The effects of shared and unshared exposure uncertainty on risk estimation via proportional hazards models in occupational cohorts

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

[1] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PRP-HOM/SRBE/LEPID, Fontenay-Aux-Roses, 92262, France, sabine.hoffmann@irsn.fr*

[2] *EA 4064 , Faculté de Pharmacie de Paris, Université Paris Descartes, 4 avenue de l'Observatoire 75006 Paris, chantal.guihenneuc@parisdescartes.fr*

[3] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PRP-HOM/SRBE/LEPID, Fontenay-Aux-Roses, 92262, France, sophie.ancelet@irsn.fr*

Résumé. Les erreurs de mesure d'exposition constituent l'une des sources d'incertitude les plus importantes dans les études épidémiologiques. Lorsqu'elles ne sont pas ou mal prises en compte, ces incertitudes d'exposition peuvent mener à des estimateurs de risque biaisés ainsi qu'à une déformation des relations exposition-risque. Dans les cohortes professionnelles, les techniques d'évaluation de l'exposition peuvent changer au cours du temps conduisant à des structures d'erreurs de mesure complexes. Bien que l'impact des erreurs de mesure non-partagées soit désormais bien établi en épidémiologie, celui des erreurs partagées sur plusieurs années de suivi d'un même individu ou partagées par plusieurs individus reste très mal connu. Dans ce contexte, l'objectif est de présenter les résultats d'une étude par simulations conduite afin d'analyser et de comparer l'impact de sources d'incertitudes d'expositions partagées et non-partagées sur l'estimation du risque et de la forme de la relation exposition-risque dans les études de cohortes professionnelles. Les résultats montrent qu'une incertitude d'exposition partagée sur plusieurs années de suivi d'un même individu conduit à des biais plus élevés ainsi qu'à une déformation plus sévère de la relation exposition-risque qu'une incertitude d'exposition non-partagée ou partagée par plusieurs individus. Cette étude souligne l'importance de faire une caractérisation détaillée des erreurs de mesure d'exposition - partagées et non partagées - potentiellement présentes dans une étude de cohorte professionnelle lorsque l'objectif est de prendre en compte ses erreurs de mesure dans les estimations du risque.

Mots-clés. Analyse de survie, erreurs de mesure, épidémiologie professionnelle, incertitude d'exposition, simulations, relation exposition-risque

Abstract. Exposure measurement error is arguably one of the most important sources of uncertainty in epidemiological studies. When exposure uncertainty is not or only poorly accounted for, measurement error can lead to biased risk estimates and a distortion of

the shape of the exposure-response relationship. In occupational cohort studies, changes in the methods of exposure assessment can often lead to complex structures of exposure measurement error. While there is a large body of literature on the impacts of unshared measurement error on statistical inference, the impacts of uncertainty components, which are shared for several individuals or which are shared for several years of the same individual remain largely unknown. In this simulation study, the aim was to compare the effects of shared and unshared sources of exposure uncertainty on risk estimation and the shape of the exposure-response curve in occupational cohort studies. Exposure uncertainty shared within individuals (i.e., shared for several years of exposure for an individual) caused more bias in risk estimation and a more severe distortion of the exposure-response curve than unshared exposure uncertainty or exposure uncertainty shared between individuals. The results of the present study underline the importance of making a careful characterisation of shared and unshared exposure uncertainty in observational studies if the aim is to account for its potential impacts on statistical inference.

Keywords. Measurement error, Survival Analysis, Occupational epidemiology, Exposure uncertainty, Simulation Study

1 Introduction

Exposure measurement error is arguably one of the most important sources of uncertainty in epidemiological studies. It is widely acknowledged that when it is not or only poorly accounted for, measurement error can lead to biased risk estimates, a distortion of the shape of the exposure-response relationship and a loss in statistical power (Carroll (2006)). Accounting for exposure measurement error can be daunting, however, because error characteristics tend to be complex in epidemiological studies.

In occupational cohort studies, for instance, one is usually interested in the association between the time until diagnosis or time until death by a certain disease and cumulative exposure to a certain chemical or physical agent. The analysis of this association may require the specification of a proportional hazards model where cumulative exposure is treated as a time-dependent variable. Owing to the time-dependent nature of cumulative exposure, the exposure history of a worker may be collected using different strategies according to the period of exposure. Changes in the methods of exposure assessment can create rather complex patterns of exposure uncertainty, where the type and magnitude of measurement error can vary over time. Uncertainty components that are shared for several years of the same individual or between individuals have received growing attention in recent years (Greenland (2015), Kromhout (2002), Simon (2006)), but the effects of these error components on risk estimation remain largely unknown. The aim of the present simulation study is to compare the effects of shared and unshared uncertainty in cumulative exposure on risk estimation and on the shape of the exposure-response

relationship in proportional hazards models.

