Estimation consistante du nombre de clusters non vides dans les modèles de mélange bayésiens par régression sur profils d'exposition. Application en épidémiologie des rayonnements ionisants

Julie Fendler [1] & Sophie Ancelet [1] & Chantal Guihenneuc [2]

[1] *IRSN, PSE-SANTE/SESANE/LEPID, France,*
emails: julie.fendler@irsn.fr ; sophie.ancelet@irsn.fr
[2] *Université de Paris Cité, BioSTM - UR 7537, Paris, France,*
email: chantal.guihenneuc-jouyaux@u-paris.fr

Résumé. Depuis sa définition par Wild (2005), l'exposome est de plus en plus mis en avant pour son rôle dans l'apparition et le développement de maladies multifactorielles telles que le diabète, le cancer ou l'asthme. Cependant, les données d'exposome sont des données de grande dimension, souvent composées de variables fortement corrélées entre elles (multi-colinéarité). Pour l'analyse de données multi-colinéaires, des méthodes statistiques spécifiques sont nécessaires. Parmi elles, les modèles de mélange bayésiens par régression sur profils d'exposition (BPRM). Ces modèles hiérarchiques permettent de former des groupes d'individus ayant des profils d'exposition similaires à plusieurs facteurs de risque et d'estimer un risque sanitaire pour chacun de ces groupes. La répartition des individus en groupes et l'estimation du risque sanitaire associé se font conjointement sous le paradigme bayésien. Dans les modèles BPRM, le sous-modèle permettant l'attribution de chaque individu à un groupe repose sur un processus de Dirichlet latent, pour lequel l'estimation du nombre de groupes non vides est connue pour être inconsistante. Notre cas d'application demandant d'estimer un nombre interprétable de groupes non vides en présence de données faiblement informatives, nous proposons d'adapter l'algorithme MCMC d'inférence desmodèles BPRM afin de faciliter l'estimation du paramètre de concentration du processus de Dirichlet latent, dans une situation où le signal dans les données est faible. Enfin, nous adaptons aux modèles BPRM et nous comparons différentes méthodes de post-traitement afin de permettre la consistance du nombre estimé de groupes non vides.

Mots-clés. Modèles de mélange, Multi-colinéarité, Processus de Dirichlet, Inférence bayésienne, Analyse de survie, Epidémiologie

Abstract. Since its definition by Wild (2005), the exposome has been increasingly highlighted for its role in the onset and development of multifactorial diseases such as diabetes, cancer and asthma. However, exposome data are large-scale data composed of variables that are often highly correlated with each other (multi-colinearity). Specific statistical methods are needed to deal with multi-colinear data. These include Bayesian Profiles Regression Mixture models (BPRM). These hierarchical models allow to identify clusters of individuals with similar exposure profiles to several risk factors and to estimate a

health risk associated to each cluster. Clustering and risk estimation are carried out jointly under the Bayesian paradigm. In BPRM models, the sub-model used to assign each individual to a cluster is based on a latent Dirichlet process, that is well-known to provide an inconsistent estimation of the number of non-empty clusters. Since our application requires the estimation of an interpretable number of non-empty clusters in the presence of poorly informative data, we propose to adapt the MCMC algorithm used to fit the BPRM models in order to make easier the estimation of the concentration parameter of the latent Dirichlet process, when the data signal is weak. Finally, we adapt to BPRM models and we compare various post-processing methods in order to allow consistency in the estimated number of non-empty groups.

Keywords. Mixture models, Multi-colinearity, Dirichlet process, Bayesian Inference, Survival Modeling, Epidemiology

1 Introduction

Nous nous intéressons aux méthodes statistiques permettant de traiter le problème dit de multicolinéarité, posé par la prise en compte simultanée de variables explicatives fortement corrélées entre elles dans des modèles de régression multiples. Ce problème se rencontre notamment dans les études épidémiologiques, dont l'un des enjeux actuel est de tenir compte au mieux de l'exposome, défini par Wild (2005), lorsqu'on estime des risques de maladies multifactorielles telles que le diabète, le cancer ou l'asthme.

