La gestion des données manquantes dans les études observationnelles publiées dans cinq journaux de renom : pratique et manière de rapporter les résultats

Pierre Blavier

UFR DE SANTE DE ROUEN NORMANDIE

ANNEE 2021 **N°**

THESE POUR LE
DOCTORAT EN MEDECINE

Spécialité Santé Publique et Médecine Sociale

Par
BLAVIER Pierre
NE LE 20/12/1994 A Mont Saint Aignan

PRESENTEE ET SOUTENUE PUBLIQUEMENT LE 25/10/2021

La gestion des données manquantes dans les études observationnelles publiées dans cinq journaux de renom : pratique et manière de rapporter les résultats

PRESIDENT DE JURY : Professeur Jacques BENICHOU
DIRECTEUR DE THESE : Docteur Thibaut PRESSAT LAFFOUILHERE
MEMBRES DU JURY : Professeur Stéfan DARMONI
Docteur Joël LADNER

ANNEE UNIVERSITAIRE 2020 - 2021

U.F.R. SANTÉ DE ROUEN

DOYEN : **Professeur Benoît VEBER**

ASSESSEURS : **Professeur Loïc FAVENNEC**

Professeur Agnès LIARD

Professeur Guillaume SAVOYE

I - **MEDECINE**

PROFESSEURS DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mr Frédéric **ANSELME**	HCN	Cardiologie
Mme Gisèle **APTER**	Havre	Pédopsychiatrie
Mme Isabelle **AUQUIT AUCKBUR**	HCN	Chirurgie plastique
Mr Jean-Marc **BASTE**	HCN	Chirurgie Thoracique
Mr Fabrice **BAUER**	HCN	Cardiologie
Mme Soumeya **BEKRI**	HCN	Biochimie et biologie moléculaire
Mr Ygal **BENHAMOU**	HCN	Médecine interne
Mr Jacques **BENICHOU**	HCN	Bio statistiques et informatique médicale
Mr Olivier **BOYER**	UFR	Immunologie
Mme Sophie **CANDON**	HCN	Immunologie
Mr François **CARON**	HCN	Maladies infectieuses et tropicales
Mr Philippe **CHASSAGNE**	HCN	Médecine interne (gériatrie)
Mr Moïse **COEFFIER**	HCN	Nutrition
Mr Vincent **COMPERE**	HCN	Anesthésiologie et réanimation chirurgicale
Mr Jean-Nicolas **CORNU**	HCN	Urologie
Mr Antoine **CUVELIER**	HB	Pneumologie
Mr Jean-Nicolas **DACHER**	HCN	Radiologie et imagerie médicale
Mr Stéfan **DARMONI**	HCN	Informatique médicale et techniques de communication
Mr Pierre **DECHELOTTE**	HCN	Nutrition
Mr Stéphane **DERREY**	HCN	Neurochirurgie

Mr Frédéric **DI FIORE**	CHB	Cancérologie
Mr Fabien **DOGUET**	HCN	Chirurgie Cardio Vasculaire
Mr Jean **DOUCET**	SJ	Thérapeutique - Médecine interne et gériatrie
Mr Bernard **DUBRAY**	CHB	Radiothérapie
Mr Frank **DUJARDIN**	HCN	Chirurgie orthopédique - Traumatologique
Mr Fabrice **DUPARC**	HCN	Anatomie - Chirurgie orthopédique et traumatologique
Mr Eric **DURAND**	HCN	Cardiologie
Mr Bertrand **DUREUIL**	HCN	Anesthésiologie et réanimation chirurgicale
Mme Hélène **ELTCHANINOFF**	HCN	Cardiologie
Mr Manuel **ETIENNE**	HCN	Maladies infectieuses et tropicales
Mr Thierry **FREBOURG**	UFR	Génétique
Mr Pierre **FREGER** *(surnombre)*	HCN	Anatomie - Neurochirurgie
Mr Jean François **GEHANNO**	HCN	Médecine et santé au travail
Mr Emmanuel **GERARDIN**	HCN	Imagerie médicale
Mme Priscille **GERARDIN**	HCN	Pédopsychiatrie
M. Guillaume **GOURCEROL**	HCN	Physiologie
Mr Dominique **GUERROT**	HCN	Néphrologie
Mme Julie **GUEUDRY**	HCN	Ophtalmologie
Mr Olivier **GUILLIN**	HCN	Psychiatrie Adultes
Mr Claude **HOUDAYER**	HCN	Génétique
Mr Fabrice **JARDIN**	CHB	Hématologie
Mr Luc-Marie **JOLY**	HCN	Médecine d'urgence
Mr Pascal **JOLY**	HCN	Dermato – Vénéréologie
Mme Bouchra **LAMIA**	Havre	Pneumologie
Mme Annie **LAQUERRIERE**	HCN	Anatomie et cytologie pathologiques
Mr Vincent **LAUDENBACH**	HCN	Anesthésie et réanimation chirurgicale
Mr Hervé **LEFEBVRE**	HB	Endocrinologie et maladies métaboliques
Mr Thierry **LEQUERRE**	HCN	Rhumatologie
Mme Anne-Marie **LEROI**	HCN	Physiologie
Mr Hervé **LEVESQUE**	HCN	Médecine interne
Mme Agnès **LIARD-ZMUDA**	HCN	Chirurgie Infantile
Mr Pierre Yves **LITZLER**	HCN	Chirurgie cardiaque
M. David **MALTETE**	HCN	Neurologie
Mr Christophe **MARGUET**	HCN	Pédiatrie
Mme Isabelle **MARIE**	HCN	Médecine interne
Mr Jean-Paul **MARIE**	HCN	Oto-rhino-laryngologie
Mr Loïc **MARPEAU**	HCN	Gynécologie - Obstétrique
Mr Stéphane **MARRET**	HCN	Pédiatrie

Mme Véronique **MERLE**	HCN	Epidémiologie
Mr Pierre **MICHEL**	HCN	Hépato-gastro-entérologie
M. Benoit **MISSET** *(détachement)*	HCN	Réanimation Médicale
Mr Marc **MURAINE**	HCN	Ophtalmologie
Mr Christian **PFISTER**	HCN	Urologie
Mr Jean-Christophe **PLANTIER**	HCN	Bactériologie - Virologie
Mr Didier **PLISSONNIER**	HCN	Chirurgie vasculaire
Mr Gaëtan **PREVOST**	HCN	Endocrinologie
Mr Jean-Christophe **RICHARD** *(détachement)*	HCN	Réanimation médicale - Médecine d'urgence
Mr Vincent **RICHARD**	UFR	Pharmacologie
Mme Nathalie **RIVES**	HCN	Biologie du développement et de la reproduction
Mr Horace **ROMAN** *(détachement)*	HCN	Gynécologie - Obstétrique
Mr Jean-Christophe **SABOURIN**	HCN	Anatomie – Pathologie
Mr Mathieu **SALAUN**	HCN	Pneumologie
Mr Guillaume **SAVOYE**	HCN	Hépato-gastrologie
Mme Céline **SAVOYE–COLLET**	HCN	Imagerie médicale
Mme Pascale **SCHNEIDER**	HCN	Pédiatrie
Mr Lilian **SCHWARZ**	HCN	Chirurgie Viscérale et Digestive
Mr Michel **SCOTTE**	HCN	Chirurgie digestive
Mme Fabienne **TAMION**	HCN	Thérapeutique
Mr Luc **THIBERVILLE**	HCN	Pneumologie
Mr Hervé **TILLY** *(surnombre)*	CHB	Hématologie et transfusion
M. Gilles **TOURNEL**	HCN	Médecine Légale
Mr Olivier **TROST**	HCN	Anatomie -Chirurgie Maxillo-Faciale
Mr Jean-Jacques **TUECH**	HCN	Chirurgie digestive
Mr Benoît **VEBER**	HCN	Anesthésiologie - Réanimation chirurgicale
Mr Pierre **VERA**	CHB	Biophysique et traitement de l'image
Mr Eric **VERIN**	Les Herbiers	Médecine Physique et de Réadaptation
Mr Eric **VERSPYCK**	HCN	Gynécologie obstétrique
Mr Olivier **VITTECOQ**	HC	Rhumatologie
Mr David **WALLON**	HCN	Neurologie
Mme Marie-Laure **WELTER**	HCN	Physiologie

MAITRES DE CONFERENCES DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mme Najate **ACHAMRAH**	HCN	Nutrition
Mme Elodie **ALESSANDRI-GRADT**	HCN	Virologie
Mme Noëlle **BARBIER-FREBOURG**	HCN	Bactériologie – Virologie
Mr Emmanuel **BESNIER**	HCN	Anesthésiologie - Réanimation
Mme Carole **BRASSE LAGNEL**	HCN	Biochimie
Mme Valérie **BRIDOUX HUYBRECHTS**	HCN	Chirurgie Vasculaire
Mr Gérard **BUCHONNET**	HCN	Hématologie
Mme Mireille **CASTANET**	HCN	Pédiatrie
Mme Nathalie **CHASTAN**	HCN	Neurophysiologie
M. Vianney **GILARD**	HCN	Neurochirurgie
Mr Serge **JACQUOT**	UFR	Immunologie
Mr Joël **LADNER**	HCN	Epidémiologie, économie de la santé
Mr Jean-Baptiste **LATOUCHE**	UFR	Biologie cellulaire
M. Florent **MARGUET**	HCN	Histologie
Mme Chloé **MELCHIOR**	HCN	Gastroentérologie
M. Sébastien **MIRANDA**	HCN	Chirurgie Vasculaire
Mr Thomas **MOUREZ** *(détachement)*	HCN	Virologie
Mr Gaël **NICOLAS**	UFR	Génétique
Mme Muriel **QUILLARD**	HCN	Biochimie et biologie moléculaire
Mme Laëtitia **ROLLIN**	HCN	Médecine du Travail
Mme Pascale **SAUGIER-VEBER**	HCN	Génétique
M. Abdellah **TEBANI**	HCN	Biochimie et Biologie Moléculaire
Mme Anne-Claire **TOBENAS-DUJARDIN**	HCN	Anatomie
Mr Julien **WILS**	HCN	Pharmacologie

