Méthodes de gestion des données manquantes

Antoine Vilotitch, biostatisticien au CHU Grenoble Alpes

Rentrée de la méthodologie, Lyon 2024

Introduction / Définition : qu'est-ce qu'une donnée manquante ?

Définition : on parle de donnée manquante (DM) « lorsque la valeur d'une variable d'intérêt n'est pas mesurée ou enregistrée ».

Comment éviter les données manquantes à priori ? Comment les prendre en charge à postériori ?

« Guideline on missing data in confirmatory clinical trials », European Medicine Agency, 2009.
« Missing data in clinical research: a tutorial on multiple imputation », Austin PC et al, Canadian Journal of Cardiology, 2021

Introduction / Avant et pendant l'étude.

Le meilleur moyen de gérer les données manquantes, c'est de ne pas en avoir :

- Choisir des critères minimisant le risque de données manquantes : passation d'un questionnaire par un ARC vs un auto-questionnaire à renvoyer, choix d'un questionnaire adapté (plus court, plus ciblé, etc…)
- Adapter le recueil des données : si possible avoir un professionnel (ARC par exemple) qui récupère les informations, récupérer les données lors d'une visite plutôt que par téléphone, etc…
- Travailler avec des durées de suivi raisonnables : plus le critère est proche de l'inclusion, moins le risque est grand de perdre le patient.
- Avoir des protocoles de traitement clairs : l'information sur l'implication du patient dans l'étude doit être la plus claire possible, afin de maximiser sa motivation à rester dans celle-ci.

Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health. 2011 Dec;14(8):1101-8. doi: 10.1016/j.jval.2011.06.003. Epub 2011 Aug 2. PMID: 22152180.

Introduction / Avant et pendant l'étude.

RJ Little et al : « The Prevention and Treatment of Missing Data in Clinical Trials », NEJM, 2012.

Table 1. Eight Ideas for Limiting Missing Data in the Design of Clinical Trials.

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study.
Include a run-in period in which all patients are assigned to the active treatment, after which only those who tolerated and adhered to the therapy undergo randomization.
Allow a flexible treatment regimen that accommodates individual differences in efficacy and side effects in order to reduce the dropout rate because of a lack of efficacy or tolerability.
Consider add-on designs, in which a study treatment is added to an existing treatment, typically with a different mechanism of action known to be effective in previous studies.
Shorten the follow-up period for the primary outcome.
Allow the use of rescue medications that are designated as components of a treatment regimen in the study protocol.
For assessment of long-term efficacy (which is associated with an increased dropout rate), consider a randomized withdrawal design, in which only participants who have already received a study treatment without dropping out undergo randomization to continue to receive the treatment or switch to placebo.
Avoid outcome measures that are likely to lead to substantial missing data. In some cases, it may be appropriate to consider the time until the use of a rescue treatment as an outcome measure or the discontinuation of a study treatment as a form of treatment failure.

Introduction / Avant et pendant l'étude.

Table 2. Eight Ideas for Limiting Missing Data in the Conduct of Clinical Trials.

Select investigators who have a good track record with respect to enrolling and following participants and collecting complete data in previous trials.
Set acceptable target rates for missing data and monitor the progress of the trial with respect to these targets.
Provide monetary and nonmonetary incentives to investigators and participants for completeness of data collection, as long as they meet rigorous ethical requirements.[15,16]
Limit the burden and inconvenience of data collection on the participants, and make the study experience as positive as possible.
Provide continued access to effective treatments after the trial, before treatment approval.
Train investigators and study staff that keeping participants in the trial until the end is important, regardless of whether they continue to receive the assigned treatment. Convey this information to study participants.
Collect information from participants regarding the likelihood that they will drop out, and use this information to attempt to reduce the incidence of dropout.
Keep contact information for participants up to date.

Introduction / Données manquantes particulières.

Si la données est manquante en fin d'étude, il faut se poser la question de possibilités logiques pour « boucher certains trous » :

- Le décès implique des données manquantes : les suivis post décès sont manquants. On peut *à priori* décider de certaines règles. Par exemple considérer que le décès est un synonyme d'échec pour l'outcome d'intérêt.