2 The motivating study

Radon is a noble and radioactive gas, resulting from the decay of uranium 238. It is considered to be the second cause of lung cancer after smoking. The French cohort of uranium miners, a prospective cohort of 5086 uranium miners, provides relevant epidemiological data to quantify the association between occupational radon exposure and lung cancer mortality. Previous analyses of this association have assumed an unshared error structure in the cohort (Hoffmann (2017)). Furthermore, an attenuation of the exposure-response curve for high exposure values has been observed. Therefore, it is important to ascertain whether this attenuation could be due to a component of shared exposure uncertainty that is not accounted for.

3 Methods

We conducted a simulation study to assess the effects of shared and unshared measurement error on statistical inference in proportional hazards models. To mimic the exposure conditions of a "true" occupational cohort, we used information on annual radon exposure of the French cohort of uranium miners as basis for all simulations.

3.1 Models used for data generation

Disease models To generate failure times, we considered two alternative proportional hazards models $\mathcal{D}_1$ and $\mathcal{D}_2$ to describe the association between instantaneous hazard rate of death by lung cancer of miner i at age t, $h_i(t)$ and his cumulative radon exposure until time t, $X_i^{\text{cum}}(t)$. They are given by

$$\boldsymbol{\mathcal{D}_1} : h_i(t) = h_0(t)(1 + \beta X_i^{\text{cum}}(t))$$

and

$$\boldsymbol{\mathcal{D}_2} : h_i(t) = h_0(t) \exp(\beta X_i^{\text{cum}}(t))$$

where $h_0(t)$ denotes the baseline hazard of lung cancer mortality at age t. $\mathcal{D}_1$ represents an excess hazard ratio (EHR) model, which is commonly used to describe the association between cancer mortality and exposure to radon and to other sources of ionising radiation. $\mathcal{D}_2$, on the other hand, is the more classical form of the Cox proportional hazards model. Cumulative radon exposure $X_i^{\text{cum}}(t)$ is a time-varying variable as it represents the sum over all annual exposure values that worker i in group j received before time t: $X_i^{\text{cum}}(t) = \sum_{u \leq t} X_{ij}(u)$.

Measurement models When modelling the association between true $X_{ij}(t)$ and observed $Z_{ij}(t)$ exposure of worker i at time t where worker i belongs to group j, one commonly distinguishes Berkson and classical measurement error. Multiplicative Berkson error can be expressed by model

$$\boldsymbol{\mathcal{M}_1}: X_{ij}(t) = Z_j(t) \cdot U_{ij}(t),$$

where $Z_{ij}(t) = Z_j(t)$ for all workers of group j and $E(U_{ij}(t)|Z_j(t)) = 1$, which implies that $E(X_{ij}(t)|Z_j(t)) = Z_j(t)$.
Multiplicative classical measurement error can be expressed by model

$$\boldsymbol{\mathcal{M}_2}: Z_{ij}(t) = X_{ij}(t) \cdot U_{ij}(t),$$

where $E(U_{ij}(t)|X_{ij}(t)) = 1$. In contrast to model $\boldsymbol{\mathcal{M}_1}$, the variability of observed exposure is bigger than the variability of true exposure in $\boldsymbol{\mathcal{M}_2}$.
We will assume in both models that log transformed measurement errors $log(U_{ij}(t))$ are independent and normal random variables with mean $-\frac{\sigma^2}{2}$ and variance σ^2, i.e., $\log(U_{ij}(t)) \sim \mathcal{N}(-\frac{\sigma^2}{2}, \sigma^2)$[1]. In particular, $U_{ij}(t)$ and $U_{i'j}(t')$ are independent for $i \neq i'$ and $t \neq t'$ in both $\mathcal{M}_1$ and $\mathcal{M}_2$. Under this independence assumption, measurement error is considered as unshared.
To describe exposure uncertainty components that are shared between or within workers, one can adapt model $\boldsymbol{\mathcal{M}_1}$ and $\boldsymbol{\mathcal{M}_2}$ by modifying the assumptions on the structure of measurement error $U_{ij}(t)$. In order to describe measurement error shared for all subjects that belong to group j, we can rewrite model $\boldsymbol{\mathcal{M}_1}$ and $\boldsymbol{\mathcal{M}_2}$ with a measurement error term $U_j(t)$ which does not depend on subject i, but only on group j and time t. Consequently, the same error component is presumed for all subjects belonging to group j (hence the term "shared between workers") and $U_j(t)$ and $U_{j'}(t')$ are independent if $j \neq j'$ or $t \neq t'$.
Similarly, to describe Berkson or classical measurement error that is shared for several years of the same worker, we can adapt model $\boldsymbol{\mathcal{M}_1}$ and $\boldsymbol{\mathcal{M}_2}$, respectively, by assuming in this case that the measurement error term $U_{ij}(t)$ neither depends on time t nor on group j: $U_{ij}(t) = U_i \ \forall j \forall t$. The same error component is supposed for all years of exposure of worker i and U_i and $U_{i'}$ are independent if $i \neq i'$ (hence the term "shared within workers").
Besides these measurement models in which we considered only one exposure period for all exposure years, we also considered more complex models with three exposure periods characterised by different types and magnitudes of measurement error. These models can be supposed to more realistically reflect changes in the methods of exposure assessment, which are typically encountered in occupational cohort studies.

