Modélisation des données d'enquêtes cas-cohorte par imputation multiple : Application en épidémiologie cardio-vasculaire

Résumé

Les estimateurs pondérés généralement utilisés pour analyser les enquêtes cas-cohorte ne sont pas pleinement efficaces. Or, les enquêtes cas-cohorte sont 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 manquant au hasard (MA) peuvent être pertinentes, en particulier, l'imputation multiple, qui utilise toute l'information disponible et permet d'approcher l'estimateur du maximum de vraisemblance partielle.

Cette méthode est fondée sur la génération de plusieurs jeux plausibles de données complétées prenant en compte les différents niveaux d'incertitude sur les données manquantes. Elle permet d'adapter facilement n'importe quel outil statistique disponible pour les données de cohorte, par exemple, l'estimation de la capacité prédictive d'un modèle ou d'une variable additionnelle qui pose des problèmes spécifiques dans les enquêtes cas-cohorte. Nous avons montré que le modèle d'imputation doit être estimé à partir de tous les sujets complètement observés (cas et non-cas) en incluant l'indicatrice de statut parmi les variables explicatives.

Nous avons validé cette approche à l'aide de plusieurs séries de simulations: 1) données complètement simulées, où nous connaissions les vraies valeurs des paramètres, 2) enquêtes cas-cohorte simulées à partir de la cohorte PRIME, où nous ne disposions pas d'une variable de phase-1 (observée sur tous les sujets) fortement prédictive de la variable de phase-2 (incomplètement observée), 3) enquêtes cas-cohorte simulées à partir de la cohorte NWTS, où une variable de phase-1 fortement prédictive de la variable de phase-2 était disponible. Ces simulations ont montré que l'imputation multiple fournissait généralement des estimateurs sans biais des risques relatifs. Pour les variables de phase-1, ils approchaient la précision obtenue par l'analyse de la cohorte complète, ils étaient légèrement plus précis que l'estimateur calibré de Breslow et coll. et surtout que les estimateurs pondérés classiques. Pour les variables de phase-2, l'estimateur de l'imputation multiple était généralement sans biais et d'une précision supérieure à celle des estimateurs pondérés classiques et analogue à celle de l'estimateur calibré. Les résultats des simulations réalisées à partir des données de la cohorte NWTS étaient cependant moins bons pour les effets impliquant la variable de phase-2 : les estimateurs de l'imputation multiple étaient légèrement biaisés et moins précis que les estimateurs pondérés. Cela s'explique par la présence de termes d'interaction impliquant la variable de phase-2 dans le modèle d'analyse, d'où la nécessité d'estimer des modèles d'imputation spécifiques à différentes strates de la cohorte incluant parfois trop peu de cas pour que les conditions asymptotiques soient réunies.

Nous recommandons d'utiliser l'imputation multiple pour obtenir des estimations plus précises des risques relatifs, tout en s'assurant qu'elles sont analogues à celles fournies par les analyses pondérées.

Nos simulations ont également montré que l'imputation multiple fournissait des estimations de la valeur prédictive d'un modèle (C de Harrell) ou d'une variable additionnelle (différence des indices C, NRI ou IDI) analogues à celles fournies par la cohorte complète.

Mots clés: Enquêtes cas-cohorte, estimateurs pondérés, imputation multiple, capacité prédictive.

Modeling of case-cohort data by multiple imputation: application to cardio-vascular epidemiology

Abstract

The weighted estimators generally used for analyzing case-cohort studies are not fully efficient. However, case-cohort surveys are a special type of incomplete data in which the observation process is controlled by the study organizers. So, methods for analyzing Missing At Random (MAR) data could be appropriate, in particular, multiple imputation, which uses all the available information and allows to approximate the partial maximum likelihood estimator.

This approach is based on the generation of several plausible complete data sets, taking into account all the uncertainty about the missing values. It allows adapting any statistical tool available for cohort data, for instance, estimators of the predictive ability of a model or of an additional variable, which meet specific problems with case-cohort data. We have shown that the imputation model must be estimated on all the completely observed subjects (cases and non-cases) including the case indicator among the explanatory variables.

We validated this approach with several sets of simulations: 1) completely simulated data where the true parameter values were known, 2) case-cohort data simulated from the PRIME cohort, without any phase-1 variable (completely observed) strongly predictive of the phase-2 variable (incompletely observed), 3) case-cohort data simulated from de NWTS cohort, where a phase-1 variable strongly predictive of the phase-2 variable was available. These simulations showed that multiple imputation generally provided unbiased estimates of the risk ratios. For the phase-1 variables, they were almost as precise as the estimates provided by the full cohort, slightly more precise than Breslow *et al*. calibrated estimator and still more precise than classical weighted estimators. For the phase-2 variables, the multiple imputation estimator was generally unbiased, with a precision better than classical weighted estimators and similar to Breslow *et al*. calibrated estimator. The simulations performed with the NWTS cohort data provided less satisfactory results for the effects where the phase-2 variable was involved: the multiple imputation estimators were slightly biased and less precise than the weighted estimators. This can be explained by the interactions terms involving the phase-2 variable in the analysis model and the necessity of estimating specific imputation models in different strata not including sometimes enough cases to satisfy the asymptotic conditions.

We advocate the use of multiple imputation for improving the precision of the risk ratios estimates while making sure they are similar to the weighted estimates.

Our simulations also showed that multiple imputation provided estimates of a model predictive value (Harrell's C) or of an additional variable (difference of C indices, NRI or IDI) similar to those obtained from the full cohort.

Keywords: Case-cohort surveys, weighted estimators, multiple imputation, predictive ability.