Spring 2017

Improved adjustment for covariate measurement error in radon studies: alternatives to regression calibration

Keyla Pagán-Rivera
University of Iowa

IMPROVED ADJUSTMENT FOR COVARIATE MEASUREMENT ERROR IN RADON STUDIES: ALTERNATIVES TO REGRESSION CALIBRATION

by

Keyla Pagán-Rivera

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics in the
Graduate College of The
University of Iowa

May 2017

Thesis Supervisor: Associate Professor Brian J. Smith

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Keyla Pagán-Rivera

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Biostatistics at the May 2017
graduation.

Thesis Committee: Brian J. Smith, Thesis Supervisor

Joseph E. Cavanaugh

Jacob J. Oleson

Mary K. Cowles

R.W. Field

ACKNOWLEDGMENTS

Earning a graduate degree from the Biostatistics Department at The University of Iowa is a great accomplishment. For six years, I have been able to grow professionally and personally. The graduate school journey was tough and I need to thank all the people that helped me along the way.

First, I need to thank my parents. Thank you for supporting me even though you might have not understood what I was doing or why I was moving to this place call Iowa. Thanks for not holding me back and for being the great parents you are. I love you! Thanks to my sisters and other relatives for being my personal cheerleaders and for celebrating my (our) accomplishments. Thanks to my friends (old and new) for your support through the not-so-easy times, for your advice when I was confused, and for celebrating with me the good things that life gave us. I am really thankful for our friendship.

A big thank you to the Biostatistics Department at The University of Iowa, you people are just great! Thanks to Dr. Gideon K. Zamba for giving me that last push that I needed to come here six years ago. To Terry Kirk for always making sure that we had everything on time and organized. To Dr. Joe Cavanaugh for our few but important conversations. Maybe for you these were just regular talks, but they inspired me to not give up. The collaborative environment among students, faculty and staff here is something that I hope I can find wherever I go. To all of you, thanks for making this Puerto Rican feel welcome when she was so far from home.

This semester was less stressful thanks to the Graduate College and the Ballard Seashore Fellowship. Being awarded this fellowship allowed me to focus all my energies on finishing my dissertation.

Another group of people that deserve my gratitude are the staff of the Biostatistics Unit at the College of Dentistry. Working there as a research assistant for over four years helped me gain invaluable experience. Thanks to Dr. Deborah V. Dawson for all her help and guidance while I was working there.

To my dissertation committee members, thanks for the knowledge you shared with me through our discussions on epidemiological topics and in the classroom. Without all of your help, I would not be able to accomplish this last step in the PhD program. A special thanks to my advisor, Dr. Brian J. Smith. Thanks for being so patient with me and for all the guidance during the dissertation and job hunting process.

ABSTRACT

Measurement error is a type of non-sampling error that could attenuate the effect of a risk factor on an outcome variable if no correction is made. Therefore, an effect might not be detectable, even if there is one. If a classical error type is present, then the power of the analysis will be lowered or a bigger sample size will be needed in order to maintain the desirable power. Thus, a correction should be made before drawing any conclusions from the analysis. The regression calibration and simulation extrapolation methods are some of the available methods developed to deal with this kind of problem.

This dissertation proposes a Bayesian method that uses a hierarchical approach to jointly model true radon exposure (measurement error model) and its effect on lung cancer (excess odds model). This method takes subject-specific characteristics into account when making the correction, and uses random effects when missing data are present. We carried out a simulation study in order to compare this method to the regression calibration and simulation extrapolation (SIMEX). Different scenarios were simulated and the simulated data were analyzed with the three methods. This is the first time that these three methods have been compared in the context of radon risk assessment.

The simulation results showed that the proposed Bayesian method had a consistent coverage through out the scenarios. However, the SIMEX method had the lowest bias and mean squared error and, most of the time, its coverage was the closest to the nominal coverage of 95%. The regression calibration was the fastest method to be implemented, but it was outperformed by the other methods.

The dissertation finalizes by performing individual and pooled analyses using data from five case-control North America radon studies (Iowa, Missouri, Winnipeg, Connecticut, and Utah/South Idaho). The data from each study were analyzed

individually, first without making any correction, and then using the three correction methods. Finally, the data were combined and the methods were applied to this bigger sample. To the best of our knowledge, regression calibration and SIMEX have not been implemented using this combined dataset.

PUBLIC ABSTRACT

Knowledge obtained from epidemiological studies can impact policies regarding public health and, as a result, improve the quality of life for individuals in society. When we understand the relationship between potential risk factors and disease outcomes, we can take steps to minimize the exposure to such risks. The assessment of this exposure may be susceptible to measurement error, which can bias risk estimates, often toward the null hypothesis of no effect. If no effect is found due to measurement error bias, then no action will be taken to prevent the exposure.

This dissertation uses Bayesian methods to assess the relationship between risk factors and disease status, while correcting for measurement error. Specifically, the focus is to estimate the lung cancer risk associated with radon exposure, which is measured with error. We conducted simulation studies to compare the proposed method to two existing methods that have been used to correct for measurement error: regression calibration and simulation extrapolation (SIMEX). This is the first time the performance among these three methods has been compared in the context of radon risk assessment.

Our simulation results on a simple case scenario suggest that correction for measurement error is necessary to remove bias that causes attenuation of lung cancer risk estimates and even suggests a protective effect of radon exposure. Therefore, it is important to correct for measurement error before analyzing radon study data and drawing conclusions from it.

A combined dataset containing information from five North American radon studies (Iowa, Missouri, Winnipeg, Connecticut, and Utah/South Idaho) will be used to apply the proposed Bayesian methods to correct for measurement error in a pooled analysis. For comparison, the same data will be analyzed using regression

calibration and SIMEX. To date, these data had not been analyze using either of the three methods mentioned above.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

In observational studies, the investigator does not control an individual's exposure to possible risk factors. Therefore, these kinds of studies are susceptible to sampling and nonsampling errors. Some of these errors can be controlled whereas others are unavoidable. This data driven research will use two frequentist methods to correct for one type of nonsampling error, and then a Bayesian method will be proposed as an alternative to those frequentist methods. The three methods will be applied to a pooled dataset to investigate the relationship between radon exposure and lung cancer. Due to uncertainties in measuring radon, for the Bayesian method we use a hierarchical approach to jointly model true radon exposure (measurement error model) and its effect on lung cancer (excess odds model). This method takes subject-specific characteristics into account when making the correction. We also compare this method to two existing frequentist methods: regression calibration and simulation extrapolation (SIMEX).

1.1 Measurement Error

Measurement error is a type of nonsampling error that may be present when collecting data to explore the relationship between possible risk factors and disease outcomes in epidemiological studies (Bäverstam and Swedjemark, 1991; Lubin et al., 1995). It has been shown that the results of an analysis will depend on the error type, structure, size (Carroll et al., 2006) or the variable chosen to represent the predictor (Heid et al., 2004). These errors tend to affect the risk estimates in the sense that the effect could be attenuated (Fuller, 2009; Carroll et al., 2006; Lubin et al., 1995). Therefore, a correction for such attenuation could lead to an increased magnitude of the effect (Cook and Stefanski, 1994). However, this is not always the case since the effect of the measurement error depends on other factors such as the model under

consideration and the measurement error joint distribution (Carroll et al., 2006). Regardless, the effects of that bias could be corrected. It is important to note that while making such correction, extra variability is added to the parameter estimates. Therefore, a trade-off between bias and variability should be kept in mind. The variability in the measurement error and the explanatory measured variable could vary from study to study. This will lead to different parameter estimates and bias among studies (Heid et al., 2002). Thus, it is important to include a method to account for that variability in the dataset.

1.1.1 Classical vs Berkson Error

There are different types of measurement error but some of the more general ones are classical and Berkson, the first one being the most common. The classical error models are a subset of a wider classification, error models. Assume that we are interested in measuring true radon exposure (X) but can only measure exposure with some error (W), and let U be that measurement error. Then, the classical measurement error can be defined as follows (Carroll et al., 2006):

$$W = X + U, \quad E(U|X) = 0 \tag{1.1}$$

where $E(W) = X$, and the variance of U could be constant or have a variance component structure. On the other hand, if we assume that the variability of X is higher than the variability of W, then we have the Berkson measurement error model. That is, we assume that the true exposure is equal to the observed one plus some measurement error as below.

$$X = W + U, \quad E(U|W) = 0. \tag{1.2}$$

In some instances, we cannot observe the true exposure X. However, we could model the distribution of X by using regression calibration, the variable measured

with error (W), and the variables measured without error ($\boldsymbol{Z}_1$). Then, the model will be the following,

$$X = \beta_0 + \beta_1 W + \beta_2^\top \boldsymbol{Z}_1 + U, \quad E(U|W, \boldsymbol{Z}_1) = 0. \tag{1.3}$$

If we assume that $\beta_i = 0$ for $i = 0, 2$, and $\beta_1 = 1$, then we are modeling the distribution of X by using Berkson error models (Carroll et al., 2006). Note that in the classical measurement error, we are interested in modeling the distribution of $W|X, \boldsymbol{Z}_1$; whereas in the Berkson we are interested in modeling the distribution of $X|W, \boldsymbol{Z}_1$.

Suppose now that we want to fit a model to assess the relationship between an outcome Y and an explanatory variable X, after adjusting for covariates measured without error $\boldsymbol{Z}_2$. Estimates for this model can not be obtained since we cannot observe X. Therefore, we would like to find nearly unbiased estimators for the parameters when modeling the relationship between Y and the covariates $\boldsymbol{Z}_2$ and W. Since W is the variable measured with error, bias will be introduced if we use W instead of X and make no correction for that extra error (Carroll et al., 2006). Hence, we will determine which type of error we are dealing with and make a correction, rather than fitting the model without any kind of correction being made.

There is a graphical way to examine if the type of error assumed is correct. For a classical additive error, if we assume that the error U is symmetric and has constant variance on X, then the means and standard deviations of W for each individual are uncorrelated. Thus, we could plot the individual means against their standard deviations (see Figure 1.1) and look for no trends (Carroll et al., 2006). Also, we can distinguish between classical and Berkson error depending on how the covariates are measured. If a covariate is measure for an individual then the error is classical. On the other hand, if the variable is measured for a group and the

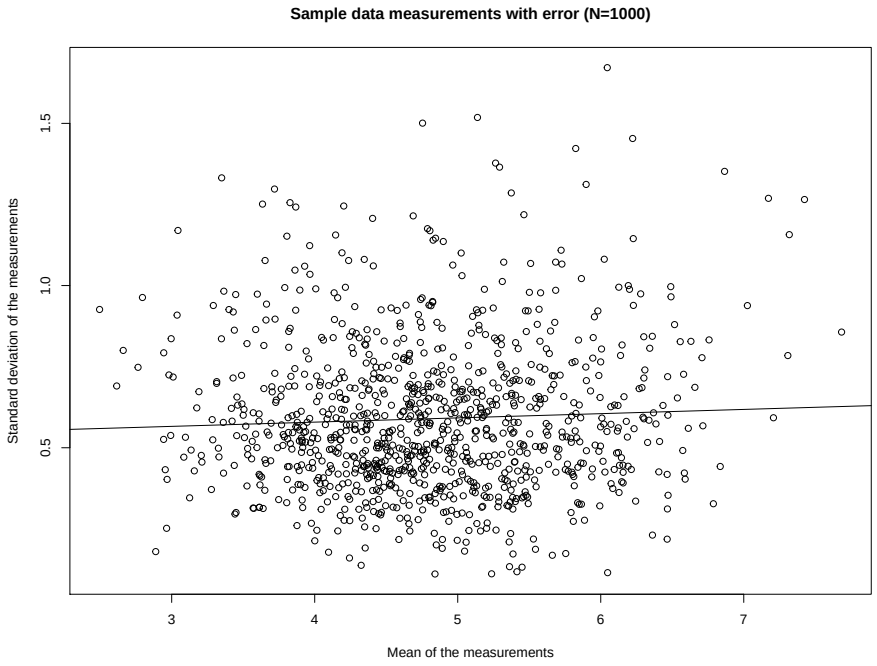


Figure 1.1: Classical error illustration

exposure for individuals varies, then the error is Berkson (Carroll et al., 2006).

1.1.2 Missing Values

Another problem when collecting and analyzing data measured with error is the missing data situation. It has been mentioned the importance of taking into account the type of missing data when performing the data analysis. In particular, a simulation study showed that when dealing with measurement error, if missing values are present, the power of the study decreases and therefore it is harder to find a significant association (Lubin et al., 1995). Different methods have been used in order to deal with this missing data problem while correcting for measurement error. For example, regression calibration assigns the population mean to those missing data points if we can assume that the data are missing at random (Fearn et al., 2008).

1.1.3 Power and Sample Size

The variability added when using a covariate measured with error will impact the power of the analysis in the sense that the power will be lowered, especially when the error is classical (Carroll et al., 2006). In fact, in a simulation study Lubin et al. (1995) showed that the power of a study dealing with measurement error decreased significantly. Analogously, if the power decreases with measurement error, the necessary sample size to maintain the power will increase (Carroll et al., 2006). Moreover, the fact that usually individuals live in more than one home during their lifetime also reduces this power.Thus, studies showed that residential mobility and measurement error will increase the necessary sample size in order to achieve the desired power (Lubin et al., 1990, 1995). Carroll et al. (2006) showed through simulations that when conducting power calculations we should assume Berkson measurement errors since we will be assuming a higher variability for X and therefore, if the error is classical, we will be overestimating the power.

1.2 Excess Odds Model

Many studies assessing the relationship between indoor radon exposure and lung cancer have used the excess odds model (or linear odds model) when analyzing the data (Darby et al., 2006; Krewski et al., 2006; Allodji et al., 2012). The model is the following:

$$\text{odds}_i = \exp(\alpha^\top \boldsymbol{Z}_{2i}) \times (1 + \eta w_i) \tag{1.4}$$

where i is the subject, $\alpha^\top$ are the parameters for the possible variables measured without error that we might want to adjust for ($\boldsymbol{Z}_{2i}$), and η is the parameter associated with the radon exposure effect. Note that if there is no adjustment by covariates, $\alpha^\top \boldsymbol{Z}_{2i}$ reduces to α_0. Also note that η is the excess odds ratio per unit of exposure to radon. That is, it represents the linear relationship between the odds

of lung cancer development and the exposure to radon.

1.3 Motivating Example

Radon is a natural gas and its isotope radon-222 is a decay product of radium and uranium (Council, 1988). It decays into a series of radioisotopes that emit alpha particles with the potential to damage lung cells and lead to lung cancer. Radon has a half-life of 3.8 days and its concentration can increase in low ventilation areas, such as mines and basements. In fact, miners were the first labor force to be associated with lung cancer, and there have been studies on the effect of radon exposure and lung cancer among miners (Radford and Renard, 1984; Lubin et al., 1995). There have also been studies on the effect of residential radon exposure and lung cancer, which have measurement error in their limitations (Darby et al., 1998, 2006; Krewski et al., 2006). Some researchers have shown that the inconsistencies among the results from residential radon and lung cancer studies can be due to errors in exposure assessment and missing values. Furthermore, they conclude that single case-control studies would not be able to estimate radon risk from indoor measurements, and even pooled studies might not be able to assess it either (Lubin et al., 1995).

In order to correct for the bias produced by measurement error, some researchers have proposed the use of regression calibration when working with radon exposure and lung cancer data (Lubin et al., 2005; Fearn et al., 2008). This correction method comes with its own limitations: it uses population distributions information to correct for the bias and does not take into account individual characteristics. Others have proposed and use simulation extrapolation (SIMEX), which also has its downside, the computer time could be extremely time consuming (Cook and Stefanski, 1994). This dissertation investigates the possibility of including subject specific information when correcting for measurement errors, while exploring

the relationship between indoor radon exposure and the risk of developing lung cancer. In order to accomplish this goal, we will perform a simulation study and analyze a pooled dataset from five different North America case-control studies (Iowa, Missouri, Winnipeg, Connecticut and Utah/South Idaho). The dataset from New Jersey was not included in the analyses since there were some discrepancies in the data that were not able to clarify. The dataset from Missouri-II was not included since the measurements were from the first five years prior to diagnosis or initial contact. The focus of the analysis will be to explain the effects of radon on lung cancer after correcting for measurement error. The North America studies that will be used for the pooled analysis are described in the next subsections. The distributions of the raw data for each site are presented in violin plots in Figure 1.2. These plots can be described as a combination of box plots and kernel density plots. In summary, the black box in the plot is the interquartile range and the median is the white dot, as in the traditional box plot. Then, a rotated kernel density is added at each side of the box plot. These raw data will be used in the error model to correct for measurement error. However, since there is a 5-year latency period for radiogenic cancer (Council, 1988), the measurements from homes that fell in the first five years prior to diagnosis or initial contact are not included in the risk model. Moreover, measurements from homes that were lived in by the individuals more than 30 years prior diagnosis or interview were also excluded. This exposure time window (ETW) was chosen following the work of Krewski et al. (2006).

1.3.1 Iowa

The Iowa Radon Lung Cancer Study (IRLCS) recruited 413 newly diagnosed female lung cancer patients and 614 age-matched controls from 1993 to 1996 (Table 1.1). Also, lung cancer patients should have lived in the same house for 20 consecutive years and be in the 40-84 age range. These gender, age and residency criteria

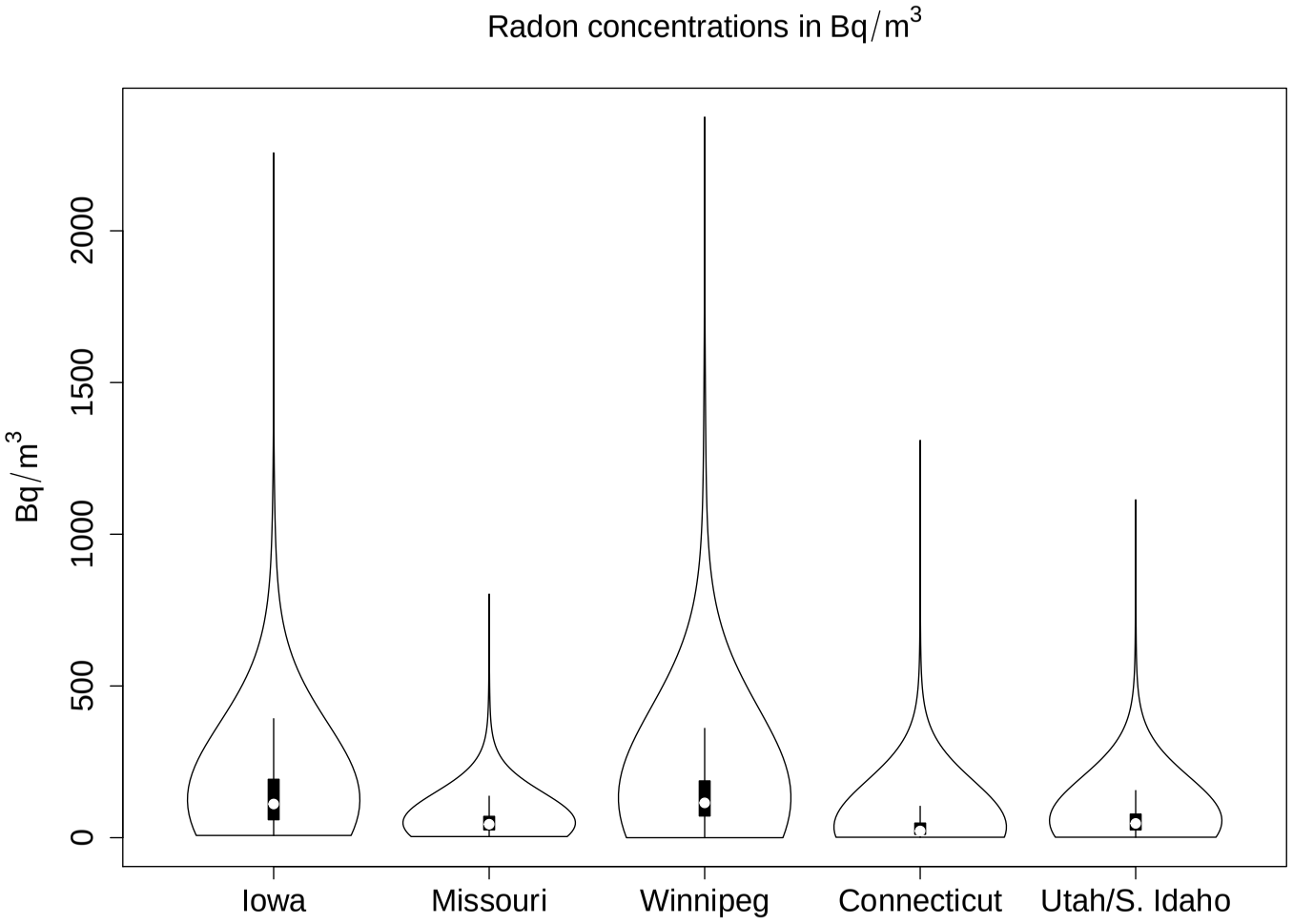


Figure 1.2: Distribution of the radon concentration in Bq/m^3

were also required for the controls, as well as no prior primary invasive lung cancer at recruitment time. In order to measure radon exposure, alpha-track detectors were placed on each level of the home and left in there for a year. Information regarding which room of the home was used to place the detectors was also collected. An interview was used to gather information about the time spent at home and in other buildings, as well as age, smoking history and education level (Field et al., 2000). Figure 1.2 shows the distribution of the radon concentrations in Becquerel per cubic meter.