- On peut, dans certains cas, utiliser une variable annexe pour suivre un lien logique permettant de récupérer une information : le sexe est manquant, mais le patient a un antécédent de cancer de la prostate.

- Il existe des règles publiées pour gérer les données manquantes en question : pour l'échelle SF36 : « Il est conseillé d'estimer les valeurs manquantes pour tous les individus ayant répondu à plus de la moitié des questions de l'échelle ; par la moyenne des réponses du même patient aux autres questions de l'échelle »

Introduction / Les types de données manquantes.

- **MCAR** « Missing completely at random » (*Manquant totalement aléatoirement*) :

La probabilité d'absence est la même pour toutes les observations, les DM ne sont pas liées aux données

Ex : un capteur qui grille lors de l'enregistrement du CJP

- **MAR** « Missing at random » (*Manquant aléatoirement*) :

La probabilité d'absence est liée à une ou plusieurs autres variables observées, mais pas à la variable en question

Ex : les patients jeunes sont plus difficilement joignables, moins disponibles pour venir aux visites de suivi

- **MNAR** « Missing not at random » (*Manquant non aléatoirement*) :

La probabilité d'absence dépend de la variable en question

Ex : les patients les plus graves ne veulent plus, ou ne peuvent plus poursuivre le suivi

Définition : l'imputation, c'est quoi ?

Face aux données manquantes, il y a deux approches :

- Faire avec : raisonner avec la notion de « cas complets », c'est-à-dire discuter des résultats pour les patients sans données manquantes sur les variables utilisées dans les analyses
- Imputer : travailler sur un moyen de remplacer la donnée, par une ou plusieurs valeurs supposée « proche », ou des infos permettant de conclure de façon robustes.

Imputation simple : une valeur seulement

- Unconditional mean Imputation :

  pour un critère quantitatif, est attribué aux DM la valeur moyenne/médiane de l'échantillon

- Conditional mean Imputation :

  idem mais au sein d'un sous-échantillon auquel appartiennent les DM

Imputation simple : LOCF exemple

Last Observation Carried Forward.

Pour les données longitudinales, la dernière valeur disponible est assignée aux valeurs manquantes

SUBJID	VISITE	EVALUATION	EVA_LOCF
01-001	1	32	32
01-001	2	43	43
01-001	3	54	54
01-001	4		54
01-001	5	33	33
01-002	1		
01-002	2	24	24
01-002	3	35	35
01-002	4	52	52
01-002	5		52
01-003	1	43	43
01-003	2	44	44
01-003	3		44
01-003	4		44
01-003	5	31	31

Imputation simple : LOCF exemple

Une utilisation en baisse dans la littérature : fréquence d'apparition dans les publications sur PUBMED

