



HAL
open science

Bayesian hierarchical models to account for complex measurement error when estimating a disease risk: Application to the French cohort of uranium miners

Julie Fendler, Sophie Ancelet, chantal guihenneuc

► To cite this version:

Julie Fendler, Sophie Ancelet, chantal guihenneuc. Bayesian hierarchical models to account for complex measurement error when estimating a disease risk: Application to the French cohort of uranium miners. Journées de statistique de le SFDS 2023, Jul 2023, Bruxelles (Belgique), France. irsn-04403066

HAL Id: irsn-04403066

<https://irsn.hal.science/irsn-04403066>

Submitted on 23 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bayesian hierarchical models to account for complex measurement error when estimating a disease risk: Application to the French cohort of uranium miners.

Julie Fendler¹, Sophie Ancelet¹, Chantal Guihenneuc²

¹ Institut de Radioprotection et de Sûreté Nucléaire (IRSN), France

² Université de Paris Cité, BioSTM-UR 7537, France

email : julie.fendler@irsn.fr

July, 6, 2023

Overview

- 1 Introduction
- 2 Bayesian hierarchical models to account for measurement errors
- 3 Simulation study : Impact of model misspecification on risk estimates
- 4 Application : Risk of death by lung cancer due to radon exposure in the French cohort of uranium miners
- 5 Conclusion

1.1 Exposure measurement error in epidemiology

- Measuring environmental exposure is an error-prone process.
- Exposure measurement error is one important source of uncertainty in risk estimates but it is rarely accounted for in epidemiology.
- Ignoring measurement error may cause [Carroll et al, 2006; Keogh et al, 2020]:
 - a bias in risk estimates;
 - an inadequate estimation of the associated confidence intervals;
 - a distortion of the shape of the exposure-response relationship [Hoffmann et al, 2017];
 - a loss in statistical power.

⇒ It is important to account for exposure uncertainty in risk estimation [ICRP103, 2007; UNSCEAR, 2012].

1.2 Different types of exposure measurement error

Let us define for an individual i at time t , $X_i(t)$ its true exposure (unobserved), $Z_i(t)$ its observed exposure (error-prone) and $U_i(t)$ the measurement error.

Classical error

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

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

vs.

Berkson error

$$X_i(t) = Z(t) \times U_i(t)$$

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

2.1 The Bayesian hierarchical approach for exposure measurement error correction : general principle

- **Step 1** : building and combining conditionally independent probabilistic submodels [Richardson and Gilks, 1993]
 - **Disease submodel**: to link the disease outcome - here a time to event - and the true exposures \Rightarrow a survival model;
 - **Measurement error submodel**: to link the true and the observed exposures
 - **Measurement submodel** : to describe the measurement error process (Berkson or classical error)
 - **Exposure submodel** : to describe the probability distribution of the true exposure (for classical error only)
- **Step 2** : fitting in one step the full hierarchical model using a Bayesian learning algorithm
 - allows to **jointly estimate** corrected exposures and risk coefficients (and their associated uncertainty);

2.2 The disease submodel

The instantaneous hazard rate of a given disease event for an individual i at time t is modeled as follows :

$$\forall t \in [0, +\infty[, h_i(t; \beta) = h_0(t) \cdot g(\beta; X_i(t))$$

- β is the unknown **risk coefficient** associated to a (possibly cumulative) environmental exposure $X_i(t)$.
- h_0 is the baseline hazard function : **Weibull inspired function**
 - $h_0 : t \mapsto \frac{\alpha}{\mu} \left(\frac{t}{\mu}\right)^{\alpha-1} \xrightarrow{\text{reparameterisation}} h_0 : t \mapsto \xi t^{\alpha-1}$ ($\xi > 0$ and $\alpha > 1$)
- g is the hazard ratio function. Two structures are compared:
 - **Excess Hazard Ratio (EHR) structure** \Rightarrow linear exposure-response relationship : $g : (\beta, X) \mapsto 1 + \beta X$
 - **Cox structure** \Rightarrow log-linear exposure-response relationship : $g : (\beta, X) \mapsto \exp(\beta X)$