Dans le cas d'une multicolinéarité trop prononcée, l'utilisation d'un modèle de régression linéaire standard incluant simultanément plusieurs facteurs de risque (i.e., variables explicatives) corrélés n'est pas adaptée: les coefficients de régression estimés sont instables et peuvent changer de signe en fonction du sous-ensemble de variables explicatives sélectionné mais aussi du sous-échantillon d'individus statistiques considéré. L'imprécision des estimations se traduit par une variance augmentée ne permettant souvent plus de conclure à des associations significatives entre la variable réponse et les variables explicatives. Dans la pratique, ce problème est souvent surmonté en sélectionnant uniquement une variable explicative (celle influençant le plus la variable réponse) parmi les variables corrélées disponibles. Cette solution ne permet ni l'étude des effets de l'ensemble des variables explicatives d'intérêt ni l'étude de probables effets sanitaires conjoints (i.e., synergiques ou antagonistes).

Les modèles de mélange bayésiens par régression sur profils d'exposition (modèles BPRM) proposés par Molitor *et al.* (2010) permettent de répondre au problème de multicolinéarité ci-dessus. Ces modèles hiérarchiques permettent de former des groupes d'individus ayant des profils d'exposition similaires à plusieurs facteurs de risque et d'estimer un risque sanitaire pour chacun de ces groupes. La

répartition des individus en groupes et l'estimation du risque sanitaire associé se font conjointement sous le paradigme bayésien. Un post-traitement (i.e., mené après l'inférence) est indispensable à l'identification et la caractérisation des clusters.

Dans les modèles BPRM, le sous-modèle permettant l'attribution de chaque individu à un groupe repose sur un processus de Dirichlet latent, pour lequel l'estimation du nombre de groupes non vides est connue pour être inconsistante (Miller et Harrisson (2013)). Par ailleurs, l'inférence bayésienne des modèles BPRM, et notamment du paramètre de concentration du processus de Dirichlet latent, à l'aide d'un algorithme MCMC standard, pose des difficultés (Liverani *et al.* (2015); Belloni *et al.* (2020)).

Dans ce travail, le cas d'application traité (cf. section 2) présente deux spécificités : 1) il demande d'estimer un nombre interprétable de clusters non vides; 2) il demande d'estimer un risque sanitaire faible en présence de données potentiellement peu informatives. Dans ce contexte, nous proposons tout d'abord d'adapter l'algorithme d'inférence bayésienne proposé par Liverani *et al.* (2015) afin de faciliter l'estimation du paramètre de concentration du processus de Dirichlet latent, dans une situation où le signal dans les données est potentiellement faible. De plus, nous proposons d'adapter les méthodes de post-traitement proposées par Wade et Ghahramani (2018) et Guha *et al.* (2019) pour les processus de Dirichlet aux modèles BPRM, afin de permettre la consistance du nombre estimé de clusters non vides. Enfin, nous comparons ces deux méthodes à celles proposées par Molitor *et al.* (2010) et Belloni *et al.* (2020). A notre connaissance, aucune étude n'a été menée afin de comparer ces différentes méthodes.

2 Cas d'étude

La population des mineurs d'uranium est une population de référence pour étudier les effets sanitaires d'une exposition chronique à faibles doses et à différentes sources de rayonnements ionisants (RI). En effet, dans le cadre de leur activité professionnelle, les mineurs d'uranium sont simultanément exposés au radon et ses descendants à vie courte (appelé simplement radon par la suite), aux poussières d'uranium et aux rayonnements gamma. En pratique, les effets sanitaires associés à ces expositions sont généralement étudiés de manière monofactorielle pour chaque source de RI. Ces estimations servent ensuite de base à la définition de normes de radio-protection.

Dans ce travail, nous cherchons à estimer et caractériser le risque de décès par cancer du poumon dans la cohorte française post-55 des mineurs d'uranium, en tenant compte de leur exposition simultanée au radon, aux rayonnements gamma et aux poussières d'uranium. Cette cohorte inclut 3377 mineurs d'uranium embauchés après le 31 décembre 1955. Ils ont été suivis en moyenne pendant 37 ans et 130 cas de décès par cancer du poumon ont été observés. De précédents

travaux ont montré que ces expositions sont fortement corrélées entre elles et qu'on est bien dans un contexte de multicolinéarité prononcée (Belloni *et al.* (2020); Vacquier et al. (2011)). Chacune de ces trois sources d'exposition aux RI a été associée, de manière individuelle, à une augmentation du risque de décès par cancer du poumon (Rage *et al.* (2015)). Cependant l'impact d'une co-exposition simultanée à l'ensemble de ces sources est encore mal caractérisé.