PROFESSEUR AGREGE OU CERTIFIE

Mr Thierry **WABLE**	UFR	Communication
Mme Mélanie **AUVRAY-HAMEL**	UFR	Anglais

ATTACHE TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE à MI-TEMPS

Mme Justine **SAULNIER**	UFR	Biologie

II - PHARMACIE

PROFESSEURS DES UNIVERSITES

Mr Jérémy **BELLIEN** (PU-PH)	Pharmacologie
Mr Thierry **BESSON**	Chimie Thérapeutique
Mr Jean **COSTENTIN** (Professeur émérite)	Pharmacologie
Mme Isabelle **DUBUS**	Biochimie
Mr Abdelhakim **EL OMRI**	Pharmacognosie
Mr François **ESTOUR**	Chimie Organique
Mr Loïc **FAVENNEC** (PU-PH)	Parasitologie
Mr Jean Pierre **GOULLE** (Professeur émérite)	Toxicologie
Mme Christelle **MONTEIL**	Toxicologie
Mme Martine **PESTEL-CARON** (PU-PH)	Microbiologie
Mr Rémi **VARIN** (PU-PH)	Pharmacie clinique
Mr Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

MAITRES DE CONFERENCES DES UNIVERSITES

Mme Cécile **BARBOT**	Chimie Générale et Minérale
Mr Frédéric **BOUNOURE**	Pharmacie Galénique
Mr Thomas **CASTANHEIRO MATIAS**	Chimie Organique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Camille **CHARBONNIER (LE CLEZIO)**	Statistiques
Mme Elizabeth **CHOSSON**	Botanique
Mme Marie Catherine **CONCE-CHEMTOB**	Législation pharmaceutique et économie de la santé
Mme Cécile **CORBIERE**	Biochimie
Mme Nathalie **DOURMAP**	Pharmacologie
Mme Isabelle **DUBUC**	Pharmacologie
Mme Dominique **DUTERTE- BOUCHER**	Pharmacologie
Mr Gilles **GARGALA** (MCU-PH)	Parasitologie
Mme Nejla EL **GHARBI-HAMZA**	Chimie analytique
Mme Marie-Laure **GROULT**	Botanique
Mr Chervin **HASSEL**	Biochimie et Biologie Moléculaire

Mme Maryline **LECOINTRE**	Physiologie
Mme Hong **LU**	Biologie
Mme Marine **MALLETER**	Toxicologie
M. Jérémie **MARTINET** (MCU-PH)	Immunologie
M. Romy **RAZAKANDRAINIBÉ**	Parasitologie
Mme Tiphaine **ROGEZ-FLORENT**	Chimie analytique
Mr Mohamed **SKIBA**	Pharmacie galénique
Mme Malika **SKIBA**	Pharmacie galénique
Mme Christine **THARASSE**	Chimie thérapeutique
Mr Frédéric **ZIEGLER**	Biochimie

PROFESSEURS ASSOCIES

Mme Cécile **GUERARD-DETUNCQ**	Pharmacie officinale
Mme Caroline **BERTOUX**	Pharmacie

PAU-PH

M. Mikaël **DAOUPHARS**

PROFESSEUR CERTIFIE

Mme Mathilde **GUERIN**	Anglais

ASSISTANTS HOSPITALO-UNIVERSITAIRES

Mme Alice **MOISAN**	Virologie
M. Henri **GONDÉ**	Pharmacie

ATTACHES TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE

Mme Soukaina **GUAOUA-ELJADDI**	Informatique
Mme Clémence **MEAUSOONE**	Toxicologie

ATTACHE TEMPORAIRE D'ENSEIGNEMENT

Mme Ramla **SALHI**	Pharmacognosie

LISTE DES RESPONSABLES DES DISCIPLINES PHARMACEUTIQUES

Mme Cécile **BARBOT**	Chimie Générale et minérale
Mr Thierry **BESSON**	Chimie thérapeutique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Elisabeth **CHOSSON**	Botanique
Mme Marie-Catherine **CONCE-CHEMTOB**	Législation et économie de la santé
Mme Isabelle **DUBUS**	Biochimie
Mr Abdelhakim **EL OMRI**	Pharmacognosie
Mr François **ESTOUR**	Chimie organique
Mr Loïc **FAVENNEC**	Parasitologie
Mr Michel **GUERBET**	Toxicologie
Mme Martine **PESTEL-CARON**	Microbiologie
Mr Mohamed **SKIBA**	Pharmacie galénique
Mr Rémi **VARIN**	Pharmacie clinique
M. Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

III – MEDECINE GENERALE

PROFESSEUR MEDECINE GENERALE

Mr Jean-Loup **HERMIL** (PU-MG)	UFR	Médecine générale

MAITRE DE CONFERENCE MEDECINE GENERALE

Mr Matthieu **SCHUERS** (MCU-MG)	UFR	Médecine générale

PROFESSEURS ASSOCIES A MI-TEMPS – MEDECINS GENERALISTE

Mr Pascal **BOULET**	UFR	Médecine générale
Mr Emmanuel **LEFEBVRE**	UFR	Médecine Générale
Mme Elisabeth **MAUVIARD**	UFR	Médecine générale
Mr Philippe **NGUYEN THANH**	UFR	Médecine générale
Mme Yveline **SEVRIN**	UFR	Médecine générale

MAITRE DE CONFERENCES ASSOCIE A MI-TEMPS – MEDECINS GENERALISTES

Mme Laëtitia **BOURDON**	UFR	Médecine Générale
Mme Elsa **FAGOT-GRIFFIN**	UFR	Médecine Générale
Mr Emmanuel **HAZARD**	UFR	Médecine Générale
Mme Lucile **PELLERIN**	UFR	Médecine générale

ENSEIGNANTS MONO-APPARTENANTS

PROFESSEURS

Mr Paul **MULDER** (phar)	Sciences du Médicament
Mme Su **RUAN** (med)	Génie Informatique

MAITRES DE CONFERENCES

Mr Sahil **ADRIOUCH** (med)	Biochimie et biologie moléculaire (Unité Inserm 905)
Mme Gaëlle **BOUGEARD-DENOYELLE** (med)	Biochimie et biologie moléculaire (UMR 1079**)**
Mme Carine **CLEREN** (med)	Neurosciences (Néovasc)
M. Sylvain **FRAINEAU** (med)	Physiologie (Inserm U 1096)
Mme Pascaline **GAILDRAT** (med)	Génétique moléculaire humaine (UMR 1079)
Mr Nicolas **GUEROUT** (med)	Chirurgie Expérimentale
Mme Rachel **LETELLIER** (med)	Physiologie
Mr Antoine **OUVRARD-PASCAUD** (med)	Physiologie (Unité Inserm 1076)
Mr Frédéric **PASQUET**	Sciences du langage, orthophonie
Mme Anne-Sophie **PEZZINO**	Orthophonie
Mme Christine **RONDANINO** (med)	Physiologie de la reproduction
Mr Youssan Var **TAN**	Immunologie
Mme Isabelle **TOURNIER** (med)	Biochimie (UMR 1079)

DIRECTEUR ADMINISTRATIF : M. Jean-Sébastien **VALET**

HCN - Hôpital Charles Nicolle	*HB - Hôpital de BOIS GUILLAUME*
CB - Centre Henri Becquerel	*CHS - Centre Hospitalier Spécialisé du Rouvray*
CRMPR - Centre Régional de Médecine Physique et de Réadaptation	*SJ – Saint Julien Rouen*

Par délibération en date du 3 mars 1967, la faculté a arrêté que les opinions émises dans les dissertations qui lui seront présentées doivent être considérées comme propres à leurs auteurs et qu'elle n'entend leur donner aucune approbation ni improbations.

Sommaire

La gestion des données manquantes dans les études observationnelles publiées dans cinq journaux de renom : pratique et manière de rapporter les résultats

Introduction

Avant-Propos

La recherche biomédicale est un domaine de recherche scientifique regroupant toute activité dont la finalité est de contribuer à un progrès dans la connaissance de la biologie humaine, des maladies et de leurs traitements. Cette recherche peut être interventionnelle (e.g. essai clinique randomisé sur les anti-diabétiques oraux) ou observationnelle (e.g. étude des patients exposés au tabac). Les avancées de la recherche scientifique se diffusent via la publication d'articles scientifiques qui relatent le déroulement et les résultats d'études ou d'expériences. Celles-ci s'appuient sur l'analyse de données dont la qualité est primordiale pour comprendre et interpréter au mieux les résultats. La connaissance basée sur la recherche biomédicale ou basée sur les preuves (*evidence based medicine*) s'appuie notamment sur la répétition des études, indissociable de la complétude du rapport permettant une bonne reproductibilité par les pairs.

La forme d'un article scientifique est standardisée (Introduction, Methods, Results, and Discussion, ou plan IMRAD). L'introduction donne les éléments de contexte aboutissant à la question de recherche. Vient ensuite une partie matériel & méthodes, qui explique d'où viennent les données, quelles données sont recueillies, de quelle manière, la temporalité du recueil, quelles données seront analysées et de quelle manière. Ensuite vient la partie dédiée aux résultats, mettant en avant les résultats les plus importants (résultats de l'analyse principale) tout en expliquant leur sens. S'en suit une partie dite de discussion, où les résultats sont nuancés, les forces, faiblesses limites de l'étude sont expliquées et où les perspectives de recherche future sont présentées. Une brève conclusion, rappelant les principaux résultats et nuances principales de l'étude apparait parfois. Des appendices de l'étude existent parfois, détaillant certains points de l'étude (détails des analyses statistiques, résultats secondaires, autres). Les articles sont soumis à des revues scientifiques, qui font relire ces articles par des chercheurs du même domaine (revue par les pairs), avant de les approuver et de les publier, permettant leur diffusion.

Cette thèse s'intéresse à la manière de rapporter les méthodes et les résultats en lien

avec les données manquantes ainsi qu'aux pratiques relatives à leur gestion. Tout d'abord les concepts de **données** et de **données manquantes** seront définis. Ensuite, les différents mécanismes de production des données manquantes, l'impact et les méthodes de gestion seront rappelés. Enfin, la problématique de notre étude sera introduite.

