Prédiction dynamique d'événements à partir de multiples marqueurs longitudinaux par « model averaging »

Hélène Jacqmin-Gadda[1], Taban Baghfalaki[1], Reza Hashemi[2]

[1]Inserm, Research Center U1219, Univ. Bordeaux, ISPED, F33076 Bordeaux, France
[2] Department of Statistics, Razi University, Kermanshah, Iran

Résumé :

Au cours des dernières années, de nombreux travaux en biostatistique ont porté sur le développement d'outils de prédiction dynamique d'événements de santé à partir de mesures répétées de marqueurs ou facteurs de risque. Ces outils constituent une contribution essentielle au développement de la médecine personnalisée car ils permettent de prédire le risque individuel de survenue d'un événement dans une fenêtre de temps à partir de l'ensemble des informations collectées jusqu'au temps courant. Les modèles conjoints pour un ou plusieurs temps d'événements (potentiellement compétitifs) et les données répétées de marqueurs longitudinaux est la méthode privilégiée dans ce domaine. Ces modèles combinent donc des modèles mixtes et des modèles de survie liés soient par des effets aléatoires partagés soit par des classes latentes. Bien que l'estimation fréquentiste ou Bayésienne de ces modèles soient dorénavant possibles grâce à différents logiciels, l'estimation de modèles conjoints incluant de nombreux marqueurs longitudinaux reste un défi majeur. Quelle que soit l'approche, l'estimation conjointe devient impossible lorsque le nombre de marqueurs est trop grand en raison du nombre trop élevé d'effets aléatoires et de paramètres à estimer et à l'imprécision du calcul numérique des intégrales de grandes dimensions.
Dans ce travail, nous proposons d'estimer les prédictions dynamiques individuelles basées sur les mesures répétées de multiples marqueurs par une moyenne pondérée des prédictions estimées à partir de modèles conjoints ne comportant chacun qu'un marqueur. Les poids peuvent dépendre du temps de prédiction. Ils sont estimés en minimisant le score de Brier dépendant du temps. Bien que le temps de calcul global puisse être long, cette approche est toujours réalisable (et aisément parallélisable), même lorsque le nombre de prédicteurs longitudinaux est très élevé. Ses avantages et limites sont évalués par divers scénarios de simulations et comparer aux prédictions du modèle multi-marqueurs dans les scénarios où il est estimable. Cette méthode est utilisée pour prédire le risque de décès dans la cohorte de personnes âgées 3C à partir de 17 prédicteurs longitudinaux et ses capacités prédictives sont comparées à celles de plusieurs approches alternatives.

Abstract :

In recent years, much work in biostatistics has focused on the development of dynamic prediction tools for health events, based on repeated measurements of markers or risk factors. These tools make an essential contribution to the development of personalized medicine, as they enable us to predict the individual risk of an event occurring within a time window, based on all the information collected up to the current time. Joint models for one or more (potentially competing) event times and repeated measures of longitudinal markers are the preferred method in this field. These models thus combine mixed models and survival models linked either by shared random effects or by latent classes. Although frequentist or Bayesian estimation of these models is now possible thanks to various open-access packages, estimation of joint models including numerous longitudinal markers remains a major challenge. Whatever the approach, joint estimation becomes impossible when the number of markers is too large, due to the excessive number of random effects and parameters to be estimated, and to the imprecision of the numerical calculation of high-dimensional integrals.

In this work, we propose to estimate individual dynamic predictions based on repeated measurements of multiple markers by a weighted average of predictions estimated from joint models each comprising only one marker. Weights may depend on prediction time. They are estimated by minimizing the time-dependent Brier score. Although the overall computation time can be long, this approach is still feasible (and easily parallelizable), even when the number of longitudinal predictors is very high. Its advantages and limitations are evaluated in various simulation scenarios and compared with the predictions of the multi-marker model in those scenarios where it may be estimated. This method is used to predict the risk of death in the 3C elderly cohort from 17 longitudinal predictors, and its predictive abilities are compared with those of several alternative approaches