

# Bayesian hierarchical approach to account for radon exposure measurement error when estimating the risk of death by lung cancer in an occupational cohort study

Julie Fendler

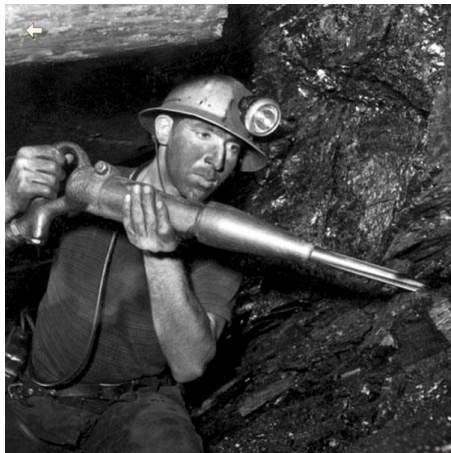
(IRSN PSE-SANTE/SESANE/LEPID)

Encadrante : Sophie ANCELET (IRSN PSE-SANTE/SESANE/LEPID)

Directrice de thèse : Chantal GUIHENNEUC (Université Paris Cité)

18 novembre 2022

# Motivations de l'étude



Source : André De Marles, «Surhomme mineur de  
Guy DUBOIS»

- Le **radon** est un gaz radioactif qui est la première source naturelle de rayonnements ionisants
- Le radon est un facteur de risque certain de **cancer du poumon**
- La population des **mineurs d'uranium** est une population de référence pour étudier les effets sanitaires d'une **exposition chronique au radon à faibles doses**.

# Contexte : les erreurs de mesure en épidémiologie

Les **erreurs de mesure** dans les études épidémiologiques sont :

- omniprésentes
- une des sources majeures d'**incertitude**

Lorsque non prises en compte, elles entraînent [Hoffmann et al., 2017, Belloni, 2021] :

- un **biais** dans l'estimation du risque
- une **perte de puissance** statistique
- une déformation de la relation dose-réponse d'intérêt

⇒ **Objectif** : Obtenir une estimation corrigée du risque de décès par cancer du poumon chez les mineurs d'uranium français qui sont chroniquement exposés au radon à faibles doses

# Sommaire

- 1 Les données
- 2 Le sous-modèle de maladie
- 3 Le sous-modèle d'erreurs de mesure
- 4 Le modèle hiérarchique bayésien
- 5 Les résultats

# LES DONNEES

# La cohorte française des mineurs d'uranium

- $N = 5086$
- Mineurs ayant travaillé minimum 1 an dans **les mines françaises d'uranium** du groupe CEA-Cogema
- Début de la période d'inclusion : 1er janvier 1946
- Date de point : 31 décembre 2014

## La cohorte française des mineurs d'uranium (2)

Statut vital	N(%)
Vivant	2580 (50.7)
Décédé (Cancer du poumon)	268 (5.3)
Décédé (Autres causes)	2196 (43.2)
Perdu de vue	42 (0.8)

Table – Statuts vitaux au 31 décembre 2014 dans la cohorte française des mineurs d'uranium

# Evolution de l'exposition annuelle au radon dans la cohorte

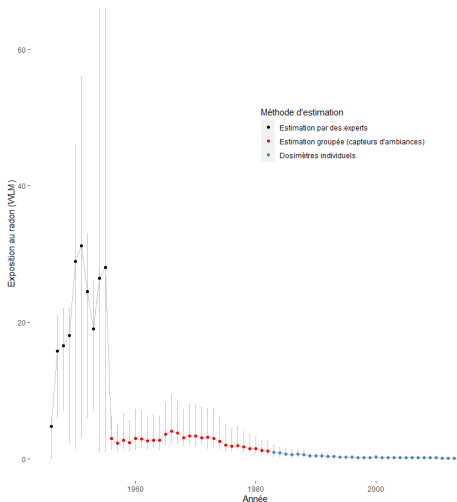


Figure – Exposition annuelle moyenne (et quartiles) au radon chez les mineurs d'uranium exposés en WLM (Working Level Month)



# Les méthodes de correction des erreurs de mesure

- Plusieurs méthodes fréquentistes existent pour la correction des erreurs de mesure (régression calibration, simulation extrapolation SIMEX)
  - **Manque de flexibilité** pour la prise en compte d'erreurs de mesure complexes (différents types, hétéroscédastiques)
  - Deux **étapes d'apprentissage disjointes** pour estimer l'exposition corrigée et le coefficient de risque
- Approche alternative considérée : **les modèles hiérarchiques bayésiens**
  - Approche **flexible** pour prendre en compte des erreurs de mesure complexes
  - **Estimations jointes** des expositions corrigées et du coefficient de risque
  - Permet de prendre en compte des **informations supplémentaires** à travers des lois *a priori* informatives.