A – Données (définition dans cette thèse)

Une variable correspond à chaque paramètre mesuré d'une observation (e.g. sexe, âge, présence de pathologie, valeur biologique). Chacune de ces variables peut prendre plusieurs valeurs (e.g. homme ou femme pour la variable sexe, tout nombre entier supérieur à 0 pour la variable âge). Chaque observation (patient, séjour, visite) est décrite avec une valeur par variable. Une donnée est donc la valeur d'une variable, pour une observation. La somme des données constitue un jeu de données.

Tableau schématisant un ensemble de données complètes

	Sexe	Age	Poids (en kg)	Statut tabagique
Patient 1	Femme	32	62	Fumeur actif
Patient 2	Homme	78	50	Ancien fumeur
Patient 3	Homme	19	115	Non fumeur
Patient 4	Femme	40	74	Non fumeur

Dans le tableau ci-dessus, 4 variables sont recueillies (sexe, âge, poids, statut tabagique). Ces 4 variables sont recueillies pour 4 patients (il y a donc 4 observations). Une valeur est attribuée à chaque variable observée.

B – Données manquantes

Une donnée manquante correspond au fait que, pour un couple variable-observation, la valeur est manquante. Une variable en particulier peut présenter de nombreuse valeurs manquantes (e.g. question intrusive sur la vie personnelle du sujet ou sur des sujets intimes de la vie privée). Un individu (observation) peut présenter beaucoup de données manquantes (e.g. questionnaire non terminé). Lorsqu'on parle des données manquantes dans un sens plus général, il s'agit donc de l'ensemble des valeurs manquantes. Le tableau suivant donne un

exemple de répartition possible des données manquantes (classiquement représentées par l'acronyme NA pour Not Available, non disponible).

Tableau schématisant un ensemble de données

	Sexe	Age	Poids	Statut tabac
Patient 1	Femme	NA	NA	NA
Patient 2	Homme	78	50	NA
Patient 3	Homme	19	115	Non fumeur
Patient 4	Femme	NA	74	Non fumeur

Chaque étude étant différente, la proportion de données manquantes est donc très variable d'un jeu de données à l'autre. La situation idéale est celle où il n'y a aucune donnée manquante, cependant, cette situation est rare en réalité [1].

Schématiquement, plus le nombre de patients ainsi que de variables croissent, plus le nombre de données manquantes croît. Cela dépend de la difficulté à recueillir la variable mais aussi du mécanisme, parfois plus complexe que le simple hasard, causant la donnée manquante.
Plusieurs raisons peuvent aboutir à la présence de données manquantes. Au sein d'une même étude, les causes d'une donnée manquante peuvent varier.

C - Mécanismes de production des données manquantes

Selon Rubin, les mécanismes de production de données manquantes peuvent être classés en 3 catégories [2].

1) Données manquantes de façon complètement aléatoire (Missing completely at random, MCAR)

Ce sont les variables pour lesquelles la donnée manquante est indépendante de toute autre variable (observable ou non), et apparaissant de façon aléatoire. Ces données manquantes n'entrainent donc pas de biais. C'est par exemple un patient dont l'information sur la consommation de drogue intraveineuse est manquante du simple fait que son questionnaire ait

été égaré.

2) Données manquantes de façon non aléatoire (Missing not at random, MNAR)

Ce sont les variables pour lesquelles la valeur de la donnée manquante est liée à sa vraie valeur. Par exemple, dans une étude s'intéressant à la consommation de drogues intraveineuses, il est envisageable que les personnes consommatrices de ces produits soient moins susceptibles de délivrer cette information du fait qu'un certain tabou existe concernant ces consommations.
Ces données manquantes exposent au risque de biais de sélection.

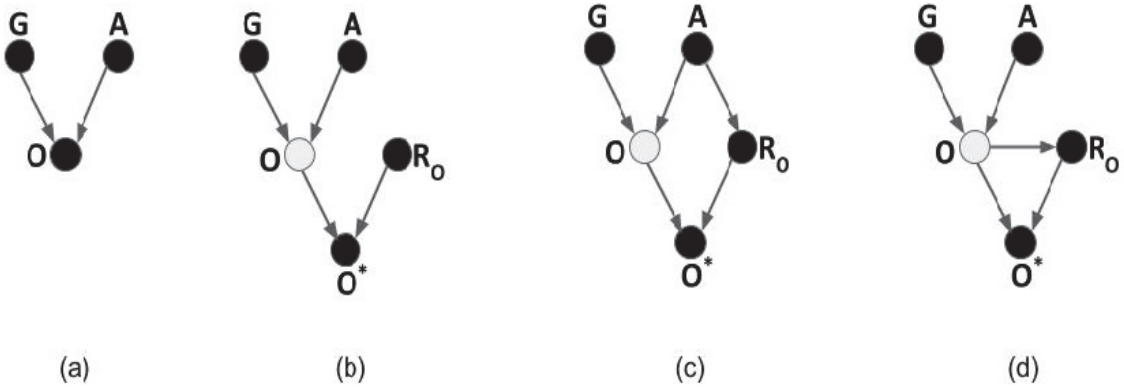
3) Données manquantes de façon aléatoire (Missing at random, MAR)

Cet intitulé est trompeur. En effet, ce sont des données manquantes de façon non aléatoire, mais où leur absence peut être attribuée à une variable pour laquelle on a une information complète. Poursuivons avec l'exemple des drogues intraveineuses : il est possible que les femmes soient moins susceptibles de délivrer cette information.
Ce type de données manquantes peut engendrer un léger biais, relativement contrôlable dans la mesure où l'on a une information sur la donnée manquante grâce à une autre variable.

Ces mécanismes de production de données manquantes sont des mécanismes probabilistes. C'est à dire que chaque valeur manquante d'une variable n'est pas systématiquement dûe au mécanisme supposé, mais seulement qu'un mécanisme domine. En reprenant les exemples vus précédemment, si la donnée relative à la consommation de drogues intraveineuses est manquante chez une femme, il est donc possible que ce soit du fait de sa consommation, du fait de son sexe, ou bien simplement du fait du hasard.

Ces différents types de données manquantes peuvent être représentés graphiquement [3].

Graphiques schématisant les mécanismes de production de données manquantes



(a) Absence de donnée manquante

(b), (c) & (d) Processus de donnée manquante MCAR, MAR et MNAR respectivement.

G représente la variable genre, A l'âge, O l'obésité. R0 représente un mécanisme de donnée manquante. O* représente l'information disponible sur la variable obésité, dépendant de l'obésité et du mécanisme R0.

Les flèches représentent l'existence d'une relation entre les variables. Ainsi, l'absence de flèche entre G et A indique leur indépendance.

Supposons que les variables G et A n'ont pas de données manquantes. La variable O ne dépend que de G et de A. Dans le graphe (a), il n'y a pas de mécanisme de production de données manquantes, il s'agit donc d'une représentation où les données sont complètes.

Dans les graphes suivants, il existe un mécanisme de donnée manquante. L'information O* est donc incomplète. Dans le graphe (b), le mécanisme en cause est indépendant de toute variable. Il s'agit donc d'une situation de donnée manquante complètement au hasard.

Dans le graphe (c), le mécanisme en cause dépend d'une autre variable, pour laquelle on a une information complète (variable A). Il s'agit d'une situation de donnée manquante de façon aléatoire.

Enfin dans le graphe (d), le mécanisme en cause dépend de la variable pour laquelle les données sont incomplètes (variable O). Il s'agit d'une situation de donnée manquante de façon non aléatoire.

D – Impacts de la présence de données manquantes

Les données manquantes exposent à deux risques principaux [4] :

- Quel que soit le mécanisme de production des données manquantes, la baisse de

quantité d'information aboutit à une baisse de puissance statistique.

- Exposition au risque de biais de sélection. Imaginons une étude s'intéressant à l'état de santé de patients à 3 mois d'une opération. Il est possible que les patients ayant le meilleur état de santé ne reconsultent pas à 3 mois, du fait de leur bon état de santé. Ainsi, l'étude risque de considérer un état de santé globalement moins bon que ce qu'il n'est en réalité. Cependant, il est possible que ce soient justement les patients dans le moins bon état de santé qui ne reconsultent pas (du fait d'une hospitalisation autre, d'un décès, d'un changement de médecin, par exemple) et donc conduire à une autre conclusion.

Tout cela peut aboutir à des résultats biaisés. Pour pallier les potentiels biais, différentes méthodes ont été développées pour gérer ces données manquantes.

E – Méthodes de gestion des données manquantes

Il existe de nombreuses méthodes pour gérer ces données manquantes. La meilleure méthode reste de prévenir l'apparition de ce problème en planifiant au mieux l'étude et en étant vigilant au recueil des données [5]. On peut par exemple réduire le nombre de visites, ne recueillir que les données nécessaires, utiliser un formulaire de saisie clair, standardiser le recueil d'information, entraîner les investigateurs et collaborateurs au recueil de données. Cependant, cela n'est parfois pas suffisant ou est difficile à mettre en place. En effet, dans les études observationnelles, l'investigateur ne rencontre pas nécessairement le patient. Les informations peuvent être recueillies par divers moyens (comme un questionnaire, le dossier médical, d'autres bases de données). Cela s'avère d'autant plus problématique si l'on est dans le cadre d'une étude rétrospective. Effectivement, dans cette situation, certaines données intéressantes dans le cadre de l'étude peuvent ne pas avoir été recueillies par le passé, et donc nécessairement conduire à une situation de donnée manquante. Par exemple, pour une étude s'intéressant au lien entre l'indice de masse corporelle (IMC) et la présence d'infection nosocomiale chez les patients dans un service donné, le poids de certains patients peut ne pas avoir été mesuré et donc l'IMC sera manquant pour ces patients.

Un grand nombre de méthodes ont donc été développées pour gérer ces données manquantes dans les analyses. Se pose alors la question de quelle méthode adopter.