3 Méthode

3.1 Description du modèle BPRM

Un modèle de mélange bayésien par régression sur profils d'exposition (BPRM) peut être vu comme un modèle hiérarchique composé de trois sous-modèles :

- un sous-modèle de maladie qui décrit l'association entre la survenue d'une pathologie (exprimée, par exemple, par l'âge au décès par cancer du poumon d'un mineur) et un profil d'exposition. La force de cette association est quantifiée avec un coefficient de risque associé à chaque profil et un risque instantané de base de développer la pathologie considérée.
- un sous-modèle d'exposition qui décrit la distribution de probabilité des variables d'exposition (continues et discrètes) dans chaque groupe
- un sous-modèle d'attribution qui décrit la répartition des individus dans les différents groupes

3.1.1 Sous-modèle de maladie

Dans ce travail, le sous-modèle de maladie considéré est un modèle de survie. Nous adaptons aux modèles BPRM, le modèle exposition-risque linéaire - appelé modèle en excès de risque instantané (EHR) - classiquement utilisé en épidémiologie des RI.

La variable réponse E_i est l'âge (en jours) à la survenue de l'événement considéré d'un individu i. Cette variable est censurée à droite (individu perdu de vue, décès par autre cause, fin de suivi) et tronquée à gauche (car l'échelle de temps considérée est l'âge). On note Z_i l'âge de l'individu i à la censure. On observe $Y_i = min(E_i, Z_i)$ et δ_i^Y l'indicateur binaire de non-censure (=1 si l'événement étudié est observé pour l'individu i; =0 si l'individu i est censuré avant la survenue de l'événement étudié).

Le risque instantané de l'individu i au temps t, noté $h_i(t)$, est défini par:

$$h_i(t) = h_0(t) \cdot (1 + \beta_{C_i})$$

C_i désigne le label de groupe (inconnu) auquel appartient l'individu i et β_c l'excès de risque instantané (inconnu) de survenue de l'événement étudié, associé au groupe c. $h_0(t)$ désigne le risque instantané de base de survenue de l'événement considéré au temps t. Nous supposons une fonction de risque instantanée de base h_0 de type Weibull :

$$h_0 : t \mapsto \xi t^{\nu-1}$$

définie par deux paramètres inconnus: un paramètre d'échelle $\xi > 0$ et un paramètre de forme $\nu > 1$ permettant d'assurer que le risque instantané de base soit positif et croissant avec le temps.

3.1.2 Sous-modèle d'exposition

Le sous-modèle d'exposition décrit la distribution de probabilité des variables d'exposition (continues et discrètes) dans un groupe $c \in \mathbb{N}^*$.

Pour notre cas d'étude, les différentes variables d'exposition considérées pour la caractérisation des groupes et profils d'exposition des mineurs d'uranium français sont les suivantes :

- l'exposition au radon du mineur i, X_i^R, cumulée au cours de sa carrière (variable continue);
- l'exposition aux rayonnements gamma du mineur i, X_i^G, cumulée au cours de sa carrière (variable continue);
- l'exposition aux poussières d'uranium du mineur i, X_i^P, cumulée au cours de sa carrière (variable continue);
- le poste de travail J_i du mineur i. Cette variable est un proxy pour les conditions d'exposition ainsi que d'éventuelles autres expositions professionnelles (variable catégorielle à 5 modalités);
- l'âge à la première exposition A_i du mineur i (variable continue);
- la localisation de la mine M_i du mineur i. Cette variable est une proxy des conditions de travail du mineur car en fonction du type de sol, les techniques d'exploitation du minerai diffèrent (variable catégorielle à deux modalités);
- la durée d'exposition T_i du mineur i (variable continue).