[1]This parametrisation ensures that $E(U_{ij}(t)|Z_j(t)) = 1$ and $E(U_{ij}(t)|X_{ij}(t)) = 1$.

Data generation We adapted a method proposed by Hendry (2014) to generate failure times as a function of time-varying covariates. We generated failure times for disease model $\boldsymbol{\mathcal{D}_1}$ and $\boldsymbol{\mathcal{D}_2}$ with values $\beta = 2$ and $\beta = 5$ for the true risk coefficient. Moreover, we compared the impact of big and moderate measurement error, corresponding to values for the variance of measurement error of $\sigma^2_\epsilon = 0.9$ and $\sigma^2_\epsilon = 0.1$, respectively.

3.2 Statistical inference

We conducted Bayesian inference via a Metropolis Hastings algorithm, which was developed and tested in Python (version 2.7). To study the effects of measurement error on risk estimation and on the shape of the exposure-response relationship, measurement error was not accounted for in risk estimation. We estimated the relative bias of the posterior median and the coverage rate of 95% credible intervals. When measurement error was generated according to more complex models with three exposure periods, we investigated the possibility of measurement error to induce a non-linear exposure-response relationship via EHR and Cox models based on natural cubic splines and via piecewise-linear disease models. Finally, we studied the effects of different error structures on disease model choice when inference was not accounted for measurement error.

4 Results

Exposure uncertainty shared within individuals (i.e., shared for several years of exposure for an individual) caused more bias in risk estimates and smaller coverage rates than unshared exposure uncertainty. Uncertainty shared between individuals, on the other hand, resulted in comparable relative bias and coverage rates in risk estimation in proportional hazard models as unshared exposure uncertainty. We observed a substantial attenuation in the exposure-response relationship for high exposure values when the first exposure period was characterised by exposure uncertainty shared among several years of the same worker when data were generated according to the Cox model, but not when data were generated according to the EHR model.

5 Discussion

The results of the present study underline the importance of making a careful characterisation of shared and unshared exposure uncertainty in observational studies if the aim is to account for its potential impacts on statistical inference. In particular, when exposure data with varying degrees of precision and different amounts of sharing are used in an epidemiological study, one should be aware of the distortions in the shape of the exposure response relationship this may induce. To obtain corrected risk estimates, it is important to use statistical methods that allow for complex patterns of shared and

unshared measurement error. As measurement error shared within individuals appears to have more impact on risk estimation than unshared error components or error components shared between individuals, it is important to correctly specify these error components as such and to account for the fact that the type of exposure uncertainty may vary over time. Up to our knowledge, there is currently no possibility to use classical methods, such as regression calibration or simulation extrapolation to handle these complex patterns of measurement error. In our view, the Bayesian hierarchical approach is the most promising framework in this context (Hoffmann (2017)). It is arguably the most flexible approach to account for exposure uncertainty and corrected parameter estimates can be obtained by Markov Chain Monte Carlo sampling. Additionally, the integration of prior knowledge on unknown parameters available from previous studies or in the form of expert knowledge can lead to more precise risk estimates and help to avoid overfitting, thereby increasing the replicability of findings.

Acknowledgements This work was partially supported by AREVA NC, in the framework of a bilateral agreement between IRSN and AREVA NC. AREVA NC had no role in study design, analysis, or interpretation.

Bibliographie

[1] Carroll, R. J. and Ruppert, D. (2006), *Measurement error in nonlinear models - A modern perspective*, Chapman & Hall, Boca Raton.
[2] Greenland, S., Fischer H. J., Kheifets L. (2016) Methods to Explore Uncertainty and Bias Introduced by Job Exposure Matrices, *Risk Analysis*, 36 (1), 74 - 82.
[3] Kromhout, H. (2002), Design of measurement strategies for workplace exposures, *Occupational and Environmental Medicine*, 59, 349 - 354.
[4] Simon, S. L. Hoffmann, F. O., Hofer E. (2015), The two-dimensional Monte Carlo: a new methodological paradigm for dose reconstruction for epidemiological research, *Radiation Research*, 183, 27 - 41.
[5] Hoffmann, S. Rage E., Laurier D., Laroche P., Guihenneuc C., Ancelet S. (2017), 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*, published ahead of print.
[6] Hendry, D. J. (2014), Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers, *Statistics in Medecine*, 33, 436 - 454.