# Prise en compte des erreurs de mesure à travers un modèle hiérarchique bayésien

Les modèles hiérarchiques bayésiens proposés s'écrivent comme la combinaison de deux sous modèles :

- Le **sous-modèle de maladie**
- Le **sous-modèle d'erreurs de mesure**

⇒ Les sous-modèles sont combinés entre eux en utilisant des hypothèses d'indépendances conditionnelles.

⇒ Tous les paramètres inconnus sont estimés conjointement

## LE SOUS-MODELE DE MALADIE (MODELE DE SURVIE)

# Les variables observables

Soient :

- $W_i$  l'âge en jours du mineur  $i$  au moment de son décès par cancer du poumon (variable censurée à droite et tronquée à gauche) ;
- $C_i$  l'âge en jours du mineur  $i$  au moment de la censure ;

On observe :

- $Y_i = \min(W_i, C_i)$
- $\delta_i$  indicateur de non-censure du mineur  $i$  ( $\delta_i = 1$  si  $W_i \leq C_i$ )

On aimerait observer :

- $X_i^{cum}(t)$  l'exposition cumulée laguée à 5 ans au radon du mineur  $i$  au temps  $t$  (en WLM)

# Le sous-modèle de maladie

On modélise le risque instantané de décès par cancer du poumon d'un individu  $i$  au temps  $t$ .

## Le modèle

$$\forall t \in [0, +\infty[, h_i(t; \beta) = h_0(t) \cdot g(\beta; X_i^{cum}(t)) \quad (1)$$

où  $h_0(t)$  est la fonction de risque instantané de base.

$\beta$  est le **coefficient de risque** inconnu d'intérêt associé à l'exposition cumulée.

# Modélisation de $g$

Modélisation du ratio de risque instantané :

- Le modèle de Cox :  $g(\beta, X) \mapsto e^{\beta \cdot X}$
- Le modèle en EHR\* :  $g(\beta, X) \mapsto 1 + \beta \cdot X$ 
  - Contrainte :  $\beta > -\frac{1}{X}$

\*EHR : Excess Hazard Ratio ou Excès de risque instantané

## Modélisation de $h_0(t)$

Modélisation de la fonction de risque de base :

- $h_0 : t \mapsto \xi t^{\alpha-1}$  avec  $\xi > 0$  (paramètre d'échelle) et  $\alpha > 0$  (paramètre de courbure) inconnus

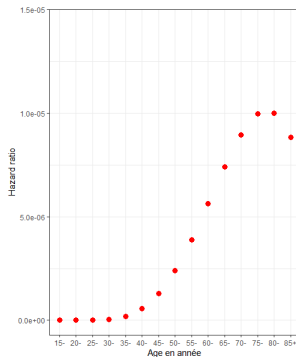
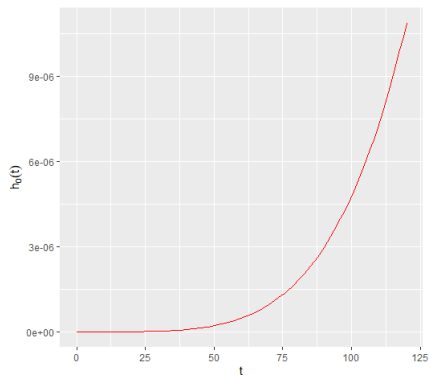
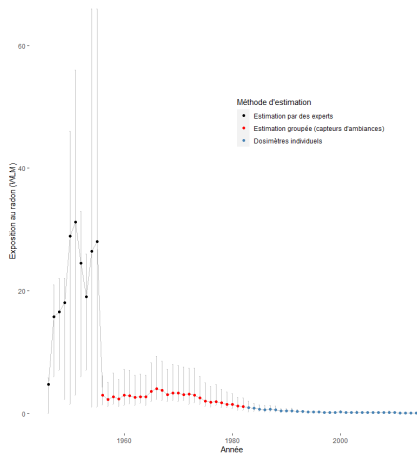


