure où les 3 types théoriques de données manquantes sont emboîtés et où, par conséquent, l'application d'une méthode valable pour un type de processus est également valable pour un type plus simple (par exemple, une méthode applicable aux données M.A.R. est applicable aux données M.C.A.R.).

Tableau 2. **Pourcentages de données problématiques repérées dans les 19 variables pré-sélectionnées avant toute gestion (% av.) et après cold-deck et déductions (% ap.), ventilés selon la typologie pratique.**

Variables	Données aberrantes		Données incohérentes		Données non rapportées		Données non reconnues		Données problématiques	
	% av.	% ap.	% av.	% ap.	% av.	% ap.	% av.	% ap.	% av.	% ap.
Ancienneté du statut marital actuel					5,6	5,5	0,1	0,1	5,7	5,6
Taille	1,9	1,2			0,4	0,4			2,3	1,6
Poids	0,2	0,2	1,0	0,8	1,9	1,9			3,1	2,9
Tabagisme présent ou passé					2,9	2,9	0,1	0,1	3,0	3,0
Allaitement dans le passé			0,1	0,0	2,1	1,2	0,2	0,0	2,4	1,2
Âge aux premières règles					0,9	0,9	2,0	2,0	2,9	2,9
Nombre de partenaires sexuels			0,2	0,2	4,2	4,2	0,1	0,1	4,3	4,3
Nombre de fausses couches spontanées ou de grossesse extra-utérines			0,2	0,2	15,3	7,7	0,1	0,1	15,6	8,0
Méthodes contraceptives avant la première grossesse					9,8	9,6			9,8	9,6
Utilisation de la pilule avant la première grossesse					9,8	0,0			9,8	0,0
Méthodes contraceptives entre les grossesses					22,5	16,4			22,5	16,4
Utilisation de la pilule entre les grossesses					22,5	0,0			22,5	0,0
Méthodes contraceptives après la dernière grossesse					13,8	7,7			13,8	7,7
Utilisation de la pilule après la dernière grossesse					13,8	0,0			13,8	0,0
Recours à un traitement contre la stérilité de la femme ou de son conjoint					2,1	1,7			2,1	1,7
Types de traitement de la stérilité pour la femme					0,3	0,3			0,3	0,3
Nombre d'interruptions volontaires de grossesse			0,8	0,8	13,5	5,9	0,1	0,1	14,4	6,8
Âge lors de la première grossesse	1,9	0,3			1,1	1,1	0,1	0,1	3,1	1,5
Nombre d'enfants nés vivants					2,3	2,3	0,1	0,1	2,4	2,4

Lors de cette étape, dans la mesure où il nous était impossible de départager les hypothèses M.N.A.R. et M.A.R. pour une variable explicative donnée, puisque nous ne disposions pas de ses vraies valeurs, nous sommes restés prudents et avons introduit une réflexion sur les comportements des femmes face aux questions, argumentée autant que possible par la littérature. Ainsi, plusieurs des 19 variables (par exemple le nombre de partenaires sexuels ou le poids) ont été classées parmi les variables à données problématiques de type M.N.A.R. parce que nous suspections des comportements de non-réponse volontaire liés aux valeurs extrêmes. Seules 3 variables ont été classées parmi celles à données problématiques pouvant être gérées comme des données M.AR. ou M.C.AR. : le tabagisme présent ou passé, l'allaitement dans le

passé et l'âge aux premières règles. Le fait de fumer ou d'avoir fumé n'obéirait pas à un processus de données M.N.A.R. : la consommation est dans certaines études sur-estimée [8], dans d'autres sous-estimée [9-12]. La question de l'allaitement n'était pas susceptible de donner lieu à un comportement de non-réponse dépendant de sa vraie valeur. Enfin, dans le cas de l'aménorrhée primaire, la probabilité pour que l'âge aux premières règles manque pourrait dépendre de la réponse cachée. Mais dans la mesure où ce phénomène est très rare, l'hypothèse M.N.A.R. a été écartée. Quant à la distinction entre les hypothèses M.A.R. et M.C.A.R., la modélisation de la probabilité pour que la donnée soit manquante en fonction d'autres variables recueillies a montré l'existence d'associations statistiques significatives ; par conséquent, les 3 variables tabagisme présent ou passé, allaitement dans le passé et âge aux premières règles ont été classées comme variables à données M.A.R.