1 - Méthodes de délétion

Avec ces méthodes, les patients présentant des données manquantes sont retirés de l'analyse.

Une des méthodes les plus courantes [6] est la méthode dite de **délétion de cas** (ou exclusion de cas). Cette méthode consiste simplement à retirer de l'analyse les patients ayant une donnée manquante. On parle alors d'**analyse en cas complets**. Cette méthode ne biaise pas l'estimation du paramètre si l'on a des données MCAR [7]. On a cependant une perte de puissance d'autant plus importante qu'il y a de patients avec des données manquantes.
Une méthode proche est la méthode de **délétion par paire** où l'on va retirer le patient des analyses uniquement s'il présente une donnée manquante pour la variable étudiée. Le patient sera analysé si l'on s'intéresse aux variables pour lesquelles les données sont observées. Cette méthode reste généralement moins pénalisante que l'analyse en cas complets. Elle implique par contre d'avoir des analyses avec des échantillons de différentes tailles et avec une erreur standard variable.
Par défaut les modèles et logiciels statistiques appliquent des analyses en cas complets.

2 - Méthodes d'imputation

On parle d'imputation simple lorsque la donnée manquante va être remplacée une seule fois, par une unique valeur.

2.1 - Imputation simple

Catégorie "donnée manquante" : En appliquant cette méthode, le fait qu'une donnée soit manquante sera considéré comme une variable de réponse catégorielle. Ainsi, lors de l'emploi de modèles multivariés (modèles de régression multiples par exemple), la variable de réponse sera analysée en fonction du fait que des données soient manquantes dans les variables d'ajustement.

Imputation par la moyenne : Lorsque l'on applique cette méthode, la moyenne des valeurs de la variable va remplacer les données manquantes de cette même variable. Cette méthode s'appuie sur le fait qu'il est raisonnable, dans le cadre d'une distribution normale, de supposer qu'une observation de la variable prise aléatoirement soit proche de la moyenne. Cependant, si les données manquantes ne le sont pas de façon complètement aléatoire, cela peut renforcer un biais déjà existant. Cette méthode n'apporte donc aucune information supplémentaire et a pour effet de maximiser l'effectif à analyser, tout en diminuant l'erreur standard [8]. Des variantes de cette méthode sont l'imputation par le mode ou la médiane.

Dernière observation reportée [9] : Cette méthode est aisément applicable dans le cas

d'études longitudinales, où les mêmes variables sont recueillies à différents moments d'une prise en charge, par exemple à 1 mois, 3 mois, 6 mois. Elle consiste simplement à remplacer la donnée manquante par la dernière valeur observée de cette variable chez le patient. Elle suppose donc qu'on considère qu'entre un instant t et l'instant t+1, la valeur n'a pas été modifiée, ce qui n'est pas toujours vrai. Cela produit donc un biais dans l'estimation du paramètre et sous-estime la variabilité du résultat.
Des variantes de cette méthode existent, où l'on va remplacer les données manquantes par la valeur recueillie à des instants autres. Une de ces variantes est celle de la **prochaine observation reportée**, où la valeur manquante à un instant t sera remplacée par la valeur à un instant t+1.

Imputation avec méthode de régression : Avec cette technique, on va estimer la valeur de la donnée manquante, en se basant sur d'autres variables présentes (via un modèle de régression). Cette estimation va remplacer la donnée manquante. Cette méthode diminue également l''erreur standard. Cette méthode peut donner lieu à la production de valeurs qui peuvent ne correspondre à aucune observation réelle.

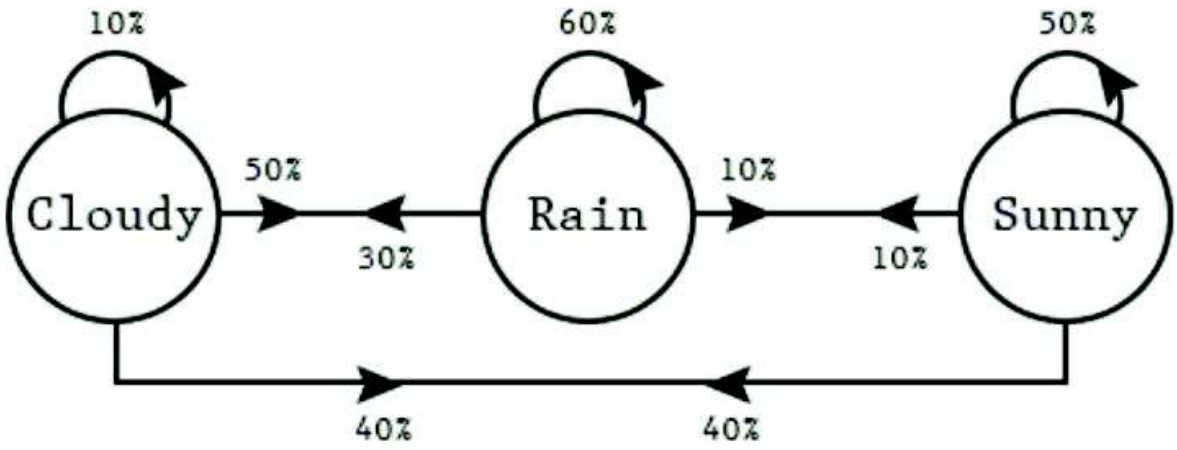
Hot deck imputation [10] : Pour appliquer cette méthode, pour chaque patient présentant une donnée manquante pour la variable d'intérêt, un sous-échantillon de patients similaires est utilisé (en fonction de la correspondance présentée par certaines variables entre le patient ayant la donnée manquante et les autres). Ensuite, une valeur présentée parmi un patient sélectionné aléatoirement dans ce sous-échantillon sera imputée à la donnée manquante.

2.2 - Imputations multiples

On parle d'imputation multiple lorsque la donnée manquante va être remplacée par plusieurs valeurs possibles. Ce qui offre l'avantage de conserver la variabilité et l'incertitude des valeurs.

Markov chain Monte-Carlo (MCMC) [11]
Cette méthode s'appuie sur 2 éléments : les chaînes de Markov et les méthodes de Monte-Carlo.
Les chaînes de Markov sont des chaînes de probabilités où l'événement à un temps t + 1 ne dépend que de l'état à l'instant t.

Schéma d'une chaîne de Markov



On voit dans ce schéma que la probabilité qu'il y ait de la pluie le lendemain d'un jour j ne dépend que de l'état de la météo du jour j, et non de la météo des jours antérieurs.

Avec les méthodes de Monte-Carlo, on va simuler plusieurs échantillons aléatoires, indépendants les uns des autres, à partir de données disponibles.
Les méthodes de Markov Chain Monte-Carlo reposent sur ces 2 principes. On va simuler plusieurs échantillons (méthode Monte-Carlo), où les valeurs de l'échantillon à l'étape n + 1 ne dépendent que de l'échantillon à l'étape n (chaine de Markov). Après un nombre d'itérations suffisamment élevé les échantillons suivent la distribution recherchée [12].

Une méthode s'appuyant sur les algorithmes de MCMC est la **Multiple imputation by chained equation (MICE)** ou **imputation multiple par équations chaînées** [13].
Dans le premier temps de cette méthode, on va imputer les données manquantes d'une première variable avec des valeurs aléatoirement sélectionnées parmi les observations avec des valeurs disponibles pour cette variable.
Ensuite, la 2ème variable sera imputée en utilisant un modèle de régression s'appuyant sur cette première variable imputée. Puis la 3ème sera imputée en utilisant un modèle de régression s'appuyant sur ces 2 variables imputées. Une fois toutes les variables imputées (échantillon Xi), il est possible de réitérer l'opération, en s'appuyant sur les données précédemment imputées. L'échantillon Xi+1 ne dépend donc que de l'échantillon Xi (Chaines de Markov).
Les variables utilisées pour effectuer ces imputations itératives peuvent être sélectionnées, ou bien représenter l'ensemble des variables disponibles. Ces variables composent la matrice de

prédiction.
Le processus est répété sur plusieurs jeux de données distincts. Les imputations de chaque jeu de données sont alors totalement indépendantes les unes des autres (Monte Carlo).
Les analyses sont ensuite effectuées sur les jeux de données simulés puis le résultat moyen de ces analyses est utilisé.

3 – Méthode de pondération

Pondération par la probabilité inverse [15**]**
Lors de l'application de cette méthode, seuls les cas complets sont considérés. La probabilité de ne pas présenter de donnée manquante va être calculée et inversée. Puis cette probabilité inverse va être utilisée pour pondérer les patients et une analyse en cas complet est effectuée. Ainsi, un patient ayant une faible probabilité d'avoir une donnée complète (10% de chance par exemple) sera pondéré plus fortement lors de l'analyse (1/10% soit un facteur 10). Cela permet une meilleure représentativité des patients ayant une forte probabilité d'avoir une donnée manquante.

D'autres méthodes existent, certaines faisant appel à des algorithmes utilisés en intelligence artificielle (apprentissage automatique) comme la méthode des k plus proches voisins [16].

Il existe donc différents mécanismes de production de données manquantes et encore plus de méthodes pour les gérer lors des analyses.

F – Introduction du sujet

Dans les études interventionnelles, la gestion des données manquantes est dans le meilleur des cas prévue et décrite dans le protocole de recherche. Le plus souvent il y a un protocole avec un recueil prospectif actif qui est censé minimiser la quantité de données manquantes. De plus, la tendance actuelle concernant les analyses d'essai clinique consiste à faire une analyse en intention de traiter, qui nécessite la complétude des données de suivi des patients [17].

Dans les études observationnelles, il n'y a pas forcément de protocole produit en amont de l'étude, donc on ne sait pas nécessairement quelles analyses sont prévues et comment est anticipée la gestion des données. Ainsi, la méthode décrite dans les articles est le seul moyen de savoir ce qui a été fait. Cela nécessite donc une manière exigeante de rapporter les

méthodes et les résultats afin de pouvoir comprendre au mieux les analyses.

