Using a structural Bayesian approach to account for measurement error: An application to radiation epidemiology

Sabine Hoffmann [1] & Sophie Ancelet [2] & Pierre Laroche [3] & Chantal Guihenneuc [4]

[1] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PRP-HOM/SRBE/LEPID, Fontenay-Aux-Roses, 92262, France, sabine.hoffmann@irsn.fr*
[2] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PRP-HOM/SRBE/LEPID, Fontenay-Aux-Roses, 92262, France, sophie.ancelet@irsn.fr*
[3] *AREVA, Direction Santé, la Défense, France*
[4] *EA 4064 , Faculté de Pharmacie de Paris, Université Paris Descartes, 4 avenue de l'Observatoire 75006 Paris, chantal.guihenneuc@parisdescartes.fr*

Résumé. Le problème des erreurs de mesure sur les variables explicatives est un sujet important dans beaucoup de domaines de recherche, dont l'épidémiologie, la biologie et l'économétrie. Ignorer ces erreurs de mesure peut conduire à une perte de puissance statistique ainsi qu'à des estimateurs biaisés des paramètres et de l'incertitude associée. Dans cette étude épidémiologique basée sur une cohorte prospective de mineurs d'uranium, nous proposons une approche structurelle bayésienne, basée sur une combinaison de modèles probabilistes conditionnellement indépendants, pour étudier l'association entre une exposition professionnelle au radon et la mortalité par cancer du poumon en prenant en compte les erreurs de mesure de type Berkson. Un algorithme MCMC de type Metropolis-Hastings a été implémenté sous Python pour mener l'inférence bayésienne du modèle proposé. Une étude par simulations suggère que l'approche bayésienne présentée conduit à une diminution substantielle du biais induit par les erreurs de mesure. Sur les données de la cohorte, on observe une augmentation du risque de mortalité par cancer du poumon associé à l'exposition au radon avec la prise en compte des erreurs de mesure. Des futures analyses seront nécessaires afin notamment de comparer les performances de l'approche bayésienne, adoptée dans cette étude, avec des méthodes fonctionnelles et d'étudier l'impact d'une mauvaise spécification de la distribution des vraies expositions.

Mots-clés. Erreurs de mesure, Approche structurelle, Statistique bayésienne, Analyse de survie, Radon, Cancer du poumon

Abstract. The problem of measurement error affecting predictor variables arises in many research areas, such as epidemiology, biology and econometrics. Ignoring this measurement error can lead to a loss of power and biased point and interval estimates of parameters.

In this epidemiological study, conducted on a prospective cohort of uranium miners, we propose a structural Bayesian approach based on conditional independence models to assess the association between occupational radon exposure and lung cancer mortality while taking into account Berkson exposure measurement error. Bayesian inference is conducted via an adaptive Metropolis-Hastings algorithm implemented in Python.
A simulation study suggests that this Bayesian approach leads to a substantial reduction in the bias caused by exposure measurement error. When the proposed methodology is applied to the cohort, one observes an increase in the risk estimate for lung cancer mortality associated with cumulated radon exposure.
More research is needed to compare the performance of this structural Bayesian approach with functional methods and to study the robustness of the proposed method concerning misspecifications of the distribution of true exposures.

Keywords. Measurement error, Structural approach, Bayesian statistics, Survival analysis, Radon, Lung Cancer

1 Introduction

In many research areas, such as epidemiology, biology and econometrics, predictor variables can often only be measured with substantial error. Ignoring the resulting measurement error when conducting statistical inference can lead to a loss of statistical power and biased point and interval estimates on unknown quantities of interest. Despite these known consequences, covariate measurement error is often ignored in regression models.
When accounting for this measurement error, one distinguishes between functional and structural methods. Structural approaches assume that the distribution of the true, unknown covariate X belongs to a known class of parametric distributions. In functional methods, on the other hand, the true values of the predictor X are either regarded as unknown fixed constants or only minimal assumptions about their distribution are made (Carroll (2006)). If the probability distribution of the latent variables is correctly specified, structural methods tend to outperform functional methods in terms of efficiency (Schafer (1996), Küchenhoff (1997)). Furthermore, functional methods, such as regression calibration, often use disjoint steps to estimate the true values of the predictor variable and of the unknown parameters, implying that the uncertainty associated with the former cannot be taken into account in the latter.
Despite their advantages, structural methods are rarely used because their implementation in a frequentist context generally requires the integration of the likelihood over the distribution of latent variables. This can be achieved via an Expectation-Maximization algorithm, but confidence intervals are not easily obtained and bootstrapping or similar methods have to be used. In this study we opt for a Bayesian structural approach based on conditional independence models to account for exposure measurement error in a sur-