1.3.2 Missouri

The Missouri study recruited 618 non-smoker (never smoked or former smoker) white women with lung cancer falling in the 30-84 age range from 1986 to 1991 (Table 1.1). Demographic information such as occupational and smoking history, diet and lung cancer information was collected through a telephone interview. From those initially interviewed, radon measurements were collected from 538 individuals using year-long alpha-track detectors (Alavanja et al., 1994). In instances where the lung cancer patient had died or was unable to participate in the interview, a next-of-kin was interviewed. A total of 1587 age-matched controls were selected from the Missouri Department of Revenue and Health Care Financing Administration. There were 1402 of those individuals that responded to a phone interview. Of those, 1183 completed an interview and had year-long alpha-track detectors in at least one home. These detectors were placed in the bedroom and kitchen for one year. The homes were lived in by the subject for at least one year out of the previous 5 to 30 years. To measure the variation within individual homes and to measure the impact of changing seasons, a few homes had detectors placed in the bedroom, kitchen and basement with readings every three months. The distribution for the radon concentrations in Becquerel per cubic meter can be found in Figure 1.2. As

Table 1.1: Disease status frequencies for each North American study

Study	Cases	Controls	Total
Iowa	413	614	1027
Missouri	618	1402	2020
Winnipeg	738	738	1476
Utah	511	862	1373
Connecticut	963	949	1912

in the Iowa study, the lower bound of 5 to 30 years of exposure was chosen because of the lung cancer latency period shown in miners studies (Hornung and Meinhardt, 1987; Lubin et al., 1994). Researchers decided on the 30 year upper bound based on a miner study which suggested that the risk of lung cancer decreases with time (Council, 1988), and based on the fact that radon measurement estimates become more inaccurate with time.

1.3.3 Winnipeg

Between 1983 and 1990, a case-control study investigating the relationship between indoor radon exposure and lung cancer risk was conducted in Winnipeg, Canada (Letourneau et al., 1994). This city was chosen because it had the highest indoor radon levels among the cities monitored by McGregor et al. (1980). A total of 738 individuals with lung cancer within the 35-80 age range were recruited. The Winnipeg telephone directory was used to identify and recruit 738 age and sex matching controls (Table 1.1). The researchers conducting the study tried to place an alpha-track detector at each of the homes (within the Winnipeg metropolitan area) lived in by the individuals for at least one year. If the individual passed away

during the study process, a proxy respondent was interviewed. There were some case-controls pairs that were excluded due to lung cancer status (not primary lung cancer), controls were not properly matched, and three subjects with mesothelioma thought that it was not due to radon exposure. Radon dosimetry was assessed by placing alpha-track detectors in one or two rooms at each house. After six months, the detectors were retrieved and replaced with new detectors. The average of those measurements was used as an estimate of the radon exposure over a 1-year period. The violin plot in Figure 1.2 shows the distribution of the radon measurements.

1.3.4 Utah/South Idaho and Connecticut

Cancer registries and medical records reviews helped identify lung cancer subjects for this study between 1989 and 1992 (Sandler, 1999). There were 963 cases from Connecticut and 511 from Utah/South Idaho between the ages of 40 and 70 years (Table 1.1). Individuals who smoke pipe or cigars smokers were excluded. Telephone screenings and Medicare recipients listings were used to identify the controls. These controls were matched by gender, age and smoking status. If the individual with lung cancer (case) was from Utah and either a never-smoker or nonsmoker during the past 10 years, then two matching controls were selected for that case. Otherwise, one control per case was identified. There were a total of 949 controls from Connecticut and 862 from Utah/South Idaho. Medical history and information about previous houses occupied for at least a year was obtained via interview. The majority of subjects were white, and there were fewer females than males on each study (49% in Connecticut and 34% Utah/South Idaho were females). Detectors were placed in the bedroom and another room on the lowest home level for one year. Homes that had more than two levels above ground, homes where the subject lived for less than a year, mobile homes without a permanent foundation or institutions were excluded from the analysis.

CHAPTER 2
METHODS

In Chapter 1 we introduce the problem, the goal and the datasets that will be used in this data-driven problem. This chapter introduces two existing frequentist methods to correct for measurement error (regression calibration and SIMEX) as well as the proposed Bayesian method and some background on Bayesian techniques. Through the chapter we talk about models with and without adjustment by covariates. The variables that were used when fitting the adjusted model were sex, and categories of age (under 50, 50-54, 55-59, 60-64, 65-69, 70-74, 75 years or more), smoking duration (nonsmoker, 1-24, 25-34, 35-44, 45 years or more) and smoking rate (nonsmoker, 0.5-9.5, 9.5-19.5, 19.5-29.5, 29.5 or more averaged cigarettes smoked per day during the active smoking period). Frequencies for these variables by study are presented in Table 4.2.

2.1 Regression Calibration

Regression calibration is a non-iterative method used to correct point and interval estimates obtained from regression models. It is a correction for bias due to measurement error in one or more continuous variables (Spiegelman et al., 1997). If we were working with generalized linear models, then this method could be convenient. On the other hand, it performs poorly for highly nonlinear models if no modifications are made (Carroll et al., 2006). This method has been used to correct for measurement error when working with lung cancer or with nutritional data (Spiegelman et al., 1997; Fearn et al., 2008). It is implemented by using the sample characteristics (mean and variability) to correct the error.

The goal of regression calibration is to model $X|W$, but with the use of Bayes

theorem, we can start by modeling $W|X$ (Carroll et al., 2006). That is,

$$f_{X|W}(x|w) = \frac{f_{W|X}(w|x)f_X(x)}{\int f_{W|X}f_X(x)dx} \tag{2.1}$$

where X is the true exposure variable, W is the observed variable measured with error, $f_{X|W}(x|w), f_{W|X}(w|x)$ and $f_X(x)$ are the densities of $X|W, W|X$ and X, respectively. Then we can use the observed measurements and the covariates measured without error ($\boldsymbol{Z}_1$) to obtain an estimated calibration function $E(X|\boldsymbol{Z}_1, W)$ and a Berkson model when working with classical error measurements. In fact, Carroll et al. (2006) defined the regression calibration as "classical measurement error made into a Berkson model" and showed how the observed measurements can be used to find the best linear predictors of $X|W$.

In other words, there is a relationship between the classical and Berkson error models described in Chapter 1, and regression calibration models. That is, suppose that we want to assess the relationship between an exposure variable X and an outcome variable Y, after adjusting for variables measured without error $\boldsymbol{Z}_2$. Assume now that instead of observing X, we observe W. Carroll et al. (2006) described the regression calibration algorithm as follow. First, estimate $m_X(\boldsymbol{Z}_1, W, \zeta)$, which is the regression of X given $\boldsymbol{Z}_1$, W and the parameter ζ. Second, substitute X with $m_X(\boldsymbol{Z}_1, W, \hat{\zeta})$ and run the analysis as usual (note that we use $\hat{\zeta}$ as an estimator of ζ). Third, the bootstrap or sandwich method could be used in order to correct the standard errors for the estimation of ζ. Therefore, instead of modeling the distribution of $Y|X, \boldsymbol{Z}_2$, we will model $Y|m_X(\boldsymbol{Z}_1, W, \hat{\zeta}), \boldsymbol{Z}_2$. We could use a validation sample or an unbiased instrument in a subset of the data in order to estimate $m_X(\boldsymbol{Z}_1, W, \zeta)$.

Fearn et al. (2008) compared the performance of regression calibration with a more complicated approach, Monte Carlo integration. They conclude that regression calibration performs well when analyzing data that include an explanatory

variable with measurement error and a binary outcome. Also, they mentioned that regression calibration would reproduce the correct regression function in simple linear regression and provide a good approximation for nonlinear models.

In our case, we use the measured radon concentrations to estimate $E(X|\boldsymbol{Z}_1, W)$. Based on the literature (Fearn et al., 2008), we can assume that the log transformed true and measured radon concentrations (X and W, respectively), as well as the conditional distribution, follow a normal distribution. That is,

$$\begin{aligned} x &\sim N(\mu_X, \sigma_X^2) \\ w|x &\sim N(x, \sigma_W^2) \\ x|w &\sim N(\mu_{X|W}, \sigma_{X|W}^2) \end{aligned} \tag{2.2}$$

where $\mu_{X|W} = \left(\frac{w}{\sigma_W^2} + \frac{\mu_X}{\sigma_X^2}\right)\left(\frac{1}{\sigma_W^2} + \frac{1}{\sigma_X^2}\right)^{-1}$ and $\sigma_{X|W}^2 = \left(\frac{1}{\sigma_W^2} + \frac{1}{\sigma_X^2}\right)^{-1}$. Since the lung cancer incidence is low, estimates for these parameters can be obtained using the measured log radon concentrations (w) on the controls (Fearn et al., 2008). Specifically, the sample mean and variance of W will be used to estimate μ_X and σ_X^2, and a mixed model on the radon exposure (measurement error model) will be used to estimate σ_W^2. That model can be defined as follow:

$$w_{ijk} \sim N(\beta^\top \boldsymbol{Z}_{1ijk} + \boldsymbol{\gamma}_{j(i)}, \sigma_W^2) \tag{2.3}$$

where $\beta^\top$ are the parameters associated with the variables measured without error ($\boldsymbol{Z}_{1ijk}$) for subject i at home j and measurement k. The parameters associated with the random effects $\boldsymbol{\gamma}_{j(i)}$, will have different meanings depending on the study data analyzed. For the Iowa dataset, these are random effects for time period j within subject i. For all other studies, there will be a nested home within subject effect. Moreover, in the Iowa study, the home and subject random effect will be the same since subjects lived in the same home. Therefore, a home- or subject-specific random effect (γ_i) will also be added to the model. The variability and dependence introduced by estimating the parameters using the sample data will be small since

the sample sizes are big (Fearn et al., 2008). In the application of this method, the parameter estimators from the mixed model will have two purposes. First, the fixed effects will be used for the imputation process when missing radon measurements are present. Second, and following Fearn et al. (2008), the parameter estimates will be used in the following equation

$$E(\exp(x)|w) = \exp(\mu_{X|W} + 0.5\sigma^2_{X|W}) \tag{2.4}$$

which will be used to correct the individual's radon measurements. Once we correct the radon measurements, we compute a single corrected measurement per subject. These corrected measurements will be used to fit the following risk model (excess odds model)

$$\text{odds}_i = \exp(\alpha^\top \boldsymbol{Z}_{2i}) \times (1 + \eta w_i^c) \tag{2.5}$$

where $\alpha^\top$ are the parameters associated with the variables measured without error ($\boldsymbol{Z}_{2i}$) that we might want to adjust for in the risk model, η is the parameter associated with radon exposure and w_i^c are the corrected measurements (one measurement per subject i). Note that if there were no adjustment by covariates, $\alpha^\top \boldsymbol{Z}_2$ will reduce to α_0.

2.2 Simulation Extrapolation - SIMEX

Similar to regression calibration, simulation extrapolation (SIMEX) is used to reduce bias due to measurement error. It was proposed by Cook and Stefanski in 1994 and it works well for additive measurement error as well as for any other error models (Carroll et al., 2006). SIMEX uses simulations to establish the effect of measurement error in a parameter estimate. According to Cook and Stefanski (1994), this method is applicable when the measurement error variance can be well estimated or it is known. The authors said that the method is convenient in the sense

that users who might not feel comfortable with the modeling theory surrounding error models could implement it. This method could also be used when the model is new and the traditional methods to implement such a model are not available yet. In fact, when Cook and Stefanski (1994) developed this method, they did it so they could find an option to fit nonstandard generalized linear measurement error models.

In order to understand the method, let us suppose that we are dealing with classical error model,

$$W = X + U, \quad E(U|X) = 0 \tag{2.6}$$

where W is the variable measured with error, X is the true radon concentration, and U is the measurement error. We could fit a model and compute naive estimators that will depend on the outcome variable Y and the variable measured with error W. However, what we would prefer is to correct for the bias introduced by the measurement error in the parameter estimates for the model that assesses the relationship between Y and W. Cook and Stefanski (1994) proposed the creation of groups of contaminated datasets which would have progressively larger measurement errors as $\boldsymbol{\lambda} = 0 < \lambda_1 < \lambda_2 < ... < \lambda_B$ increases. Each contaminated dataset m in the group will have a variable created using the original measurements W and the computer-generated extra errors U as in the equation below

$$W_m(\lambda) = W + \lambda^{1/2} U_m, \quad \lambda \geq 0, \quad m = 1, 2, ..., M \tag{2.7}$$

where $U_m \sim N(0, 1)$ are mutually independent, and independent of Y, W, X, and any other variable measured without error, $\boldsymbol{Z}$. Note that the variance of $W|X$ is σ_U^2, whereas $\text{var}\big(W_m(\lambda)|X\big) = (1 + \lambda)\sigma_U^2 = (1 + \lambda)\text{var}(W|X)$. Therefore, instead of having a measurement error variance of σ_U^2, each of the M datasets in a λ group will have $(1 + \lambda_b)\sigma_U^2$ error variance (Carroll et al., 2006). For $\lambda = 0$ we have the original

dataset and as a result, $\text{var}(W_m|X) = \sigma_U^2$. Note that we could express Equation 2.7 as

$$W_m(\lambda) = X + (1+\lambda)^{1/2} U_m, \quad \lambda \geq 0, \quad m = 1, 2, ..., M \tag{2.8}$$

If $\lambda = 0$ then $W_m = X + U_m = W$, and when $\lambda = -1$ then $W_m = X$. Therefore, the measurement error variance remained as $(1 + \lambda_b)\sigma_U^2$.

The SIMEX algorithm can be describe as follows. First, simulate $m = 1, 2, ..., M$ datasets for each value of λ from the original data, each with additional measurement error added as in 2.7. Second, compute the parameter estimates from those M datasets as usual. Repeat steps one and two multiple times and average the naive estimates for each λ_b. Finally, use regression to extrapolate to the case where $\lambda = -1$ (no measurement error). In order to construct confidence intervals, bootstrap could be use to obtain the standard errors.

This method can be used to correct for measurement error in radon dosimetry (Allodji et al., 2012) as well as in atomic bomb survivors dosimetry (Allodji et al., 2015). In both cases, the method was compared to the regression calibration described in the previous section. Allodji et al. (2015) concluded that SIMEX could be a good alternative to the regression calibration. Even though Allodji et al. (2012) found that for some cases the simulation extrapolation method works slightly better than the regression calibration, they showed that in other cases it could overestimate the results. One of the limitations of the method is that it is computing intensive. Moreover, Allodji et al. (2015) found that the correction done by applying the SIMEX method depends not only on the range of λ but also on the extrapolation function used. Additionally, they found that a good trade off between effectiveness and implementation time can be found if we generate 20 contaminated datasets for each value of λ.

In our case, we will implement SIMEX in the study of the relationship between

indoor radon exposure and the risk of developing lung cancer. Since we cannot observe the true radon exposure (X), we cannot fit the risk model and obtain the true parameter estimate η. However, we will use the measured radon exposure (W) to fit the risk model 2.5 and obtain a naive parameter estimate $\hat{\eta}$. By using $M = 20$ and $\lambda = 0, 0.4, ..., 1.6, 2$, we obtained 20 of those $\hat{\eta}$ for each value of λ. These parameters were then averaged for each λ case and the resulting estimate, $\hat{\eta}(\lambda)$, was obtained. The resulting six pairs of λ and $\hat{\eta}(\lambda)$ were used to fit a regression model to extrapolate to the case where there is no measurement error ($\lambda = -1$). For this implementation of SIMEX, the estimation of σ_U^2 was based on the results from the linear mixed models fitted for each of the five North America studies. As with the regression calibration correction, these mixed models were also used to impute values when we are dealing with missing values.

2.3 Bayesian Inference

A method to correct for measurement error when the explanatory variable is measured with error, while jointly fitting the risk model, is proposed in this section. This method will correct for measurement error (error model) and assess the relationship between indoor radon exposure and the risk of developing lung cancer (risk model) in one process. In order to accomplish this, we used a Bayesian hierarchical approach and Bayesian mixed models to model the true radon exposure and its effect on lung cancer status. This section presents an introduction to Bayesian inference, followed by some background on Bayesian models, and finally the proposed model is explained.

2.3.1 Background

In Bayesian statistics, the parameters are treated as a random quantity and the sample data as non-random observed quantities. Inference is made based on

probability models from which we make conclusion about the parameter (θ) or unobserved data ($\tilde{Y}$), based on the observed data (Y_{obs}) and the explanatory variables ($\boldsymbol{Z}$) (Gelman et al., 2014). In order to make inferences, we use Bayes rule to specify the posterior distribution

$$p(\theta|Y_{obs}) = \frac{p(\theta, Y_{obs})}{p(Y_{obs})} = \frac{p(Y_{obs}|\theta)p(\theta)}{p(Y_{obs})} \tag{2.9}$$

where $p(Y_{obs})$ is the marginal distribution of Y_{obs}, $p(\theta)$ is the prior distribution of θ, $p(Y_{obs}|\theta)$ is the sampling distribution, and $p(\theta, Y_{obs})$ is the joint distribution.

A Bayesian model can be divided into two stages: the probability model and the prior distribution (Cowles, 2013). In the first stage, the statistician identifies a probability distribution for the data. In the second stage we incorporate the prior knowledge that we have about any unknown parameters. If the information about the parameter is limited, then we can specify vague priors. On the other hand, if we know more about the parameter and we want to incorporate that knowledge, then we specify more informative priors. Regardless, the prior distribution should at least reflect the range of the parameter. For example, if we want to estimate a proportion, the prior distribution should be chosen such that the fact that the parameter lies between zero and one is taken into account. In cases where the posterior distribution is of the same family as the prior distribution, the prior is known as a conjugate prior. In other words, let $\mathcal{F}$ be a class of sampling distributions $p(Y_{obs}|\theta)$ and let $\mathcal{P}$ be a class of prior distributions for θ, we say that $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$p(\theta|Y_{obs}) \in \mathcal{P} \text{ for all } p(\cdot|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P} \tag{2.10}$$

Another type of prior is the improper prior, those with distributions that do not integrate to one. Special attention needs to be paid when using improper priors since they do not always lead to proper posterior distributions (Gelman et al., 2014). In summary, the Bayesian data analysis can be outlined in three steps:

1. Specify the joint probability model, $p(\theta, Y_{obs})$
2. Compute the posterior distribution, $p(\theta|Y_{obs})$
3. Perform model diagnosis

There are some differences between frequentist and Bayesian statistics. One of them is the interpretation of probability. In the frequentist approach, it is assumed that if an experiment is repeated multiple times under the same conditions and the results are used to compute a $(1-\alpha)100\%$ confidence interval, then the confidence interval will include the true parameter value $(1-\alpha)100\%$ of the time. On the other hand, Bayesian statistics uses credible intervals. In that case, a $(1-\alpha)100\%$ credible interval is interpreted as having $(1-\alpha)100\%$ probability that the true parameter value lies within that interval given the observed data.

An advantages of Bayesian analysis is that it allows us to incorporate prior information (prior distribution) when fitting a model, and to fit hierarchical models that are difficult or sometimes even impossible to fit using frequentist approaches. Other advantages are that the inferences are conditional on the observed data, and probability statements can be made about quantities of interest. Moreover, estimation and asymptotic theories, and special adjustment for multiple comparisons are not needed since the analyses follow directly from the posterior distributions. On the other hand, there are disadvantages to this approach: it can be computationally intensive and has been criticized for being subjective when choosing the priors.

2.3.2 Bayesian Models

The most common probability distributions used in frequentist modeling can also be used when fitting Bayesian models. In the Bayesian modeling scenario, we could use the same distribution or a combination of them to specify the prior, marginal or sampling distribution. The posterior distribution could also be part of

these widely known distributions, which would make it easier to compute posterior statistics. However, sometimes the sampling distribution is not so common or the posterior distribution does not have a closed form. Other times, we are dealing with a model that involves more than one or two parameters. In those cases, inference about the parameters can still be done.