Figure – Risque de décès par cancer du poumon dans la population générale masculine entre 1968 et 2005

# LE SOUS-MODELE D'ERREURS DE MESURE



# Evaluation des expositions annuelles au radon dans la cohorte française des mineurs d'uranium



- 3 méthodes d'évaluation des expositions différentes
- Les erreurs de mesure sur la période 1983-2001 sont négligées

# Notations

Soient

- $X_i(t)$ , la variable représentant **l'exposition "vraie"** de l'individu  $i$  au temps  $t$  (inconnue)
- $Z_i(t)$ , **l'exposition observée** de l'individu  $i$  au temps  $t$  (valeur erronée de  $X_i(t)$ )
- $U_i(t)$ , **l'erreur** entachant  $Z_i(t)$

# Quelques définitions

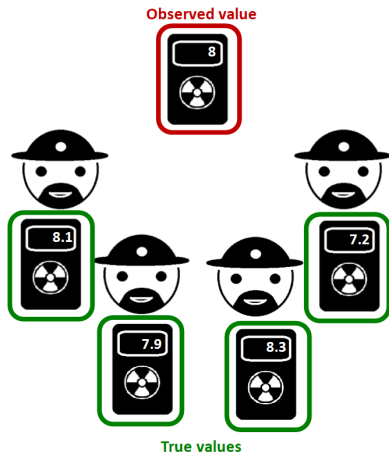


Figure – Erreur Berkson :

$$X_i(t) = Z_i(t) \cdot U_i(t)$$

$$Z_i(t) \perp\!\!\!\perp U_i(t)$$

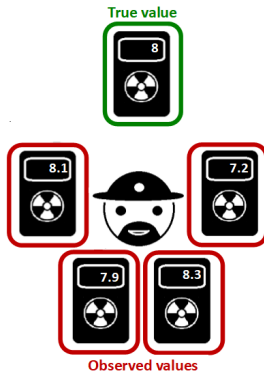


Figure – Erreur classique :

$$Z_i(t) = X_i(t) \cdot U_i(t)$$

$$X_i(t) \perp\!\!\!\perp U_i(t)$$

## Sous-modèle $\mathcal{M}1$ : première hypothèse de modélisation des erreurs de mesure

Pour un individu  $i$  travaillant dans la mine  $m$  le modèle s'écrit :

$$\begin{cases} X_{im}^1(t) = Z_{im}^1(t) \cdot U_i^1(t) & \text{pendant la période 45-55} \\ X_{im}^2(t) = \underbrace{Z_{im}^2(t)}_{\text{valeur estimée}} \cdot \underbrace{U_i^2(t)}_{\text{erreur}} & \text{pendant la période 56-82} \\ X_{im}^3(t) = Z_{im}^3(t) & \text{après 83} \end{cases}$$

$\forall i \in \llbracket 1, N \rrbracket, \forall k \in \{1, 2\},$

$$\mathbf{U}_i^k = (U_i^k(t_1), \dots, U_i^k(t_{ik}))^T \sim \mathcal{LN}(-\frac{\sigma_k^2}{2} \mathbf{1}_{t_{ik}}, \sigma_k^2 \Gamma_{t_{ik}}), E[\mathbf{U}_i^k] = \mathbf{1}_{t_{ik}}$$

$$\mathbf{1}_{t_{ik}} = (1, \dots, 1)^T \in \mathbb{R}^{t_{ik}} \text{ et } \Gamma_{t_{ik}} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}, \rho \in [0; 1[.$$

$$\sigma_1 = 0.93, \sigma_2 = 0.39 \text{ [Allodji et al., 2012]}$$

## Sous-modèle $\mathcal{M}2$ : hypothèse de modélisation plus fine des erreurs de mesure sur la période 45-55

Le deuxième modèle pour un individu  $i$  travaillant dans la mine  $m$  au temps  $t$  s'écrit :

$$\left\{ \begin{array}{ll} \bullet Z_m^1(t) = \zeta_m^1(t) \cdot U_m(t) & 1945-55 \\ \bullet X_{im}^1(t) = \zeta_m^1(t) \cdot T_{im}(t) \cdot U_i^1(t) & \text{si } Z_m^1(t) \text{ est connu} \\ \bullet X_{im}^1(t) = Z_m^1(t) \cdot U_i^1(t) & \text{sinon} \\ \bullet X_{im}^2(t) = Z_{im}^2(t) \cdot U_i^2(t) & 1956-82 \\ \bullet X_{im}^3(t) = Z_{im}^3(t) & \text{après 1983} \end{array} \right.$$

$$\forall t, \log(U_m(t)) \sim \mathcal{N}\left(-\frac{\sigma_*^2}{2}, \sigma_*^2\right)$$

