Analyse d'enquêtes cas-cohorte par imputation multiple

Helena Marti, Michel Chavance

Analyse d'enquêtes cas-cohorte par imputation multiple

Helena Marti & Michel Chavance

INSERM-U780 Recherches en épidémiologie et statistiques,
16 avenue Paul Vaillant-Couturier, 94807 Villejuif, France
helena.marti-soler@inserm.fr
Tel +33 (0)1 45 59 50 63, Fax +33 (0)1 45 59 51 69

Résumé

Les estimateurs pondérés utilisés en analyse des études cas-cohorte sont parfois peu efficaces. Or, l'enquête cas-cohorte peut aussi être vue comme un cas particulier de données incomplètes et des méthodes d'analyse pour données incomplètes peuvent être pertinentes, en particulier, l'imputation multiple. Cette approche est basée sur la génération de plusieurs jeux plausibles de données complètes, prenant en compte l'incertitude sur les données manquantes. Si le modèle d'imputation est correctement défini, l'estimateur de l'imputation multiple est non-biaisé. Nous avons montré qu'un modèle d'imputation correct peut être estimé à partir des données complètes (cas et témoins) en utilisant la variable indicatrice des cas comme variable explicative. Nous avons simulé des enquêtes cas-cohorte dont les sous-cohortes étaient sélectionnées par un tirage uniforme ou stratifié. L'imputation multiple et les estimateurs pondérés fournissaient des estimations non-biaisés. Les estimations de l'imputation multiple étaient légèrement plus précises que celles obtenues par l'analyse pondérée. Pour les variables de phase-1, l'augmentation relative des écart-type de l'analyse pondérée par rapport à l'imputation multiple variait de 8 à 39%. Pour les variables de phase-2, l'augmentation relative variait de 3 à 24%. Ainsi, l'imputation multiple, qui utilise toutes les données disponibles et fournit une approximation du maximum de l'estimateur de la vraisemblance partielle, est une bonne alternative à l'estimateur pondéré.

Mots-clés: Enquêtes cas-cohorte, imputation multiple.

Abstract

The weighted estimators used for analyzing case-cohort studies are not fully efficient. Case-cohort studies can be seen as a special type of incomplete data, and methods for analyzing incomplete data could be appropriate, in particular, multiple imputation. This approach is based on the generation of several plausible complete data sets, taking into account the uncertainty about missing values. When the imputation model is correct, it reflects appropriately the distribution of the incomplete variables, respectively among the cases and controls, and the multiple imputation estimator is unbaised. We have shown that a correct imputation model can be estimated from the fully observed data (cases and controls) using the case status as an

explanatory variable. Using simulations, case-cohort data, with subcohort selected by uniform or stratified sampling, were analyzed. Multiple imputation and weighted estimators provided unbiased estimators. The multiple imputation estimators were slightly more precise than those obtained with weighted analysis, especially when the phase-2 variable was linked to the event occurrence. For phase-1 variables, the relative standard deviation increase for the weighted-analysis estimators compared to that of multiple imputation, varied from 8 to 39%. For phase-2 variables, that relative increase ranged from 3 to 24%. Thus, multiple imputation which uses all the available data, gives an approximation of the maximum partial likelihood estimator is a good alternative to weighted analysis.

Keywords: Case-cohort designs, multiple imputation.

1 Introduction

Les études de cohorte sont de plus en plus souvent utilisées en épidémiologie parce qu'elles sont plus faciles à interpréter en termes de causalité. Généralement les maladies étudiées ont une incidence très faible, et la puissance dépend du nombre de cas. Les études de cas-cohorte ainsi que les études cas-témoins emboîtées dans une cohorte permettent d'en réduire le coût au prix d'une perte minime d'efficacité (Langholz 1990).