Hierarchical Bayesian models are an example of the more complicated models mentioned before. These models include two or more parameters that are somehow connected by the structure of the problem. They are useful in more realistic scenarios where the data have a hierarchy and the (wrongful) use of a non-hierarchical models could not fit large datasets accurately or could over-fit the data (Gelman et al., 2014). An example of the usefulness of this kind of models is the work by Zhang et al. (2008), where they used a hierarchical approach to fit a Cox model to double censored HIV data, and model all parameters simultaneously. According to Cowles (2013), we could describe three stages for the hierarchical models: 1) define the likelihood, 2) define the priors on the parameters in the likelihood and, 3) define hyperpriors. The third stage is needed because some of the priors on stage two might not have a fixed number/value for their parameter. Therefore, the unknown parameters in the priors will require a prior distribution. Those parameters are known as hyperparameters and their distributions as hyperpriors. In general, hierarchical models are in the form of

$$\begin{aligned} Y_{obs}|\theta,\phi &\sim p(Y_{obs}|\theta,\phi) \\ \theta|\phi &\sim p(\theta|\phi) \\ \phi &\sim p(\phi) \end{aligned} \tag{2.11}$$

where ϕ is the hyperparameter and $p(\phi)$ is the hyperprior with fixed numerical parameter(s).

Once the model is defined, we could use a Markov chain Monte Carlo (MCMC) algorithm like Gibbs or Metropolis-Hastings sampler to fit the model. MCMC is a

computational approach that uses approximate distributions to draw values of θ, and at each iteration the draws are corrected such that we can approximate the posterior distribution better in the next draw. For this purpose, we use a Markov chain in which the next draw only depends on the previous one. In other words, we start at an initial point θ^0 and then draw a new value θ^t from the approximate distribution Q, we repeat this process multiple times such that $Q_t(\theta^t|\theta^{t-1})$. If we are able to sample directly from a full conditional distributions with known form, then we could use the Gibbs sampler algorithm. Otherwise, we could use Metropolis-Hastings if sampling from the conditional distribution is difficult. Metropolis-Hastings is also useful when the model parameters are unconstrained (Gelman et al., 2014).

After fitting the model, a convergence assessment of the Markov chains must be performed to corroborate that the resulting sequence converges to the posterior distribution. Trace plots and Gelman and Rubin diagnostic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) are some of the tools that can be used to achieve that. We will also want to obtain a small MCMC standard error (MCSE). The early iterations when the chains have not converged yet, must be discarded and not used for the inferences. These iterations are part of the burn-in sequence. After the chain has converged, the model performance needs to be checked. Similar to the frequentist approach, there are test statistics that could be used for this purpose. In the case of Bayesian statistics, Bayesian p-values can be use to compare the test statistics from replicated data to the ones from the observed data. These replicated data are future observable (or that could have been observed) quantity obtained using the parameters from the fitted model. The Bayesian p-value (Gelman et al., 1996) can be defined as follows:

$$p_B = Pr\big(T(Y^{rep}, \theta) > T(Y, \theta)|Y\big) \tag{2.12}$$

where p_B is the p-value, T is the test statistic and Y^{rep} is the replicated data.

Contrary to the frequentist p-value, we do not want to obtain a small p_B. More specifically, we would like to obtain a Bayesian p-value that is as close to 0.5 as possible. This p-value could be computed by using the simulations obtained from the posterior predictive distribution.

Inferences made using Bayesian models can be done after assessing the Markov chain convergence and model diagnosis. The resulting posterior distributions can be used to obtain posterior means, credible intervals, highest posterior density intervals, and to make predictions.

2.3.3 Proposed Model

A Bayesian hierarchical model was fitted in order to correct the variable measured with error (indoor radon exposure), and to explore the association between that risk factor and the risk of developing lung cancer. The correction part of the model was specified using a mixed model and the study-specific information available. In general, that section of the model can be expressed as follows:

$$\begin{aligned} \ln Bq_{ijk} &\sim N\left(\mu_{ijk}, \frac{1}{\tau_{\text{lnBq}}}\right) \\ \mu_{ijk} &= \beta^\top \boldsymbol{Z}_{1ijk} + \gamma_{j(i)} \\ \tau_{\text{lnBq}} &\sim Gamma(0.001, 0.001) \\ \beta_p &\sim N(0, 0.0001) \quad \text{for } p = 0, 1, ..., P \\ \gamma_{j(i)} &\sim N\left(0, \frac{1}{\tau}\right) \\ \tau &\sim Gamma(0.001, 0.001) \end{aligned} \tag{2.13}$$

where lnBq_{ijk} is the natural logarithm of the radon exposure measurement in Becquerel per cubic meter (lnBq/m^3) for individual i at home j and measurement k, which follows a Normal distribution centered in μ_{ijk} and with measurement error precision equal to τ_{lnBq}. The mean μ_{ijk} for each study will be computed based

on the available information for each study. It will depend on the variables measured without error $\boldsymbol{Z}_1$ (rooms, disease status or home level), the intercept (β_0) and parameters associated with those fixed effects $\beta_1, ..., \beta_P$, and the random effects (home-specific or home within subject nested effect) γ. The second stage of this hierarchical model includes the priors to the parameters in the likelihood. We use vague Normal distributions centered around 0 and with small precisions for the fixed effects parameters. For the random effects, we used Normal distributions centered around 0 with precisions equal to τ. The precision of the Normal distribution of lnBq, has a vague Gamma prior distribution with shape and scale parameters equal to 0.001. The third stage of the hierarchical model assigned vague Gamma hyperpriors to the precision τ.

Similar to the approach of Zhang et al. (2008), we deal with the missing data problem within the MCMC process. That is, rather than imputing values when missing measurements are present as with the frequentist methods (regression calibration and SIMEX), this Bayesian model assigns random effects to any missing home during the 6-30 years exposure time window (ETW). In our case, we used the available measurements to obtain parameter values for Model 2.13 and random effects to compute the corrected measurements for those missing data points. This ETW was chosen following the work of Krewski et al. (2006) and it is explained in more detail in Chapter 1. There is however a similarity between regression calibration and this method, the parameter from the error model will be use to correct the radon measurements. Also, once the measurements are corrected, they will be transformed back to the original scale (Bq/m^3) and averaged in order to obtain an exposure time window (ETW) measurement of them. This process was done individually for each study since the information available for one study might not be consistent with the information in other studies. The single corrected ETW

measurement was used to specify the risk model (excess odds model) as follows:

$$\begin{aligned} \text{STATUS}_i &\sim Bernoulli(\pi_i) \\ \pi_i &= \frac{\text{odds}_i}{1+\text{odds}_i} \\ \text{odds}_i &= \exp(\alpha^\top \boldsymbol{Z}_{2i}) \times (1+\eta \text{Bq}_i^c) \\ \alpha_g &\sim N(0, 0.001) \quad \text{or} \\ \alpha_g &\sim Unif(a,b) \quad \text{for } g = 0,1,...,G \\ \eta &\sim N(0,0.001) \prod_{i=1}^{n} I_{[(1+\eta \text{Bq}_i^c)>0]} \end{aligned} \tag{2.14}$$

where STATUS_i is the lung cancer status (coded as 1 for disease and 0 for non-disease) for individual i, π_i is the probability of developing lung cancer, $\alpha^\top = (\alpha_0, \alpha_1, ..., \alpha_G)$ are the parameters associates with the variables measured without error ($\boldsymbol{Z}_2$) that might be included in the adjusted model (age, smoking history, gender) and η is the parameter associated with the corrected radon measurement (Bq_i^c). In most cases, the prior distributions for the $\alpha^\top$ parameters are Normal distributions centered around 0 and with small precisions, but for some cases the priors might be Uniform distributions. This restriction on the prior bounds is done so that the MCMC error could be lowered quicker. The prior distribution for the parameter associated with the corrected radon exposure, η, is also a Normal distribution centered around 0 and with small precision. This prior distribution has a constraint, that the resulting values of $(1+\eta \text{Bq}_i^c)$ must be positive. This constraint is necessary in order to obtain valid odds estimates, that is $0 \leq \text{odds} < \infty$ (Chu et al., 2011). This Bayesian method jointly models the error and risk models so that all parameters are modeled simultaneously.

Model convergence will be assessed by examining the trace plots, computing Gelman and Rubin diagnoses and by examining the MCMC error. For the model diagnosis, we will use a goodness-of-fit test for the error model. The test statistic

for this tests can be defined as follow:

$$T(\text{lnBq}, \eta) = \sum_{i=1}^{n} \frac{\big(\text{lnBq}_i - E(\text{lnBq}_i|\eta)\big)^2}{\text{var}(\text{lnBq}_i|\eta)} \tag{2.15}$$

Whereas for the risk model diagnosis we used the concordance index (c-index). This c-index has been used before to assess the predictive capacity of a Bayesian model with binary outcomes and it can be defined as $C = P(\pi_D > \pi_{\bar{D}}|\text{STATUS})$. We can compute the c-index (Hanley and McNeil, 1982; Souza and Migon, 2004) within the MCMC iterations using the the equation below for each iteration

$$C = \frac{1}{n_{pairs}} \sum_{r=1}^{n_D} \sum_{s=1}^{n_{\bar{D}}} C_{rs} \tag{2.16}$$

where n_{pairs} is the number of discordant pairs, n_D is the number of cases, $n_{\bar{D}}$ is the number of controls, and C_{rs} is either 1 if $\pi_{rD} > \pi_{s\bar{D}}$ or 0 if $\pi_{rD} < \pi_{s\bar{D}}$ for each combination of cases and controls.

This proposed Bayesian method allows us to use one big process in order to correct for measurement error, obtain a single ETW variable per individual and, assess the relationship between indoor radon exposure and the risk of developing lung cancer. That is, we will be modeling all model parameters simultaneously, while applying the exposure time window inclusion/exclusion criteria. Moreover, this model has the flexibility of being implemented for a single study or for a pooled dataset that contains all the studies, while still implementing a study-specific correction for measurement and computing a study-specific ETW, but keeping the risk model intact.

CHAPTER 3
SIMULATIONS

A simulation study was performed to compare the three methods: regression calibration, simulation extrapolation (SIMEX) and the proposed Bayesian method. There were multiple scenarios considered and different methods to compare the performance and efficiency of the methods. Within each scenario, the parameter estimate, mean squared error (MSE), bias, and coverage probabilities were computed. Robustness of the methods was tested by varying the sample size, risk factor effect, and the number of studies.

3.1 Single Study Scenario

The first scenario deals with a simple model were there is only one study, an even number of cases and controls, and where each individual has two measurements per home in a total of two homes. True radon concentrations for subject i at home j were generated using the following model:

$$\begin{aligned} x_{ijk} &= \beta_0 + \beta_1 \text{ROOM}_{ijk} + \gamma_{1j(i)} \\ \gamma_{1j(i)} &\sim N(\gamma_{0i}, \sigma_1^2) \\ \gamma_{0i} &\sim N(0, \sigma_0^2) \end{aligned} \tag{3.1}$$

where x_{ijk} are the true log radon concentrations for subject i at home j and measurement k, β_0 is the intercept, β_1 is the room effect (coded as 0 for "ROOM 1" and 1 for "ROOM 2"), γ_{0i} is the subject-specific random effect, and $\gamma_{1j(i)}$ is the random effect for each home within each individual. The true parameter values were set as follows: $\beta_0 = 4.4812$, $\beta_1 = 0.6301$, $\sigma_0^2 = 0.5659$, and $\sigma_1^2 = 0.0188$. These choices were made based on analysis of observed radon concentrations. Simulated measurements were averaged over home and then over subject, and then transformed in order to obtain a single measurement per subject in the original scale (Bq/m^3). That single radon measurement (x_i^a) along with the intercept, $\alpha_0 = -0.4704$, and

the parameter associated with the radon effect, $\eta = 0.0005$ (or $\eta = 0.05$ per 100 Bq/m^3), were used in Equation 3.2 to compute the disease status for each subject.

$$p(\pi_i = 1) = \frac{\exp(\alpha_0) \times (1 + \eta x_i^a)}{1 + \big(\exp(\alpha_0) \times (1 + \eta x_i^a)\big)} \tag{3.2}$$

Measurement error was introduced to the model by adding independent samples from $N(0, \sigma^2 = 0.0643)$ to the simulated true log radon measurement in order to obtain the variable measured with error (W). Samples were generated big enough so that each of them will have 500 cases and 500 controls. A total of 1200 samples were generated in approximately 30 minutes using a PC with Intel(R) Core(TM) i5-3570 CPU @ 3.40GHz. For the Bayesian models, a burn-in of 5k iterations was discarded and an additional 15k iterations were used for the inferences.

The next step was to analyze the generated data by applying regression calibration, SIMEX and the proposed Bayesian method to the data measured with error on each of the samples. The R (R Core Team, 2016) software was used for data generation as well as for applying the regression calibration and SIMEX methods. The Just Another Gibs Sampler (JAGS) software (Plummer, 2003) was used in combination with the R package `rjags` (Plummer, 2016) for the Bayesian analysis. The maximum likelihood estimator (MLE) was obtained for the regression calibration and SIMEX analyses, whereas the posterior mean was obtained for the Bayesian analysis.

This simulation scenario was later modified so that the real parameter value for η was 0.14 per 100 Bq/m^3 in order to explore how the methods will compare if the true exposure effect was higher. The effect of the sample size was also studied by reducing the total number of subjects to half (from $n = 1000$ to $n = 500$). The true parameter value for this scenario was $\eta = 0.14$ per 100 Bq/m^3. The results from the three methods are compared in Table 3.1.

3.1.1 Regression Calibration

Using only the disease free subjects, the sample mean and variance were computed. That same subsample was used to fit the mixed model 3.3 in order to obtain the parameter estimates that will be used when implementing the regression calibration. The mixed model fitted was the following:

$$w_{ijk} \sim N(\beta_0 + \beta_1 \text{ROOM}_{ijk} + \gamma_{0i} + \gamma_{1j(i)}, \sigma^2_W) \tag{3.3}$$

where w_{ijk} is the log radon variable measured with error for subject i at home j with measurement k, β_1 is the parameter associated with the room (fixed) effect, γ_{0i} is a subject-specific random effect, and $\gamma_{1j(i)}$ is a random effect for each home within each individual. The "observed" values (w_{ijk}) were substituted by the corrected ones using the following equation:

$$E(\exp(x)|w) = \exp(\mu_{X|W} + 0.5\sigma^2_{X|W}) \tag{3.4}$$

where $x \sim N(\mu_X, \sigma^2)$ is the true log radon measurement, $w|x \sim N(x, \sigma^2_W)$ is the log observed radon measured with error, $x|w \sim N(\mu_{X|W}, \sigma^2_{X|W})$ is the conditional distribution, $\mu_{X|W} = \left(\frac{w}{\sigma^2_W} + \frac{\mu_X}{\sigma^2}\right)\left(\frac{1}{\sigma^2_W} + \frac{1}{\sigma^2}\right)^{-1}$, and $\sigma^2_{X|W} = \left(\frac{1}{\sigma^2_W} + \frac{1}{\sigma^2}\right)^{-1}$. After correcting the observed radon measurements, an average was computed in order to obtain one measurement per subject (w^c_i). The excess odds model 3.5 was fitted using those single corrected measurements for each subject. This excess odds model can be defined as follows:

$$\text{odds}_i = \exp(\alpha_0)(1 + \eta w^c_i) \tag{3.5}$$

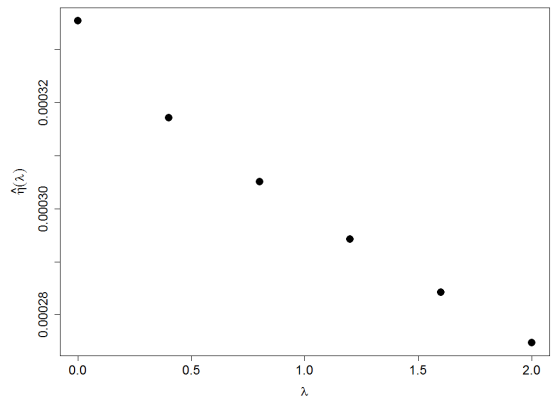
where η is the parameter associated with the single corrected radon measurement w^c_i, and α_0 is the intercept. The point estimate, as well as coverage, MSE, and bias can be seen in Table 3.1.

3.1.2 SIMEX

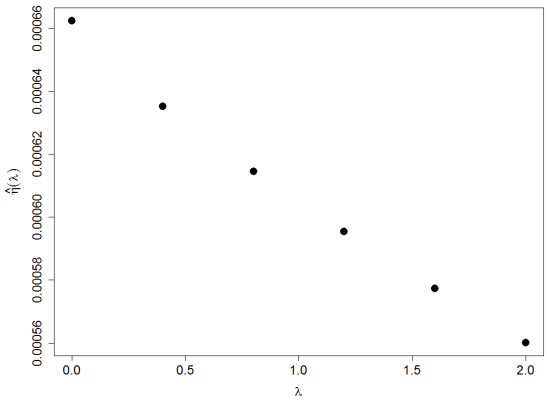
The SIMEX method was applied to the 1200 generated datasets in order to correct for measurement error. For each of those generated datasets, 20 contaminated datasets were created for each value of $\lambda = 0, 0.4, ..., 1.6, 2$. As the value of λ increased so did the amount of error added to each group of 20 new set of datasets. At each case, the error was added to the individual measurements w_{ijk} in the simulation part of the SIMEX method. The equation used to accomplished this was the following:

$$w_{mijk}(\lambda) = w_{ijk} + \lambda^{1/2} u_{mijk}, \quad \lambda \geq 0, \quad m = 1, ...20 \tag{3.6}$$

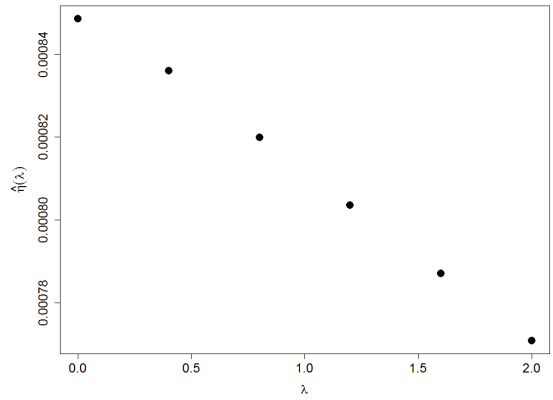
where $u \sim N(0, \sigma_U^2 = 0.06446)$ is the additional error added to each observation for each of the 20 contaminated datasets. After adding that extra error, the multiple observations by subject were exponentiated and then averaged, which lead to a single measurement per home in the original scale (Bq/m^3). Then, an average over the homes was computed so that a single measurement per subject was obtained (w_i^c). Using w_i^c, the risk model 3.5 was fitted for each contaminated dataset. An average of the 20 parameter estimates for each λ, $\hat{\eta}(\lambda)$, was computed and then used for the extrapolation part of the SIMEX algorithm. A simple linear regression (see Figures 3.1a, 3.1b, 3.1c) using λ and $\hat{\eta}(\lambda)$ was used to extrapolate to the case of no measurement error ($\lambda = -1$). A total of 1k bootstrap replicates were used to obtain the standard errors needed to compute the confidence intervals. The computing time needed to run this set of simulations was immense. To be more specific, it took almost 10 days for the 1200 generated datasets to be analyzed using this method on two Linux machine (Intel Xeon(R) CPU E3-1280 V2 @ 3.60GHz 8). Most of that computing time was due to the bootstrap part of the analysis. The results from this section of the simulations and running time for a single dataset can be found in Table 3.1.



