

ANALYSE DE DONNEES LONGITUDINALES INCOMPLETES DANS UN ESSAI CLINIQUE EN NUTRITION

Anissa Elfakir ^{1,2}, Sébastien Marque ¹ & Célestin C. Kokonendji ²

¹ *Danone Nutricia Research,
Département Life Science, Etudes cliniques, et Biométrie
RD 128 Avenue de la Vauve, 91 767 Palaiseau Cedex, France
anissa.elfakir@danone.com ; sebastien.marque@danone.com*

² *Université de Franche-Comté, UFR Sciences et Techniques,
Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS-UFC
16 route de Gray, 25 030 Besançon Cedex, France
celestin.kokonendji@univ-fcomte.fr*

Résumé. La modélisation et l'inférence statistique en présence de données incomplètes en recherche clinique sont complexes. La transition de méthodes simples vers des méthodes plus appropriées est encouragée. En pratique, le statisticien doit s'entendre avec un clinicien sur la définition d'une données incomplète, des objectifs, des hypothèses, choisir un logiciel, et maîtriser les modèles et méthodes d'estimation. Des avancées récentes pourraient permettre de dépasser ces difficultés. L'importance de définir précisément les objectifs de l'étude et les hypothèses cliniques avant toute modélisation statistique a été soulignée. Un cadre général pour le choix des paramètres à estimer et des estimateurs a été proposé. Les exemples concernent principalement le domaine pharmaceutique. Pour souligner l'importance de l'adaptation aux spécificités de la nutrition, nous avons décidé de ré-analyser un essai clinique longitudinal investiguant le potentiel bénéfique d'un produit nutritionnel sur un critère continu. Cette ré-analyse visait à évaluer l'influence des données incomplètes sur l'estimation de l'efficacité du produit en conditions contrôlées et en vie réelle. Pour chaque objectif, les paramètres à estimer, hypothèses cliniques et de mécanisme d'apparition de données manquantes ont été explicités et discutés. L'éclairage apporté par cette analyse servira à optimiser les futures études cliniques en nutrition.

Mots-clés. Données manquantes, efficacité, inférence, modélisation, paramètre à estimer

Abstract. Incomplete data analysis in clinical research is challenging. The shift from well-established simple methods to modern advanced methods is encouraged. In practice, the statistician has to agree with a clinician on the definition of incomplete data, objectives and hypotheses, choose the appropriate software, and master the models and estimation methods. Recent advances could help overcome these limitations. The importance of a clear definition of the objectives of the study and clinical hypotheses before any statistical modeling was emphasized. A general framework for the choice of "estimands", i.e. parameters to be estimated, and estimators has been proposed. Most examples relate to drug development. In order to highlight the importance of an adaptation to the specificities of nutrition, we have decided to re-analyze a longitudinal clinical trial investigating the potential health benefit of a nutritional product with a continuous outcome. The re-analysis was designed to assess the influence of missing data on the estimate of efficacy and effectiveness. For each objective, "estimands", clinical hypotheses and missing data mechanism were explained and discussed. The insight provided by this analysis will be used to optimise future clinical trials.

Keywords. Efficacy, effectiveness, estimand, inference, missing data, modeling

1 Introduction

La modélisation et l'inférence statistique en présence de données incomplètes ou manquantes en recherche clinique présentent un réel challenge. Le cas particulier de la recherche clinique en nutrition a été décrit entre autre par Fitzmaurice (2008). Les études sont souvent longitudinales et conduites sur des périodes pouvant aller de quelques semaines à plusieurs mois, voire années. Les effets produits investigués sont de petits tailles, multifactoriels et présentent une forte variabilité inter-sujet. Dans ces essais, le processus de collecte et de gestion des données est généralement bien maîtrisé, mais inévitablement des données incomplètes sont observées principalement sur les paramètres d'intérêt mesurés en cours d'étude.