2.3 The measurement error submodel

For each individual i and each calendar period p (i.e., successive time points characterized by the same magnitude of exposure measurement error) :

Structure 1 : Only correlated Berkson error terms

$$X_i^p(t) = Z^p(t) \cdot \underbrace{U_i^p(t)}_{\text{Berkson error term}}$$

Structure 2: A mixture of classical and correlated Berkson error terms

$$Z^p(t) = \zeta^p(t) \cdot \underbrace{U^p(t)}_{\text{classical error term}}$$

$$X_i^p(t) = \zeta^p(t) \cdot \underbrace{U_i^p(t)}_{\text{Berkson error term}}$$

2.4 Modelling the error components

Measurement submodel :

- For the individual and period specific **Berkson error terms**, a multivariate log-normal probability distribution is assumed :

$$\mathbf{U}_i^p = (U_i^p(t_0), \dots, U_i^p(t_{ip})) \sim \mathcal{LN} \left(-\frac{\sigma_p^2}{2} \mathbb{1}_{t_{ip}}, \sigma_p^2 \Sigma_i^p(\rho) \right) \text{ with}$$

$$\Sigma_i^p(\rho) = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}, \rho \in [0; 1[, \sigma_p > 0$$

- For the period specific **classical error terms**:

$$U^p(t) \sim^{i.i.d} \mathcal{LN} \left(-\frac{\sigma_*^2}{2}, \sigma_*^2 \right)$$

Exposure submodel :

$$\zeta^p(t) \sim^{i.i.d} \mathcal{LN}(\mu_\zeta, \sigma_\zeta^2)$$

2.5 Prior probability distributions

- $\beta \sim \mathcal{N}(0, 10^6)$ [left-sided truncated at 0 for the EHR disease model to guarantee $h_i > 0$]
- $\alpha' = \alpha - 1 \sim \mathcal{G}(0.01, 0.01)$
- $\xi \sim \mathcal{G}(1, 1)$
- $\mu_\zeta \sim \mathcal{N}(0, 100)$
- $\sigma_\zeta \sim \mathcal{IG}(0.001, 0.001)$

Remark: When no pair of values $(X_i^P(t), Z^P(t))$ is available, σ_p and ρ must be fixed \Rightarrow Information required!

2.6 Bayesian inference

- Adaptive Metropolis-Within-Gibbs algorithms have been developed in Python 3.8 to approximate the joint posterior distribution of :
 - $\theta_0 = (\beta, \alpha, \xi)$ for the uncorrected Cox and EHR disease models
 - $\theta_1 = (\beta, \alpha, \xi, \mathbf{U})$ for the hierarchical model with only correlated Berkson error terms (Structure 1)
 - $\theta_2 = (\beta, \alpha, \xi, \mu_\zeta, \sigma_\zeta, \zeta, \mathbf{U})$ for the hierarchical model with a mixture of classical and correlated Berkson error terms (Structure 2)

3.1 Impact of disease submodel misspecification

- Generate failure times for **2000 individuals** according to the **uncorrected** Cox or EHR disease model with time-varying covariates
- Generate **100 replicated datasets** for each scenario :
 - true $\alpha = 5.5$; true $\xi = 8.8 \times 10^{-25}$;
- Bayesian inference **without accounting for measurement error**

Simulation Model	true β	Estimation Model	Mean relative bias on HR*	Coverage rate (95%)
Cox model	2	Cox model	0.04	0.95
EHR model	5	EHR model	0.03	0.96
Cox model	0.4	EHR model	0.01	0.23
EHR model	0.5	Cox model	0.01	0.47

$$*HR = \begin{cases} 1 + \beta & \text{for an EHR model} \\ \exp(\beta) & \text{for a Cox model} \end{cases}$$

3.2 Impact of measurement error submodel misspecification

Design of the simulation study :

