IRSN
INSTITUT
DE RADIOPROTECTION
ET DE SÛRETÉ NUCLÉAIRE

PROPOSITION DE SUJET DE STAGE

Entreprise et laboratoire d'accueil

Institut de Radioprotection et de Sûreté Nucléaire (IRSN)
Laboratoire d'épidémiologie des rayonnements ionisants (LEPID)
31 avenue de la Division Leclerc - BP 17
F-92262 Fontenay-aux-Roses Cedex

En collaboration avec l'Unité Mixte de Recherche « Mathématique et Informatique Appliquée » de AgroParistech/INRA

Contacts

Sophie Ancelet **sophie.ancelet@irsn.fr**
Tel : 01.58.35.79.89
Eric Parent **eric.parent@agroparistech.fr**

Intitulé du sujet :

Calibration bayésienne de protocoles d'études de cohorte et calcul de puissance pour l'estimation d'un risque sanitaire en épidémiologie des rayonnements ionisants

Contexte :

Au sein du Laboratoire d'épidémiologie des rayonnements ionisants (LEPID) de l'IRSN, des études de cohorte sont régulièrement mises en place puis mises à jour afin, notamment, d'estimer les effets sanitaires potentiellement induits par une exposition aux rayonnements ionisants (RI). Ces études de cohorte s'intéressent principalement à des expositions faibles aux RI et à des risques radio-induits associés eux-mêmes potentiellement faibles. Ainsi, elles n'ont pas toujours la puissance statistique nécessaire pour mettre en évidence, si elle existe, une association entre la variable réponse d'intérêt (e. g. nombre de décès, délai de survenue d'une pathologie donnée) et l'exposition aux RI. Cette limite conduit à une mise en doute des (futurs) résultats d'analyse alors que le coût budgétaire engendré par le recueil de données observationnelles longitudinales est souvent élevé.

En épidémiologie des RI, une approche standard utilisée pour augmenter la puissance statistique des études de cohorte est de construire des cohortes internationales combinant les données épidémiologiques issues de plusieurs cohortes nationales afin d'augmenter le nombre d'individus. Une autre approche consiste à mettre à jour régulièrement les cohortes existantes afin d'augmenter le temps de suivi des individus. Enfin, des calculs de puissance statistique sont généralement réalisés dans un cadre fréquentiste avec, pour objectif spécifique, de mettre en évidence un accroissement fixé de l'incidence d'une pathologie chez les sujets d'une cohorte par rapport à la population générale [1]. Néanmoins, à notre connaissance, aucun outil statistique n'est actuellement disponible pour :

a) calculer la puissance d'une étude de cohorte à nombre de sujets et temps de suivi préfixés

b) calibrer un protocole d'étude de cohorte (e.g., nombre de sujets, temps de suivi...) permettant d'atteindre une puissance minimale fixée (typiquement 80%) ou optimale sous contrainte budgétaire

quand l'*objectif spécifique* est de mettre en évidence, si elle existe, une association entre une variable réponse et une exposition aux RI en univers incertain (e.g., statuts vitaux, taux de base, valeurs des facteurs de risque inconnus) et à partir de modèles dose-réponse standards dans le domaine (e.g., régression de Poisson et modèle de survie à hasards proportionnels en excès de risque (instantané)).

Bien que les travaux de Little et al. (2010) [2] aient récemment abordé le problème, ceux-ci se focalisent uniquement sur la question a), ne considèrent pas les modèles dose-réponse classiquement utilisés en épidémiologie des RI et se placent dans un cadre statistique fréquentiste ne permettant pas de tenir compte de l'incertitude sur les nombreuses quantités intervenant dans la calibration des protocoles d'étude. Quelques logiciels [3] ont été proposés pour la calibration de plans d'expérience optimaux dans le cadre de