Le besoin de faire évoluer la pratique de méthodes simples et bien établies vers des modèles plus appropriés tenant mieux compte de l'incertitude liée aux données manquantes est bien reconnu en théorie. En pratique l'application de ces principes s'avère complexe pour le statisticien. Il n'existe pas de consensus scientifique sur la « meilleure » méthodologie à adopter. Les conditions d'acceptabilité d'une méthode dépendent du contexte spécifique étudié, du type de données manquantes (monotone/non-monotone), des hypothèses faites sur les mécanismes d'apparition, de la nature (quantitatif, qualitatif, semi-quantitatif), ou encore du rôle des variables considérées (variable d'intérêt ou variable explicative). La première étape consiste à convaincre un clinicien du besoin de définir clairement les objectifs, hypothèses, et ce qu'est une donnée incomplète dans le cadre spécifique de l'étude considérée. Un choix est ensuite à faire entre investir dans un logiciel spécialisé implémentant la méthode avancée qui serait la mieux adaptée ou se satisfaire d'une méthode un peu plus standard disponible dans son logiciel habituel. Enfin l'analyse et l'interprétation correctes des résultats nécessitent une bonne compréhension des propriétés et hypothèses sous-jacentes des modèles et méthodes d'estimation.

Des avancées récentes dans les domaines de la recherche académique et réglementaires, et du développement logiciel pourraient permettre de surmonter ces limitations (Burzykowski et al., 2010 ; Carpenter & Kenward, 2008 ; Carpenter et al., 2013 ; Chan et al., 2013 ; Committee for Medicinal Product for Human Use (CHMP), 2010 ; Mallinckrodt & Kenward, 2009 ; Mallinckrodt et al., 2013, 2012 ; Molenberghs & Kenward, 2007 ; National Research Council, 2010 ; O'Neill & Temple, 2012 ; Ratitch & O'Kelly, 2011 ; Royston & White, 2011 ; van Buuren & Groothuis-Oudshoorn, 2011 ; White et al., 2007, 2011ab).

L'une des principales avancées étant l'affirmation dans le rapport du National Research Council (2010) qu'une définition claire des objectifs de l'étude, des hypothèses cliniques, et de leurs implications devrait précéder toute modélisation statistique. Un cadre général pour le choix des paramètres à estimer (« estimands » en anglais) et des estimateurs a ensuite été proposé par Mallinckrodt et al. (2012, 2013). Carpenter et al. (2013) ont formalisé les notions de paramètres à estimer « de jure » et « de facto ». Pour le moment, les applications présentées concernent principalement le domaine pharmaceutique.

Dans le domaine de la nutrition, les méthodes statistiques ont été étudiées par exemple par Groenwold et al. (2012), mais jusqu'à présent peu d'illustrations complètes d'analyse d'étude ont été proposées.

Nous présentons ici une description sommaire du travail effectué, sans développer pour raisons de confidentialité, puis nous terminons en esquissant des perspectives.

2 Description du travail proposé

Pour souligner l'importance de l'adaptation aux spécificités de la recherche clinique en nutrition, et d'améliorer un plan de développement clinique, nous avons décidé de reprendre l'analyse du critère principal d'une étude précédente réalisée avec des méthodes de traitement des données manquantes simples. L'étude considérée est un essai clinique randomisé contrôlé longitudinal investiguant le potentiel bénéfique santé d'un produit nutritionnel sur l'évolution d'un paramètre d'intérêt continu mesuré à différentes reprises pendant l'étude.

Cette ré-analyse visait à évaluer l'influence des données incomplètes sur l'estimation de l'efficacité du produit en conditions contrôlées (ECC) et en vie réelle (EVR). Pour chaque objectif, les paramètres à estimer, les hypothèses sous-jacentes cliniques et de mécanisme d'apparition de données manquantes ont été explicités et discutés. Les premières étapes de ce travail ont consisté à déterminer la question de recherche, la population d'intérêt, ainsi qu'à proposer une adaptation du principe de l'Intention de Traiter (ITT), et des concepts d'ECC et d'EVR au contexte nutritionnel. L'ECC étant définie comme exprimant l'effet du produit dans le scénario idéal où tous les sujets respecteraient le protocole (« efficacy » en anglais ; hypothèse « de juré ») et l'EVR ou l'« effectivité », comme exprimant l'effet du produit dans les conditions réelles de l'essai (« effectiveness » en anglais, hypothèse « de facto »). Les estimateurs et modèles d'analyses ont été choisis en se basant sur le mécanisme d'apparition de données manquantes le plus plausible pour chaque objectif.

Afin de faciliter l'implémentation et la diffusion de ces nouvelles méthodes, les logiciels généralistes couramment utilisés dans notre domaine, SAS, R et Stata, ont été privilégiés autant que possible pour la réalisation de ce travail.