Pour certaines mines,  $\zeta_m^1(t_k) = \zeta_m^1(t_l)$ ,  $k \neq l \Rightarrow U_m(t_k) = U_m(t_l)$

$T_{im}(t)$  est le temps de travail du mineur  $i$  dans la mine  $m$  au temps  $t$ .

## Sous-modèle $\mathcal{M3}$ : Un modèle plus parcimonieux

Hypothèse : l'erreur de mesure Berkson sur chaque période est constante lorsque le poste et la mine sont inchangés

$$\left\{ \begin{array}{ll} \bullet Z_m^1(t) = \zeta_m^1(t) \cdot U_m(t) & 1946-55 \\ \bullet X_{im}^1(t) = \zeta_m^1(t) \cdot T_{im}(t) \cdot U_i^1 & \text{si } Z_m^1(t) \text{ est connu} \\ \bullet X_{im}^1(t) = Z_m^1(t) \cdot T_{im}(t) \cdot U_i^1 & \text{sinon} \\ \bullet X_{im}^{21}(t) = Z_{im}^{21}(t) \cdot U_i^2(q_{ip}(t)) & 1956-74 \\ \bullet X_{im}^{22}(t) = Z_{im}^{22}(t) \cdot U_i^3(q_{ip}(t)) & 1975-77 \\ \bullet X_{im}^{23}(t) = Z_{im}^{23}(t) \cdot U_i^4(q_{ip}(t)) & 1978-82 \\ \bullet X_{im}^3(t) = Z_{im}^3(t) & \text{après 1983} \end{array} \right.$$

$\forall i \in \llbracket 1, N \rrbracket, \forall k \in \{1, 21, 22, 23\},$

$$\mathbf{U}_i^k \sim \mathcal{LN}\left(-\frac{\sigma_k^2}{2}, \sigma_k^2\right)$$

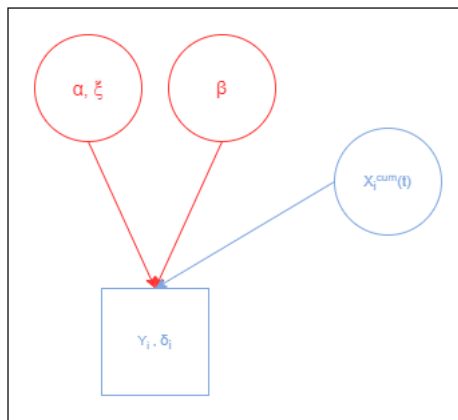
avec  $\sigma_{21} = 0.47, \sigma_{22} = 0.42, \sigma_{23} = 0.33$  [Allodji et al., 2012].