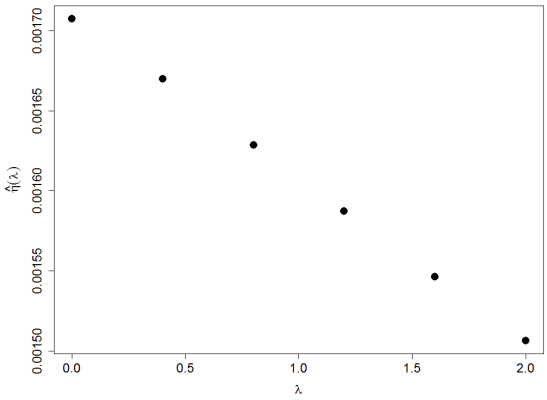
(a) $\eta = 0.05, n = 1\text{k}$



(b) $\eta = 0.14, n = 1\text{k}$



(c) $\eta = 0.14, n = 500$



(d) 2 sites

Figure 3.1: SIMEX plots for all simulation scenarios

3.1.3 Bayesian Model

Finally, the proposed Bayesian method was used to jointly correct for measurement error and assess the relationship between the risk factors and disease outcome for each of the 1200 generated datasets. In contrast to the frequentist methods used in the previous subsections, this method includes all individuals (cases and controls) from the beginning. The first part of this method is the error model which was specified as:

$$\begin{aligned} w_{ijk} &\sim N\left(\mu_{ijk}, \frac{1}{\tau_W}\right) \\ \mu_{ijk} &= \beta_0 + \beta_1 \text{STATUS}_i + \beta_2 \text{ROOM}_{ijk} + \gamma_{0i} + \gamma_{1j(i)} \\ \beta_p &\sim N(0, 0.0001) \quad \text{for } p = 0, 1, 2 \\ \gamma_{0i} &\sim N\left(0, \frac{1}{\tau_{\gamma_0}}\right) \\ \gamma_{1j(i)} &\sim N\left(0, \frac{1}{\tau_{\gamma_1}}\right) \\ \tau_g &\sim Gamma(0.001, 0.001) \quad \text{for } g = \gamma_0, \gamma_1, w \end{aligned} \tag{3.7}$$

where β_0 is the intercept, β_1 and β_2 are the parameters associated with the fixed effects (disease status and room, respectively), γ_{0i} is a subject-specific random effect, $\gamma_{1j(i)}$ is the random effect for home j within individual i, the variable "STATUS" was coded as 0 for disease free subjects and as 1 for the subjects with cancer, and the variable "ROOM" was coded as 0 for "ROOM 1" and 0 for "ROOM 2". The β parameters had vague Normal prior distributions, whereas the γ parameters had Normal distributions centered around 0 with τ hyperparameters for the precisions. The hyperpriors used for the precisions were Gamma distributions with shape and scale parameter equal to 0.001. Therefore, τ_W, τ_{γ_0} and τ_{γ_1} are the precisions for the distributions of w, γ_{0j}, and $\gamma_{1j(i)}$, respectively. The next step was to obtain corrected measurements by using the parameters from Model 3.7. The corrected

measurements k from each home j for subject i were computed as follows

$$w_{ijk}^{c} = \begin{cases} \exp\big(\beta_0 + \beta_1 \text{STATUS}_i + \gamma_{0i} + \gamma_{1j(i)}\big) & \text{for room 1} \\ \exp\big(\beta_0 + \beta_1 \text{STATUS}_i + \beta_2 + \gamma_{0i} + \gamma_{1j(i)}\big) & \text{for room 2} \end{cases} \tag{3.8}$$

These corrected measurements were then averaged in order to obtain a corrected measurement per subject for each of the two homes. Then, an average of those two corrected measurements was computed to obtain a single measurement per subject. This resulting single corrected measurement was used in the following risk model:

$$\begin{aligned} \text{STATUS}_i &\sim Bernoulli(\pi_i) \\ \pi_i &= \frac{\text{odds}_i}{1 + \text{odds}_i} \\ \text{odds}_i &= \exp(\alpha_0) \times (1 + \eta w_i^c) \\ \alpha_0 &\sim N(0, 0.0001) \\ \eta &\sim N(0, 0.0001) \prod_{i=1}^{n} I_{[(1+\eta w_i^c)>0]} \end{aligned} \tag{3.9}$$

where π_i is the probability of developing lung cancer, α_0 is the intercept, and η is the parameter associated with the corrected radon exposure effect. The joint error and risk model required a burn-in of 5k iterations and an additional 15k iterations for the inferences. The posterior mean was obtained (Table 3.1). It took approximately a week to obtain all the results from this simulation, using two Linux machines and seven out of the eight cores running in parallel (Intel Xeon(R) CPU E3-1280 V2 @ 3.60GHz). The running time for a single dataset analysis can be found on Table 3.1.

3.1.4 Comparison

Bayesian models that had MCMC error greater than 5% of the posterior mean, or that had Gelman and Rubin estimates greater than 1.2 were classified as models that did not converge or did not generate enough iterations from the posterior

distribution, and therefore were excluded from the analysis. In order to be consistent, only samples where the Bayesian model converged were used in the regression calibration, SIMEX and the Bayesian analysis. A total of 1k samples and their respective analyses were used for the method comparison (200 were excluded). In general, a limitation for the SIMEX algorithm is the computational time needed to obtain the standard errors for the confidence intervals, whereas regression calibration was the quickest to make the correction. The results are shown in Table 3.1.

The first scenario had a total of 1k subjects (500 cases and 500 controls) and a true parameter value of $\eta = 0.05$ per 100 Bq/m^3. If we focus on the point estimate, the SIMEX algorithm had the smallest bias and thus its parameter estimate is the closest to the real parameter values. The regression calibration method was the one with the larger bias and MSE. Moreover, the regression calibration method seems to be overestimating the results. The proposed Bayesian method is barely underestimating coverage at 93% and has the parameter estimate and bias values in between the respective results for the other two methods.

The results when the real parameter value was $\eta = 0.14$ per 100 Bq/m^3 and the sample size remained with a 1k subjects, showed that the SIMEX method was the method that did the best correction; its parameter estimate was the closest to the real value, the coverage was the same as the nominal coverage of 95%, and the MSE and bias are the lowest. The regression calibration method had the highest bias and MSE, whereas the Bayesian method had a similar coverage as when the true value was $\eta = 0.05$ per 100 Bq/m^3.

If we keep the true effect as $\eta = 0.14$ per 100 Bq/m^3 but reduce the sample size to half (250 cases and 250 controls), then we can see that the three methods perform well in terms of coverage. Similar to the first scenario, the results for the point estimate, bias and MSE of the Bayesian method are in between the respective

Table 3.1: Simulation results per 100 Bq/m^3

Scenario	Method[a]	$\hat{\eta}$	Coverage[b]	Bias	MSE	Time[c]
$\eta = 0.05$, $n = 1000$	Regression Calibration	0.13	0.997	0.08	0.0295	0.92
	SIMEX	0.06	0.982	0.01	0.0052	1766.43
	Bayesian (post. median)	0.08	0.928	0.03	0.0075	2610.87
	Bayesian (post. mean)	0.09	0.928	0.04	0.0087	2610.87
$\eta = 0.14$, $n = 1000$	Regression Calibration	0.31	0.988	0.17	0.0966	0.81
	SIMEX	0.15	0.951	0.01	0.0115	1797.63
	Bayesian (post. median)	0.18	0.931	0.04	0.0171	2503.17
	Bayesian (post. mean)	0.19	0.931	0.05	0.0199	2503.17
$\eta = 0.14$, $n = 500$	Regression Calibration	0.37	0.942	0.23	0.2923	0.41
	SIMEX	0.16	0.945	0.02	0.0268	1190.93
	Bayesian (post. median)	0.22	0.939	0.08	0.0444	1052.10
	Bayesian (post. mean)	0.25	0.939	0.11	0.0528	1052.10
$\eta = 0.14$, $n = 1000$, 2 studies	Regression Calibration	0.38	0.983	0.24	0.2988	0.98
	SIMEX	0.16	0.939	0.02	0.0231	2342.28
	Bayesian (post. median)	0.21	0.933	0.07	0.0386	2248.74
	Bayesian (post. mean)	0.24	0.933	0.10	0.0447	2248.74

[a] The Bayesian results are presented first using the posterior median and then using the posterior mean.
[b] Coverage based on the 95% confidence interval for regression calibration and SIMEX, and 95% credible interval for the Bayesian method.
[c] Running time in seconds for a single dataset using a PC with Intel(R) Core(TM) i5-3570 CPU @ 3.40GHz

results from the frequentist methods, and the respective results from the regression calibration correction are the largest.

Overall we can see that the Bayesian methods' results in terms of coverage, parameter estimate and MSE were consistent regardless of the sample size or true

exposure effect. The parameter estimates for this method remained bigger than the true parameters regardless of the scenario. The MSE and bias remained small, and the coverage was never more than 3% below the 95% nominal coverage. The regression calibration results had the largest MSE and bias at each scenario, and the coverage was close to the nominal coverage only once. The SIMEX correction method seems to work well for most of the cases but it was the one that consistently took the longest computer time to run. A summary of this results can be found on Table 3.2.

3.2 Combined Studies Scenario

A scenario that is more consistent with the application presented in this dissertation was also recreated. For this case, data from two studies were generated, a correction was made for each study individually, and then the corrected measurements were pooled and used when fitting a risk model. As in the previous scenario, each individual is assumed to have lived in two different homes and the measurements were taken in two rooms from each home. The true radon concentrations were generated from Model 3.1. For individuals in study 1, the parameters were set as follows: $\beta_0 = 5.1443$, $\beta_1 = -0.6040$, $\sigma_0^2 = 0.5607$ and $\sigma_1^2 = 0.01938$. The true radon concentrations for study 2 were created using these values for the parameters: $\beta_0 = 4.2467$, $\beta_1 = -0.4984$, $\sigma_0^2 = 0.09352$ and $\sigma_1^2 = 0.4565$. Measurement error was added to each observation by sampling from $N(0, \sigma_g^2)$ for $g = 1, 2$, where $\sigma_g^2 = 0.06279$ for study 1 and $\sigma_g^2 = 0.09204$ for study 2. Once the four true and observed radon measurements per subject were created, the true exposure was exponentiated to simulate the original radon scale (Bq/m^3). The measurements were averaged over home and then over subject which resulted in a single measurement per subject (x_i^a). These single measurements were used in the following equation in

order to determine the disease probability of the subject.

$$p(\pi_i = 1) = \frac{\exp(\alpha_0 + \alpha_1 \text{STUDY}_i) \times (1 + \eta x_i^a)}{1 + \big(\exp(\alpha_0 + \alpha_1 \text{STUDY}_i) \times (1 + \eta x_i^a)\big)} \tag{3.10}$$

where $\alpha_0 = -0.8156$, $\alpha_1 = 0.3836$, $\eta = 0.14$ per 100 Bq/m^3, and STUDY is an indicator variable equal to 1 for "STUDY 1" and equal to 0 for "STUDY 2". These parameter values were chosen based on an analysis of observed radon concentrations. The disease status for each subject was obtained by using a Binomial distribution with the probabilities computed earlier. We generate enough observations so that each study could have 250 cases and 250 controls for a total of 1k subjects coming from two different studies. There were 1400 datasets generated with these characteristics. For each dataset, regression calibration, SIMEX and the proposed Bayesian methods were used to correct for measurement error. The resulting MLE, posterior mean, MSE, and bias were then used to compare the methods.

3.2.1 Regression Calibration

The parameter estimates needed to apply the regression calibration correction were computed using only the controls for each study. For this purpose, the mixed model 3.3, the sample mean and variance were used. After obtaining the parameter estimates for each study, Equation 3.4 was used to correct the measurements for all individuals within each study. The corrected measurements for each subject were then averaged over home and then over individuals, and a single corrected measurement was obtained for each subject. Then, the following risk model was fitted

$$\text{odds}_i = \exp(\alpha_0 + \alpha_1 \text{STUDY}_i)(1 + \eta w_i^c) \tag{3.11}$$

where α_0 and α_1 are the intercept and parameter associated with the study effect (respectively), and η is the parameter associated with the radon exposure effect.

The MLE for η, coverage, bias, and MSE resulting from this part of the simulation are presented in Table 3.1.

3.2.2 SIMEX

In order to implement the SIMEX method for this scenario, λ was set as a sequence from 0 to 2 in increments of 0.4, and the number of contaminated datasets for each value of λ was set to 20. Each contaminated observation was computed as follows

$$w_{mijk}(\lambda) = w_{ijk} + \lambda^{1/2} u_{mijk}, \quad \lambda \geq 0, \quad m = 1, ...20 \tag{3.12}$$

where $U \sim N(0, \sigma_U^2 = 0.06446)$ for individuals in study 1, and $U \sim N(0, \sigma_U^2 = 0.09110)$ for individuals in study 2 (parameter values based on previous analyses of observed radon concentrations). Once that extra error was added to the dataset, a single measurement per subject in the original scale (Bq/m^3) was calculated. That single corrected measurement was then used in model 3.11 to obtain the parameter estimates for η. This process was repeated for each of the 20 contaminated datasets at each level of λ. The average of those 20 parameter estimates was then computed and $\hat{\eta}(\lambda)$ was obtained for each λ. A simple linear regression (see Figure 3.1d) using λ and $\hat{\eta}(\lambda)$ was then fitted in order to extrapolate to the case of $\lambda = -1$, i.e., no measurement error. One thousand bootstrap replicates were used to compute the standard errors needed to construct confidence intervals (Table 3.1).

3.2.3 Bayesian

The proposed Bayesian model was fitted using all available data from both simulated studies. The model jointly corrects for measurement error for each study

individually, computes a single corrected measurement for each individual, and combines the data used to specify the risk model. The error model used was the following:

$$\begin{aligned}
w^l_{ijk} &\sim N\left(\mu^l_{ijk}, \frac{1}{\tau^l_w}\right) \quad \text{for } l = \text{Study 1 or Study 2} \\
\mu^l_{ijk} &= \beta^l_0 + \beta^l_1 \text{STATUS}_j + \beta^l_2 \text{ROOM}_{ijk} + \gamma^l_{0i} + \gamma^l_{1j(i)}, \\
\beta_p &\sim N(0, 0.0001) \text{ for } p = 0, 1, 2 \\
\gamma^l_{0j} &\sim N\left(0, \frac{1}{\tau^l_0}\right) \\
\gamma^l_{1j(i)} &\sim N\left(0, \frac{1}{\tau^l_1}\right) \\
\tau^l_g &\sim Gamma(0.001, 0.001) \quad \text{for } g = 0, 1, w
\end{aligned} \tag{3.13}$$

From the error model, the study-specific parameters were used to correct the variable measured with error, W, as in Equation 3.8. The corrected measures were averaged and the resulting single corrected measurement was used in the following risk model.

$$\begin{aligned}
\text{STATUS}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \frac{\text{odds}_i}{1 + \text{odds}_i} \\
\text{odds}_i &= \exp(\alpha^l_0) \times (1 + \eta w^c_i) \quad \text{for } l = \text{Study1, Study2} \\
\alpha_0 &\sim N(0, 0.0001) \\
\eta &\sim N(0, 0.0001) \prod_{i=1}^{n} I_{[(1+\eta w^c_i)>0]}
\end{aligned} \tag{3.14}$$

This joint error and risk model was fitted for each of the 1400 generated datasets. The resulting posterior means were computed. The bias, coverage, and MSE were also obtain for this method (Table 3.1).

3.2.4 Comparison

It appears that when the exposure effect was $\eta = 0.14$ per 100 Bq/m^3 and there was another study added, the SIMEX correction outperforms the other two methods. The bias and MSE were the lowest, and the coverage was the closest

Table 3.2: Simulation summary

Method	RC	SIMEX	Bayesian
Small bias		X	X
Quick implementation	X		
Prior information			X
Joint model			X
Small variance		X	X

to the nominal coverage of 95%. The Bayesian method was not too bad, with a coverage of 93% and, bias and MSE values smaller than the ones for the regression calibration method. A summary of the overall comparison of the three methods can be found on Table 3.2.

CHAPTER 4
DATA ANALYSIS

This chapter presents the application of the methods presented in Chapter 2 and compared in Chapter 3. The data used for this implementation were described in Chapter 1. As a review, data from five case-control North America studies (Iowa, Missouri, Winnipeg, Connecticut and Utah/South Idaho) were used. Indoor radon exposure is the variable measured with error and lung cancer status the outcome. Only radon measurements taken with alpha-track detectors were used in the analysis. Also, following the work of Krewski et al. (2006), individuals with missing smoking or home information were excluded from the final analyses (Table 4.1).

In general, frequentist and Bayesian mixed models were fitted depending on the available data for each study. Then, regression calibration, SIMEX and the proposed Bayesian method were used to obtain parameter estimates and confidence/credible intervals for the effect of indoor radon exposure while trying to correct for measurement error. Since there is a latency period for radiogenic cancer with a minimum of 5 years, the first five years prior to disease diagnosis or prior interview were excluded from the analyses (Council, 1988). Therefore, a 6-30 year window was used to determine which measurements were kept when fitting the excess odds model 4.1 (Krewski et al., 2006). This window also allowed us to put weight on each home measurement. That is, if a subject lived in a house for 10 years out of the 25 in this window, then that home measurement will have more weight than a home in which the individual lived for only a year. Hence, the measurements used in the risk model were calculated based on an exposure time window (ETW) of the radon measurements.

Table 4.1: Inclusion/exclusion criteria

			Frequency of exclusion							
	Original sample size[a]		No smoking data		No home data[b]		Removed for other reasons[c]		Final sample size	
Study	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Iowa	413	614							413	614
Missouri	618	1402	7	1	84	232			527	1169
Winnipeg	738	738	9	4	20	13	32	8	677	713
Utah/South Idaho	511	862					15	13	496	849
Connecticut	963	949					68	34	895	915

[a] Based on the available demographic information and annual radon concentration

[b] No home data for the 6-30 year window.

[c] The datasets for those studies had a "flag" variable indicating which subjects should be removed. For Winnipeg, that variable indicated that the radon exposure was less than 300 days. In the case of Connecticut and Utah/South Idaho, the flag variable indicated duplicates that should be removed. Moreover, there were individuals with demographic information but no raw radon measurements

The data analysis started by using the MIXED procedure in the SAS software (SAS Institute Inc, 2012) to fit a mixed model and obtain parameter estimates for the implementation of regression calibration and SIMEX. The fixed effects estimates, along with the radon exposure measurements available, were used for the imputation process when dealing with missing values. After the imputation process was completed, the results from the mixed model, the sample mean and sample variance were used for the regression calibration correction using Equation 3.4. The raw measurements were substituted by those corrected ones. Then, a single measurement per subject for each of the available homes was computed by taking the average of the corrected measurements. For the SIMEX application, we used R (R Core Team, 2016) to generate 20 new datasets with additional measurement error for each value of $\lambda = 0, 0.4, ..., 1.6, 2$. Also, 1k bootstrap replicates were computed to obtain the standard errors needed to construct the confidence intervals.

Only homes in which the subject lived 6-30 years prior to the interview or lung cancer diagnosis were included in the analysis. Note that for Iowa this was not necessary since the study only included individuals who lived in the same residence for the past 20 consecutive years. An average over all remaining homes was taken in order to compute a single corrected measurement per subject. This measurement was used in an excess odds model to assess the relationship between the risk of developing lung cancer and indoor radon exposure. The excess odds model that was fitted for regression calibration and SIMEX is defined as follows

$$\text{odds}_i = \exp(\alpha^\top \boldsymbol{Z}_{2i})(1 + \eta \text{Bq}_i^c) \tag{4.1}$$

where η is the parameter associated with the single corrected radon exposure measurement for subject i (Bq_i^c), and $\alpha^\top$ are the intercept and parameters associated with the variables measured without error $\boldsymbol{Z}_{2i}$ that we might want to adjust for.

These covariates are sex, categories of age (under 50, 50-54, 55-59, 60-64, 65-69, 70-74, 75 years or more), smoking duration (nonsmoker, 1-24, 25-34, 35-44, 45 years or more) and smoking rate (nonsmoker, 0.5-9.5, 9.5-19.5, 19.5-29.5, 29.5 or more averaged cigarettes smoked per day during the active smoking period). Frequencies for these variables by site are shown in Table 4.2. According to Carroll et al. (2006), the parameters from the regression calibration in one study might not be transportable to another. Hence, for the pooled analysis, we had a study-specific correction process and then combined all corrected measurements before fitting the risk model. In this case, a variable for the study effect was added to the risk models.

Additionally, a Bayesian hierarchical model was fitted in order to jointly fit the measurement error model and the excess odds model. That is, the model will simultaneously correct the variable measured with error (indoor radon exposure), while exploring the association between that risk factor and the probability of developing lung cancer. The model is divided into two parts, the first uses the available information in each study dataset in order to correct individual measurements. For this, fixed and random effects were incorporated into the model. While the frequentist methods used imputation when missing data was present, this Bayesian method assigns a random effect for each missing observation. All parameters from the mixed model were used to correct the multiple measurements per individual. Those corrected measurements were transformed back to the original scale, Becquerel per cubic meter (Bq/m^3), and averaged to obtain a single measurement per subject per home. The measurements for the homes that were part of the 6-30 year window were then averaged and an exposure time window (ETW) measurement per subject was obtained. The excess odds model was specified using those single

corrected measurements as follow:

$$
\begin{aligned}
\text{STATUS}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \frac{\text{odds}_i}{1+\text{odds}_i} \\
\text{odds}_i &= \exp(\alpha^\top \boldsymbol{Z}_{2i}) \times (1+\eta \text{Bq}_i^c) \\
\alpha_g &\sim N(0, 0.0001) \quad \text{or} \\
\alpha_g &\sim Unif(a,b) \quad \text{for } g = 0,1,...,G \\
\eta &\sim N(0, 0.0001) \prod_{i=1}^{n} I_{[(1+\eta \text{Bq}_i^c)>0]} \quad \text{or} \\
\eta &\sim Unif(-2,2) \prod_{i=1}^{n} I_{[(1+\eta \text{Bq}_i^c)>0]}
\end{aligned}
\tag{4.2}
$$

where π_i is the probability of developing lung cancer, $\alpha^\top$ are the intercept and parameters associated with the covariates measured without error (Table 4.2), and η is the parameter associated with the corrected radon exposure (could have an Uniform or Normal prior depending on the study). As in the frequentist risk model, the covariates used for the adjusted model were sex, and categories of age (under 50, 50-54, 55-59, 60-64, 65-69, 70-74, 75 years or more), smoking duration (nonsmoker, 1-24, 25-34, 35-44, 45 years or more) and smoking rate (nonsmoker, 0.5-9.5, 9.5-19.5, 19.5-29.5, 29.5 or more averaged cigarettes smoked per day during the active smoking period). The Bayesian data analysis was performed using the Just Another Gibbs Sampler (JAGS) software (Plummer, 2003) with the R package `rjags` (Plummer, 2016). The `coda` package (Plummer et al., 2006) was used to assess the convergence of the MCMC chains and perform model diagnosis. The burn-in period and the number of iterations that were kept for the inferences varied between models. For all the models fitted, trace plots and Gelman and Rubin diagnosis (Gelman and Rubin, 1992) were used to assess model convergence. The posterior mean as well as the 95% credible interval were obtained. Goodness-of-fit test were performed for model diagnosis purposes. Since the resulting plots were similar for

all studies, only the Iowa one is presented. The Bayesian p-values for these tests were computed and the concordance index was obtained to assess the performance of the risk model.