Imputation simple : Worst Case exemple

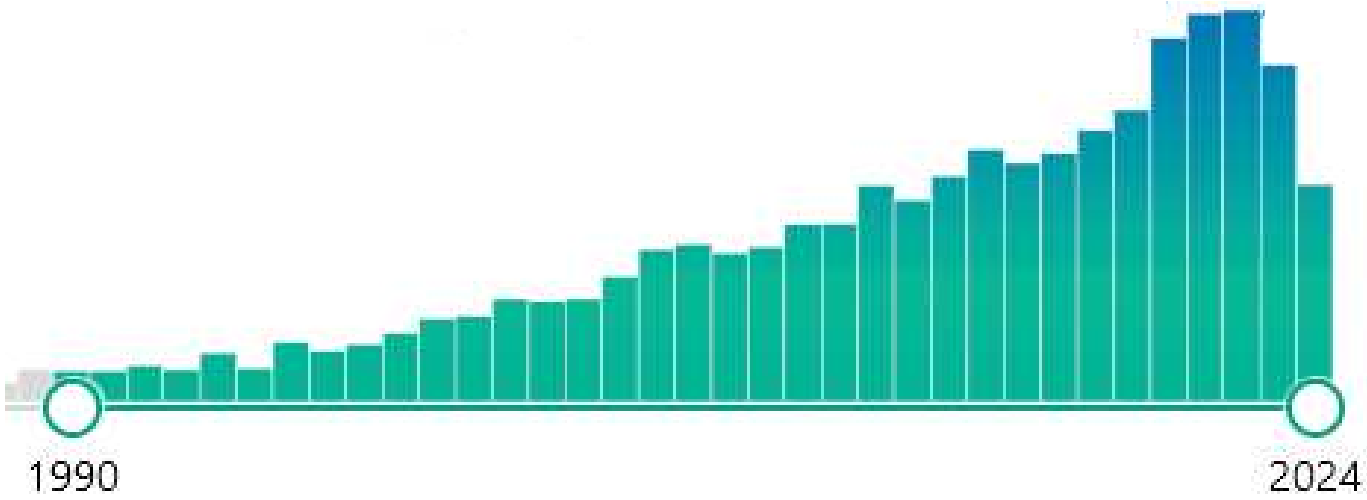
Dans le cas d'un outcome en échec/réussite, les données manquantes du groupe intervention sont assignées en « échec », les résultats du groupe contrôle en « réussite ».

On se place dans le « pire » scénario, afin d'être conservateur pour le résultat significatif d'un test statistique.

A noter :
- Il existe une variante combinant le LOCF et le Worst Case pour les données continues : Worst Observation Carried Forward
- On peut définir un « Best Case » qui serait le pendant optimiste du Worst Case

Imputation simple : Worst Case exemple

Augmentation globale de son utilisation dans les papiers : fréquence d'apparition dans PUBMED



Imputation simple : par régression

Simple :

- Les données manquantes sont estimées via le résultat d'une régression expliquée par un set de variables sélectionnées.

Stochastique :

- Une variance est ajoutée sur le résultat obtenu

Imputation simple : limitations

- Imputation par la moyenne :
  - N'est « valide » qu'avec des données réparties de façon symétrique
  - Réduit mécaniquement l'écart-type
  - Par déduction fausse les tests de moyennes, ou les régressions basées sur la moyenne que l'on pourrait faire en analyse
- Worst-case :
  - Conservateur pour un test simple (Fisher, Student, Wilcoxson)
  - Mais n'est pas informatif pour rendre compte d'une ampleur d'effet, et d'un intervalle de confiance
  - Parfois (souvent !) trop conservateur

Dettori JR, Norvell DC, Chapman JR. The Sin of Missing Data: Is All Forgiven by Way of Imputation? Global Spine J. 2018 Dec;8(8):892-894. doi: 10.1177/2192568218811922. Epub 2018 Dec 3. PMID: 30560043; PMCID: PMC6293424.

Imputation simple : limitations

- LOC-F :
  - Ne semble valide que si les données sont réparties de façon identiques entre les dernières données disponibles et les données censées être manquantes : hypothèses très forte et peu réaliste.
  - Son utilisation est souvent peu documentées dans les papiers où la méthode est utilisée, alors que c'est potentiellement un biais sur les résultats présentés

- Regression :

Comme pour l'imputation sur la moyenne, les données suivent une ligne de régression et la variance est gommée.

Le cas stochastique attendue, mais la répartition du bruit sur une seule salve d'imputation est difficile à maitriser

Lachin JM. Fallacies of last observation carried forward analyses. Clin Trials. 2016 Apr;13(2):161-8. doi: 10.1177/1740774515602688. Epub 2015 Sep 22. PMID: 26400875; PMCID: PMC4785044.

Imputation multiple : plusieurs jeux de données

Le principe “simpliste” est de considérer que l’on réalise N fois une imputation simple par régression stochastique, et de faire des analyses globales réunissant les résultats des N échantillons créés.

En créant plusieurs jeux de données, le risque d’une distribution atypique de la variance ajoutée aux régressions stochastiques est limité.

Pourquoi l'imputation multiple ?