Le guide de bonnes pratiques STROBE [18] précise les éléments essentiels devant figurer dans les articles. Concernant les données manquantes, il est spécifié que la manière de gérer les données manquantes doit apparaître dans la partie méthode : "*Statistical methods 12 (c) : Explain how missing data were adressed*" et que le nombre de patients présentant des données manquantes pour chaque variable d'intérêt doit figurer dans la partie résultats : "*Descriptive data 14 (b) : Indicate number of participants with missing data for each variable of interest.*". Ces recommandations étant peu précises, d'autres auteurs ont proposé des recommandations supplémentaires spécifiques aux données manquantes [19] : "*how many individuals were excluded because of missing data*", "*Describe the type of analysis used to account for missing data (eg, multiple imputation), and the assumptions that were made (eg, missing at random)*".

Les articles étudiant la production scientifique et s'intéressant aux données manquantes retrouvent que les articles mentionnent insuffisamment la présence de données manquantes [20] ainsi que leur mode de gestion. Cependant, ces études s'appuient sur des articles anciens ou sur des études non observationnelles.

La gestion des données manquantes dans les études observationnelles publiées dans cinq journaux de renom : pratique et manière de rapporter les résultats

P Blavier, T Pressat Laffouilhère

Introduction

La construction de modèles multivariés est très courante dans les études observationnelles, à la fois pour répondre à des questions pronostiques, diagnostiques, de causalité ou encore de prédiction des événements. Cependant ces études, notamment celles se basant sur des données rétrospectives, peuvent comporter un nombre de données manquantes qui varie fortement d'une étude à l'autre. Cela impacte directement la qualité des analyses, particulièrement lorsqu'il y a recours à des modèles multivariés. En effet, si les données manquantes ne sont pas gérées dans les analyses, elles peuvent entraîner un biais de l'estimation de l'effet des variables ainsi qu'une perte de puissance statistique [4].

Le biais dans l'estimation dépend du mécanisme de production des données manquantes. Le plus fréquemment, trois mécanismes de production des données manquantes sont décrits [2]. Premièrement, les données manquantes de façon complètement aléatoire (missing completely at random, MCAR). Dans cette situation, l'existence d'une donnée manquante est indépendante de toute autre variable (observable ou non) et apparaît de façon aléatoire. Viennent ensuite les données manquantes de façon non aléatoire (missing not at random, MNAR), où l'existence d'une donnée manquante dépend de la valeur réelle de cette donnée. Enfin les données manquantes de façon aléatoire (missing at random, MAR) sont les données pour lesquelles leur absence peut être attribuée à une variable ne présentant pas de données manquantes.

Les données MCAR n'entraînent pas de biais dans l'estimation de l'effet des variables. Les données MAR peuvent entraîner un biais qui peut être atténué en ayant recours à des méthodes d'imputation adaptées, dans la mesure où l'on dispose d'une information sur la variable à l'origine de ce mécanisme. Enfin, les données MNAR entraînent un biais de sélection, ce qui impacte l'estimation de l'effet des variables inclues dans les modèles.

Une analyse de cas complets consiste à exclure de l'analyse toute observation présentant au-moins une donnée manquante. Cette méthode altère peu les analyses si les données manquantes sont peu nombreuses et à plus forte raison si elles sont manquantes de façon complètement aléatoire [7].

Avec les méthodes d'imputation simple, la donnée manquante est remplacée par une unique valeur (exemple : imputation par la moyenne, le mode ou la médiane; dernière observation reportée; hot deck imputation). Ces méthodes sont à risque de renforcer le biais de sélection et de renforcer artificiellement la précision de l'analyse en ne prenant pas en compte la variabilité des valeurs [21]. Elles peuvent être efficaces s'il y a peu de données manquantes et si celles-ci sont manquantes de façon aléatoire ou de façon complètement aléatoire.

Enfin, avec les méthodes d'imputation multiple, la donnée manquante va être remplacée par plusieurs valeurs possibles (simulation de plusieurs jeux de données), avant de faire la moyenne de ces simulations et imputer la donnée manquante par cette moyenne. Une des méthodes d'imputation multiple, s'appuyant sur la méthode dite de Markov Chain Monte Carlo (MCMC) [11] est l'imputation multiple par équations chaînées (MICE) [13]. Ces méthodes rendent compte de la variabilité des valeurs possibles et sont donc préférables, et ce d'autant plus s'il y a de nombreuses données manquantes. Elles sont adaptées aux situations de données manquantes de façon non aléatoire ou de façon aléatoire [22].

Si les données sont MNAR, alors généralement le modèle d'imputation multiple est invalide car il suppose des données MAR. Cela peut être corrigé mais est rarement fait en pratique. [22, 23]**.**

Il existe aussi des méthodes de pondération, avec lesquelles les données disponibles sont pondérées afin d'attribuer une plus grande importance à certaines valeurs. La pondération par la probabilité inverse est une de ces méthodes. Elle consiste en une analyse en cas complets où les cas sont pondérés par l'inverse de leur probabilité d'être un cas complet [15].

Cependant, il n'existe pas une méthode de référence pour choisir la manière de gérer les données manquantes. La diversité des méthodes existantes laisse autant de possibilités d'adapter les analyses pour obtenir un résultat souhaité (pratiques de p-hacking). A plus forte raison lors de l'utilisation de modèles multivariés, en modifiant le nombre de sujets analysés en fonction des covariables sélectionnées. De plus, lors de l'utilisation de méthodes d'imputation multiple pour gérer les données manquantes, les variables sélectionnées pour imputer les données manquantes ont leur importance et influencent les résultats de l'imputation [19].

Il est donc crucial d'être complet et précis lors de la rédaction afin de rapporter toutes les informations concernant la proportion de données manquantes ainsi que la méthode de gestion de celles-ci. Le guide des bonnes pratiques Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) [18] précise bien que la gestion des

données manquantes est un des éléments importants devant figurer dans la partie méthode des articles scientifiques "*12(c) explain how missing data were addressed*", ainsi que la proportion de données manquantes par variable doit être précisée dans la partie résultats "*Descriptive data 14 (b): Indicate number of participants with missing data for each variable of interest."*. Ces recommandations ont par la suite été enrichies par d'autres auteurs [19].

De précédentes études constatent que la plupart des articles ne rapportent pas la présence de données manquantes [24, 25] et qu'elles sont le plus souvent exclues des analyses [24, 26]. Cependant, il s'agit d'études s'intéressant à des articles publiés depuis plusieurs années, dans des journaux de qualité hétérogène, concernant divers domaines et dont l'exhaustivité du recueil est incertaine. Une mise à jour de l'état de l'art concernant la manière de rapporter les informations relatives aux données manquantes serait intéressante. L'objectif principal de ce travail est donc d'évaluer les pratiques et la manière de rapporter les informations relatives aux données manquantes, dans les études publiées récemment (2018 et 2019) dans des journaux à haut facteur d'impact. L'objectif secondaire est d'étudier l'impact de la présence d'un auteur biostatisticien sur la mention de l'existence de données manquantes dans l'étude et sur les pratiques de gestion des données manquantes.

Méthode

Inclusion des articles

La sélection des articles a été réalisée à partir des sommaires en ligne de cinq journaux médicaux majeurs. Ces cinq journaux correspondent aux journaux ayant le plus grand facteur d'impact dans la catégorie "medecine, general & internal" selon le Journal Citation Reports de 2016 [27] : New England Journal of Medicine (N Engl J Med), The British Medical Journal (BMJ), The Lancet (LANCET), The Journal of the American Medical Association (JAMA) et Annals of Internal Medicine (Ann Intern Med), aussi nommés les 'big five'.

Des articles originaux publiés entre janvier 2017 et décembre 2019 inclus, comprenant des études observationnelles avec des modèles statistiques multivariés ont été sélectionnés par un des auteurs de cette étude (T.PL). Seules des études de santé observationnelles sur les sujets humains ont été considérées. Il s'agit d'études incluant des humains (patients, volontaires sains, praticiens), chez qui des données de santé ont été mesurées, mais sans intervention. Cela inclut les études transversales, cas-témoin, cohortes prospectives et

rétrospectives, études quasi-expérimentales, avec ou sans groupe contrôle. Seules les études utilisant au moins un modèle nécessitant la sélection de covariables ont été incluses. Les études sur l'économie, la génétique, ainsi que les méta-analyses, revues systématiques et études épidémiologiques descriptives ont été exclues. Seules les études publiées en 2018 et 2019 seront considérées dans cet article afin de dresser un état de l'art le plus récent possible.

Variables recueillies

Les parties méthodes et résultats ont été lues afin de recueillir les différentes variables. Les appendices ont également été lus à la recherche d'informations complémentaires.

Le nom du journal et l'année de publication ont été recueillis. La présence de données manquantes a été catégorisée comme suit : (i) Présence : l'article ou l'appendice mentionne la présence de données manquantes; (ii) Non concerné : l'article ou l'appendice précise l'absence de données manquantes; (iii) Non mentionné : il n'y a pas mention de données manquantes ni dans l'article ni dans les appendices.

Le choix des variables a été fait en se basant sur les critères du STROBE [18] ainsi que la recommandation proposée par Sterne *et al*. [19].

En cas de présence de données manquantes :

Pour chaque article, l'indication de la proportion (ou quantité) de données manquantes par variable a été recueillie [19], sans que cela n'indique forcément le nombre de patients présentant des données manquantes.

La (ou les) méthode(s) de gestion des données manquantes a (ont) été recueillie(s) en distinguant analyse principale et analyse de sensibilité. Une liste de méthodes préalablement sélectionnées par une lecture de la littérature [21, 28, 29] a été constituée : (i) Sujets exclus (analyse en cas complets); (ii) Imputation simple (catégorie "donnée manquante", mode, médiane, moyenne, dernière observation reportée, hot deck imputation, non spécifiée); (iii) Imputation multiple (MICE, MCMC, non spécifiée); (iv) Inconnue (aucune gestion des données manquantes n'est mentionnée). Cette liste est non limitative et non exhaustive. Si une méthode non listée était retrouvée elle était donc recueillie. Ces catégories ne sont pas mutuellement exclusives. En effet, deux variables d'une même étude pouvaient avoir une méthode de gestion des données manquantes différente (e.g. une variable A gérée avec une méthode d'exclusion et une variable B gérée avec une méthode d'imputation simple).