Table 4.2: Outcome variable and covariates included in some models

Variable	Iowa	Missouri	Winnipeg	Utah/South Idaho	Connecticut
Status					
Cases	413	527	677	496	895
Controls	614	1169	713	849	915
Sex					
Male	-	-	917	891	926
Female	1027	1696	473	454	884
Age[a]					
under 50	23	107	112	84	179
50-54	59	30	138	114	171
55-59	117	146	213	173	234
60-64	164	145	271	249	346
65-69	216	267	306	297	394
70-74	207	284	258	262	285
75+	241	717	92	166	201
Smoking duration[a]					
non-smoker	470	1428	266	253	102
1-24	91	163	210	190	246
25-34	82	68	228	166	369
35-44	187	31	327	306	494

[a] in years
[b] averaged cigarettes smoked per day during the active smoking period

Table 4.2: Outcome variable and covariates included in some models

Variable	Iowa	Missouri	Winnipeg	Utah/South Idaho	Connecticut
45+	197	6	359	430	599
Smoking rate[b]					
non-smoker	470	1428	266	253	102
0.5-9.5	113	61	132	97	195
9.5-19.5	259	89	357	465	680
19.5-29.5	113	77	458	350	484
29.5+	72	41	177	180	349

[a] in years
[b] averaged cigarettes smoked per day during the active smoking period

After correcting the raw data using the three methods, we fitted risk models using the average annual radon concentration in the living area. This is the data as used by Krewski et al. (2006). Moreover, using the raw/uncorrected measurements, an ETW was calculated and excess odds models were fitted using the living area concentrations from the Krewski et al. (2006) analysis. The results from those models were compared to the results obtained from the risk models using all the correction methods mentioned before.

4.1 Iowa

Information regarding radon exposure measurements in the home levels and rooms, and at different time periods was used to analyze the data from the Iowa study. Remember that one of the inclusion criteria for this study stated that individuals must be females and must have lived in the same residence for at least 20 consecutive years. Since this dataset only includes females and one home, the sex

variable and the 6-30 year time window were not necessary when fitting the risk models, and the random effects γ_i could be interpreted as subject- or home-specific random effect.

4.1.1 Regression Calibration

Following the regression calibration algorithm, the mixed model 4.3 was fitted using only the information from lung cancer free patients. In this model, lnBq_{ijk} is the natural logarithm of the observed radon measurement k for subject i at time period j in Becquerel per cubic meter (Bq/m^3), β_0 is the intercept, β_1 and β_2 are fixed effects for the home levels one and two, respectively (coded as indicator variables and with level 0 as the reference), γ_{0i} is a random effect for each home, and $\gamma_{j(i)}$ is a random effect for each time period within each individual. That is, if two measurements for the same subject were taken over six months apart, then that new measurement was part of a new time period. This will take into account any possible time variation within the same home.

$$\text{lnBq}_{ijk} \sim N(\beta_0 + \beta_1 \text{LEVEL1}_{ijk} + \beta_2 \text{LEVEL2}_{ijk} + \gamma_{0i} + \gamma_{1j(i)}, \sigma^2_{\text{lnBq}}) \qquad (4.3)$$

The fixed effects' results obtained from the linear mixed model 4.3 were used to impute values when a measurement was missing for a given level in a given year. Other results from the mixed model (Table 4.3), and the controls' sample mean and variance were used with Equation 3.4 to compute the corrected measurements. Only the corrected measurements from levels where bedroom or living room measurements were taken, were kept and used in the remaining part of the analysis. Then, the mean of the corrected measurements over a given time period was computed, followed by the mean over each subject. This process allowed us to have a single corrected measurement per subject. The means for this single radon measurement per subject by disease status are presented in Table 4.5 and a box plot of

Table 4.3: Results obtained from the mixed model

Study	$\hat{\mu_X}$	$\hat{\sigma}^2_W$	$\hat{\sigma}^2$	σ^2_U
Iowa	4.66	0.74	0.66	0.06446
Missouri	3.76	0.64	0.55	0.09204
Winnipeg	4.79	0.66	0.59	0.07751
Connecticut	3.21	1.24	1.04	0.20410
Utah/South Idaho	3.85	0.71	0.62	0.09310

the distribution can be seen in Figure 4.4. Finally, the risk model was fitted with and without adjustment by covariates using the corrected measurements. Table 4.4 shows the results for these models. The models were adjusted by categories of age, smoking duration and smoking rate.

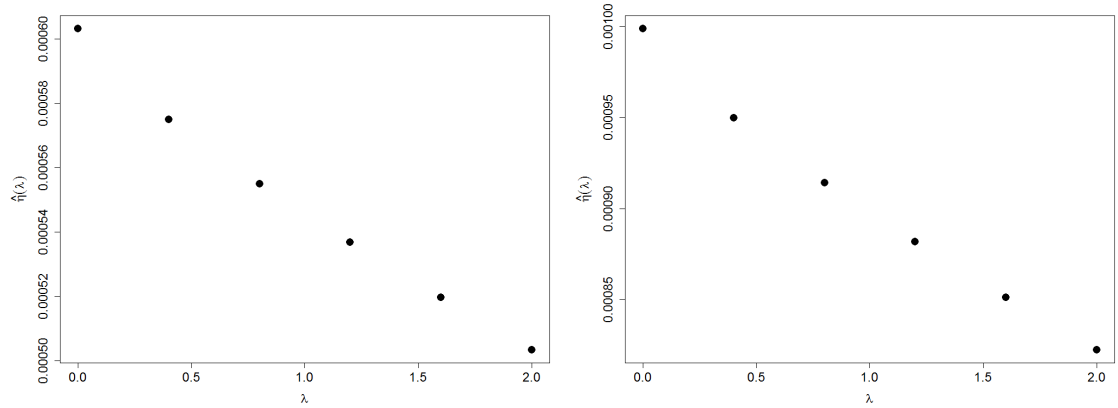
4.1.2 SIMEX

In order to implement the SIMEX method, the results from the mixed model 4.3 were used. The imputation process was the same as the one used for regression calibration. Once the missing data problem was solved, 20 contaminated datasets for each value of $\lambda = 0, 0.4, ..., 1.6, 2$ were created using the following equation

$$\text{lnBq}_{mijk}(\lambda) = \text{lnBq}_{ijk} + \lambda^{1/2} u_{mijk}, \quad \lambda \geq 0, \quad m = 1, ..., 20 \tag{4.4}$$

where $U \sim N(0, \sigma^2_U = 0.06446)$ is the additional error added to each of the observations in the 20 datasets for a specific value of λ. The value for σ^2_U was selected as the estimated residual error from the mixed model. Note that, each contaminated dataset had the same structure as the original data. Therefore, in order to obtain a single measurement per subject, average over the floors that had a bedroom or a living room detector was computed. Then, for each dataset, the naive parameter

estimate for the radon effect was computed using model 4.1. An average of the 20 naive parameter estimates for each λ was computed and one $\lambda(\hat{\eta})$ was obtained for each specific λ. These averaged estimates were then used in a simple linear model in order to extrapolate to the case with no measurement error, i.e. $\lambda = -1$ (Figure 4.1a). Bootstrap was used to obtain the standard errors that were use to compute the confidence intervals. The R package `boot` (Canty and Ripley, 2016; Davison and Hinkley, 1997) was used for this purpose. The resulting parameter estimate and confidence interval for the risk model are shown in Table 4.4.



(a) model with no adjustment (b) model with adjustment

Figure 4.1: SIMEX plots for the Iowa models

The above process used to create the contaminated datasets was repeated and models adjusted by covariates were fitted using those contaminated datasets. Dummy variables for the categories of the age, smoking rate and smoking duration variables were created and used in the model. The relationship between the λ's and the parameter estimates obtained from case, seemed to be linear (Figure 4.1b). Therefore, a simple linear regression was fitted to extrapolate the results to the case where there is no measurement error in the model adjusted by covariates. Results for this model can also be found in Table 4.4.

Table 4.4: Excess odds model results

	Unadjusted[a]		Adjusted by covariates[b]	
Study	$\hat{\eta}$	CI/CrI[c]	$\hat{\eta}$	CI/CrI[c]
Iowa				
Annual concentration	0.141	(0.001, 0.391)	0.253	(0.016, 0.773)
Uncorrected	0.046	(-0.049, 0.225)	0.082	(-0.054, 0.394)
Regression Calibration	0.057	(-0.058, 0.273)	0.122	(-0.057, 0.533)
SIMEX	0.065	(-0.114, 0.244)	0.108	(-1.736, 1.738)
Bayesian[d]	0.105 (0.093)	(-0.022, 0.299)	0.194 (0.164)	(-0.015, 0.576)
Missouri				
Annual concentration	-0.049	(-0.162, 0.200)	-0.056	(-0.162, 0.192)
Uncorrected	-0.050	(-0.163, 0.296)	-0.077	(-0.163, 0.254)
Regression Calibration	-0.059	(-0.236, 0.374)	-0.091	(-0.236, 0.327)
SIMEX	-0.071	(-0.416, 0.272)	-0.098	(-0.394, 0.201)
Bayesian[d]	0.024 (-0.010)	(-0.188, 0.434)	-0.005 (-0.038)	(-0.197, 0.377)

[a] per 100 Bq/m^3
[b] Models adjusted by gender, and categories of age, smoking duration and smoking rate.
[c] Confidence intervals (CI) for the frequentist methods and credible intervals (CrI) for the Bayesian method.
[d] Posterior mean (posterior median)

Table 4.4: Excess odds model results (continued)

	Unadjusted[a]		Adjusted by covariates[b]	
Study	$\hat{\eta}$	CI/CrI[c]	$\hat{\eta}$	CI/CrI[c]
Winnipeg				
Annual concentration	-0.029	(-0.049, 0.026)	-0.023	(-0.049, 0.057)
Uncorrected	-0.032	(-0.047, 0.011)	-0.025	(-0.046, 0.041)
Regression Calibration	-0.046	(-0.063, 0.010)	-0.040	(-0.063, 0.039)
SIMEX	-0.035	(-0.077, 0.006)	-0.031	(-0.444, 0.383)
Bayesian[d]	-0.033 (-0.036)	(-0.065, 0.019)	-0.025 (-0.030)	(-0.064, 0.044)
Connecticut				
Annual concentration	-0.043	(-0.198, 0.244)	-0.024	(-0.195, 0.304)
Uncorrected	-0.045	(-0.116, 0.105)	-0.046	(-0.117, 0.113)
Regression Calibration	-0.048	(-0.217, 0.254)	-0.088	(-0.244, 0.209)
SIMEX	-0.020	(-0.203, 0.136)	-0.056	(-0.348, 0.207)
Bayesian[d]	-0.050 (-0.064)	(-0.174, 0.157)	-0.051 (-0.067)	(-0.178, 0.170)

[a] per 100 Bq/m^3
[b] Models adjusted by gender, and categories of age, smoking duration and smoking rate.
[c] Confidence intervals (CI) for the frequentist methods and credible intervals (CrI) for the Bayesian method.
[d] Posterior mean (posterior median)

Table 4.4: Excess odds model results (continued)

	Unadjusted[a]		Adjusted by covariates[b]	
Study	$\hat{\eta}$	CI/CrI[c]	$\hat{\eta}$	CI/CrI[c]
Utah/South Idaho				
Annual concentration	-0.123	(-0.191, 0.130)	-0.064	(-0.191, 0.274)
Uncorrected	-0.049	(-0.153, 0.177)	0.029	(-0.137, 0.373)
Regression Calibration	-0.011	(-0.214, 0.482)	0.151	(-0.171, 0.977)
SIMEX	-0.022	(-0.565, 0.521)	0.093	(-2.215, 2.401)
Bayesian[d]	0.146	(-0.144, 0.733)	0.519 (0.096)	(-0.069, 1.911)
Combined				
Annual concentration	-0.010	(-0.044, 0.046)	-0.007	(-0.044, 0.059)
Uncorrected	-0.026	(-0.046, 0.014)	-0.020	(-0.045, 0.030)
Regression Calibration	-0.033	(-0.061, 0.027)	-0.030	(-0.061, 0.039)
SIMEX	-0.025	(-0.306, 0.255)	-0.022	(-0.371, 0.320)
Bayesian[d]	-0.019 (-0.022)	(-0.053, 0.030)	-0.017 (-0.021)	(-0.054, 0.039)

[a] per 100 Bq/m^3
[b] Models adjusted by gender, and categories of age, smoking duration and smoking rate.
[c] Confidence intervals (CI) for the frequentist methods and credible intervals (CrI) for the Bayesian method.
[d] Posterior mean (posterior median)

4.1.3 Bayesian

A Bayesian mixed effects model using all available data was fitted for this part of the analysis. In order to account for the time changes, a home-specific random effect is included in the model. As in the frequentist approaches, if the measurements for the same home were taken six or more months apart, that new measurement was considered in a different time period. Unlike the frequentist approaches, the random effects were also used when dealing with missing values. The measurement error model specified was the following:

$$\begin{aligned}
\text{lnBq}_{ijk} &\sim N\left(\mu_{ijk}, \frac{1}{\tau}\right) \\
\mu_{ijk} &= \beta_0 + \beta_1 \text{LEVEL1}_{ijk} + \beta_2 \text{LEVEL2}_{ijk} + \beta_3 \text{STATUS}_i + \gamma_{0i} + \gamma_{1j(i)} \\
\beta_p &\sim N(0, 0.0001) \text{ for } p = 0, 1, 2, 3 \\
\gamma_{0i} &\sim N\left(0, \frac{1}{\tau_0}\right) \\
\gamma_{1j(i)} &\sim N\left(0, \frac{1}{\tau_1}\right) \\
\tau &\sim Gamma(0.001, 0.001) \\
\tau_0 &\sim Gamma(0.001, 0.001) \\
\tau_1 &\sim Gamma(0.001, 0.001)
\end{aligned} \tag{4.5}$$

where β_0 is the overall intercept, β_1 and β_2 are the parameters associated with the effects of home levels compared to level 0 (reference), β_3 is the parameter associated with the disease status, γ_{0i} is the home-specific random effect, and $\gamma_{1j(i)}$ is the random effect for time period j within home i. While specifying the models, precisions were specified rather than variances. Therefore, τ, τ_0 and τ_1 are the precisions for the lnBq distribution and for the prior distributions for γ_{0i} and $\gamma_{1j(i)}$, respectively. The variables "LEVEL1", "LEVEL2", and "STATUS" are indicator variables coded as 1 if the measurement was from level 1, level 2 or from an individual with lung cancer, respectively, and coded as 0 otherwise.

Once again, only measurements from a level where a bedroom or living room

measurement was taken, were kept for the remaining part of the analysis. Therefore, using the parameters from model 4.5, Equation 4.6 (below) was applied to obtain a single corrected radon measurement per subject per time period.

$$\mathrm{Bq}_{ij}^{c} = \exp\big(\beta_0 + \beta_3 \mathrm{STATUS}_i + \gamma_{0i} + \gamma_{1i(j)}\big) \times \big(q_0 + q_1 \exp(\beta_1) + q_2 \exp(\beta_2)\big) \quad (4.6)$$

In the previous equation, q_0, q_1 and q_2 are indicator variables for the presence of detectors at a specified home level. That is, for LEVEL_k $(k = 0, 1, 2)$,

$$q_k = \begin{cases} 0, & \text{if } \mathrm{LEVEL}_k \text{ had no measurements in the bedroom or living room} \\ 0.5, & \text{if } \mathrm{LEVEL}_k \text{ had measurements, either in the bedroom or living room} \\ 1, & \text{if } \mathrm{LEVEL}_k \text{ had measurements in the bedroom and in the living room} \end{cases}$$

An average over time period was computed in order to obtain a single corrected radon exposure measurement per subject. Bayesian posterior summary statistics for the distributions of the corrected measurements are shown in Table 4.7. The posterior means of these corrected measurements were used as the corrected measurements per subject. The average of these measurements by disease status can be found in Table 4.5. A box plot showing the distribution of the corrected measurements' posterior means is in Figure 4.4. This information about indoor radon exposure, now corrected for measurement error, was used in the specification of the risk model without covariates as follow

$$\begin{aligned} \mathrm{STATUS}_i &\sim Bernoulli(\pi_i) \\ \pi_i &= \frac{\mathrm{odds}_i}{1 + \mathrm{odds}_i} \\ \mathrm{odds}_i &= \exp(\alpha_0) \times (1 + \eta \mathrm{Bq}_i^c) \\ \alpha_0 &\sim N(0, 0.0001) \\ \eta &\sim N(0, 0.0001) \prod_{i=1}^{n} I_{[(1+\eta \mathrm{Bq}_i^c) > 0]} \end{aligned} \quad (4.7)$$

The burn-in period for this joint model was 5k iterations and an additional 80k iterations were used for the inferences (Figure 4.3a). Figure 4.2 shows the goodness-of-fit test performed to assess the measurement error model diagnosis. The Bayesian

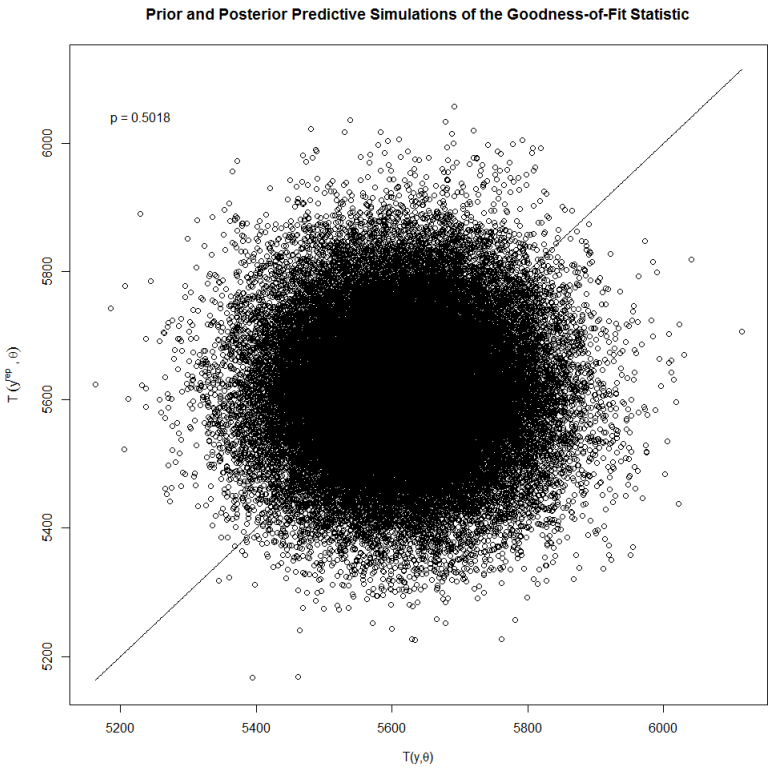


Figure 4.2: Goodness-of-fit test for the Iowa measurement error model

p-value for that test was $p = 0.5018$ which indicates good fit and the c-index for the risk model was 0.52 with a credible interval of (0.49, 0.51) which indicates that the model is not discriminating well (Table 4.6). For the model adjusted by covariates, dummy variables were created to indicate that an individual belongs to certain age, smoking duration and smoking rate category. The following risk model was specified

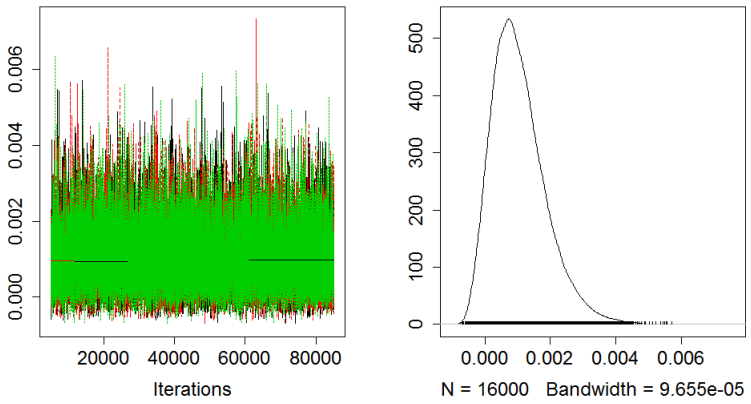
using those categories.

$$
\begin{aligned}
\text{STATUS}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \frac{\text{odds}_i}{1+\text{odds}_i} \\
\text{odds}_i &= \exp(\alpha_0 + \alpha_1 I_{[50\leq age_i\leq 54]} + \alpha_2 I_{[55\leq age_i\leq 59]} + \\
&\quad \alpha_3 I_{[60\leq age_i\leq 64]} + \alpha_4 I_{[65\leq age_i\leq 69]} + \alpha_5 I_{[70\leq age_i\leq 74]} + \\
&\quad \alpha_6 I_{[age_i\geq 75]} + \alpha_7 I_{[9.5<rate_i\leq 19.5]} + \alpha_8 I_{[19.5<rate_i\leq 29.5]} + \\
&\quad \alpha_9 I_{[rate_i>29.5]} + \alpha_{10} I_{[1\leq dur_i\leq 24]} + \alpha_{11} I_{[35\leq dur_i\leq 44]} + \\
&\quad \alpha_{12} I_{[dur_i\geq 45]}) \times (1+\eta \text{Bq}_i^c) \\
\alpha_g &\sim N(0, 0.0001) \quad \text{for } g = 0,2,4,7,8,9,10,11,12 \\
\alpha_g &\sim Unif(-2,2) \quad \text{for } g = 1,3,5,6 \\
\eta &\sim Unif(-2,2) \prod_{i=1}^{n} I_{[(1+\eta \text{Bq}_i^c)>0]}
\end{aligned}
\tag{4.8}
$$

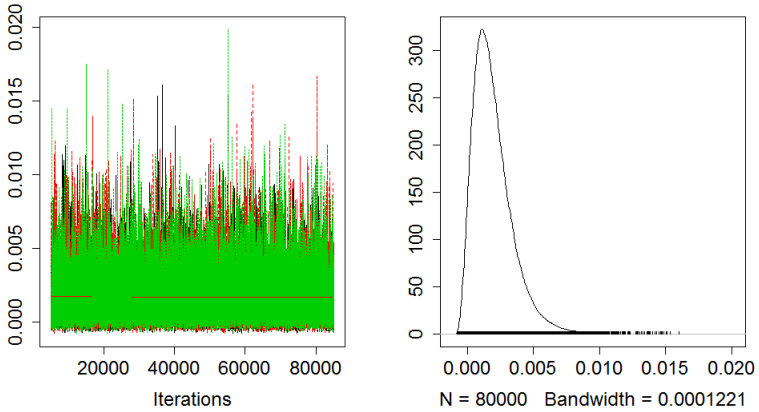
There were some α parameters that, even after convergence of the chains, were taking longer to reduce the MCMC error. For those parameters, the prior bounds were restricted with uniform priors. The same was applied to the η prior. The burn-in period and the number of iterations kept for inferences were also 5k and 80k, respectively (Figure 4.3b). The discrimination abilities of this model are better than the ones from the model without adjustment by covariates. The c-index for this model was 0.85 with a 95% credible interval of (0.85, 0.86). The results for the risk estimates and c-index can be found in Table 4.4 and 4.6, respectively.