l'analyse de données longitudinales mais ces derniers concernent essentiellement la recherche clinique pour laquelle les variables du plan d'expérience sont contrôlables expérimentalement. Enfin, bien que la calibration de plans d'expérience optimaux pour la mise en place d'études expérimentales (e.g., essais cliniques) ait fait l'objet de nombreux travaux de recherche ces dernières années [4] [5], cela est nettement moins le cas pour la calibration de protocoles d'étude pour données observationnelles longitudinales telles que celles rencontrées en épidémiologie. Une des différences majeures est que les mesures d'exposition ainsi que les valeurs des différents facteurs de risque susceptibles d'entrer en jeu dans l'occurrence de la pathologie d'intérêt ne sont pas contrôlables par l'épidémiologiste. Par ailleurs, contrairement aux études expérimentales, un certain nombre de facteurs incontrôlables et inhérents à tout processus observationnel peuvent venir affaiblir le protocole d'étude initial (e.g., perdus de vue, données manquantes, sources de biais…). Ainsi, des travaux de recherche semblent nécessaires dans ce contexte afin de développer des outils statistiques mieux appropriés et permettant d'envisager rapidement le passage de la preuve de concept à l'implémentation opérationnelle.

Mission :

L'objectif du stage est de proposer et mettre en œuvre une première approche statistique bayésienne pour le calcul de puissance statistique (question a)) et la calibration de protocoles d'études de cohorte (question b)) en épidémiologie des RI. L'approche bayésienne est en effet particulièrement reconnue pour la prise de décisions en univers incertain [6]. La question a) se traduira, en particulier, par le calcul d'un (de) critère(s) bayésien(s) de sélection de modèle (e.g., BIC, facteurs de Bayes,…) [7]. La question b) nécessitera par ailleurs de définir puis d'optimiser [8] en univers incertain une fonction d'utilité adaptée au problème et au modèle dose-réponse considéré. Les travaux de Chaloner et Verdinelli [5] pourront servir de base à la réflexion. Dans le cadre de ce stage, le stagiaire se concentrera sur une classe choisie de modèles dose-réponse. Enfin, la méthodologie développée pourra être appliquée à des cas d'études de cohorte du LEPID :

- La cohorte enfant-scanner : elle a pour objectif d'étudier le risque potentiel de cancer radio-induit (e.g., leucémie, tumeur cérébrale, lymphome) après exposition au scanner dans l'enfance. 110 000 enfants ont été suivis en moyenne 8 ans entre 2000 et 2013. Une dose à l'organe est disponible pour tous les enfants.

- La cohorte Coccinelle : elle a pour objectif d'étudier le risque potentiel de cancer radio-induit (e.g., leucémie, tumeur cérébrale, lymphome) après exposition à des cathétérismes cardiaques pédiatriques. 16 000 enfants ont été suivis en moyenne 8 ans entre 2000 et 2016. Une dose à l'organe est disponible pour une partie des enfants.

- La cohorte française des travailleurs d'EDF : elle a pour objectif d'améliorer les connaissances sur les pathologies susceptibles de se développer tardivement après une exposition chronique et à faibles doses aux RI (e.g., cancer du poumon, leucémie…). 30000 agents statutaires ayant travaillé au moins un an entre 1968 et 2014 sont suivis (moyenne 27 ans). Des données d'exposition sont disponibles pour la cohorte.

- Le projet SPACE : il vise à mettre en place, pour la première fois en France, une cohorte de personnels navigants d'Air France, afin d'étudier la mortalité de cette population en relation avec son exposition professionnelle aux rayonnements cosmiques. La cohorte n'a pas encore été mise en place. Environ 43000 personnels navigants embauchés depuis les années 70 par Air France pourraient être intégrés dans l'étude. Aucune information concernant les expositions aux RI n'a été recueillie.

Résultats attendus :

Le stagiaire devra fournir :

- une **estimation de la puissance statistique** relative à une ou, si le temps le permet, plusieurs des 4 études de cohorte précédemment décrites quand l'objectif spécifique est de mettre en évidence des risques radio-induits de l'ordre de ceux observés au sein de la cohorte des bombardements d'Hiroshima-Nagasaki [9] ou d'autres études publiées.

- une **estimation des composantes-clés du protocole d'étude de cohorte** (e.g., combien de sujets ? Quel temps de suivi ?...) qui permettrait d'atteindre une puissance minimale fixée (typiquement 80%) ou optimale sous contrainte budgétaire pour un ou plusieurs des 4 cas d'étude ci-dessus.