vival context. This approach, even if it allows a conceptually simple and straightforward implementation of the conditional reasoning inherent in measurement error problems, is rarely used to account for measurement error in survival models. It offers a flexible and coherent framework to incorporate all sources of uncertainty as well as the available prior information on unknown quantities in statistical inference and estimated credible intervals are automatically available through Markov Chain Monte Carlo methods.

2 The motivating study: Analyzing lung cancer risk among uranium miners

Radon, a noble and radioactive gas, was classified as a pulmonary carcinogen in humans by the International Agency for Research on Cancer in 1988. The French cohort of uranium miners, followed up by the laboratory of epidemiology of the IRSN, provides relevant epidemiological data to quantify the association between occupational radon exposure and lung cancer mortality. This prospective cohort comprises 5086 uranium miners. The average time of follow-up is 35 years. At the end of follow-up, 211 miners had died of lung cancer. Previous studies on this cohort have shown an increased risk of lung cancer mortality associated with cumulative radon exposure. However, these analyses mainly assumed true exposure to be known, whereas exposure assessment in the French cohort of uranium miners is often associated with substantial measurement error. Indeed, annual exposures between 1946 and 1956 were retrospectively reconstructed by a group of experts based on environmental measurements performed in the mines and information concerning the place and type of work of each miner. In 1956, measurements of ambient radon gas concentration at work sites were introduced to estimate the monthly exposure of each miner. Finally, starting in 1983, personal dosimetry was used to record the potential alpha energy of radon decay products.
In earlier studies on the cohort, three functional methods of measurement error correction were used to obtain a measurement error corrected estimate for the association between radon exposure and lung cancer risk (Allodji (2012)). These functional methods, which where mainly applied to simulated data, could only achieve a partial reduction of the bias induced by measurement error. Our study thus aims at obtaining an error-corrected estimate of the risk of lung cancer mortality associated with cumulated radon exposure in the French cohort of uranium miners using a structural Bayesian approach.

3 Models and notation

In order to account for measurement error, we build a Bayesian model based on conditional probabilistic independence models (Richardson & Gilks (1993)) using the assumption that exposure measurement is non differential, e.g. $P(Y_i|X_i, Z_i) = P(Y_i|X_i)$, where Y_i is the

outcome, X_i is the true value of the predictor variable and Z_i is the surrogate variable. In the case of a Berkson error model, which we will assume in the following, two sub-models may be distinguished and linked: the disease model and the measurement model.

3.1 Disease model

The outcome of interest here is the age at death by lung cancer T_i of miner i in days. We assume that the instantaneous hazard rate of death by lung cancer of miner i at time t, denoted $h_i(t)$, follows the proportional hazard structure

$$h_i(t) = h_0(t)\varphi(X_i^{\text{cum}}(t), V_{i1}(t), ..., V_{ip}(t), \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters, h_0 is the baseline hazard function, which is assumed to be the same for all miners and $\varphi(X_i^{\text{cum}}(t), V_{i1}(t), ..., V_{ip}(t), \boldsymbol{\theta})$ is a positive term expressing how the hazard varies in relation to the predictor variables.
$X_i^{\text{cum}}(t)$ is the true and unknown cumulated exposure to radon of miner i at time t, lagged by 5 years to allow for a minimal latency period of 5 years between an exposure to radon and the expression of a radio-induced risk. $V_{i1}(t), \ldots, V_{ip}(t)$ are potentially time-varying effect modifying variables. We adopt a piecewise constant hazard model with four time intervals for which baseline hazard is assumed to be constant. Rather than T_i, we only observe the couple (Y_i, δ_i) where $Y_i = \min(T_i, C_i)$ and δ_i is the indicator function $\mathbb{1}_{[T_i \leq C_i]}$. The censoring variable C_i is assumed to be non informative and independent of T_i. We considered and compared the fitting abilities of several exposure-risk relationships expressed by $\varphi(X_i^{\text{cum}}(t), V_{i1}(t), ..., V_{ip}(t), \boldsymbol{\theta})$ based on different structures and predictor variables.