4.1.4 Comparison

The regression calibration and Bayesian methods work by correcting individual observations, whereas the SIMEX method corrects the parameter estimate. Therefore, Figure 4.4 compares the box plots for the ETW of the observed measurements (no correction) and the ETW for the corrected measurements using regression



(a) model with no adjustment



(b) model with adjustment

Figure 4.3: Trace and density plots for the η node in the Iowa models

calibration and the Bayesian method, but there is no box plot for the SIMEX correction. For the uncorrected and regression calibration measurements, the average of the measurements for each individual are plotted, whereas for the Bayesian method we used the posterior mean for each subject's corrected measurement distribution. The box plot comparison shows that there is less variability when correcting the measurements using regression calibration. Other than that, the distributions look similar to each other. This is consistent with the means calculated by status and

presented in Table 4.5. In this table, the sample size for the annual radon concentration was smaller due to the fact that two subjects were missing that information. In the cases of the uncorrected ones, there were four subjects for whom observed measurements were not in the bedroom of living room and therefore were excluded for this calculations. These four subjects however, had radon exposure measurements in other rooms and those were used for the imputation and later for the correction.

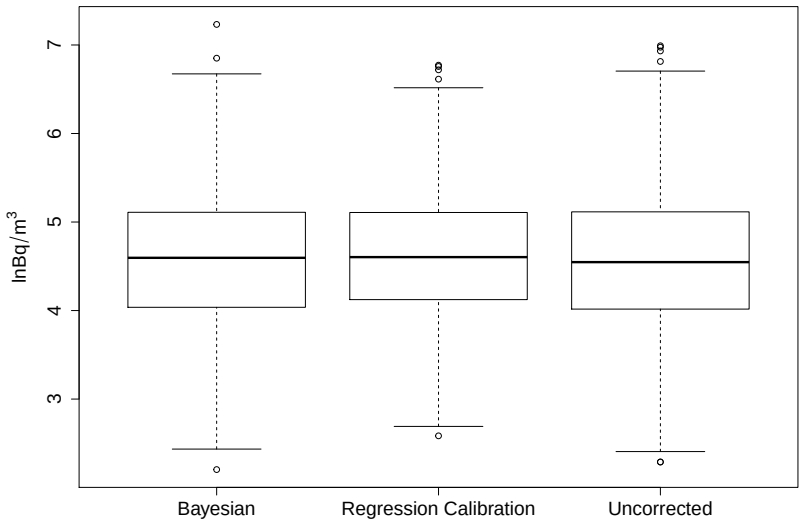


Figure 4.4: Distributions for the corrected and uncorrected measurements of lnBq in the Iowa sample

By analyzing the results from the risk models we can conclude the following. The point estimates from the model without adjustment by covariates were closer to zero (no effect) than the point estimates for the models adjusted by covariates. All confidence and credible intervals computed using the raw data include the value of 0 which implies a lack of statistical significance for the effect of indoor radon exposure on lung cancer risk. However, the confidence interval for the annual radon

concentration does show an effect. Thus, we can assume that the way in which the exposure is computed might play a role in the results of the analyses. If we compare the three methods used in this dissertation, the Bayesian correction leads to a higher point estimate, whereas the two frequentist methods have point estimates closer to the one obtained using the uncorrected measurements. The parameter estimates and c-index results showed that the adjustment by covariates was necessary.

Table 4.5: ETW of the corrected and uncorrected radon concentration[a] in Bq/m^3

Study	N	Cases	Controls	All subjects
Iowa[b]				
Annual concentration	1025	136.77	120.91	127.3
Uncorrected Measurements	1023	130.01	123.84	126.3
Regression Calibration	1027	128.38	123.05	125.2
Bayesian	1027	134.44	124.67	128.6
Missouri				
Annual concentration	1696	61.33	62.51	62.14
Uncorrected Measurements	1696	61.30	62	33.75
Regression Calibration	1696	56.44	56.99	56.82
Bayesian	1696	58.51	59.40	59.12
Winnipeg				
Annual concentration	1427	136.93	148.2	142.6
Uncorrected Measurements	1390	159.11	175.84	167.7
Regression Calibration	1390	150.94	163.87	157.6
Bayesian	1390	157.69	175.94	167.1

[a] One measurement per person.
[b] There were 4 individuals that had no measurements in the bedroom or living room.
[c] There were individuals with homes outside of the 6-30 year window. These subjects' measurements are not included in the uncorrected measurements.

Table 4.5: ETW of the corrected and uncorrected radon concentration[a] in Bq/m^3 (cont.)

Study	N	Cases	Controls	All subjects
Connecticut[c]				
Annual concentration	1912	32.07	32.7	32.38
Uncorrected Measurements	1564	46.33	48.93	47.67
Regression Calibration	1810	38.12	38.8	38.46
Bayesian	1810	45.14	48.09	46.63
Utah/South Idaho[c]				
Annual concentration	1373	55.42	58.03	57.06
Uncorrected Measurements	1153	65.10	67.19	66.45
Regression Calibration	1345	59.05	59.17	59.13
Bayesian	1345	68.43	67.02	67.54

[a] One measurement per person.
[b] There were 4 individuals that had no measurements in the bedroom or living room.
[c] There were individuals with homes outside of the 6-30 year window. These subjects' measurements are not included in the uncorrected measurements.

4.2 Missouri

The Missouri study provided information about radon exposure for multiple rooms (kitchen, bedroom, basement and other rooms) in multiple homes for almost all the individuals. Also, there were various types of detectors used to collect the data but only alpha-track detector measurements were used for the analyses. The information from this study was used to fit a mixed model and obtain the parameter estimates needed to make the correction to the multiple measurements per subject. According to the article written by Krewski et al. (2006), individuals missing smoking information as well as individuals missing home information for the 6-30 year

window should be removed. Therefore, we removed 8 individuals that had no smoking data and 316 individuals that had no home information (see Table 4.1). For this study, all participants were women and therefore, we did not adjust by sex in the risk models.

4.2.1 Regression Calibration

All available data for individuals that did not have lung cancer were used in order to fit a mixed model to get parameter estimates for the imputations when missing data were presented, for the regression calibration and SIMEX application. The mixed model used was

$$\text{lnBq}_{ijk} \sim (\beta_0 + \beta_1 \text{BA}_{ijk} + \beta_2 \text{BR}_{ijk} + \beta_3 \text{OT}_{ijk} + \gamma_{j(i)}, \sigma^2_{\text{lnBq}}) \tag{4.9}$$

where lnBq_{ijk} is the natural logarithm of the indoor radon exposure measurement k for subject i at home j in Becquerel per cubic meter (Bq/m^3), BA is an indicator variable coded as 1 if the measurement was taken from the basement and 0 otherwise, BR and OT are also indicator variables for bedroom and other, respectively (kitchen is the reference category), β_0 is the intercept, β_p (for $p = 1, 2, 3$) is a fixed effect for the room where the detector was placed (basement, bedroom and other, respectively), and $\gamma_{j(i)}$ is a random effect for each home within each individual. Using the results from the mixed model, imputation was used when missing data for a given room (kitchen or bedroom) were found. No imputation was done for the missing measurements for basement and other rooms in order to be consistent with the pooled paper by Krewski et al. (2006). Therefore, the remainder of the analysis was based only on the bedroom and kitchen.

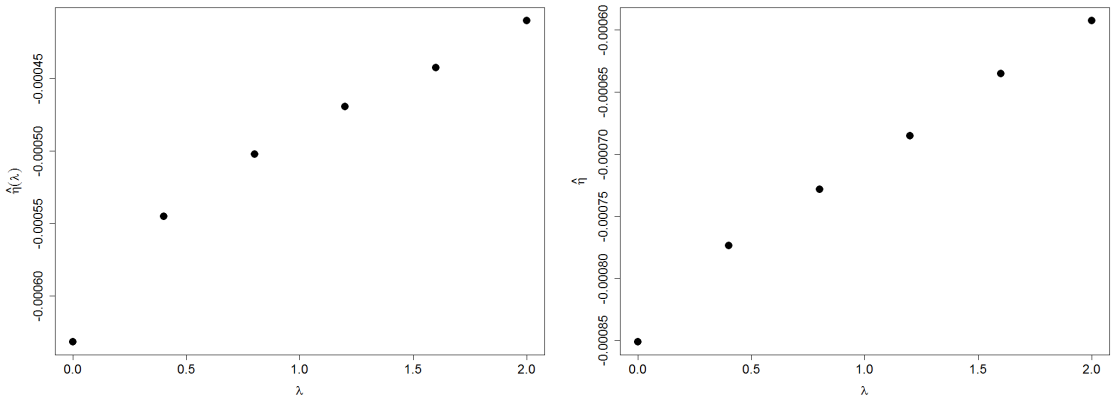
For this study, the imputation process for the bedroom and kitchen depended on different situations. If both rooms for a given house had missing measurements, then the average among all individuals that had the same disease status and the

same room was used for the imputation. In the situation where a specific room measurement was missing, then the solution for the fixed effects obtained from the mixed model together with the available radon measurements were used to impute the missing measurement. After the imputation procedure was done, the results from the mixed model (Table 4.3), along with the sample mean and variance for the radon exposure in the controls, were used in Equation 3.4. The corrected measurements obtained from this equation were transformed back to the original scale (Bq/m^3) and then used to substitute the original measurements. In cases where an individual had more than one measurement per room at a given home, an average of those corrected measurements was taken. Then, an average over room was computed in order to have one measurement per home per subject. The 6-30 year window was used to decide whether or not a home should be kept. In the case where there was a year in the 6-30 year window were there was no home assigned, the corrected radon measurement mean for each disease status was used to impute the values. An exposure time window (ETW) was computed using the home measurements, and one corrected radon measurement was obtained for each individual. The average of the radon concentrations for each disease status and overall can be found in Table 4.5. Moreover, the distribution of the radon exposure measurements is in Figure 4.7. These corrected measurements were used in the excess odds model (4.1) to assess the relationship between indoor radon exposure and the risk of developing lung cancer. The model was fitted without adjustment by covariates and then adjusted by categories of age, smoking duration and smoking rate (Table 4.4).

4.2.2 SIMEX

In order to correct for measurement error by implementing SIMEX, the results from the mixed model 4.9 were used. The imputation process for this frequentist

approach was the same one used for the regression calibration. After the imputation was done, the SIMEX algorithm was applied following the steps described in the Iowa study but now using $U \sim N(0, \sigma_U^2 = 0.09204)$ for the additional error added. One measurement was obtained as explained in the previous section, and the $\lambda(\hat{\eta})$ were obtained for the risk models with and without adjustment by covariates. Simple linear regression models were used for the extrapolation part (Figures 4.5a and 4.5b). The results for both model can be found in Table 4.4.



(a) model with no adjustment (b) model with adjustment

Figure 4.5: SIMEX plots for the Missouri models

4.2.3 Bayesian

In contrast to the mixed model fitted for the frequentist approaches, the Bayesian mixed model included radon exposure measurements for all subjects - not only the disease free. However, the factor that was consistent among models was the inclusion of all available measurements in the home for the correction part of the analysis, but with only the kitchen and bedroom ones used for the risk model.

The Bayesian error model specification was the following:

$$\begin{aligned}
\text{lnBq}_{ijk} &\sim N\left(\mu_{ijk}, \frac{1}{\tau}\right) \\
\mu_{ijk} &= \beta_0 + \beta_1 \text{BA}_{ijk} + \beta_2 \text{BR}_{ijk} + \beta_3 \text{OT}_{ijk} + \beta_4 \text{STATUS}_i + \gamma_{1j(i)} \\
\beta_p &\sim N(0, 0.0001) \text{ for } p = 0, 1, 2, 3, 4 \\
\gamma_{1j(i)} &\sim N\left(0, \frac{1}{\tau_1}\right) \\
\tau &\sim Gamma(0.001, 0.001) \\
\tau_1 &\sim Gamma(0.001, 0.001)
\end{aligned} \tag{4.10}$$

where β_0 is the overall intercept, β_p $(p = 1, 2, 3)$ are the parameters associated with the room fixed effect (basement [BA], bedroom [BR], and other [OT] respectively; kitchen is the reference category), β_4 is the parameter associated with the disease status (coded as 1 for disease and 0 for disease free), and $\gamma_{1j(i)}$ is the random effect for home j within individual i. The precisions for the distributions of lnBq and γ_1 were τ and τ_1, respectively, and they both had vague Gamma priors. The parameters from Model 4.10 were used in the following correction equation in order to obtain corrected radon exposure measurements in the original scale Bq/m^3.

$$\text{Bq}_{ij}^{c} = \frac{\exp\big(\beta_0 + \beta_4 \text{STATUS}_i + \gamma_{1j(i)}\big) \times \big(1 + \exp(\beta_2)\big)}{2} \tag{4.11}$$

These measurements were averaged over the homes that were included in the 6-30 year time window in order to obtain an ETW measurement per individual. A box plot with the posterior means of these corrected measurements can be found in Figure 4.7. Moreover, posterior summary statistics for the corrected measurements can be found on Table 4.7. A goodness-of-fit test was performed and the p-value for this test was 0.4987 which implies good fit (see Table 4.6). Then, the excess odds model part was specified without adjustment by covariates using those corrected

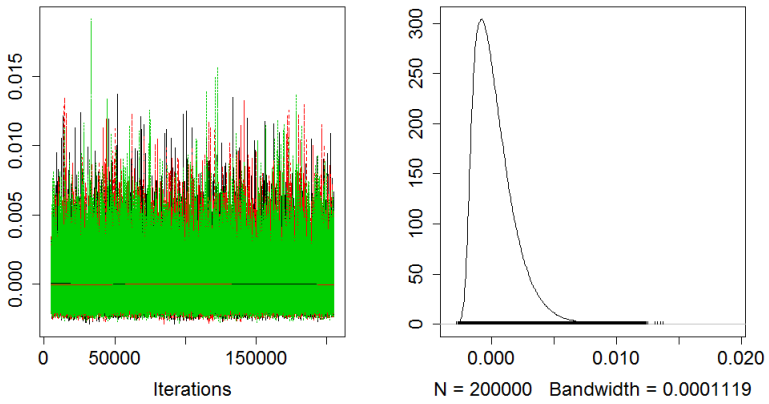
measurements (results in Table 4.4). The model specification was the following:

$$\begin{aligned}
\text{STATUS}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \frac{\text{odds}_i}{1+\text{odds}_i} \\
\text{odds}_i &= \exp(\alpha_0) \times (1+\eta \text{Bq}_i^c) \\
\alpha_0 &\sim N(0, 0.0001) \\
\eta &\sim Unif(-2,2) \prod_{i=1}^{n} I_{[(1+\text{Bq}_i^c)>0]}
\end{aligned} \tag{4.12}$$

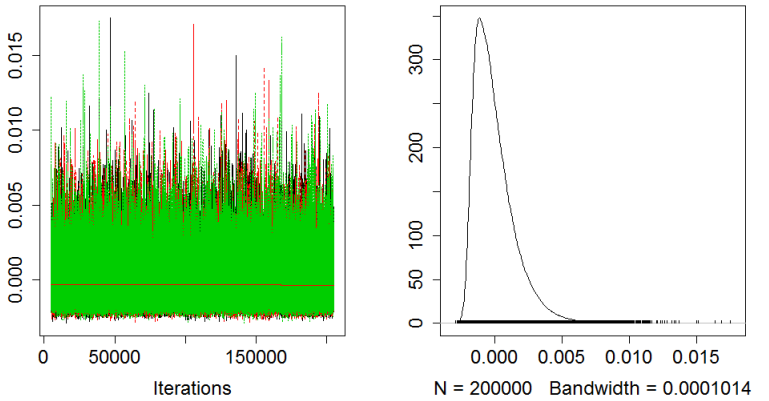
There was a burn-in period of 5k and an additional 80k iterations were used for the inferences. Figure 4.6a shows the trace and density plots for the parameter associated with the radon effect, η. The c-index for this model was 0.50 with a 95% credible interval of (0.48, 0.52), which implies that the model is not discriminating well (Table 4.6). Then, the risk model was specified with adjustment by covariates.

$$\begin{aligned}
\text{STATUS}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \frac{\text{odds}_i}{1+\text{odds}_i} \\
\text{odds}_i &= \exp(\alpha_0 + \alpha_1 I_{[50\leq age_i\leq 54]} + \alpha_2 I_{[55\leq age_i\leq 59]} + \\
&\quad \alpha_3 I_{[60\leq age_i\leq 64]} + \alpha_4 I_{[65\leq age_i\leq 69]} + \alpha_5 I_{[70\leq age_i\leq 74]} + \\
&\quad \alpha_6 I_{[age_i\geq 75]} + \alpha_7 I_{[9.5<rate_i\leq 19.5]} + \alpha_8 I_{[19.5<rate_i\leq 29.5]} + \\
&\quad \alpha_9 I_{[rate_i>29.5]} + \alpha_{10} I_{[1\leq dur_i\leq 24]} + \alpha_{11} I_{[35\leq dur_i\leq 44]} + \\
&\quad \alpha_{12} I_{[dur_i\geq 45]}) \times (1+\eta \text{Bq}_i^c) \\
\alpha_g &\sim N(0, 0.0001) \quad \text{for } g = 0, 1, ..., 12 \\
\eta &\sim N(0, 0.0001) \prod_{i=1}^{n} I_{[(1+\eta \text{Bq}_i^c)>0]}
\end{aligned} \tag{4.13}$$

Note that for this model, it was not necessary to constrain the bounds for the α or η priors. Therefore, all priors were kept as originally intended. The burn-in period was also 5k but the number of iterations needed for inferences was 200k. The trace and density plots for the η node in this model is presented in Figure 4.6b. The c-index for this model was 0.59 with a 95% credible interval of (0.58, 0.61) which



(a) model with no adjustment



(b) model with adjustment

Figure 4.6: Trace and density plots for the η node in the Missouri models

indicates that the model does not have a good discrimination ability (Table 4.6). This might be due to the fact that the study includes a relatively homogeneous group of subjects with respect to smoking (ex and never-smokers). With smoking being the biggest determinant of lung cancer, lack of variability in smoking would minimize the ability to discriminate. The results for the risk model with adjustment by covariates are shown in Table 4.4.

