Inférence bayésienne et méthodes MCMC pour l'estimation de modèles conjoints de données longitudinales non linéaires et de survie

Marion Kerioui,[1] Solène Desmée,[1] Jérémie Guedj[2]
[1] *Université de Tours, Université de Nantes, INSERM SPHERE, UMR 1246, Tours, France*
[2] *INSERM IAME, UMR 1137, F-75018 Paris, France; Université Paris Diderot, F-75018 Paris, France*

Contexte

En oncologie, le critère principal d'évaluation d'un traitement est souvent le délai de survenue du décès. Le suivi du patient peut être assuré par des mesures longitudinales de biomarqueurs qui apportent de l'information sur la réponse du patient à l'intervention ou permet de détecter de manière précoce les patients à risque. Pour le cancer de la prostate, on observe la protéine PSA (Prostate Specific Antigen) produite par les cellules de la prostate et les cellules tumorales et dont la production est supposée liée à l'évolution de la maladie. La modélisation des processus longitudinaux et de survie peut se faire en utilisant la modélisation conjointe [1]. La survenue de l'évènement est alors décrite par un modèle de survie à risque proportionnel dont la fonction de risque instantanée dépend de la cinétique du biomarqueur, elle-même décrite par un modèle à effets mixtes. La pharmacométrie permet de décrire la cinétique du biomarqueur en fonction de paramètres biologiques. Ce sont souvent des modèles non-linéaires, ce qui complexifie l'estimation des paramètres du modèle conjoint car la vraisemblance du modèle n'a pas de forme analytique.
Le logiciel d'inférence bayésienne par algorithme HMC (Hamiltonian Monte-Carlo), Stan [2], a été développé récemment pour l'estimation de modèles complexes, pour lesquels les algorithmes MCMC classiques rencontrent des difficultés de convergence. L'approche bayésienne paraît particulièrement en accord avec ce type de modèle, l'expérience d'essais cliniques passés et l'avis d'experts pouvant être pris en compte à travers la loi *a priori* des paramètres ayant un sens biologique. Stan n'a jamais été utilisé pour l'estimation de modèles conjoints non-linéaires.

Objectif

L'objectif de ce travail est de réaliser une étude de faisabilité pour l'estimation des paramètres d'un modèle conjoint non-linéaire par le logiciel Stan. Pour cela, on a recours à une étude de simulation de données de PSA et de survie, dans laquelle on évalue la capacité d'estimation de Stan en terme de précision d'estimation, de maîtrise de l'incertitude et de temps de calcul.

Méthodes

Pour le design principal de simulation, on s'appuie sur un cadre de simulation déjà utilisé précédemment pour le cancer de la prostate [3]. La cinétique du PSA est décrite par une fonction bi-exponentielle qui dépend de quatre paramètres. La fonction de risque instantanée s'exprime grâce à une fonction de Weibull et à la valeur courante du PSA. L'impact de la valeur courante du PSA sur la survie est contrôlée par un paramètre de lien dont on fait varier la force (aucun lien, lien faible, lien fort). On simule 100 jeux de données de 100 patients, avec des mesures du PSA toutes les 3 semaines durant deux ans. On simulera 3 designs supplémentaires de jeux de données, en faisant varier le nombre de patients ($N = 30$) et la fréquence de mesures du biomarqueur (toutes les 9 semaines).
On implémente la totalité du modèle en langage Stan, notamment l'intégration numérique de la fonction de risque instantanée pour expliciter la densité du processus de survie. La calibration de l'algorithme est spécifiée grâce au package R `rstan` avec 2000 itérations au total, dont 1000 itérations de chauffe et on garde une réalisation sur 5 pour éviter l'autocorrélation. Les lois *a priori* sont choisies pour être peu informatives tout en tenant compte des contraintes biologiques des paramètres.
On s'intéresse à l'estimation des paramètres de population, dont on calculera les erreurs relatives d'estimation (REE) et la différence entre l'erreur moyenne quadratique (RMSE) et l'écart-type *a posteriori* du paramètre pour chaque jeux de données de chaque design. On considère également les taux de couverture à 95% des paramètres de population.

Résultats

Pour le design principal de simulation, les paramètres longitudinaux et le paramètre de lien ne présentent pas de biais ($|\text{REE}| \leq 25\%$). On observe en revanche l'apparition d'un biais pour l'estimation des paramètres de Weibull lorsque la force du lien s'accroît (en moyenne, lien faible : REE=-4%, lien fort : REE=-20%). Les taux de couverture sont proches de 95% quel que soit le scénario, hormis pour le paramètre d'échelle de Weibull présentant un biais d'estimation. Le temps de calcul moyen pour un jeu de données de 100 patients est de 23 heures et diminue à 4 heures pour un jeu de 30 individus.

Conclusion

Ces travaux constituent, à notre connaissance, la première implémentation d'un modèle conjoint non-linéaire en Stan. Ces travaux demandent à être approfondis pour améliorer la précision d'estimation des paramètres de Weibull et réduire les temps de calcul par une meilleure calibration de l'algorithme. On considèrera également des critères d'évaluation supplémentaires. Ces travaux de stage ont été financés par le GDR Statistiques et Santé.

References

[1] D. Rizopoulos. Joint models for longitudinal and time-to-event data: With applications in R. 2012.

[2] B. Carpenter A. Gelman et al. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[3] S. Desmée et al. Nonlinear mixed-effect models for prostate-specific antigen kinetics and link with survival in the context of metastatic prostate cancer: a comparison by simulation of two-stage and joint approaches. *The AAPS journal*, 17(3), 2015.