- une **fonction d'utilité** la plus générique possible permettant de répondre à la question b)

- une **analyse critique** de la méthode proposée

Intérêts du stage:

Ce travail de stage présente trois intérêts principaux :

- un intérêt direct pour les épidémiologistes du Laboratoire d'épidémiologie des rayonnements ionisants de l'IRSN. En effet, il doit au moins permettre d'apporter de premiers éléments de réponse aux questions a) et b) auxquelles ces derniers sont régulièrement confrontés en amont de la mise en place d'études de cohorte ou de la mise à jour de cohortes existantes.

- un intérêt en santé publique et en radioprotection en contribuant indirectement à l'amélioration des connaissances sur les effets sanitaires radio-induits à faibles doses en permettant ou simplement en apportant de l'information concernant la faisabilité de la mise en place d'études de cohorte ayant une puissance statistique optimale, compte-tenu des contraintes budgétaires fixées, pour la mise en évidence de ces risques, potentiellement faibles : « there is a need to recognize that studies of high statistical power are necessary in order to be sure that health effects at these doses have not been missed » (Rapport 2012 du Comité Scientifique des Nations Unies sur les effets des rayonnements ionisants (UNSCEAR)).

- Un intérêt méthodologique en contribuant à promouvoir l'utilisation de l'approche bayésienne qui est encore très rarement utilisée en épidémiologie des rayonnements ionisants malgré sa grande souplesse pour la modélisation de phénomènes aléatoires complexes, la prise en compte de sources d'incertitude multiples et sa pertinence pour la prise de décisions en univers incertain.

Profil recherché :

Ce stage de 6 mois (à partir de février, mars ou avril 2018) est destiné à un(e) étudiant(e) de M2 ou équivalent avec spécialité statistique appliquée aux sciences du vivant ayant des bases solides en modélisation probabiliste et en statistique bayésienne ainsi qu'un intérêt pour la programmation informatique et les applications en épidémiologie/santé publique. La maitrise du langage R est indispensable. Une excellente maitrise de l'anglais est exigée et une connaissance du langage de programmation Python serait un plus.

Bibliographie :

[1] Bouyer, J., et al. (1996) Epidémiologie : Principes et méthodes quantitatives. Les éditions INSERM.

[2] Little M.P, Wakeford R., Lubin, J.H., Kendal G.M. (2010) The statistical power of epidemiological studies analyzing the relationship between exposure to ionizing radiation and cancer with special reference to childhood leukemia and natural background radiation. *Radiation Research.* 174(3): 387–402.

[3] Vazquez-Alcocer A. et al. (2014) LADES: A Software for Constructing and Analyzing Longitudinal Designs in Biomedical Research. PlosOne. 9(7)

[4] Atkinson, A. C.; Donev, A. N.; Tobias, R. D. (2007). Optimum experimental designs, with SAS. Oxford University Press. pp. 511+xvi. ISBN 978-0-19-929660-6.

[5] Chaloner K., Verdinelli I (1995) Bayesian Experimental Design: a review. Volume 10, Number 3 (1995), 273-304.. Statistical Science.

[6] Parent E., Bernier, J. (2007) Le raisonnement bayésien: modélisation et inférence. Springer. ISBN 978-2-287-33906-6

[7] Collectif BIOBAYES : Albert I, Ancelet S, David O, Denis J-B, Makowski D, Parent E, Rau A, Soubeyrand S. (2015) Initiation à la statistique bayésienne : Bases théoriques et applications en alimentation, environnement, épidémiologie et génétique. Ellipses. ISBN:9782340005013

[8] Ryan E.G., Drovandi C.C., McGree J.M. and Pettitt A.N. (2015) A review of modern computational algorithms for Bayesian optimal design. International Statistical Review.

[9] Ozasa et al. (2012) Studies of the Mortality of Atomic Bomb Survivors, Report 14, 1950-2003: an overview of cancer and noncancer diseases. *Radiation Research.* 177(3):229-243