3 Perspectives

Cette analyse a permis d'apporter un éclairage sur l'impact potentiel des données manquantes et sur l'incertitude de l'estimation obtenue en présence de données manquantes dans le cas particulier de cette étude. Il est prévu de compléter ces résultats par des analyses de sensibilité au mécanisme d'apparition de données manquantes. Ce travail servira à optimiser les futures études cliniques en nutrition, et contribuera à faciliter le processus de prise de décision en apportant une meilleure compréhension des hypothèses faites et des résultats obtenus.

Bibliographie

- [1] Burzykowski, T., Carpenter, J., Coens, C., Evans, D., France, L., Kenward, M., Lane, P., Matcham, J., Morgan, D., Phillips, A., Roger, J. & Sullivan, B. (2010), Missing data : discussion points from the PSI missing data expert group, *Pharmaceutical Statistics*, **297**, 288–297.
- [2] Carpenter, J. R. & Kenward, M. G. (2008), *Missing Data in Randomised Controlled Trials—A Practical Guide*. National Health Service Coordinating, Birmingham.
- [3] Carpenter, J. R., Roger, J. H. & Kenward, M. G. (2013), Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation, *Journal of Biopharmaceutical Statistics*, **23**, 1352–1371.

- [4] Chan, A.-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J. A., Dickersin, K., Hróbjartsson, A., Schulz, K. F., Parulekar, W. R., Krleza-Jeric, K., Laupacis, A. & Moher, D. (2013), SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials, *British Medical Journal*, **346**, e7586.
- [5] Committee for Medicinal Product for Human Use (CHMP) (2010), *Guideline on missing data in confirmatory clinical trials*, EMA/CPMP/EWP/1776/99 (Rev. 1.).
- [6] Fitzmaurice, G. (2008), Missing data: implications for analysis, *Nutrition*, **24**, 200–202.
- [7] Groenwold, R. H. H., Donders, A. R. T., Roes, K. C. B., Harrell, F. E. & Moons, K. G. M. (2012), Dealing with missing outcome data in randomized trials and observational studies, *American Journal of Epidemiology*, **175**, 210–217.
- [8] Mallinckrodt, C. H. & Kenward, M. G. (2009), Conceptual considerations regarding endpoints, hypotheses, and analyses for incomplete longitudinal clinical trial data, *Drug Information Journal*, **43**, 449–458.
- [9] Mallinckrodt, C. H., Lin, Q., Lipkovich, I. & Molenberghs, G. (2012), A structured approach to choosing estimands and estimators in longitudinal clinical trials, *Pharmaceutical Statistics*, **11**, 456–461.
- [10] Mallinckrodt, C., Roger, J., Chuang-Stein, C., Molenberghs, G., Lane, P. W., O’Kelly, M., Ratitch, B., Xu, L., Gilbert, S., Mehrotra, D. V., Wolfinger, R. & Thijs, H. (2013), Missing data: turning guidance into action, *Statistics in Biopharmaceutical Research*, **5**, 369–382.
- [11] Molenberghs, G. & Kenward, M. (2007), *Missing Data in Clinical Studies*. John Wiley & Sons, Chichester.
- [12] National Research Council (2010), *The Prevention and Treatment of Missing Data in Clinical Trials, Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education*. The National Academies Press, Washington, DC.
- [13] O’Neill, R. T. & Temple, R. (2012), The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it, *Clinical Pharmacology and Therapeutics*, **91**, 550–554.
- [14] Ratitch, B. & O’Kelly, M. (2011), Implementation of Pattern-Mixture Models Using Standard SAS / STAT Procedures, In *PharmaSUG2011*. at <http://www.pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP04.pdf>
- [15] Royston, P. & White, I. R. (2011), Multiple Imputation by Chained Equations (MICE): Implementation in Stata, *Journal of Statistical Software*, **45**, 1–20.
- [16] Van Buuren, S. & Groothuis-Oudshoorn, K. (2011), MICE: Multivariate imputation by chained equations in R, *Journal of Statistical Software*, **45**, 1–67.

- [17] White, I. R., Carpenter, J., Evans, S. & Schroter, S. (2007), Eliciting and using expert opinions about dropout bias in randomized controlled trials., *Clinical Trials*, **4**, 125–139.
- [18] White, I. R., Horton, N. J., Carpenter, J. & Pocock, S. J. (2011a), Strategy for intention to treat analysis in randomised trials with missing outcome data, *British Medical Journal*, **342**, d40.
- [19] White, I. R., Royston, P. & Wood, A. M. (2011b), Multiple imputation using chained equations: Issues and guidance for practice., *Statistics in Medicine*, **30**, 377–399.