La partie de l'article précisant la méthode de gestion des données manquantes a été recueillie (corps de l'article ou appendice).

Lorsque les données manquantes étaient gérées par l'analyse en cas complet (que ce soit dans l'analyse principale ou de sensibilité), la proportion (ou la quantité) de patients exclus des modèles multivariés a été recherchée, en précisant si celle-ci est: (i) clairement mentionnée dans l'article, (ii) clairement mentionnée dans l'appendice, (iii) si un calcul a été nécessaire à partir d'informations provenant de l'article (appendice inclus) ou (iv) si cette information n'a pas pu être retrouvée, soit car aucune information n'a été retrouvée ou lorsque celle-ci était trop imprécise.

Concernant les imputations multiples: (i) le nombre d'itérations, (ii) le nombre de jeux de données simulées, (iii) la mention de la matrice de prédiction (variables prédictives par variable à imputer) ainsi que (iv) le type de donnée manquante supposé par les auteurs (MAR, MNAR, MCAR ou non mentionnée) ont été recueillis.

Auteur biostatisticien :

Les problèmes relatifs aux données manquantes étant plus particulièrement connus des biostatisticiens, la présence d'un biostatisticien parmi les auteurs a été recherchée. Seuls les auteurs mentionnés clairement comme statisticien de l'étude ou affiliés à une structure mentionnant clairement "statistics", "biostatistics" ou "data science" ont été considérés comme effectivement statisticiens.

Concernant le recueil de la méthode de gestion des données manquantes, toutes les références ont été relues par deux auteurs (TPL et PB) en ouvert. Seul PB a lu les appendices. La décision était basée sur un consensus.

Analyses statistiques :

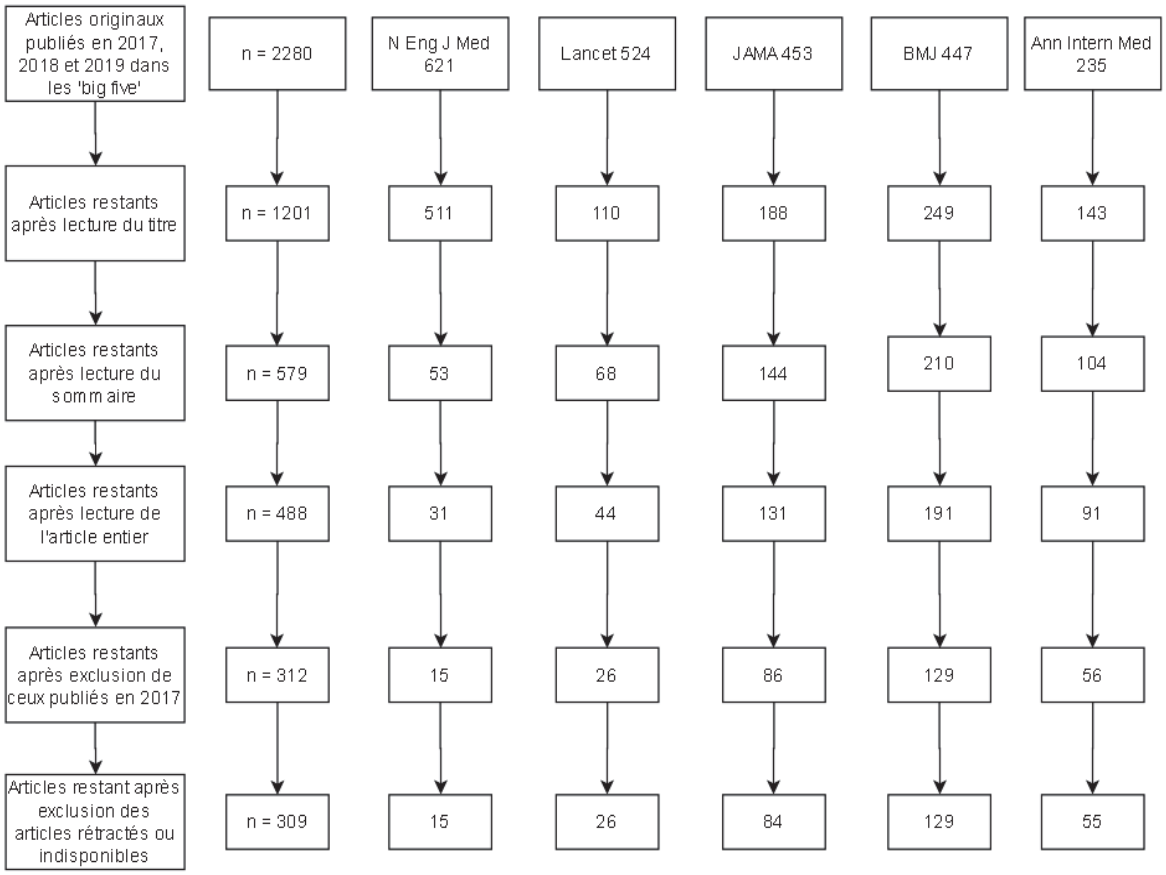
Les résultats sont présentés avec des pourcentages. Les proportions sont présentées séparément pour l'analyse principale et l'analyse de sensibilité. La comparaison de variables qualitatives a été réalisée par un test du Chi^2 ou par un test exact de Fisher. Etant donné le nombre important d'absence de mention des données manquantes, une analyse post hoc a été réalisée à la recherche du type de jeux de données utilisé (e.g. : bases médico-administratives, registres, cohorte) dans ce sous-groupe. En effet, ce type de base de données présente des données manquantes qui sont parfois imputées nativement (l'absence de notification d'un événement ne signifie pas nécessairement que cet événement n'a pas eu lieu) ou par traitement avant la mise à disposition (gestion des données manquantes par l'organisme responsable du registre).

La gestion des données et les analyses statistiques ont été réalisées avec le logiciel R (version

4.0.2 (2020-06-22) -- "Taking Off Again" Copyright (C) 2020 The R Foundation for Statistical Computing).

Résultats

Diagramme de flux



Au total, 312 articles répondaient aux critères d'inclusion, 2 ont été rétractés et 1 était indisponible. Les 309 articles restants et leurs appendices (dans la mesure où l'article mentionne la présence d'un appendice) ont été lus. Parmi les 309 articles, 15 (4,9%) ont été publiés dans le N Engl J Med, 26 (8,4%) dans le Lancet, 55 (17,8%) dans le Ann Intern Med, 84 (27,2%) dans le JAMA et 129 (41,7%) dans le BMJ.

Rapport d'information sur les données manquantes et pratiques :

Un total de 10 (3,2%) articles explicitait l'absence de données manquantes, 63 (20,4%) ne mentionnait pas les données manquantes et 236 (76,4%) mentionnait leur présence.

Articles mentionnant la présence de données manquantes *:*

Concernant les articles mentionnant la présence de données manquantes, la quantité de données manquantes par variable était indiquée dans 161 (68,2%) des cas. Les méthodes de gestion des données manquantes sont détaillées dans le tableau 1. La majorité des articles (216/236) apportait une précision sur la méthode de gestion. Le recours à des analyses de sensibilité concernant les données manquantes n'a été retrouvée que dans 36 (15,3%) articles. Parmi celles-ci, une méthode d'imputation multiple a été employée dans 75% des cas (Tableau 2). Pour 3 (1,3%) articles, la gestion des données manquantes n'était pas précisée dans l'analyse principale, mais l'était dans l'analyse de sensibilité (1 par exclusion, 1 par imputation multiple sans précision et 1 par Shared Parameters Model [30]).

Tableau 1 : Méthodes de gestion des données manquantes (pourcentages)* dans l'analyse principale	
Nombre d'articles mentionnant l'existence de données manquantes	236 (100%)
Exclusion	116 (49,2%)
Imputation simple :	55 (23,3%)
Catégorie "donnée manquante"	35 (14,8%)
Moyenne, mode, médiane	6 (2,5%)
Régression	1 (0,4%)
Dernière observation reportée	6 (2,5%)
Hot deck imputation [10]	1 (0,4%)
Imputation sans précision	6 (2,5%)
Imputation multiple :	55 (23,3%)
MICE**	24 (10,2%)
MCMC non précisée***	5 (2,1)
Méthode non précisée	26 (11%)
Gestion non précisée	20 (8,5%)
IPW†	3 (1,3%)
K plus proches voisins [16]	1 (0,4%)
Full information maximum likelihood estimation [31]	2 (0,8%)
* La somme des pourcentages peut excéder 100% du fait que certains articles utilisent différentes méthodes pour la gestion des données manquantes. Par exemple exclusion si la variable présente moins de 1% de données manquantes, méthode d'imputation multiple sinon. ** Multiple imputation by chained equation *** MCMC : Markov Chain Monte Carlo † Pondération par la probabilité inverse	

L'information sur l'existence de données manquantes ainsi que leur méthode de gestion

étaient retrouvées dans 210 articles (97,2% des cas), dans le corps de l'article. Pour 6 articles (2,8%), une de ces informations n'était retrouvée que dans l'appendice. L'appendice a permis d'éclaircir la méthode d'imputation multiple utilisée dans deux articles.

Tableau 2 : Méthodes de gestion des données manquantes (pourcentages)* dans l'analyse de sensibilité

Nombre d'études	36 (100%)
Exclusion	4 (11,1%)
Imputation simple	3 (8,3%)
Catégorie "donnée manquante"	1 (2,8%)
Moyenne, mode, médiane	1 (2,8%)
Imputation sans précision	1 (2,8%)
Imputation multiple	27 (75%)
MICE **	12 (33,3%)
MCMC***	4 (11,1%)
Méthode non précisée	10 (27,8%)
IPW†	3 (8,3%)
SPM°	1 (2,8%)

*La somme des pourcentages peut excéder 100% du fait que certains articles conduisent plusieurs méthodes de gestion de données manquantes dans les différentes analyses de sensibilité.
** Multiple imputation by chained equation
*** MCMC : Markov Chain Monte Carlo
† Pondération par la probabilité inverse
° SPM : Shared Parameters Model

Lors d'analyses en cas complets, que ce soit dans l'analyse principale (116 articles) ou dans l'analyse de sensibilité (4 articles), la proportion de patients exclus des modèles multivariés a été recherchée. Dans 19,2% des articles avec une analyse en cas complet la proportion de patients perdus dans l'analyse multivariée n'était pas précisée (Tableau 3).