Ensuite, les données problématiques (supposées M.A.R.) de ces 3 variables ont été gérées par imputation multiple. Les 3 variables concernées par ce traitement présentaient 3,0%, 1,2% et 2,9% (respectivement pour les variables tabagisme présent ou passé, allaitement dans le passé et âge aux premières règles*)* de valeurs manquantes, c'est-à-dire moins de 5% ; nous avons tout de même choisi d'utiliser la méthode que nous avions retenue pour gérer les données problématiques de type M.A.R. pour ne pas multiplier les sources de biais, même mineurs pris isolément.

Pour modéliser le logit de la probabilité du statut observationnel (donnée renseignée / donnée manquante) d'une telle variable, par exemple le tabagisme, les variables explicatives ont été entrées sous forme de classes, y compris la classe des données problématiques, les variables explicatives étant elles aussi touchées par ce phénomène. Le *tableau 3* présente les variables explicatives retenues dans le modèle du logit de la probabilité du statut observationnel de la variable tabagisme. Le modèle a été appliqué à chaque individu pour que soit calculée sa probabilité estimée de présenter une donnée manquante à la variable tabagisme. L'échantillon a alors été découpé en intervalle inter-déciles selon la probabilité estimée de chaque individu.

Tableau 3. **Liste des variables incluses dans le modèle du logit de la probabilité que la variable tabagisme soit manquante.**

Date de naissance	Âge à la première grossesse
Âge aux premières règles	Régularité des règles
Âge au premier rapport sexuel	Nombre de partenaires sexuels
Contraception avant la première grossesse et après la dernière grossesse	Stérilité
Allaitement	Activité physique[a]
Niveau d'études de la femme	Niveau d'études de son conjoint
Nombre d'années de vie commune	Statut ménopausique
Type de ménopause	Pénibilité du travail[b]
Tension artérielle	Cholestérol
Morphologie	

[a]Sports pratiqués de façon intensive ou activité vigoureuse, sports pratiqués de façon modérée ou activité modérée, à l'adolescence et actuellement.
[b]physique et nerveuse.

.

Les données manquantes ont ensuite été remplacées, dans chacune des 3 bases de données, par des données générées d'après la distribution observée sur l'intervalle inter-déciles de chaque individu. La variable tabagisme étant binaire, nous avons généré aléatoirement la réponse de chaque individu du premier intervalle inter-déciles selon une loi de Bernouilli paramétrée par le taux de réponses affirmatives observées chez les répondeurs de cet intervalle, et fait de même sur les autres intervalles. Le *tableau 4* donne les taux de réponses affirmatives propres à chaque intervalle inter-déciles.

L'imputation multiple a complété les observations de 3 942 individus, soit 4,0 % à cette étape.

Tableau 4. **Taux de réponses affirmatives à la variable tabagisme selon l'intervalle inter-déciles de la probabilité estimée que la donnée soit « non problématique ».**

Intervalles inter-déciles de la probabilité estimée que la donnée tabagisme soit non problématique	Taux de réponses affirmatives dans l'intervalle inter-déciles
[0 ; 0,9556 [	34,0
[0,9556 ; 0,9648 [	34,9
[0,9648 ; 0,9694 [	35,3
[0,9694 ; 0.9725 [	36,0
[0,9725 ; 0,9748 [	35,5
[0,9748 ; 0,9769 [	34,9
[0,9769 ; 0,9788 [	33,5
[0,9788 ; 0,9809 [	33,4
[0,9809 ; 0,9833 [	32,8
[0,9833 ; 1]	29,6

Les différents traitements ont donc permis de passer des 50,5 % d'individus exploitables à 74,6%. Enfin, puisque la modélisation de la variable « end-point » n'impliquera pas nécessairement toutes les variables présélectionnées, il est important de souligner que 89,0 % des individus avaient au plus une variable indisponible à la fin de ce premier travail de gestion des données problématiques.

Conclusions

Au-delà d'une taille critique d'échantillon, le retour à la source pour gérer les données problématiques est pratiquement impossible. Le travail présenté ici sur les données de l'étude E3N propose une gestion alternative des données problématiques. Ainsi, la méthodologie retenue a-t-elle permis de passer de la moitié des individus exploitables aux trois quarts ; soulignons que ce gain est essentiellement attribuable aux méthodes les plus simples, faisant appel à la seule logique.

L'un des points fondamentaux en matière de gestion des données manquantes est la qualification de leur processus d'existence. Malheureusement, cette qualification n'est pas toujours réalisable, faute de techniques, d'information, mais aussi de temps car cette étape peut s'avérer particulièrement longue. De plus, la pratique se heurte à la question pour le moment irrésolue de la gestion des données M.N.A.R.