Les études cas-cohorte sont réalisées en deux phases. 1) La cohorte est sélectionnée par tirage au sort. On recueille l'information de phase-1 sur tous les sujets. Une sous-cohorte est sélectionnée par tirage au sort et la cohorte entière est suivie de manière à identifier la date de survenue du ou des événements d'intérêt. 2) On recueille l'information de phase-2, plus coûteuse, sur tous les cas, qu'ils appartiennent ou non à la sous-cohorte ainsi que sur les sujets de la sous-cohorte. La sous-cohorte peut être sélectionnée par un tirage uniforme ou stratifié (Prentice 1986, Barlow 1999, Therneau 1999, Borgan 2000).

La méthode usuelle d'analyse des enquêtes cas-cohorte est l'analyse pondérée, décrite initialement par Prentice (1986). Or, l'enquête cas-cohorte peut aussi être vue comme un cas particulier de données incomplètes où le processus d'observation est contrôlé par les organisateurs de l'étude. Ainsi, des méthodes d'analyse pour données incomplètes peuvent être pertinentes, en particulier, l'imputation multiple.

L'objectif de ce travail est de mettre en œuvre l'imputation multiple pour analyser les enquêtes cas-cohorte. Nous validerons cette approche en comparant ses résultats à ceux d'un estimateur pondéré classique sur des données entièrement simulées.

2 Observations incomplètes et imputation multiple

Little et Rubin (1987) proposent de distinguer trois processus d'observation: données manquant complètement aléatoirement (MCA) lorsque la probabilité qu'une observation soit incomplète est constante; données manquant aléatoirement (MA) lorsque cette probabilité ne dépend que de valeurs observées; et données manquant non aléatoirement (MNA) lorsque cette probabilité dépend de valeurs non observées.

Les observations incomplètes posent des problèmes de biais, de précision et de puissance. Si on limite l'analyse aux seules observations complètes on s'expose à un biais de sélection pour les processus d'observation MA et MNA. Avec des données MA et une méthode d'analyse pertinente, il est possible d'effectuer des inférences correctes. Le plan cas-cohorte est l'une des rares situations où l'on peut affirmer que les données sont MA car l'observation ne dépend que du statut cas-témoins (échantillonnage uniforme) et éventuellement de variables observées (échantillonnage stratifié).

L'imputation multiple permet d'obtenir une approximation de l'estimateur du maximum de vraisemblance. Sous l'hypothèse de données MA, elle permet de : a) Corriger le biais, b) Obtenir une variance asymptotique correcte. Cette méthode repose sur la génération de plusieurs (M) jeux plausibles de données complètes, en prenant en compte tous les niveaux d'incertitude concernant les valeurs manquantes.
- On ne remplace pas les données manquantes par leur espérance mais par une valeur tirée dans la loi postulée par le modèle.
- Pour tenir compte de l'incertitude sur les paramètres du modèle d'imputation on effectue plusieurs imputations avec des paramètres tirés dans la loi asymptotique de l'estimateur obtenu à partir des observations complètes.

On obtient une estimation du paramètre d'intérêt $\hat{\theta}_m, \quad m = \{1, \ldots, M\}$ et une estimation de la variance de l'estimateur, $\widehat{V}(\hat{\theta}_m)$ pour chaque jeu de données complétées. Si le modèle d'imputation est correct, les estimateurs $\hat{\theta}_m$ sont non-biaisés. Puis, on obtient une estimation unique moyenne de ces M estimateurs, qui est aussi non-biaisée:

$$\hat{\theta}_{IM} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m$$

On peut, grâce à la multiplicité des imputations, estimer correctement la variance de cet estimateur unique, formée par deux composantes: La composante *intra-imputations* (W_{IM}) et la composante *inter-imputations* (B_{IM}) par:

$$\widehat{V}(\hat{\theta}_{IM}) = \widehat{W}_{IM} + \widehat{B}_{IM} = \frac{1}{M} \sum_{m=1}^{M} \widehat{V}(\hat{\theta}_m) + (1 + M^{-1}) \frac{\sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta}_{IM})(\hat{\theta}_m - \hat{\theta}_{IM})'}{M - 1}$$

Pour ne pas biaiser la mesure de la relation entre la réponse modélisée et la variable incomplète le modèle d'imputation doit prendre en compte cette relation. Dans une étude cas-cohorte où une variable serait incomplètement observée parmi les témoins il suffirait d'estimer le modèle d'imputation chez les seuls témoins complètement observés mais le problème est ici compliqué par le processus de censure.