Après toutes les imputations dites « déductibles », l'imputation multiple est la solution recommandée si on veut imputer.

Li et al : « Multiple imputation: a flexible tool for handling missing data », 2015. JAMA.

Heymans MW, Twisk JWR. Handling missing data in clinical research. J Clin Epidemiol. 2022 Nov;151:185-188. doi: 10.1016/j.jclinepi.2022.08.016. Epub 2022 Sep 21. PMID: 36150546.

Dans le CONSORT Statement : pas de mention de l'imputation multiple, mais l'imputation simple est décrite comme non souhaitable.

Moher et al : « CONSORT 2010 Explanation and elaboration: updated guideline for reporting parallel group randomised trials », 2010. BMJ.

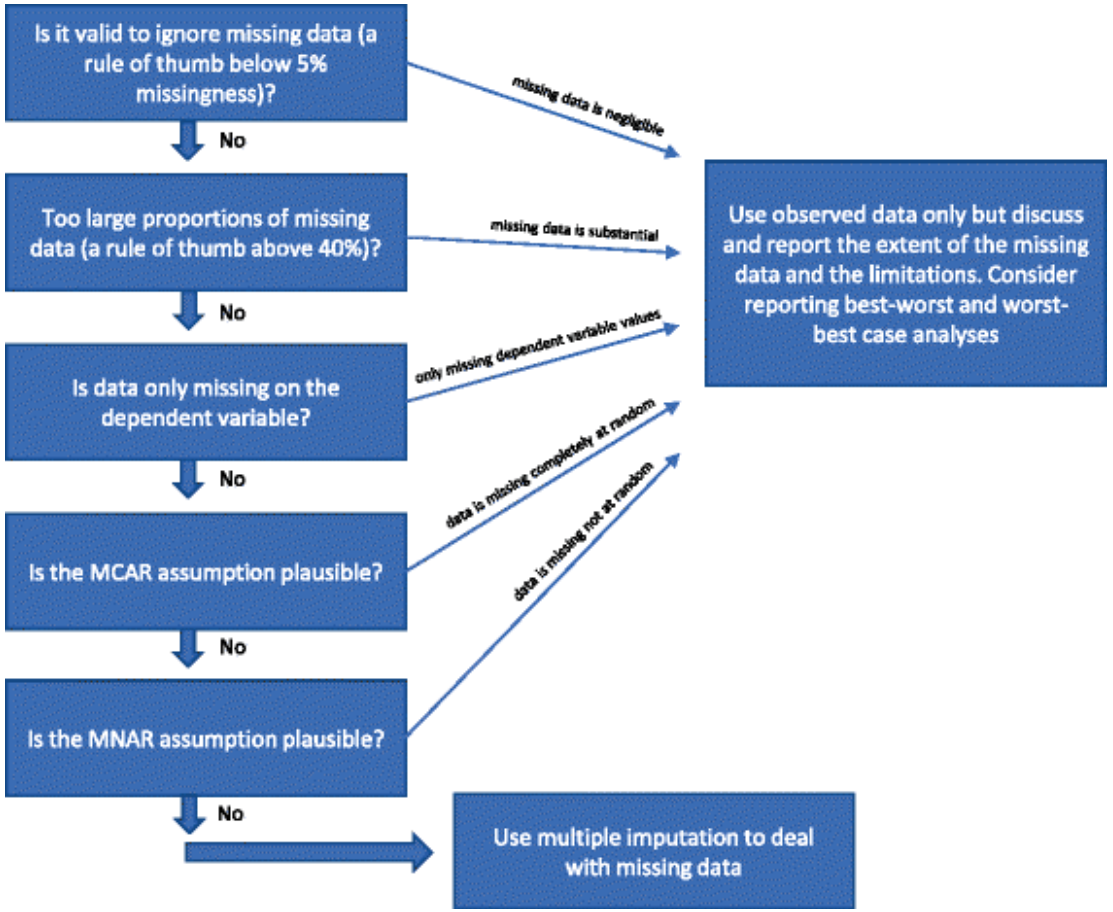
Quand et comment faire une imputation multiple ?