$q_{ip}(t)$  est la période de travail du mineur  $i$  au poste  $p$

# LE MODELE HIERARCHIQUE BAYESIEN

# Graph acyclique orienté du modèle de survie

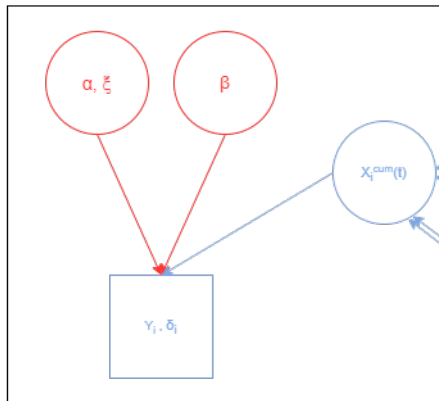
Modèle de survie



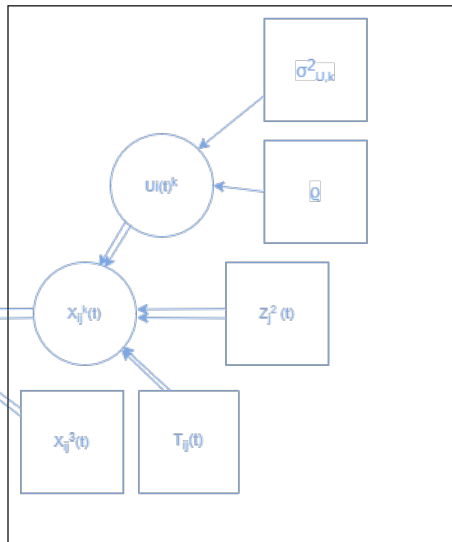


# Graph acyclique orienté du $\mathcal{M}1$

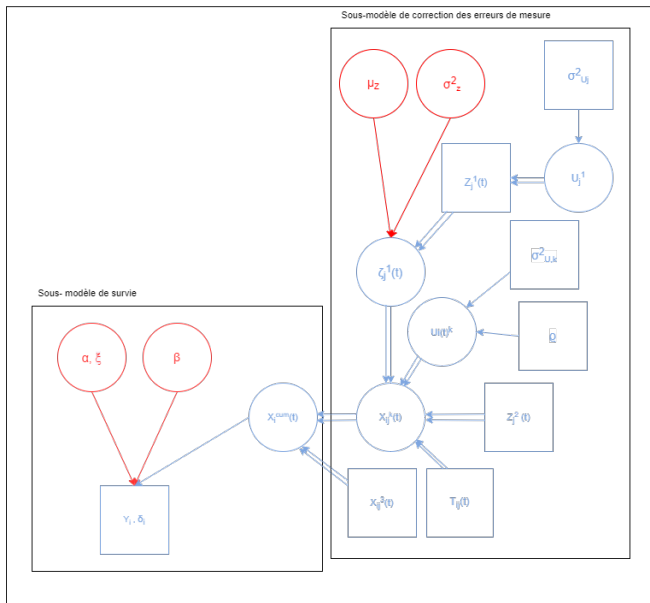
Sous-modèle de survie



Sous-modèle de correction des erreurs de mesure



# Graph acyclique orienté du $\mathcal{M}2$



# Inférence bayésienne des modèles hiérarchiques proposés

- Paramètres fixés dû à un manque d'information :  $\rho, \sigma_1, \sigma_*, \sigma_2, \sigma_{21}, \sigma_{22}, \sigma_{23}$
- Lois *a priori* :
  - loi gamma faiblement informative :  $\alpha, \xi, \tau_\zeta = \frac{1}{\sigma_\zeta}$
  - loi normale faiblement informative :  $\beta, \mu_\zeta$
- Loi *a posteriori* jointe complexe  $\theta = (\beta, \alpha, \xi, \mu_\zeta, \tau_\zeta, \zeta, \mathbf{U})$ 
  - Plus de 40000 quantités inconnues à estimer simultanément  $\Rightarrow$  temps de calcul élevé
  - Algorithme MCMC de type **Metropolis-Hasting-Within-Gibbs adaptatif** développé en Python  $\Rightarrow$  inférence sur cluster HPC

## LES RESULTATS

## Résultats du modèle $\mathcal{M}_0$ de maladie sans correction des erreurs de mesure

	HR* 100WLM	IC** 95%	WAIC***
EHR	2.06	1.60 ;2.70	6861
Cox	1.32	1.18 ;1.45	6876

Table – Résultats des différentes configurations du modèle de maladie appliqué au risque de décès par cancer du poumon

\*HR : médiane *a posteriori* du hazard ratio

\*\*IC : Intervalle de crédibilité à 95%

\*\*\*WAIC : Widely Applicable Information Criterion (plus le WAIC est petit, meilleur est le modèle)

## Impact du paramètre $\rho$

		HR* 100WLM	IC** 95%	WAIC***
EHR	$\rho = 0$	2.14	1.65 ; 2.81	6860
	$\rho = 0.2$	2.16	1.66 ; 2.86	6860
	$\rho = 0.4$	<b>2.21</b>	<b>1.70 ; 2.94</b>	<b>6858</b>
	$\rho = 0.6$	2.23	1.68 ; 2.99	6859
	$\rho = 0.8$	2.17	1.66 ; 2.89	6862
	$\rho = 0.99$	2.13	1.63 ; 2.80	6863
Cox	$\rho = 0$	1.32	1.18 ; 1.48	6971
	$\rho = 0.2$	1.33	1.18 ; 1.50	6870
	$\rho = 0.4$	<b>1.33</b>	<b>1.18 ; 1.52</b>	<b>6870</b>
	$\rho = 0.6$	1.35	1.18 ; 1.54	6872
	$\rho = 0.8$	1.35	1.20 ; 1.55	6875
	$\rho = 0.99$	1.19	1.08 ; 1.35	6886