Table 4.6: Bayesian p-values (p_B) for the error model goodness-of-fit test, and c-index with credible intervals (CrI) for the risk models with and without adjustment (adj.)

		c-index			
Study	p_B	No adj.	CrI	With adj.	CrI
Iowa	0.5018	0.51	(0.49, 0.51)	0.85	(0.85, 0.86)
Missouri	0.4987	0.50	(0.48, 0.52)	0.59	(0.58, 0.61)
Winnipeg	0.5038	0.54	(0.45, 0.57)	0.80	(0.79, 0.80)
Connecticut	0.5000	0.52	(0.46, 0.55)	0.67	(0.66, 0.67)
Utah/South Idaho	0.4995	0.51	(0.48, 0.55)	0.71	(0.70, 0.72)

4.2.4 Comparison

Box plots comparing the uncorrected measurements, the corrected ETW measurements using regression calibration, and the posterior mean of the Bayesian corrected measurements are shown in Figure 4.7. Similar to the Iowa results, the regression calibration one has the least variability. In this case, the posterior means of the Bayesian corrected measurements also have smaller variability than the uncorrected ones. The medians and interquartile range for the box plots are similar across distributions, although the IQR is wider for the uncorrected measurements.

The point estimates and the confidence and credible intervals for the risk models are presented in Table 4.4. Even though the posterior mean of η for the Bayesian model without adjustment by covariates is within the values of the confidence intervals for the other risk models, it is noticeably different from the point estimates of the other models. A similar pattern can be seen if we compare the $\hat{\eta}$ for the models adjusted by covariates. Based on the intervals, the effect of indoor radon exposure on the risk of developing lung cancer is not statistically significant.

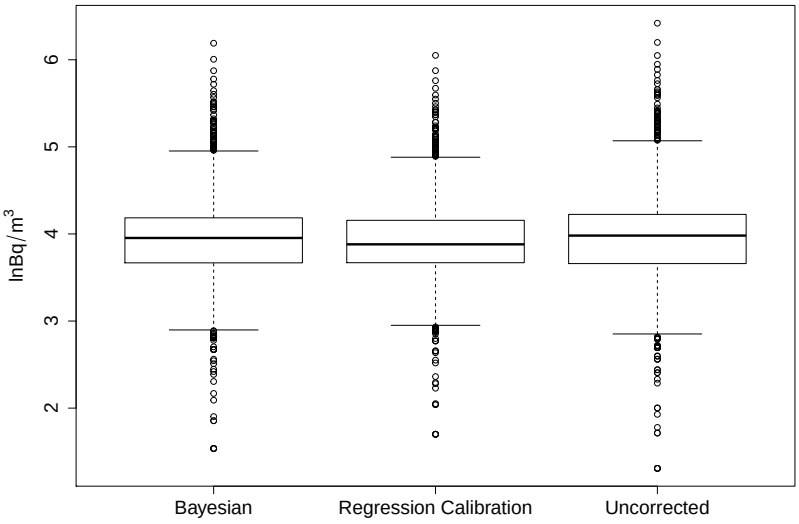


Figure 4.7: Distributions for the corrected and uncorrected measurements of lnBq in the Missouri sample

4.3 Winnipeg

The detectors were placed in the bedroom and in the basement of multiple homes for the individuals in this dataset. According to Krewski et al. (2006), individuals with missing home information for the 6-30 year window as well as individuals missing smoking information were excluded from the analysis. There were nine cases and four controls with missing smoking information, 20 cases and 13 controls with missing home information for the 6-30 year window, and 32 cases and 8 controls with raw measurements missing or exposure that was less than 300 days. After taking into account the inclusion and exclusion criteria, the final sample included 1390 subjects (see Table 4.1).

4.3.1 Regression Calibration

A mixed model was fitted using the disease free subjects. For this model, the natural logarithm of the radon exposure measurement k in Becquerel per cubic meter

(lnBq/m^3) measured by alpha-track detectors was used as our outcome variable. The room information (basement or bedroom) for subject i at home j was used as our fixed effects variable (coded as 1 for basement and 0 for bedroom), and the parameter associated with it was β_1. A random effect for each home within each individual ($\gamma_{j(i)}$) was also added into the model. In summary, the model was the following:

$$\text{lnBq}_{ijk} \sim N(\beta_0 + \beta_1 \text{ROOM}_{ijk} + \gamma_{j(i)}, \sigma^2_{\text{lnBq}}) \tag{4.14}$$

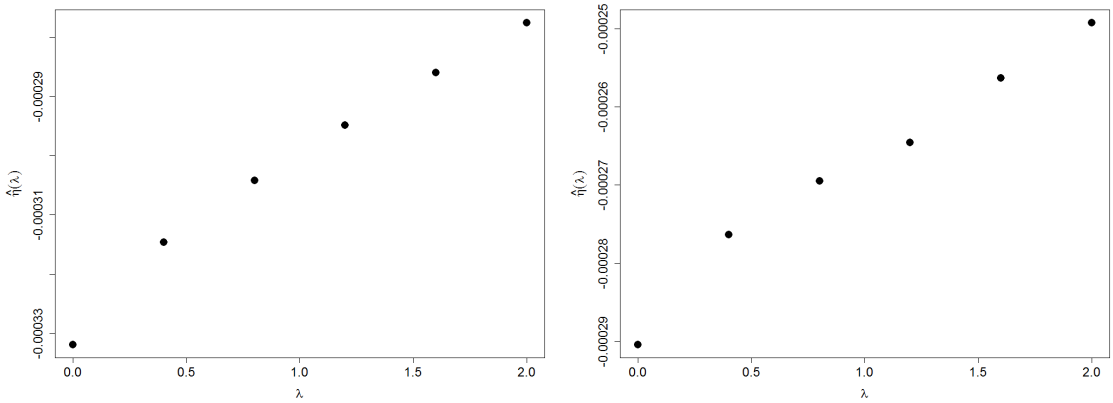
Results for this model along with the controls' sample mean and variance for the radon exposure can be found in Table 4.3. Following the suggestion from Krewski et al. (2006), the overall mean was used to impute values whenever there were missing measurements for a given home. In the cases where at least one of the rooms was measured, that available measurement and the results from the room fixed effects, were used in order to impute the missing measurement. Once the imputation was completed, we used the results in Table 4.3 as the estimates for Equation 3.4 in order to correct each radon exposure measurement. Those corrected measurements were then transformed back to the original scale (Bq/m^3). There were no individuals with multiple measurements for a given room at the same home. Therefore, the next step was to use the corrected measurements for each basement and bedroom to obtain a corrected averaged measurement for each home.

The homes that fell into the 6-30 year time window were kept for the remainder of the analysis. That is, if a home was outside of that time frame, the corresponding corrected radon measurement was discarded. Also, whenever there was a year with no home information, the corrected average for the radon measurements was used to impute that measurement. An exposure time window (ETW) measurement was obtain for each individual using the corrected radon measurements. Then, the

risk model 4.1 was fitted with and without adjustment by covariates to find the relationship between indoor radon exposure and the risk of developing lung cancer. Table 4.4 shows the point estimates as well as the confidence intervals obtained from these models.

4.3.2 SIMEX

The simulation extrapolation (SIMEX) method was also used with this data in order to correct the parameter estimate associated with the radon exposure (results in Table 4.4). For this method, the imputation process used was the same that was used for the regression calibration whenever a home or a single room measurement was missing, and the SIMEX algorithm was applied as in the Iowa data but using $U_m \sim N(0, \sigma_U^2 = 0.07751)$ as the additional error added. Then, the measurements were exponentiated and averaged to obtain a single ETW measurement and fit risk models as explained in the previous section. Simple linear regressions were used for the extrapolation part of SIMEX (see Figures 4.8a and 4.8b).



(a) model with no adjustment (b) model with adjustment

Figure 4.8: SIMEX plots for the Winnipeg models

4.3.3 Bayesian

The proposed Bayesian method was used to analyze this dataset belonging to the Winnipeg study. After applying the inclusion/exclusion criteria mentioned at the beginning of this section, all remaining observations were used to specify a Bayesian mixed model. This model included a fixed effect for room (β_1), one for the status effect (β_2), and a random effect ($\gamma_{j(i)}$) that accounts for the effect of each measurement k for home j within individual i. The β and γ parameters had vague Normal priors, and the precisions had Gamma distributions for the priors. That is,

$$\begin{aligned}
\ln\mathrm{Bq}_{ijk} &\sim N\left(\mu_{ijk}, \frac{1}{\tau}\right) \\
\mu_{ijk} &= \beta_0 + \beta_1 \mathrm{ROOM}_{ijk} + \beta_2 \mathrm{STATUS}_i + \gamma_{1j(i)} \\
\beta_p &\sim N(0, 0.0001) \text{ for } p = 0, 1, 2 \\
\gamma_{1j(i)} &\sim N\left(0, \frac{1}{\tau_1}\right) \\
\tau &\sim Gamma(0.001, 0.001) \\
\tau_1 &\sim Gamma(0.001, 0.001).
\end{aligned} \tag{4.15}$$

Using the parameters in that section of the model and Equation 4.16, the corrected radon measurements were obtained. This equation computes each corrected measurement and exponentiates the result so the final corrected measurement is in the original scale, Bq/m^3. The formula for such correction is the following:

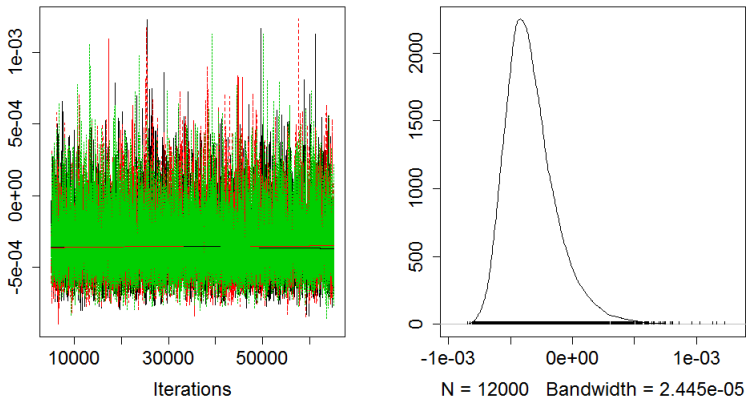
$$\mathrm{Bq}_{ij}^{c} = \frac{\exp\big(\beta_0 + \beta_2 \mathrm{STATUS}_i + \gamma_{1j(i)}\big) \times \big(1 + \exp(\beta_1)\big)}{2}. \tag{4.16}$$

The 6-30 year time window was used to compute an ETW measurement for each individual. This single corrected measurement was then used in the second part of the model, the risk model. The model specified without adjustment by covariates was the same model used for the Iowa study (model 4.7). A goodness-of-fit test was performed to assess how well the model fitted the data. The Bayesian p-value for this test was $p = 0.5038$ (see Table 4.6) which implies good fit. Posterior summary

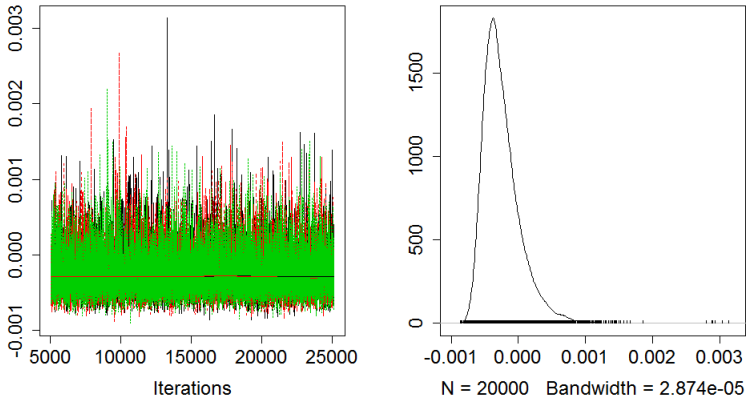
statistics for the corrected measurements are presented on Table 4.7. A burn-in of 5k iterations was discarded, and an additional 60k were kept for the inferences in this model. Figure 4.9a shows the trace and density plots for the η node. Also, the c-index for this model was 0.54 with a 95% credible interval of (0.45, 0.57) which implies that the model is not discriminating well (Table 4.6). The model adjusted by covariates was similar to the one used for the Missouri study (model 4.13). There is one difference between those model, the Missouri model did not require adjustment by sex (all subjects were female) but the Winnipeg model includes it. Therefore, the Winnipeg model with adjustment by covariates can be expressed as follow:

$$\begin{aligned}
\text{STATUS}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \frac{\text{odds}_i}{1+\text{odds}_i} \\
\text{odds}_i &= \exp(\alpha_0 + \alpha_1 I_{[50\leq age_i\leq 54]} + \alpha_2 I_{[55\leq age_i\leq 59]} + \\
&\quad \alpha_3 I_{[60\leq age_i\leq 64]} + \alpha_4 I_{[65\leq age_i\leq 69]} + \alpha_5 I_{[70\leq age_i\leq 74]} + \\
&\quad \alpha_6 I_{[age_i\geq 75]} + \alpha_7 I_{[9.5<rate_i\leq 19.5]} + \alpha_8 I_{[19.5<rate_i\leq 29.5]} + \\
&\quad \alpha_9 I_{[rate_i>29.5]} + \alpha_{10} I_{[1\leq dur_i\leq 24]} + \alpha_{11} I_{[35\leq dur_i\leq 44]} + \\
&\quad \alpha_{12} I_{[dur_i\geq 45]} + \alpha_{13}\text{SEX}_i) \times (1+\eta \text{Bq}_i^c) \\
\alpha_g &\sim N(0, 0.0001) \quad \text{for } g = 0, 1, ..., 13 \\
\eta &\sim N(0, 0.0001) \prod_{i=1}^{n} I_{[(1+\eta \text{Bq}_i^c)>0]}
\end{aligned} \tag{4.17}$$

This joint model had a burn-in period of 5k iterations and an additional 20k iterations were used to make the inferences. The results for both models are presented in Table 4.4. The trace and density plots for this model are shown in Figure 4.9b. The discrimination ability for the risk model adjusted by covariates was better (Table 4.6), c-index of 0.80 and 95% credible interval of (0.79, 0.80).



(a) model with no adjustment



(b) model with adjustment

Figure 4.9: Trace and density plots for the η node in the Winnipeg models

4.3.4 Comparison

Figure 4.10 shows the box plots for the ETW of the radon exposure measurements without correction, then with correction using regression calibration, and the posterior mean of the corrected measurements for the proposed Bayesian method. There is a remarkable difference in the distribution of the uncorrected measurement versus the distributions of the corrected ones. Similar to the distributions for the Missouri study, we can see that the variability for the uncorrected measurements is the highest, followed by the variability of the corrected measurements using the

Bayesian method, and finally the corrected measurements using regression calibration had a slightly smaller variability than the posterior mean for the Bayesian corrected measurements.

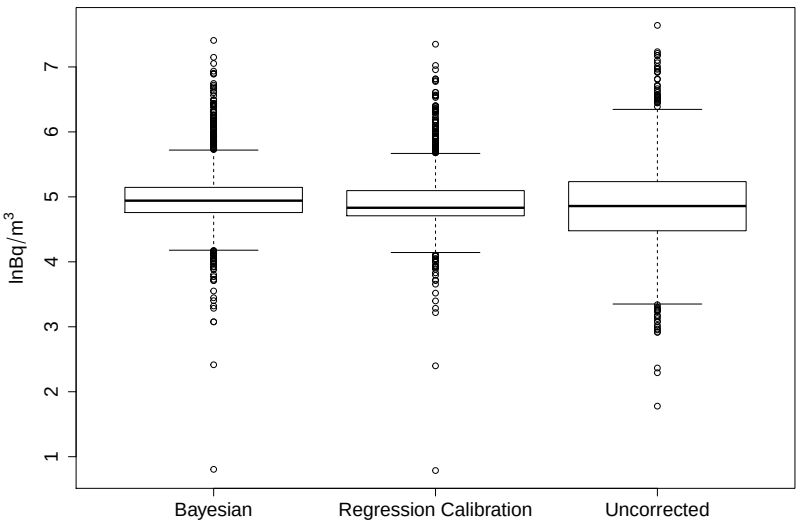


Figure 4.10: Distributions for the corrected and uncorrected measurements of lnBq in the Winnipeg sample

If we compare the point estimates from the models fitted without adjustment by covariates, we could see that the one obtained using the regression calibration correction method is the most dissimilar. Nonetheless, that point estimate is within the range of the confidence and credible intervals obtained when using no correction, or the other two correction methods. This pattern is repeated for the point estimates obtained from models fitted with adjustment by covariates. In neither case an association between indoor radon exposure and the risk of developing lung cancer was found.

4.4 Connecticut

The detectors in the Connecticut study were placed in different levels of the homes. The information available is the vertical distance from ground to the detector placement, based on flight of stairs where 6 stairs were equivalent to a half-flight. This distance was equal to one for the ground, less than one when the detector was placed below ground and greater than one when the detector was placed above ground. For the data analysis, this information was coded as follow:

- if the distance was less than or equal to 0.5, then the level will be considered as level 0
- if the distance was greater than 0.5 but less than or equal to 1.5, then the level will be 1
- if the distance was greater than 1.5, then the level will be 2.

The dataset also included some radon measurement marked as duplicates. Following the suggestion in the code book, these measurements were removed before the analysis. There were also subjects who had demographic information but did not have raw radon measurements; those subjects were excluded from the analyses. The final sample size for this study was 1810 (Table 4.1).

4.4.1 Regression Calibration

Using the natural logarithm of the radon measurements in Becquerel per cubic meter ($\ln\text{Bq/m}^3$), and the level and home information for the controls, a linear mixed model was fitted by employing the MIXED procedure in the SAS software (SAS Institute Inc, 2012). The model included a fixed component for the level variable and a random components for the subject and home information. That is,

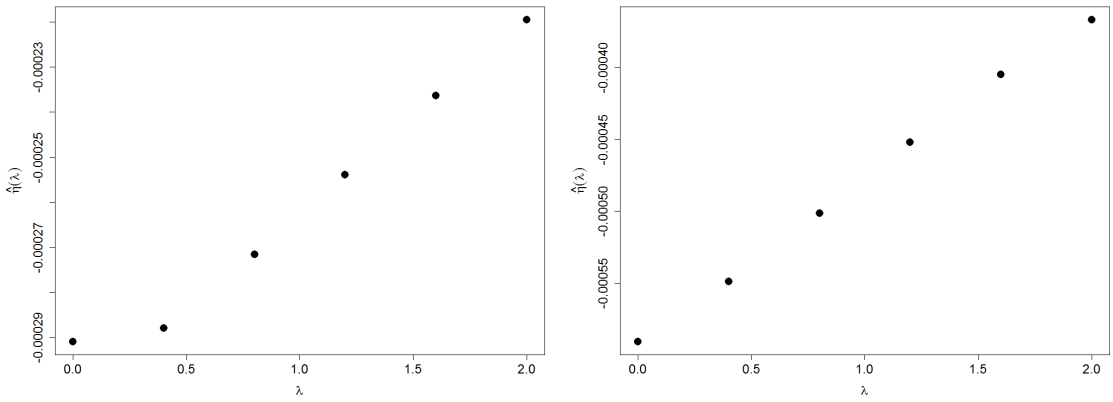
$$\ln\text{Bq}_{ijk} \sim N(\beta_0 + \beta_1\text{LEVEL1}_{ijk} + \beta_2\text{LEVEL2}_{ijk} + \gamma_{j(i)}, \sigma^2_{\ln\text{Bq}}) \tag{4.18}$$

where β_0 is the intercept, β_1 and β_2 are the level fixed effects compared to level 0, "LEVEL1" and "LEVEL2" are indicator variables coded as 1 if the measurement k was taken from level 1 or 2, respectively, and coded as 0 otherwise, and $\gamma_{j(i)}$ is the random effect for home j within individual i. The results from the fixed effects in this model as well as any available measurement, were used for the imputation when a radon measurement was missing for a given home level. According to Krewski et al. (2006) if all measurements were missing for a given house, the controls' mean was used for the imputation. However, in instances were at least one measurement per home was available, the results from the fixed effects along with the available measurement(s) were used for the imputation. The estimates from Table 4.3 were obtained from the mixed model fitted only on the controls and from the controls descriptive statistics. Those estimates were used in Equation 3.4 to obtain corrected measurements for each level, and then the corrected measurements were averaged in order to obtain a single measurement for each home. Then, the 6-30 year window was used to determine which homes were going to be kept in the analysis. An exposure time window (ETW) measurement was computed for each individual and was then used in the risk model 4.1 to assess the relationship between indoor radon exposure and lung cancer. The model was fitted first without adjustment by covariates and then adjusted by gender, categories of age, smoking duration and smoking rate. The resulting parameter estimates and confidence intervals for the excess odds can be found in Table 4.4.