Nous avons montré que la distribution de la variable de phase-2 chez les cas est translatée de la distribution chez les témoins et ne dépend pas du délai de censure. Donc, le modèle d'imputation peut être estimé sur tous les sujets, cas et non-cas, en introduisant une indicatrice des cas comme variable explicative. Le raisonnement pour une variable de phase-2 binaire est analogue.

3 Simulations

Pour comparer les propriétés des estimations obtenues par imputation multiple et par analyse pondérée nous avons réalisé des simulations. Nous avons simulé 2 variables de phase-1 : une variable binaire, Z_1, et une variable gaussienne, Z_3, observées dans la cohorte entière. Nous avons aussi simulé une variable gaussienne de phase-2,Z_2, indépendante de Z_1 mais avec un coefficient de corrélation de 0.3 avec Z_3. Le délai jusqu'à l'événement était distribué selon une loi exponentielle de paramètre $\lambda = exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$. β_1, β_2 et β_3 étaient fixés à 0 ou à $log(2)$. Le délai de censure était distribué uniformément dans l'intervalle $[0, \tau]$, dont τ était définit tel que la probabilité d'événement soit approximativement 0.01 ($\tau = 0.008$). La taille de la cohorte était de 25.000 sujets.

Nous avons simulé une variable de phase-1 prédictive de la variable Z_2, $\widetilde{Z}_2 \equiv Z_2 + \varepsilon$ où $\varepsilon \sim N(0, \sigma^2)$. La corrélation entre Z_2 et $\widetilde{Z}_2$ était approximativement de 0.7 ($\sigma^2 = 1$) pour un scénario et de 0.3 ($\sigma^2 = 9$) pour l'autre.

Nous désirions estimer l'effet de Z_2 sur la survenue de l'événement, après ajustement sur Z_3, d'abord dans le cadre d'un échantillonnage uniforme (résultats non présentés), puis d'un échantillonnage stratifié de la sous-cohorte. La cohorte était divisé en 9 strates, définies par les tertiles de $\widetilde{Z}_2$ et Z_3. L'échantillonnage de la sous-cohorte était réalisé de façon à obtenir 1000 sujets dans la sous-cohorte. Nous avons envisagé 2 niveaux de corrélation entre la variable Z_2 et $\widetilde{Z}_2$, $\rho = 0.3$ et $\rho = 0.7$ (tableau 1). Les estimations obtenues pour les variables de phase-1 ainsi que pour celle de phase-2 étaient non-biaisées, pour tous les deux niveaux de corrélation. Les écarts-type observés, pour les variables de phase-1, étaient du même ordre de grandeur pour la cohorte entière et pour l'imputation multiple, aux deux niveaux de corrélation. La dispersion observée était plus grande avec l'estimateur pondéré. Pour la variable de phase-2, les écart-type observés étaient plus grands avec l'imputation multiple que sur la cohorte entière, mais légèrement plus petits avec imputation multiple qu'avec l'analyse pondérée, notamment quand le paramètre

Table 1: Paramètres estimés (échantillonnage stratifié)