L'ensemble $(X_i^R, X_i^G, X_i^P, A_i, J_i, M_i, T_i)$ constitue le profil d'exposition du mineur i. Chaque variable d'exposition suit une loi de probabilité conditionnelle au groupe c. On note $X_i^{cat,k}$ la $k^{\text{i-ème}}$ variable catégorielle du profil d'exposition du mineur i et $X_i^{cont,k}$ la $k^{\text{i-ème}}$ variable continue du profil d'exposition du mineur i. On suppose que :

- $X_i^{cat,k}|C_i = c \sim \mathcal{M}ultinomial(\boldsymbol{p}_c^{cat,k})$;
- $X_i^{cont,k}|C_i = c \sim \mathcal{LN}(\mu_c^{cont,k}, \sigma_c^{cont,k})$

avec $\mathcal{LN}$ la loi lognormale.

3.1.3 Sous-modèle d'attribution

Le sous-modèle d'attribution répartit les individus dans les différents groupes. Il est défini par:

$$P(C_i = c) = \phi_c \qquad \text{pour tout} \qquad \text{c} \in \mathbb{N}^*$$

Le vecteur $\boldsymbol{\phi} = (\phi_c)_{c\in\mathbb{N}^*}$ definit la probabilité d'appartenance des individus à chacun des groupes c. Ce vecteur suit un processus de Dirichlet. La construction de ces poids de mélange, appelée "stick-breaking", est définie grâce aux relations suivantes:

$$\phi_c = V_c \cdot (1 - \sum_{k=1}^{c-1} \phi_k) \qquad \text{pour tout} \qquad \text{c} \in \mathbb{N}^*$$
$$\phi_1 = V_1$$

avec $(V_c)_{c\in\mathbb{N}^*}$ des variables aléatoires latentes inépendantes définies, pour tout c, par:

$$V_c \sim Beta(1, \alpha)$$

Si, en théorie, il existe une infinité de groupes d'individus différents, en pratique, il existe un nombre fini de groupes non vides C. Le vecteur $\boldsymbol{\phi}$ est donc un vecteur sparse avec $\phi_c = 0$ pour tout $c > C$. La valeur de C dépend du paramètre dit de concentration α. Plus la valeur de α est petite, plus le nombre estimé de groupes non vides est petit et vice versa.

3.2 Choix des lois *a priori*

Le Graphe Acyclique Orienté de notre modèle est donné dans la Figure 1.

Les lois *a priori* considérées sont les suivantes :

- $\beta_c \sim \mathcal{N}(0, 10^6)$, $c \in \mathbb{N}^*$;
- $\mu_c^{cont,k} \sim \mathcal{N}\left(\mu_{prior}^{(k)}, \sigma_{prior}^{2(k)}\right)$, $c \in \mathbb{N}^*$. Pour l'âge à la première exposition et la durée d'exposition, les lois *a priori* normales considérées sont centrées et de grande variance 10^6. Pour les expositions au radon, aux rayonnements gamma et aux poussières d'uranium, les lois *a priori* sont des lois normales dont les valeurs des paramètres ont été définies à partir de la distribution de ces expositions dans la cohorte allemande des mineurs d'uranium;