3.2 Measurement model

To account for uncertainty in the radon exposure measurement, we treat the true exposure $X_i(t)$ as a latent variable (i.e. one that is not observed), while $Z_i(t)$ denotes the error prone measurement of $X_i(t)$. For the time periods 1946-1955 and 1956-1982, which were characterised by radon exposure estimations by experts and by ambient measurement devices respectively, we suppose a Berkson error model.
We postulate a lognormal and multiplicative error model to represent this Berkson error in radon exposure in the French cohort of uranium miners with different error variances $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$ associated with the time periods 1946-1955, 1956-1974, 1975-1977 and 1978-1982.
Assuming $\mathbb{E}\left(X_i(t)|Z_i(t)\right) = Z_i(t)$ yields

$$X_i(t) = Z_i(t) \cdot U_i(t), \tag{2}$$

where $U_i(t)$ are independent lognormal random variables with mean $-\frac{\sigma^2[p_i(t)]}{2}$ and variance $\sigma^2[p_i(t)]$ according to the period $p_i(t)$ ($p_i(t) \in \{1, 2, 3, 4\}$) at which the measurement for

miner i at time t was taken. For the time period between 1983 and 1999, on the other hand, exposure measurement was based on personal dosimetry presumably leading to an error best described by a classical error model. Since the classical measurement error for this time period can be assumed to be much smaller in magnitude than the Berkson error for the earlier periods, we will neglect it in the following.

4 Prior choice and Bayesian inference

We used external data on the yearly lung cancer mortality rate in French males between 1968 and 2005 to specify independent and informative gamma priors $\lambda_j \sim \mathcal{G}(\alpha_{0j}, \lambda_{0j})$, $j = 1, \ldots, 4$ for the parameters of the piecewise constant model describing baseline hazard h_0. We adopted centered normal distributions with large variances (10^4) for all regression coefficients.
Sampling from the joint posterior distribution of the unknown parameters and latent variables $\boldsymbol{\theta}$ was carried out via an adaptive Metropolis Hastings Markov Chain Monte Carlo algorithm, since no standard expression of the posterior distribution $\pi(\boldsymbol{\theta}|y, Z, V_1, ..., V_p)$ was available. This algorithm was developed and tested in Python.

5 Results

A simulation study demonstrated that accounting for measurement error with the proposed Bayesian structural approach led to a substantial bias reduction in risk estimates. In our application to the French cohort of uranium miners, the model accounting for Berkson error yielded risk estimates for cumulative radon exposure which were notably higher than in the model without measurement error. There was no considerable difference in the estimates of baseline hazard.

6 Discussion

In this study on the risk of mortality by lung cancer associated with cumulative radon exposure in the French cohort of uranium miners, we built and fitted a Bayesian model in order to explicitly take the uncertainty due to exposure measurement error into account through a unique, global and coherent framework. We made the assumption that the classical error was negligible in comparison with the magnitude of the Berkson error which occurred before 1982. Due to the flexibility of the structural Bayesian approach, this classical measurement error can easily be included in the measurement model. Moreover, further simulation studies and sensitivity analyses are needed to assess the impact of our modeling choices, including choices for prior and true exposure distributions, and to

compare the performance of the proposed method with standard functional methods, such as regression calibration and simulation extrapolation.

Bibliographie

[1] Carroll, R. J. and Ruppert, D. (2006), *Measurement error in nonlinear models - A modern perspective*, Chapman & Hall, Boca Raton.
[2] Schafer, D. W. and Purdy, K. G. (1996), Likelihood analysis for errors-in-variables regression with replicate measurements, *Biometrika*, 83 (4), 813 - 824.
[3] Küchenhoff, H. and Carroll, R. J. (1997), Segmented regression with errors in predictors: semi-parametric and parametric methods, *Statistics in Medecine*, 16, 169 - 188.
[4] Allodji, R., Thiébaut, A. C. M., Leuraud, K., Rage, E., Henry, S., Laurier, D. et Bénichou, J. (2012), The performance of functional methods for correcting non-Gaussian measurement error within Poisson regression: corrected excess risk of lung cancer mortality in relation to radon exposure among French uranium miners, *Statistics in Medicine*, 31, 4428 - 4443.
[5] Richardson, S. et Gilks, W. R. (1993), A Bayesian approach to measurement error problems in epidemiology using conditional independence models, *American Journal of Epidemiology*, 138 (6), 430-442.