**Table** – Résultats du sous-modèle de maladie combiné au sous-modèle d'erreurs de mesure  $\mathcal{M}_1$  appliqué au risque de décès par cancer du poumon

# Résultat des sous-modèles $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2(\rho = 0.4)$ combinés à un sous-modèle de maladie en EHR et impact des paramètres $\sigma_1$ et $\sigma_*$

		HR* 100WLM	IC** 95%	WAIC***
$\mathcal{M}_0$		2.06	1.60 ;2.70	6861
$\mathcal{M}_1$		2.21	1.70 ;2.94	6858
$\mathcal{M}_2$	$\sigma_1 = 0.84 ; \sigma_* = 0.41$	2.38	1.78 ;3.26	6855
	$\sigma_1 = 0.84 ; \sigma_* = 0.82$	2.50	1.81 ;3.46	6854
	$\sigma_1 = 0.63 ; \sigma_* = 0.31$	2.32	1.73 ;3.13	6857

**Table –** Résultats des sous-modèles  $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2(\rho = 0.4)$  combinés à un sous-modèle de maladie en EHR appliqué au risque de décès par cancer du poumon

## Impact des erreurs de mesure sur estimation du hazard ratio ( $\rho = 0.4, \sigma_1 = 0.84, \sigma_* = 0.41$ )

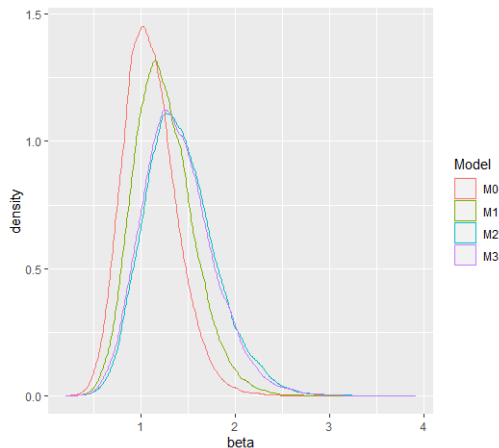


Figure – Densité *densite a posteriori* du paramètre  $\beta$  pour 100 WLM du sous-modèle de maladie combiné aux différents sous-modèles de correction des erreurs de mesure et appliqué au risque de décès par cancer du poumon



# Conclusions

- Le modèle en EHR s'ajuste mieux aux données de la cohorte française des mineurs d'uranium que le modèle de Cox
- Il y a un **impact des erreurs de mesure** sur l'exposition au radon sur l'estimation du risque de décès par cancer du poumon
- Le **risque de décès par cancer du poumon augmente** significativement après une exposition chronique à faibles doses au radon

# Limites et perspectives

- Vérification de la **robustesse de l'estimation du hazard ratio à une mauvaise spécification** du modèle d'erreurs de mesure.
- Exploration d'**autres critères bayésiens** de sélection de modèles hiérarchiques (ex : WAIC marginaux)
- Prise en compte de **facteurs modifiants** la relation dose-réponse
- Proposer des algorithmes d'inférence bayésienne moins coûteux en temps de calculs tout en permettant une bonne estimation des intervalles de crédibilité (de la variance a posteriori)

**MERCI POUR VOTRE ATTENTION !**

# Bibliographie I



Allodji, S. R., Leuraud, K., Bernhard, S., Henry, S., Bénichou, J., and Laurier, D. (2012).

Assessment of uncertainty associated with measuring exposure to radon and decay products in the french uranium miners cohort.

*Journal of radiological protection*, 32(1) :32–85.



Belloni, M. (2021).

*Approche hiérarchique bayésienne pour l'estimation du risque de cancers radio-induits en situation d'expositions professionnelles multiples et incertaines. Application aux travailleurs du cycle du combustible nucléaire.*

PhD thesis.

# Bibliographie II



Hoffmann, S., Rage, E., Laurier, D., Laroche, P., Guihenneuc, C., and Ancelet, S. (2017).

Accounting for berkson and classical measurment error in radon exposure using a bayesian structural approach in the analysis of lung cancer mortality in the french cohort of uranium.

187 :196–209.