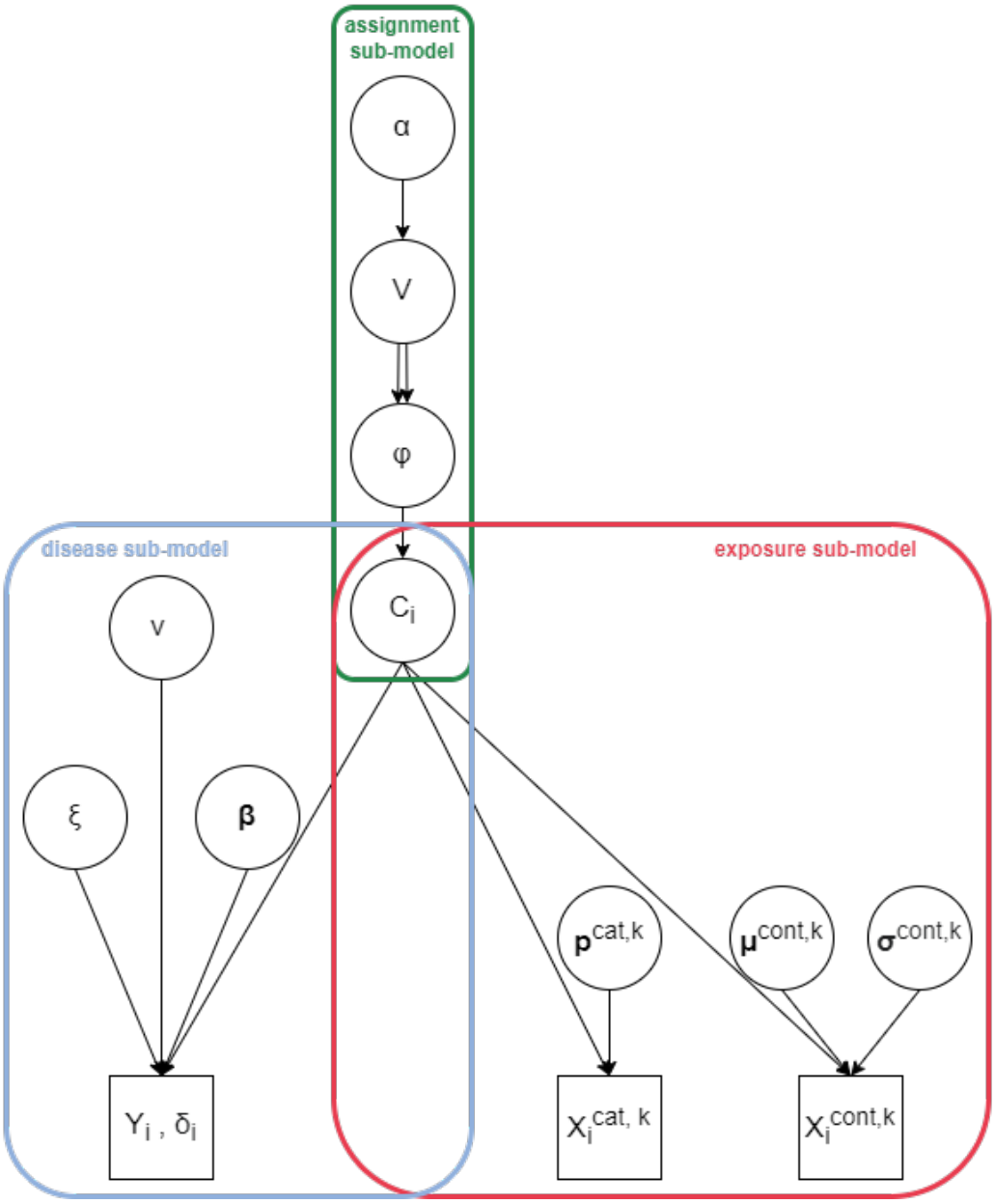

Figure 1: Graphe Acyclique Orienté du modèle de mélange bayésien par régression sur profils d'exposition.

- $\nu' = \nu - 1 \sim \mathcal{G}amma(0.001, 0.001)$;
- $\xi' = \xi \times 10^{25} \sim \mathcal{G}amma(1, 1)$;
- $\sigma_c^{cont,k} \sim \mathcal{G}amma(0.001, 0.001)$, $c \in \mathbb{N}^*$;
- $\boldsymbol{p}_c^{cat,k} \sim \mathcal{D}irichlet(0.5, 0.5, \ldots)$, $c \in \mathbb{N}^*$.
- $\alpha \sim \mathcal{G}amma(1, 2)$ tel que recommandé dans Liverani *et al.* (2015).

3.3 Inférence bayésienne

Un algorithme Monte-Carlo par Chaînes de Markov (MCMC) de type Metropolis-within Gibbs adaptatif a été implémenté en Python 3.11. Par ailleurs, le nombre de variables de groupe latentes étant infini, un slice sampler a été utilisé afin de se ramener à un nombre de variable latente fini à chaque itération, tel que proposé par Walker (2007). Enfin, la loi de probabilité *a posteriori* jointe étant potentiellement multimodale, ce qui pourra notamment être favorisé en présence de données peu informatives, l'estimation du paramètre de concentration α du processus de Dirichlet latent pose des difficultés (Belloni *et al.* (2020)). Afin de faciliter la convergence l'algorithme, nous avons couplé notre algorithme MCMC avec un algorithme de type "parallel tempering".