- Generate only correlated Berkson error terms (Structure 1) or a mixture of classical and correlated Berkson error terms (Structure 2)
 - Consider two calendar periods with different magnitudes of exposure measurement error
- Generate failure times for **2000 individuals** from the EHR model
- Generate **100 replicated datasets** for each scenario :
 - true $\beta = 5$; true $\alpha = 5.5$; true $\xi = 8.8 \times 10^{-25}$; true $\rho = 0.4$
- Bayesian inference **with or without** accounting for exposure measurement error

3.3 Impact of measurement error submodel misspecification

Impact of the error structure when ρ is fixed to 0.4

Simulation model	Estimation model	Mean relative bias on β	Coverage rate (95%)
Berkson error	Uncorrected	-0.08	0.92
	Mixture of errors	0.04	0.96
	Berkson error	0.06	0.97
Mixture of errors	Uncorrected	-0.08	0.89
	Berkson error	0.07	0.92
	Mixture of errors	0.06	0.96

Impact of a misspecification of ρ

Simulation model	Estimation model	Mean relative bias on β	Coverage rate (95%)
Mixture of errors	Mixture of errors		
$\rho = 0.2$	$\rho = 0.8$	0.13	0.86
$\rho = 0.8$	$\rho = 0.2$	0.01	0.95

3.3 Impact of measurement error submodel misspecification

Impact of the error structure when ρ is fixed to 0.4

Simulation model	Estimation model	Mean relative bias on β	Coverage rate (95%)
Berkson error	Uncorrected	-0.08	0.92
	Mixture of errors	0.04	0.96
	Berkson error	0.06	0.97
Mixture of errors	Uncorrected	-0.08	0.89
	Berkson error	0.07	0.92
	Mixture of errors	0.06	0.96

Impact of a misspecification of ρ

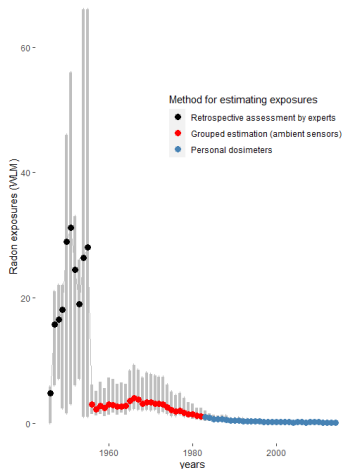
Simulation model	Estimation model	Mean relative bias on β	Coverage rate (95%)
Mixture of errors	Mixture of errors		
$\rho = 0.2$	$\rho = 0.8$	0.13	0.86
$\rho = 0.8$	$\rho = 0.2$	0.01	0.95

4.1 Risk of death by lung cancer and radon exposure in the French cohort of uranium miners

- The **French cohort of uranium miners** (N = 5086):
 - a reference population to estimate the health effects associated with low-dose exposure to radon ;
 - Measuring radon exposure is an error-prone process

Vital status	N (%)
Alive	2580 (50.7)
Death (lung cancer)	268 (5.3)
Lost of follow-up	42 (0.8)

Vital status on the 31rd December 2014



Radon exposure (in WLM) in the French cohort of

4.2 The hierarchical model

- Failure time : (Right-censored) age at death by lung cancer of miner i
- Disease submodel : EHR structure for the hazard ratio function
- Measurement error submodel : For a miner i working in mine m at year t

$$\left\{ \begin{array}{ll}
 \bullet Z_m^1(t) = \zeta_m^1(t) \cdot U_m^1(t) & \sigma_* = 0.41 \quad \mathbf{1945-55} \\
 \bullet X_{im}^1(t) = \zeta_m^1(t) \cdot T_{im}(t) \cdot U_i^1(t) & \sigma_1 = 0.84 \quad \text{if } Z_m^1(t) \text{ is known;} \\
 \bullet X_{im}^1(t) = Z_{im}^1(t) \cdot U_i^1(t) & \sigma_1 = 0.84 \quad \text{otherwise;} \\
 \\
 \bullet X_{im}^{21}(t) = Z_{im}^{21}(t) \cdot U_i^{21}(t) & \sigma_{21} = 0.47 \quad \mathbf{1956-74} \\
 \bullet X_{im}^{22}(t) = Z_{im}^{22}(t) \cdot U_i^{22}(t) & \sigma_{22} = 0.42 \quad \mathbf{1975-77} \\
 \bullet X_{im}^{23}(t) = Z_{im}^{23}(t) \cdot U_i^{23}(t) & \sigma_{23} = 0.33 \quad \mathbf{1978-82} \\
 \bullet X_{im}^3(t) = Z_{im}^3(t) & \mathbf{1983-2001}
 \end{array} \right.$$