La proportion (ou nombre) de données manquantes était clairement indiquée par les auteurs dans le corps de l'article pour 72 (60%) articles et dans l'appendice pour 15 (12,5%)

articles. Cette information a été calculée à partir d'informations recueillies en différents points de l'article (appendice compris) pour 10 (8,3%) articles. Pour 3 études, la proportion de patients exclus était supérieure à 50% (avec un maximum de 59% de patients exclus).

Tableau 3 : Proportion de patients exclus parmi les articles avec une analyse en cas complet

Proportion de patients exclus	Nombre d'études (pourcentage)
< 1%	27 (22,5%)
[1% - 5%]	28 (23,3%)
[5% - 10%]	12 (10%)
[10% - 20%]	19 (15,8%)
> 20%	11 (9,2%)
Non retrouvée	23 (19,2%)

Le recours à des méthodes d'imputation multiple, que ce soit dans l'analyse principale ou dans l'analyse de sensibilité, a été rapporté par les auteurs de 82 articles. La MICE a été utilisée dans 36 (43,9 %) études, 10 (12,2 %) utilisent une méthode MCMC non précisée et 36 (43,9 %) une méthode d'imputation multiple non précisée. Quelle que soit la méthode d'imputation multiple indiquée par les auteurs, les variables inclues pour la procédure d'imputation sont précisées dans moins d'un tiers des cas (Tableau 3).

Tableau 3 : Précisions des modèles d'imputation multiple			
	MICE 36 (100%)	MCMC 10 (100%)	Imputation multiple sans précision 36 (100%)
Précision du mécanisme de données manquantes (MAR)	10 (27,8%)	0 (0%)	2 (5,6%)
Précision du nombre d'itération	8 (22,2%)	1 (10%)	0 (0%)
Précision du nombre de jeu de données simulés	30 (83,3%)	7 (70%)	15 (41,7%)
Précision de la matrice de prédiction	9 (25%)	3 (30%)	10 (27,8%)

L'information sur le nombre d'itérations, le nombre de jeux de données simulé et la matrice de prédiction n'est retrouvée simultanément que dans 2 (5,6%) articles utilisant une MICE. Aucun des articles utilisant une gestion des données manquantes par MCMC ou par imputation multiple non précisée n'indique ces informations simultanément.

Articles ne mentionnant pas la présence ou l'absence de données manquantes :

Parmi ces 63 (20,1%) articles, 40 (63,5%) concernent des études s'appuyant sur des registres ou des bases de données médico-administratives. Les 23 (36,5%) autres articles ne concernent pas des études s'appuyant sur ce type de base de données et aucune information concernant de possibles données manquantes n'est fournie.

Impact du biostatisticien sur la mention de la présence de données manquantes et sur la méthode de gestion des données manquantes :

Parmi les 309 articles étudiés, un biostatisticien est présent parmi les auteurs dans 107 (34,6%) d'entre eux. La différence de proportion d'articles mentionnant la présence ou absence de données manquantes en fonction de la présence d'un biostatisticien parmi les auteurs a été analysée (Tableau 5). Cette différence est de 0,04 en faveur de ceux n'ayant pas

de biostatisticien parmi les auteurs (IC95% = [-0,09 ; 0.12], valeur p = 0,84).

Parmi les articles présentant des données manquantes la différence de proportion d'articles précisant la gestion des données manquantes et ceux ne le faisant pas, en fonction de la présence d'un biostatisticien est de 0,3%, p-valeur = 0,9 (différence en faveur de ceux ayant un biostatisticien).

Parmi les articles précisant la méthode de gestion des données manquantes, la différence de proportion d'articles utilisant une méthode d'exclusion et une imputation simple en fonction de la présence d'un biostatisticien sont de 5,6% et 2% (en faveur de ceux n'ayant pas de biostatisticien), p-valeurs = 0,41 et 0,77 respectivement. En revanche, concernant l'utilisation d'une méthode d'imputation multiple et la mention d'une gestion imprécise (imputation simple sans précision ou imputation multiple sans précision), ces différences sont de 5,8% et 3% (en faveur de ceux ayant un biostatisticien), p-valeurs = 0,32 et 0,53 respectivement.

Tableau 5 : Méthodes de gestion des données manquantes selon la présence ou l'absence d'un biostatisticien

Méthode de gestion des données manquantes	Nombre d'articles avec biostatisticien en auteur	Nombre d'articles sans biostatisticien en auteur
Exclus	39 (42,9%)	77 (47,8%)
Imputation simple	19 (20,9%)	36 (22,4%)
Imputation multiple	23 (25,3%)	32 (19,9%)
Autres	3 (3,3%)	3 (1,9%)
Non précisée	7 (7,7%)	13 (8,1%)
Total	91 (100%)	161 (100%)

Discussion

L'objectif de ce travail était de décrire la manière de rapporter les informations relatives aux données manquantes (leur présence ou non, leur proportion) ainsi que les différentes pratiques de gestion de ces données manquantes, dans les études observationnelles publiées dans cinq journaux ayant un facteur d'impact majeur.

Dans notre étude, la majorité (79,6%) des articles sélectionnés font mention de la présence ou de l'absence de données manquantes. Une partie seulement des articles avec des

données manquantes répondent aux critères du STROBE [18]. En cas de présence de données manquantes il faut : (i) mentionner la méthode de gestion (91,5%), (ii) préciser la quantité de données manquantes pour chaque variable d'intérêt (68,2%).

Il existe une grande hétérogénéité des méthodes de gestion des données manquantes avec 14 méthodes distinctes retrouvées et une majorité d'analyses en cas complet (120/236). D'après Sterne *et al*. [19], s'il y a recours à une analyse en cas complets, il convient de préciser le nombre de patients exclus des analyses (80,8%). Lors de l'emploi de cette méthode, il est recommandé d'indiquer précisément combien de sujets sont exclus, d'apporter une explication si possible sur la cause de ces données manquantes et d'étudier s'il y a une différence entre les sujets exclus et les autres [19]. Lorsque la proportion de patients exclus est faible (< 5%) et qu'un grand nombre de sujets sont analysés, le bénéfice d'une imputation multiple par rapport à une analyse en cas complet ne semble pas majeur [32]. En revanche, dans les situations avec un grand nombre de données manquantes, le recours à des méthodes d'imputation multiple est généralement préférable [22, 33], sans que cela soit systématique [34].

Les méthodes d'imputation multiple sont utilisées dans près d'un quart (23,3%) des études présentant des données manquantes. Leur utilisation est facilitée par les logiciels statistiques, ce qui expose à un risque de leur utilisation dans un contexte inapproprié [32, 35, 36]. D'après Sterne *et al*. [19], s'il y a recours à une méthode d'imputation multiple, il convient de fournir tous les détails relatifs à la modélisation de l'imputation (5,6%). Le nombre de mots composants l'article étant limité par les journaux, le recours à des appendices pour expliquer clairement les détails relatifs à la justification et à la mise en œuvre de la méthode d'imputation multiple serait une aide mais n'est pas une habitude retrouvée dans les appendices de cette étude. [35, 37].

La présence d'un auteur biostatisticien ne semble pas liée à une mention plus systématique de la présence ou l'absence de données manquantes, ni liée à une méthode de gestion de ces données manquantes particulière.

La présence ou l'absence de données manquantes n'est par retrouvée dans près de 20% des articles de notre étude. Ce chiffre s'élève à 35% dans la revue de la littérature de Fielding *et al*. [38] et à 55% dans celle de Hayat *et al*. [20]. Dans d'autres études traitant de la manière de rapporter l'information concernant les données manquantes, cette information n'est pas présente [24, 25].

Lorsqu'il y a présence de données manquantes, la méthode de gestion de celles-ci est mentionnée de façon très variable selon les articles, allant de 38% [39] à 84% [4], jusqu'à

91,5% dans notre étude.

La recours à une analyse en cas complets est majoritaire dans les articles de notre étude (49,2%). Karahalios *et al*.[24] retrouvent que cela concerne 66% des études de cohortes entre 2000 et 2009, Hussain *et al*. [26] retrouvent que cela concerne 60% des essais contrôlés randomisés entre 2009 et 2014. Parmi les articles de santé publique publiés en 2014, 68% de ceux précisant la méthode de gestion des données manquantes ont recours à une analyse en cas complets [20].

L'utilisation d'imputation multiple comme méthode de gestion des données manquantes était employée dans moins de 1% des articles d'essai clinique randomisé en 2005-2006 (publiés dans le Lancet, le BMJ, le N Eng J Med ou le JAMA) contre 7,8% en 2013-2014 [38].
Sur les journaux de santé publique, cela représente 12% des articles en 2014. Dans notre étude, cela concerne 23,3% des articles.

Ces différences peuvent être expliquées par plusieurs facteurs. Tout d'abord, l'ancienneté des articles a probablement un impact sur la qualité de la manière de rapporter les résultats et les pratiques. Notre étude s'intéresse aux études publiées en 2018 et 2019. Or les revues de la littérature retrouvées concernent des articles publiés entre 2000 et 2014. Les journaux desquels sont tirés les articles étudiés dans ces revues sont d'une qualité hétérogène et ont un facteur d'impact généralement moins important que les 'big five'. De plus, dans ces revues, il n'est pas précisé si les informations ont également été cherchées dans les appendices. Or dans notre étude, tous les appendices ont été lus à la recherche d'informations complémentaires. Enfin, les logiciels statistiques proposent avec le temps de plus en plus de fonctionnalités complexes telles que la possibilité de faire une imputation multiple par équations chaînées. En effet, la popularisation de ces méthodes ne peut se faire que par l'intermédiaire des outils statistiques.