Le choix des méthodes de gestion des données manquantes dépend de la taille du biais susceptible d'être induit. Ainsi, on considère généralement qu'en deçà de 5% de données manquantes il n'est pas nécessaire d'adopter des méthodes raffinées pour gérer des données manquantes, y compris à processus d'existence informatif, car le risque de biais est bas. De même on pourrait assimiler des données manquantes M.N.A.R. à des données M.A.R. (et les gérer comme telles avec un risque de biais suffisamment faible) sous la condition suivante : que le statut observationnel de la variable à données manquantes ~~(i.e. sa probabilité d'être problématique)~~ ne soit expliqué que dans une faible proportion (seuil à définir) par la variable elle-même en comparaison des autres variables recueillies. Ce pourrait être en particulier le cas lorsque les variables recueillies sont aussi nombreuses et mutuellement corrélées qu'elles le sont dans l'étude E3N.

Enfin, sur les données demeurées problématiques, nous avons eu recours aux méthodes de gestion des données manquantes pour données transversales parce qu'il s'agissait du premier questionnaire posé. D'autres possibilités s'offrent pour la gestion des données problématiques des questionnaires suivants, en ayant recours aux méthodes adaptées aux données longitudinales, telle la méthode de Diggle et Kenward [13], qui intègre l'informativité du processus d'existence des abandons.

Remerciements : Les auteurs sont redevables aux participantes E3N, à toute l'équipe E3N, plus particulièrement à V. Avenel, S. Clech, C. Le Corre, K. Le Denmat, ainsi qu'à P. Druilhet (ENSAI) et JP. Pignon (IGR). L'étude E3N est réalisée grâce au soutien de la Ligue Contre le Cancer, l'Union Européenne, la société 3M, la M.G.E.N. et l'I.N.S.E.R.M.

Références

[1] RUBIN DB. Inference and missing data : *Biometrika* 1976 , *63*: 581-92.
[2] GREENLAND S., FINKLE W. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analysis. *Am. J. Epidemiol.* 1995, *142*: 1255-64.
[3] CARON N. Les Principales Techniques de correction de la non-réponse et les modèles associés. *I.N.S.E.E. Série des Documents de Travail. 'Méthodologie statistique.*
[4] SCHUMACHER M., VACH W., BLETTNER M. Workshop on Statistical Methods for Incomplete Covariate Data in Clinical and Epidemiological Studies. *Stat. Med.* 1997, *16*: n°1/2/3.
[5] LITTLE RJA. Regression With Missing X's : A Review. *J. Am. Stat. Assoc.* 1992, *87*: 1227-37.
[6] LITTLE RJA, RUBIN DB. : Statistical Analysis With Missing Data. New York, NY : *John Wiley & Sons*, 1987.
[7] VACH W., BLETTNER M. : Logistic Regression with Incompletely Observed Categorical Covariates. Investigating the Sensitivity against Violation of the Missing at random Assumption. *Stat. Med.* 1995, *14*: 1315-29.
[8] CLARK P.I., GAUTAM S.P., HLAING W.M., GERSON L.W. : Response Error in Self-Reported Current Smoking Frequency by Black and White Established Smokers. *Ann Epidemiol* 1996, *6*: 483-9.
[9] WELLS A.J., ENGLISH P.B., POSNER S.F., WAGENKNECHT L.E., PEREZ-STABLE E.J. : Misclassification rates for current smokers misclassified as nonsmokers. *Am J Public Health* 1998, *88*: 1503-9.
[10] PEREZ-STABLE E.J. , MARIN B.V., MARIN G., BRODY D.J., BENOWITZ N.L. : Apparent underreporting of cigarette consumption among Mexican American smokers. *Am J Public Health* 1990, *80*: 1057-61.
[11] HATZIANDREU E.J., PIERCE J.P., FIORE M.C., GRISE V., NOVOTNY T.E., DAVIS R.M. : The reliability of self-reported cigarette consumption in the United States. *Am J Public Health* 1989, *79*: 1020-3.
[12] HANSEN K.S. : Validity of occupational exposure and smoking data obtained from surviving spouses and colleagues. *Am J Ind Med* 1996, *30*: 392-7.
[13] DIGGLE PJ, KENWARD MG. : Informative Drop-out in Longitudinal Analysis. *Applied Statistics* 1994, *43*: 49-93.