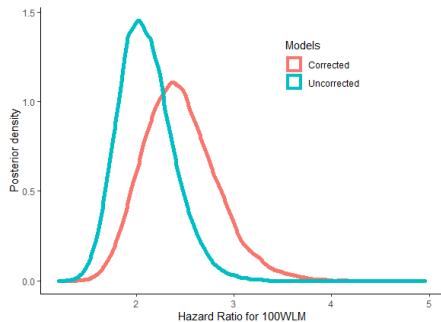
- $T_{im}(t)$ is the working time of miner i in mine m during year t
- Multivariate log-normal distributions for \mathbf{U}_i^p ($p=1,21,22,23$) with σ_p fixed according to [Allodji et al., 2012] and $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 0.99\}$.
- Univariate log-normal distributions for $U_m^1(t)$ and $\zeta_m^1(t)$.

4.3 Bayesian risk estimation and model comparison

Model	HR*	95%CI	Conditional WAIC**	Marginal WAIC
Uncorrected	2.06	[1.60;2.70]	6862	6862
Corrected ($\rho = 0.4$)	2.40	[1.78;3.26]	6855	6853

* HR : Hazard Ratio for 100 Working Level Months (i.e., $1 + \beta \times 100$)

** WAIC: Widely Applicable Information Criterion



Posterior probability distribution of the uncorrected and corrected HR for 100 WLM with $\rho = 0.4$

Conclusion

- We propose different Bayesian hierarchical models to account for complex exposure measurement error in risk estimates when working with right-censored survival data and time-varying exposure.
- Ignoring exposure measurement error when estimating a health hazard ratio (HR) from survival data may lead to an **underestimation** of both the **HR and its associated posterior variance**.
- A **misspecification of the disease submodel** seems to have the worst consequences on risk estimation.
- If ρ is known, the corrected HR **does not seem to be so sensitive** to a misspecification of the measurement error submodel
- An overestimation of ρ can have a very bad impact on risk estimation.
- The WAIC criteria potentially lack the power to discriminate between models in a weakly informative context, even in their marginal version.

References

-  Sabine Hoffmann, Estelle Rage, Dominique Laurier, Pierre Laroche, Chantal Guihenneuc, and Sophie Ancelet (2012)
Accounting for Berkson and Classical Measurement Error in Radon Exposure Using a Bayesian Structural Approach in the Analysis of Lung Cancer Mortality in the French Cohort of Uranium Miners
Radiation Research 187, 196 – 209.
-  Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, Ciprian M. Crainiceanu (2006)
Measurement Error in Nonlinear Models, 2nd Edition
Chapman and Hall/CRC
-  Ruth H. Keogh, Pamela A. Shaw, Paul Gustafson, Raymond J. Carroll, Veronika Deffner, Kevin W. Dodd, Helmut Küchenhoff, Janet A. Tooze, Michael P. Wallace, Victor Kipnis, Laurence S. Freedman (2020)
STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment
Statistics in medicine 39(16), 2197–2231.
-  The International Commission on Radiological Protection (2017)
The 2007 Recommendations of the International Commission on Radiological Protection
-  United Nations Scientific Committee on the Effects of Atomic Radiation (2012)
Sources, affects and risks of ionizing radiation
-  Sylvia Richardson and Walter R. Gilks (1993)
A Bayesian Approach to Measurement Error Problems in Epidemiology Using Conditional Independence Models
American Journal of Epidemiology 138(6) 430 – 442

Thank you for your listening !