3.4 Post-traitement

Un post-traitement est indispensable à l'identification et la caractérisation des clusters. En effet, à chaque itération de l'algorithme, une partition différente de l'ensemble des individus en groupes est estimée. Le post-traitement permet d'agréger ces différentes partitions.

Plusieurs méthodes de post-traitement sont comparées :

- La première est la méthode *ad hoc* proposée par Molitor *et al.* (2010). Elle consiste à définir les groupes grâce à un algorithme K-meloids prenant comme paramètre d'entrée la matrice de dissimilarité moyenne (i.e. moyenne des matrices de dissimilarité obtenues à chaque itération).
- La deuxième méthode est proposée par Wade et Ghahramani (2018). La partition optimale est la partition minimisant, sur l'ensemble des partitions possibles, une fonction de perte appelée "variation of information".
- La troisième méthode agrège l'algorithme Merge-Truncate-Merge de Guha *et al.* (2019) et la méthode de post-traitement proposée par Belloni *et al.* (2020). Le premier algorithme permet de regrouper à chaque itération les groupes de petites tailles créés artificiellement par le processus de Dirichlet. Cette étape permet de garantir la consistance du processus de Dirichlet quant au nombre estimé de groupes non-vides (Guha *et al.* (2019))).

La seconde étape consiste à trouver la partition dont la matrice de dissimiliarité minimise la distance des moindres carrés à la matrice de dissimilarité moyenne.

Les lois *a posteriori* des paramètres définissant chaque groupe i.e., $\theta_c = (\boldsymbol{\mu}_c^{cont}, \boldsymbol{\sigma}_c^{cont}, \boldsymbol{p}_c^{cat}))$ sont alors déduites de cette meilleure partition $\boldsymbol{z}^{best}$. Ainsi, un échantillon *a posteriori* du paramètre θ_c du groupe c est obtenu à chaque itération j par $\bar{\theta}_{c,j} = \frac{1}{n_c} \sum_{i:z_i^{best}=c} \theta_{z_i^j,j}$ avec n_c le nombre d'individus du groupe c, z_i^j le groupe auquel appartient l'individu i à l'itération j.

4 Résultats

Une étude par simulations sera présentée afin de déterminer, 1) l'impact de l'algorithme "parallel tempering", 2) l'impact de la procédure de post-processing, sur le nombre estimé de groupes non vides des modèles BPRM.
Une application du modèle proposé aux données de la cohorte post-55 des mineurs d'uranium français sera également présentée.

Bibliographie

Wild C. P. (2005), Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology, Cancer Epidemiol Biomarkers, 14(8), 1847-50

Miller J. W. and Harrison M. T. (2013), Inconstistency of Pitman-Yor process mixtures for the number of components, ArXiv, 1309.0024v1

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., & Richardson, S. (2015). PReMiuM: An R package for profile regression mixture models using Dirichlet processes. Journal of statistical software, 64(7), 1.

Molitor, J., Papathomas, M., Jerrett, M., & Richardson, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics, 11(3), 484-498.

Rage, E., Caër-Lorho, S., Drubay, D., Ancelet, S., Laroche, P., & Laurier, D. (2015). Mortality analyses in the updated French cohort of uranium miners (1946–2007). International archives of occupational and environmental health, 88(6), 717-730.

Vacquier, B., Rage, E., Leuraud, K., Caër-Lorho, S., Houot, J., Acker, A., & Laurier, D. (2011). The influence of multiple types of occupational exposure to radon, gamma rays and long-lived radionuclides on mortality risk in the French "post-55" sub-cohort of uranium miners: 1956–1999. Radiation research, 176(6), 796-806

Wade S., & Gahramani Z., (2018), Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion), Bayesian Analysis, 13(2), 556-629

Guha A., Ho N., & Nguyen X. (2019) On posterior contraction of parameters and interpretability in Bayesian mixture modeling, arXiv, 1901.05078v1

Belloni M., Laurent O., Guihenneuc C., & Ancelet S. (2020) Bayesian Profile Regression to Deal With Multiple Highly Correlated Exposures and a Censored Survival Outcome. First Application in Ionizing Radiation Epidemiology, frontiers in Public Health, 8, 557006