	Corrélation $(Z_2, \widetilde{Z}_2)$=0.7					Corrélation $(Z_2, \widetilde{Z}_2)$=0.3				
	Est	$\widehat{ET}$	ET	PR	Ratio	Est	$\widehat{ET}$	ET	PR	Ratio
$\beta_1 = 0$										
Cohorte	0.0143	0.1553	0.1568	95.0		0.0143	0.1553	0.1568	95.0	
IM	0.0143	0.1553	0.1569	95.0		0.0143	0.1553	0.1569	95.0	
BII	0.0151	0.1713	0.1763	95.3	1.10	0.0165	0.1712	0.1747	94.5	1.10
$\beta_2 = 0$										
Cohorte	0.0002	0.0618	0.0613	95.4		0.0002	0.0618	0.0613	95.4	
IM	0.0003	0.0671	0.0665	95.5		0.0003	0.0709	0.0701	96.5	
BII	0.0014	0.0703	0.0701	95.1	1.05	0.0002	0.0727	0.0717	95.8	1.03
$\beta_3 = 0$										
Cohorte	0.0022	0.0606	0.0624	93.9		0.0022	0.0606	0.0624	93.9	
IM	0.0023	0.0609	0.0625	93.8		0.0019	0.0611	0.0633	93.7	
BII	0.0015	0.0658	0.0682	94.4	1.08	0.0008	0.0651	0.0676	94.4	1.08
$\beta_1 = 0.6931$										
Cohorte	0.7133	0.1737	0.1744	95.2		0.7133	0.1737	0.1744	95.2	
IM	0.7047	0.1741	0.1745	95.5		0.6989	0.1743	0.1748	94.7	
BII	0.7177	0.2011	0.2011	95.7	1.16	0.7232	0.2022	0.2007	95.9	1.16
$\beta_2 = 0.6931$										
Cohorte	0.6940	0.0588	0.0589	95.0		0.6940	0.0588	0.0589	95.0	
IM	0.6856	0.0681	0.0691	93.9		0.6800	0.0726	0.0740	93.7	
BII	0.7040	0.0844	0.0835	95.8	1.24	0.7060	0.0866	0.0944	92.7	1.19
$\beta_3 = 0.6931$										
Cohorte	0.6955	0.0576	0.0621	92.7		0.6955	0.0576	0.0621	92.7	
IM	0.6922	0.0594	0.0626	94.4		0.6915	0.0605	0.0638	94.1	
BII	0.7069	0.0824	0.0894	94.1	1.39	0.7056	0.0837	0.0910	92.9	1.38

Est, estimation moyenne; $\widehat{ET}$, écart-type estimé moyen; ET: écart-type des estimations;
PR, % recouvrement; Ratio, ratio des ET de l'estimateur pondéré et l'imputation multiple;
IM, Modèle d'imputation: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas\_i} + \alpha_3 Z_{3i} + e_i$

était fixé à $log(2)$ et quand la corrélation entre Z_2 et $\widetilde{Z}_2$ était forte. Pour les variables de phase-1, l'augmentation relative des écart-type de l'analyse pondérée par rapport à l'imputation multiple variait de 8 à 39%. Pour les variables de phase-2, l'augmentation relative variait de 3 à 24%.

4 Conclusion

Nous avons observé des estimations plus précises avec l'imputation multiple qu'avec les estimateurs pondérés pour la variable de phase-2. Le gain en précision était plus sensible pour les variables de phase-1 que pour les variables de phase-2. Ainsi, l'imputation multiple est une bonne alternative à l'estimateur pondérée puisque ce dernier ignore l'information de phase-1 apportée par les témoins en dehors de la sous-cohorte tandis

que l'imputation multiple utilise toute l'information disponible en phase-1, dans et en dehors de la sous-cohorte. Pour la variable de phase-2, nous obtenions une estimation plus efficace que celle-lui fournie par l'analyse pondérée car l'imputation multiple fournit une approximation de l'estimateur du maximum de vraisemblance partielle, l'estimateur le plus efficace.

Bibliographie

[1] Barlow, W.E. (1999) Analysis of case-cohort designs, *Journal of Clinical Epidemiology*, 52, 1165–1172.
[2] Borgan, O., Langholz, B., Samuelsen, S.O., Goldstein, L. and Pogoda, J. (2000) Exposure stratified case-cohort designs, *Lifetime Data Analysis*, 6, 39–58.
[3] Langholz, B. and Thomas, D.C. (1990) Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology*, 131, 169–176.
[4] Little, R.J.A. and Rubin, D.B. (1987) *Statictical Analysis with Missing Data*, New York: J. Wiley & Sons.
[5] Prentice, R.L. (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73, 1–11.
[6] Therneau, T.M. and Li, H. (1999) Computing the Cox model for case-cohort design. *Lifetime Data Analysis*, 5, 99–112.