Statistical analysis of matched survival data in national Heath databases

Vanessa Chezeu*[1], Valerie Gares[2], and Jean-François Dupuy[3]

[1]Institut National des Sciences Appliquées - Rennes – Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes – France

[2]Institut de Recherche Mathématique de Rennes – Université de Rennes, Institut National des Sciences Appliquées - Rennes, École normale supérieure - Rennes, Université de Rennes 2, Centre National de la Recherche Scientifique, Institut Agro Rennes ANgers – France

[3]Institut National des Sciences Appliquées - Rennes – Institut National des Sciences Appliquées – France

Résumé

The French National Health Data System (SNDS in french) collects longitudinal health records and insurance information on most of the French population, and as such, it contains a huge amount of information pertaining to health insurance, hospitalizations, causes of death, pathologies. . . These data can be used to enrich other existing cohorts, or health registries, which allows to get a more comprehensive medical information on each patient, and thus, to improve the subsequent statistical analysis. However, patients in the SNDS and health databases are usually anonymised, and moreover, no unique patient identifier is available to match the SNDS and other registries. Fellegi and Sunter (1969) proposed a solution to this matching issue. Their probabilistic record linkage method is based on the fact that we usually have at hand some "matching variables". These matching variables are partial identifiers common to both databases (e.g., gender, postal codes, dates of the treatment. . .). They are not sufficient to match patients between two databases, but nevertheless allow to calculate "matching probabilities" for each pair of patients taken in the SNDS and the health registry of interest. These probabilities allow to link each patient in the SNDS to her/his most likely counterpart in the health registry. But these probabilities also convey some uncertainty on the matching process, and this uncertainty must be taken into account in any subsequent statistical analysis. In this talk, we propose a new method in order to take account of these errors in a survival analysis based on the Cox model. This method is based on the well-known EM algorithm for estimation in a missing-data context. We first describe the basic principles of the proposed method and then, we assess its properties via simulations. Keywords: Record linkage; Censored data; Duration data; Public health; Numerical simulations.

*Intervenant