Forces & limites

Cette étude ne s'intéresse qu'à cinq journaux ayant un facteur d'impact important et une qualité de publication reconnue. Cela n'est pas forcément représentatif de la majorité des études observationnelles publiées. Ces journaux restent cependant des exemples et sont mondialement connus pour la qualité de leur production scientifique.

De plus, il n'a été recueilli que ce qui était explicité par les auteurs. Certaines suppositions et interprétations auraient pu être faites pour compléter certaines imprécisions

(non spécifiées, inconnues ou non mentionnées). C'est le cas par exemple des imputations multiples non spécifiées et MCMC pour lesquelles il est probable que la méthode précise ayant été employée puisse être déduite à partir des informations de l'article. Cependant, toutes les informations relatives aux données manquantes délivrées par les auteurs ont été recherchées et recueillies à partir des articles et également de leurs appendices.

La présence d'un biostatisticien parmi les auteurs ne semble pas avoir d'impact majeur, ni sur la rigueur de la manière de rapporter les informations sur les données manquantes, ni sur la méthode de gestion des données manquantes. Cependant, seuls les auteurs affiliés à un service mentionnant spécifiquement "biostatistics" ont été considérés comme biostatisticiens. Il est possible que certains auteurs affiliés à des services portant la mention "epidemiology" ou "public health" aient des compétences biostatistiques et aient participé à l'étude en tant que biostatisticien, sans que cela ait été rapporté. De plus, les remerciements n'ont pas été analysés.

Parmi les articles qui ne mentionnent pas de données manquantes, la plupart (63,5%) sont des études utilisant des bases de données médico-administratives ou registres. Dans ces bases de données, l'absence de mention d'une pathologie (donnée manquante) est imputée comme équivalent à l'absence de cette pathologie. Cela conduit donc à créer des bases exemptes de données manquantes apparentes. Cela peut expliquer que les auteurs ne mentionnent pas la présence ou l'absence de données manquantes.

Enfin, il y a eu une double lecture ouverte du corps de chaque article pour établir le recueil de la méthode de gestion des données manquantes et la décision était basée sur un consensus.

Conclusion

Dans la plupart des études observationnelles de notre étude, les auteurs rapportent la présence de données manquantes ainsi que leur gestion. Cependant, la précision du nombre de données manquantes et les détails relatifs à leur gestion, notamment lors du recours à des méthodes d'imputation multiple qui se développent de plus en plus, nécessitent d'être améliorés. Divers guides de bonnes pratiques existent déjà [18, 19, 32]. Il convient donc de les appliquer de façon plus systématique. Cela permettrait une meilleure analyse et compréhension des résultats, ainsi qu'une meilleure reproductibilité de ceux-ci. Expliciter clairement que l'article est concerné par les données manquantes est un élément fondamental et devrait être ajouté aux recommandations existantes.

Bibliographie

1. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157-166. Published 2017 Mar 15. doi:10.2147/CLEP.S129785

2. Rubin DB. Inference and missind data. Biometrika 1976; 63: 581-92.

3. Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. Journal of the American Statistical Association, 1-16.

4. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Med Res Methodol. 2012;12:96. Published 2012 Jul 11. doi:10.1186/1471-2288-12-96

5. Little RJ, Cohen ML, Dickersin K, et al. The design and conduct of clinical trials to limit missing data. Stat Med. 2012;31(28):3433-3443. doi:10.1002/sim.5519

6. Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of educational research, 74(4), 525-556

7. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):568-575. doi:10.1093/aje/kwx348

8. Malhotra N. Analyzing marketing research data with incomplete information on the dependent variable. J Mark Res. 1987;24:74–84.

9. Mavridis D, Salanti G, Furukawa TA, Cipriani A, Chaimani A, White IR. Allowing for uncertainty due to missing and LOCF imputed outcomes in meta-analysis. Stat Med. 2019;38(5):720-737. doi:10.1002/sim.8009

10. Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. Int Stat Rev. 2010 Apr;78(1):40–64.

11. Schunk D. A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. AStA Adv Stat Anal. 2008 Feb;92(1):101–14. 46

12. METROPOLIS N, ULAM S. The Monte Carlo method. J Am Stat Assoc. 1949 Sep;44(247):335-41. doi: 10.1080/01621459.1949.10483310. PMID: 18139350.

13. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. Int J Methods Psychiatr Res. 2011;20(1):40-49. doi:10.1002/mpr.329

14. Rubin D.B. Wiley; 1987. Multiple imputation for nonresponse in surveys.https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316696.fmatter

15. Seaman SR, White IR.. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013;22: 278-95. 21220355

16. Malarvizhi, R., & Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. International Journal of Engineering Research and Development, 5(1), 5-7

17. Elkins MR, Moseley AM. Intention-to-treat analysis. J Physiother. 2015 Jul;61(3):165-7. doi: 10.1016/j.jphys.2015.05.013. Epub 2015 Jun 19. PMID: 26096012.

18. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008 Apr;61(4):344-9. doi: 10.1016/j.jclinepi.2007.11.008. PMID: 18313558.

19. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. Published 2009 Jun 29. doi:10.1136/bmj.b2393

20. Hayat MJ, Powell A, Johnson T, Cadwell BL. Statistical methods used in the public health literature and implications for training of public health professionals. PLoS One. 2017;12(6):e0179032. Published 2017 Jun 7. doi:10.1371/journal.pone.0179032

21. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402 47

22. White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine, 29(28), 2920–2931. doi:10.1002/sim.3944

23. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach.Statistical Methods in Medical Research2007;16(3):259--275.

24. Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R., & Simpson, J. A. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC medical research methodology, 12(1), 1-10

25. Akl, E. A., Carrasco-Labra, A., Brignardello-Petersen, R., Neumann, I., Johnston, B. C., Sun, X., ... & Alonso-Coello, P. (2015). Reporting, handling and assessing the risk of bias associated with missing participant data in systematic reviews: a methodological survey. BMJ open, 5(9), e009368.

26. Hussain, J. A., Bland, M., Langan, D., Johnson, M. J., Currow, D. C., & White, I. R. (2017). Quality of missing data reporting and handling in palliative care trials demonstrates that further development of the CONSORT statement is required: a systematic review. Journal of clinical epidemiology, 88, 81-91.

27. 2016 Journal Impact Factor. Sci Ed.

28. How to handle missing data, towards data science, A.Swalin, Jan 31,2018

29. Budhiraja P, Kaplan B, Mustafa RA. Handling of Missing Data. Transplantation. 2020 Jan;104(1):24-26. doi: 10.1097/TP.0000000000002865. PMID: 31365474.

30. Griswold ME, Talluri R, Zhu X, Su D, Tingle J, Gottesman RF, Deal J, Rawlings AM, Mosley TH, Windham BG, Bandeen-Roche K. Reflection on modern methods: sharedparameter models for longitudinal studies with missing data. Int J Epidemiol. 2021 Aug 30;50(4):1384-1393. doi: 10.1093/ije/dyab086. PMID: 34113988; PMCID: PMC8407871.

31. Enders CK. The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data. Educ Psychol Meas. 2001 Oct;61(5):713–40. 48

32. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol. 2021;134:79-88. doi:10.1016/j.jclinepi.2021.01.008

33. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019;110:63-73. doi:10.1016/j.jclinepi.2019.02.016

34. Hughes R.A., Heron J., Sterne J.A.C., Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol. 2019:dyz032

35. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol. 2015;15:30. Published 2015 Apr 7. doi:10.1186/s12874-015-0022-1

36. Mackinnon A. The use and reporting of multiple imputation in medical research - a review. J Intern Med. 2010;268(6):586–93. doi: 10.1111/j.1365-2796.2010.02274.x.

37. Ware JH, Harrington D, Hunter DJ, D'Agostino RB. Missing Data. N Engl J Med. 2012;367(14):1353–4. doi: 10.1056/NEJMsm1210043

38. Fielding S, Ogbuagu A, Sivasubramaniam S, MacLennan G, Ramsay CR. Reporting and dealing with missing quality of life data in RCTs: has the picture changed in the last decade?. Qual Life Res. 2016;25(12):2977-2983. doi:10.1007/s11136-016-1411-6

39. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004 Jul 5;91(1):4-8. doi: 10.1038/sj.bjc.6601907. PMID: 15188004; PMCID: PMC2364743.

Résumé

Introduction

Les études observationnelles comportent un nombre très variable de données manquantes, impactant la qualité des analyses. Parmi les nombreuses méthodes de gestion des données manquantes, aucune ne fait référence. L'indication de la proportion de données manquantes ainsi que leur mode de gestion sont importants. Cette étude a donc pour objectif d'évaluer les pratiques et la manière de rapporter les informations relatives aux données manquantes dans les études publiées récemment et dans des journaux à haut facteur d'impact.

Méthode

Les articles originaux publiés dans les '*big five*' entre 2018 et 2019 comprenant des études observationnelles avec des modèles statistiques multivariés ont été sélectionnés. Les informations relatives aux données manquantes ainsi qu'à la méthode de gestion de celles-ci, en se basant sur les critères du STROBE ainsi que la recommandation proposée par Sterne *et al.* ont été recueillies à partir du corps et appendices des articles.

Résultats

La majorité des articles (79,6%) mentionnent la présence ou absence de données manquantes. En cas de présence de données manquantes, la méthode de gestion est précisée dans 91,5% des cas. La méthode la plus employée (49,2%) est la méthode de délétion (analyse en cas complet). Lors de l'utilisation de méthodes d'imputation multiples, les détails relatifs à la modélisation de l'imputation ne sont retrouvés que dans 5,6% des cas.

Conclusion

La plupart des études observationnelles mentionnent la présence ou non de données manquantes. Cependant, la précision de la proportion de données manquantes et les détails relatifs à leur gestion nécessitent d'être améliorés. Le suivi systématique des guides de bonnes pratiques permettrait ces améliorations.

Mots-clefs : études observationnelles, données manquantes, analyse en cas complet, imputation, rapport d'information.