Une guideline connue (1408 citations, 252k visites) :

Jakobsen et al : « When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts », 2017. BMJ Medical Research Methodology.

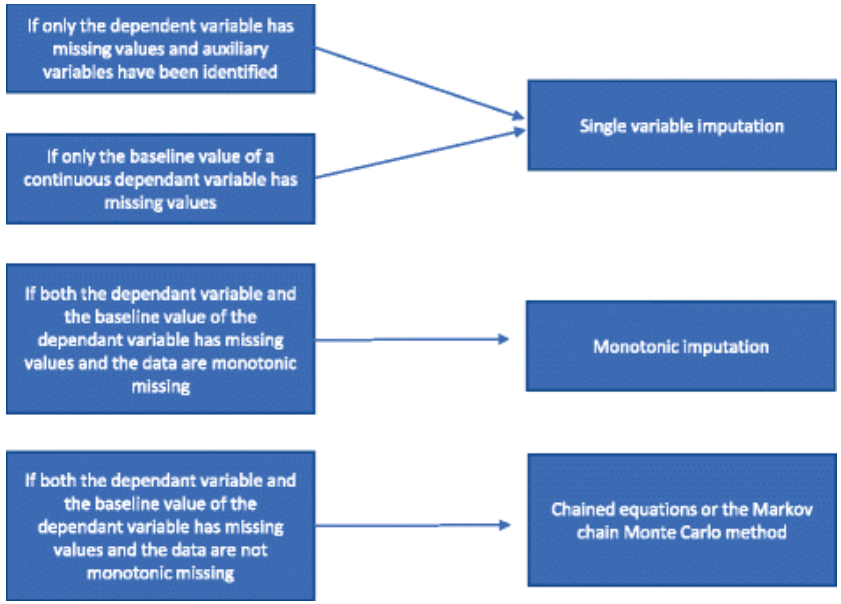
Ce papier résume et actualise un gros travail de bibliographie et propose un arbre décisionnel sur quand l'analyse en imputation multiple peut être utilisée pour gérer les données manquantes.

Jakobsen, 2017.



Flowchart: when should multiple imputation be used to handle missing data when analysing results of randomised clinical trials

Jakobsen, 2017 : quelle imputation multiple ?



Flowchart of multiple imputation

Jakobsen, 2017 : conclusions.

En conclusion :

Si on est entre 5% et 40% de DM, que plusieurs variables d'intérêt sont manquantes et qu'on peut valider que les DM sont de type MAR, alors il est possible d'envisager de traiter les DM via une imputation multiple et d'en présenter les résultats.

Dans tous les autres cas, le résultat en analyse de cas complet doit être le résultat principal, en agrémentant de worst-best best-worst analysis si besoin et éventuellement d'une imputation multiple en analyse de sensibilité (si prévue au PAS).

Dans la vraie vie

- Par expérience, 100% des RCT pour lesquels nous avons travaillé dans l'équipe statistique du CHUGA depuis 2019 ont des résultats similaires entre cas complets et imputation multiple pour le critère principal (le plus souvent NS)
- Pour des régressions multivariées d'étude de cohorte, c'est plus complexe : l'imputation multiple permet de faire ressortir des facteurs statistiquement significatifs qui ne l'étaient pas en cas complet. Une discussion appropriée est nécessaire dans ces cas là.
- Les travaux pour les différentes présentations internes ont montré un manque de transparence et d'information sur comment ont été réalisées les imputations multiples dans la littérature (seulement 60% des papiers ont les informations nécessaires).

Dans la vraie vie : quelques chiffres

<u>Travail mené en 2023 par Laurène Barret et Antoine Vilotitch :</u>

83 RCT issus de Lancet, NEJM, JAMA, entre mai et Juillet 2023.

- 18 imputations simples (22%) : 12 Worst case, 6 Worst Best, 3 LOCF
- 25 imputations multiples (30%) : 15 MICE, 10 non précisées

Merci pour votre attention !