4.4.2 SIMEX

The simulation extrapolation method (SIMEX) was the other frequentist approach used to analyze this data. In order to do that, we used the dataset that was created after imputing missing radon exposure values as in the regression calibration subsection. The SIMEX algorithm used followed the same process explained

earlier but with $U_m \sim N(0, \sigma_U^2 = 0.2041)$ as the additional error for each of the 20 datasets. The computation of the ETW measurement per subject was also the same as the one in the previous regression calibration subsection. The possible associations between λ and $\hat{\eta}(\lambda)$ for the models with and without adjustment by covariates were explored using scatter plots (Figures 4.11a, 4.11b). A simple linear regression was fitted and then used to extrapolate to the case of no error ($\lambda = -1$). Bootstrap was used to obtain the standard errors needed to create the confidence interval for the parameter estimate. The point estimates and confidence intervals for these models can be found on Table 4.4



(a) model with no adjustment (b) model with adjustment

Figure 4.11: SIMEX plots for the Connecticut models

4.4.3 Bayesian

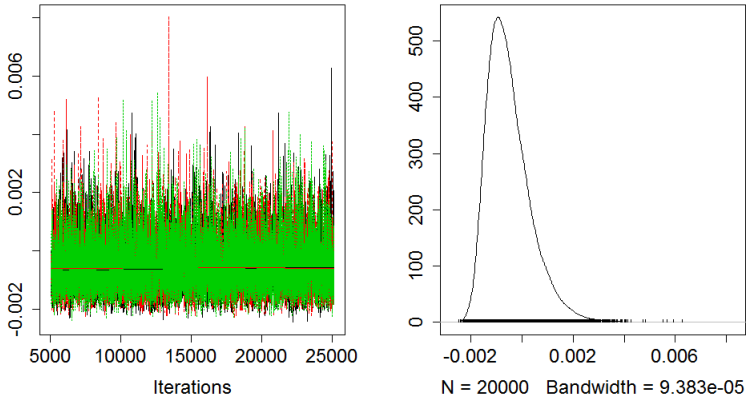
Using all available data, the following mixed model was specified and its parameters were used to obtain corrected radon measurements.

$$\begin{aligned}
\text{lnBq}_{ijk} &\sim N\left(\mu_{ijk}, \frac{1}{\tau}\right) \\
\mu_{ijk} &= \beta_0 + \beta_1 \text{LEVEL1}_{ijk} + \beta_2 \text{LEVEL2}_{ijk} + \beta_3 \text{STATUS}_i + \gamma_{1j(i)} \\
\beta_p &\sim N(0, 0.0001) \text{ for } p = 0, 1, 2, 3 \\
\gamma_{1j(i)} &\sim N\left(0, \frac{1}{\tau_1}\right) \\
\tau &\sim Gamma(0.001, 0.001) \\
\tau_1 &\sim Gamma(0.001, 0.001)
\end{aligned} \tag{4.19}$$

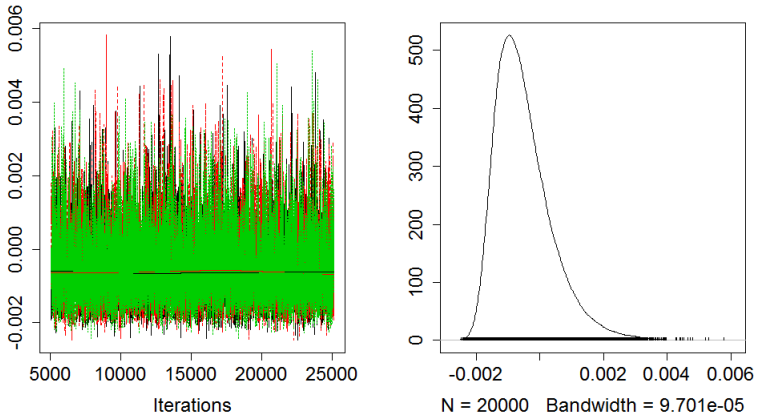
In this model, β_0 is the intercept, β_1 and β_2 are the fixed effects for the home levels, β_3 is the disease status fixed effect, and $\gamma_{1j(i)}$ is a nested random effect for each home j within each individual i. This Bayesian model used precisions rather than variances. Thus, a vague Gamma prior was used for the precisions of the lnBq and γ distributions, whereas vague Normal distributions were used for the β priors. Using the parameters from the mixed model, Equation 4.20 was used to obtain corrected radon measurements for each subject at each home.

$$\text{Bq}_{ij}^{c} = \frac{\exp\big(\beta_0 + \beta_3 \text{STATUS}_i + \gamma_{1j(i)}\big) \times \big(1 + \exp(\beta_1) + \exp(\beta_2)\big)}{3} \tag{4.20}$$

The homes that were part of the 6-30 year window were included in the ETW that was computed using the corrected measurements. Those single measurements per subject were used to specify a risk model without adjustment by covariates, which had the same structure as the model used with the Iowa and Winnipeg studies when adjustment by covariates was not taken into account (model 4.7). A burn-in of 5k iterations was discarded and an additional 20k iterations (after convergence was met) were used for the inferences. Figure 4.12a shows the trace and density plots for the η node. The results for this model can be found in Table 4.4, and the



(a) model with no adjustment



(b) model with adjustment

Figure 4.12: Trace and density plots for the η node in the Connecticut models

posterior summary statistics for the corrected measurements' distributions can be found in Table 4.7. Also, a goodness-of-fit test was conducted to assess the model performance and the p-value for this test was $p = 0.5$ (Table 4.6). The discrimination abilities of this model (Table 4.6) are not better than the ones obtained by chance (c-index = 0.52). The risk model was then specified with an adjustment by gender, categories of age, smoking rate and smoking duration. This adjusted model had the same structure as the one for the Winnipeg data (model 4.17). Once more, 20k iterations were needed for the inferences and Figure 4.12b shows the trace and

density plots for node η. The discrimination abilities of the adjusted model slightly increased as compared to the model without adjustment (Table 4.6).

4.4.4 Comparison

The distributions for the ETW uncorrected measurements, the corrected radon measurements using regression calibration and the posterior means of the corrected measurements using Bayesian models, are shown in Figure 4.13. The correction methods are shrinking the measurements' distributions, and as a consequence, the variability of those distributions are significantly smaller than the variability of the distribution for the ETW measurements without correction. It seems that the Bayesian method correction yielded a distribution with higher values for the posterior mean corrected radon measurements than the regression calibration correction. The point estimates for the models fitted without adjustment by covariates are sim-

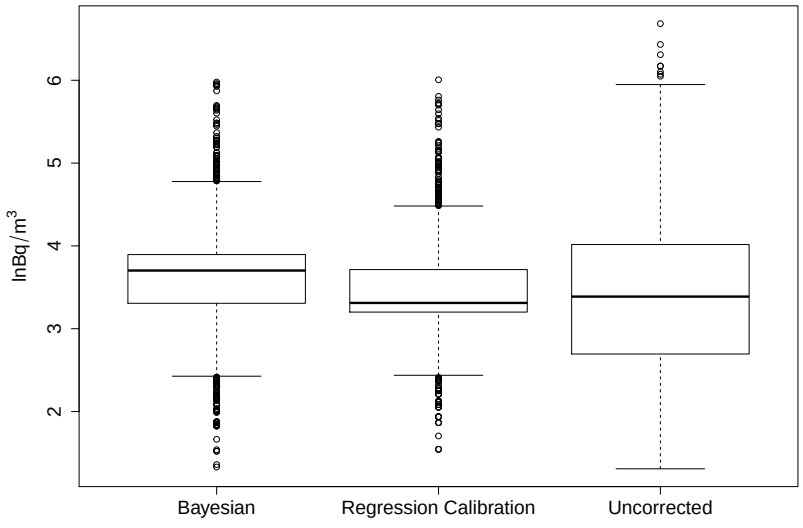


Figure 4.13: Distributions for the corrected and uncorrected measurements of lnBq in the Connecticut sample

ilar, except the point estimate for the SIMEX correction method. However, when comparing the point estimates for the models adjusted by covariates, only the ones obtained from the SIMEX and Bayesian methods are similar. Also, depending on which method we focus, the adjustment by covariates may or may not affect the estimates. However, all the point estimates are within the range of the confidence and credible intervals.

Table 4.7: Posterior Summary Statistics[a] for η

Study	Mean	Median	Std	MCSE	CrI
Iowa					
no adjustment	0.105	0.093	0.082	0.0005	(-0.022, 0.299)
with adjustment	0.194	0.164	0.155	0.0011	(-0.015, 0.576)
Missouri					
no adjustment	0.024	-0.01	0.165	0.0009	(-0.188, 0.434)
with adjustment	-0.005	-0.038	0.152	0.0007	(-0.197, 0.377)
Winnipeg					
no adjustment	-0.033	-0.036	0.021	0.0001	(-0.065, 0.010)
with adjustment	-0.025	-0.03	0.028	0.0002	(-0.064, 0.044)
Connecticut					
no adjustment	-0.050	-0.064	0.086	0.0009	(-0.174, 0.157)
with adjustment	-0.051	-0.067	0.090	0.0009	(-0.178, 0.170)
Utah/South Idaho					
no adjustment	0.146	0.096	0.232	0.0030	(-0.144, 0.733)
with adjustment	0.519	0.386	0.562	0.0136	(-0.069, 1.911)

[a] Std, MCSE, and CrI stand for standard deviation, Monte Carlo Standard Error and, credible interval, respectively

4.5 Utah/South Idaho

The Utah/South Idaho and the Connecticut datasets were part of the same study. Hence, the dataset, and therefore the data management, in this Utah/South Idaho study were similar to the Connecticut one. As a summary, the information about the absolute distance from ground was used to create a home level variable. There was home information available as well as the radon measurements in Becquerel per cubic meter taken from alpha-track detectors. Some individuals had duplicated measurements and according to the information provided, these measurements were removed. Other individuals had demographic information but no radon exposure information, these subjects were removed as well. The final sample size available was 1345 (Table 4.1).

4.5.1 Regression Calibration

The radon measurements from the Utah/South Idaho study were first corrected using regression calibration. In order to do that, the MIXED procedure form the SAS software (SAS Institute Inc, 2012) was used to fit a mixed model using only the disease free subjects (model 4.18). If a level measurement was missing, the parameter estimates for the fixed effects from that model, along with the available measurement(s) for that home, were used to impute the radon exposure value. Following the suggestion in Krewski et al. (2006), the mean of the radon concentration for the controls was used in the case where all measurements were missing for a given home. The random effects results and the controls' radon exposure measurement mean and variance were used in Equation 3.4 to obtain the corrected measurement for each level at each house. These measurements were then exponentiated and the corrected measurements in Bq/m^3 were obtained. For each home, an average using the levels' corrected measurements was computed which resulted in a single measurement per home. If those homes were lived by the subject during the 6-30

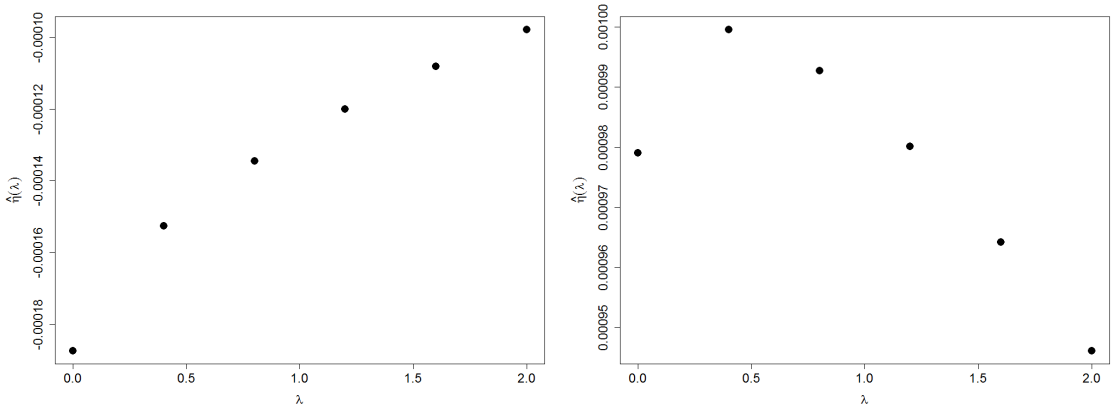
year window, then they were included in the ETW (otherwise, they were excluded). An excess odds model (4.1) was fitted with and without adjustment by covariates to determine the relationship between indoor radon exposure and the probability of developing lung cancer. The results for these models are presented in Table 4.4.

4.5.2 SIMEX

Instead of correcting each observation, the simulation extrapolation (SIMEX) method tries to correct the parameter estimate. After imputing the missing values as explained in the previous subsection, the SIMEX method was applied to the data using the Equation 4.4 for 20 datasets, $\lambda = 0, 0.4, ..., 1.6, 2$, and $\sigma_U^2 = 0.0931$. Once each radon measurement in lnBq/m^3 was contaminated with extra error, a single measurement was computed using the same procedure explained in the regression calibration subsection. A simple linear regression was used for the extrapolation when the risk model was fitted without adjustment by covariates (Figure 4.14a). On the other hand, for the SIMEX extrapolation when a risk model with adjustment by covariates was fitted, a quadratic regression was used (4.14b). The resulting parameter estimates for each model, as well as the confidence intervals are shown in Table 4.4.

4.5.3 Bayesian

The proposed Bayesian method was applied to this dataset. The Bayesian model consists of three parts. First, the mixed model 4.19 was specified using all available data. Second, Equation 4.20 was used to compute the corrected measurements and an ETW measurement was obtained for each individual based on the 6-30 year window. Third, the risk model was specified and the parameter estimate associated with the radon exposure was obtained. Similar to the Connecticut risk model without adjustment by covariates, the Utah/South Idaho risk model had the



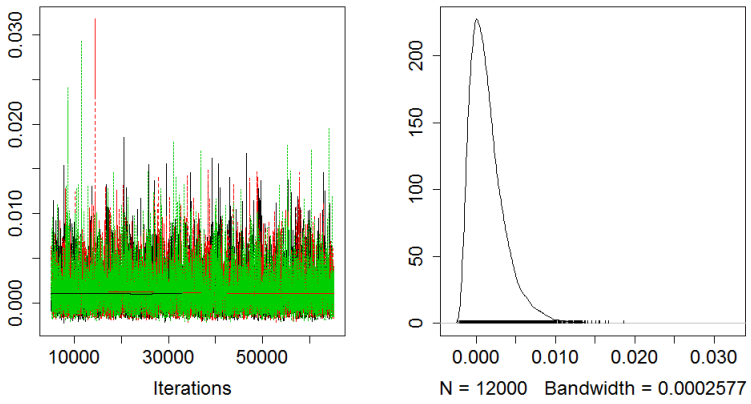
(a) model with no adjustment (b) model with adjustment

Figure 4.14: SIMEX plots for the Utah/South Idaho models

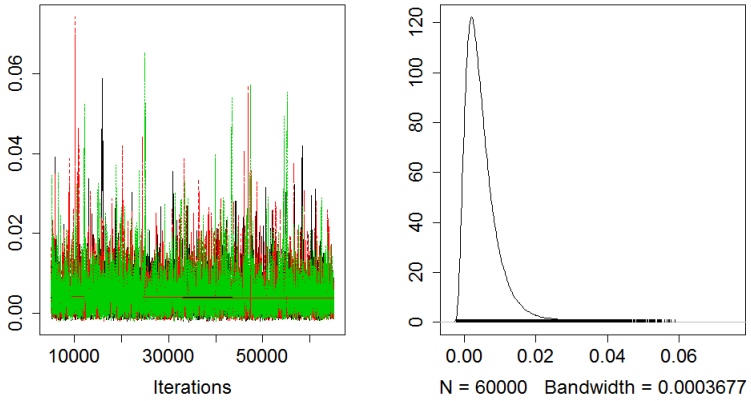
same structure as the Iowa and Winnipeg studies. Trace and density plots for this model are shown in Figure 4.15a. A goodness-of-fit test was performed and the p-value obtained for such model was $p = 0.4995$ whereas the c-index for the risk model was 0.51 (Table 4.6). The results for this joint model with and without adjustment by covariates are shown in Table 4.4. That model with adjustment by covariates was fitted in the same way as the model for the Connecticut study. Figure 4.15b shows the trace and density plots for this model, while Table 4.6 shows the c-index for the risk model. The models fitted for the Utah/South Idaho dataset (with and without adjustment by covariates) needed a burn-in of 5k iterations for convergence, and an additional 60k iterations were kept for the inferences.

4.5.4 Comparison

The box plots created using the ETW measurements using the data from the Utah/South Idaho study had a similar distribution to the ones plotted using the Connecticut study. This might be due to the fact that both studies were part of a bigger study. In general, the variability of the ETW measurements was reduced in the box plots constructed using the corrected measurements rather than the original



(a) model with no adjustment



(b) model with adjustment

Figure 4.15: Trace and density plots for the η node in the Utah/South Idaho models

ones. Once again, the posterior means for the Bayesian correction set the median at a higher value that the regression calibration correction. The resulting point estimates from the risk models fitted with and without adjustment by covariates are quite different from each other. Some of the credible and confidence intervals are narrow, whereas others are wider. Nonetheless, neither of the models show a statistically significant effect of radon exposure on the lung cancer risk.

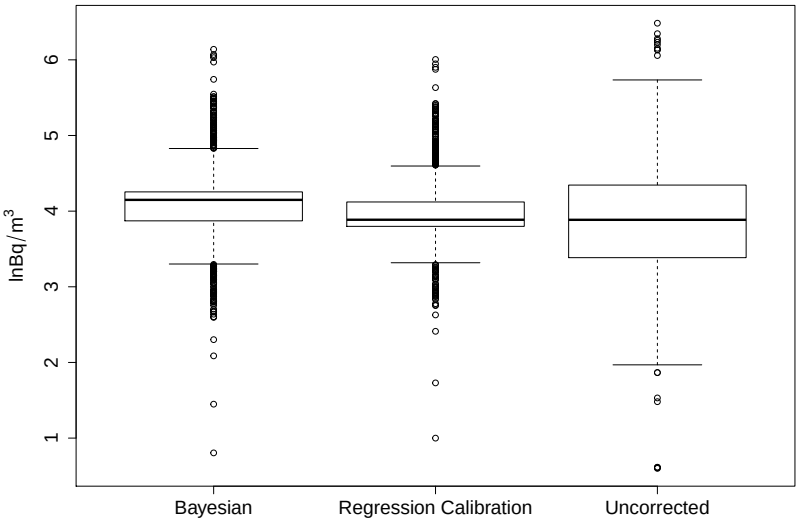


Figure 4.16: Distributions for the corrected and uncorrected measurements of lnBq in the Utah/South Idaho sample

4.6 Combined Analyses

The data used for this part of the analysis were the same used for the individual study analyses. Therefore, we were able to use the information available for each study in order to fit the error model. As a reminder, the Iowa study had home and room level information, but only one home for the individuals. The Missouri study had room information and multiple homes but no home level information. The Winnipeg study had measurements in the basement and bedroom only, and measurements at different homes. The Connecticut and Utah/South Idaho study had absolute measurements from the ground to the dosimeter, which were then coded as levels 0, 1 and 2. They were also the studies where individuals had the most mobility - multiple homes lived by them. Since there was no common ground for the exposure variable information and the variability within each study might be different among studies, the error model for this section was done independently for each study. That is, the imputation process and the correction of the radon exposure

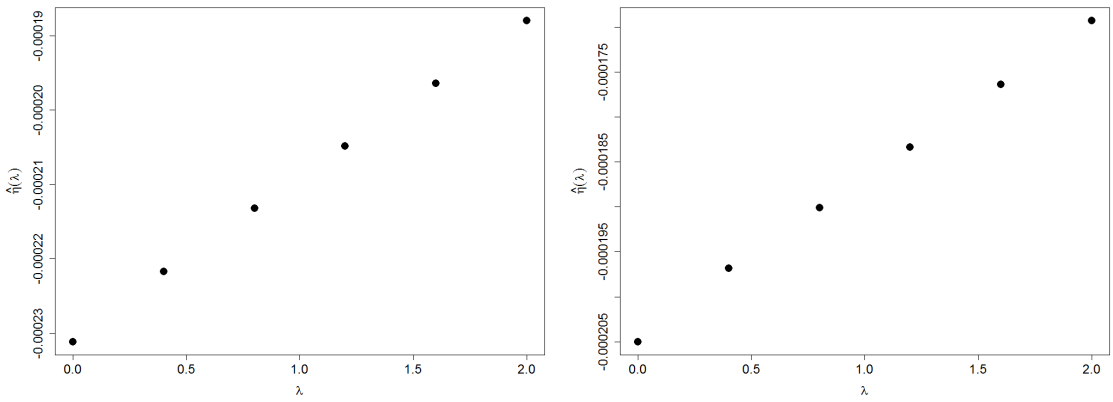
measurements were performed individually for each study. These measurements were then combined in order to obtain a pooled dataset of all the studies and fit the risk model using this bigger sample. The risk model was fitted with and without adjustment by sex, and categories of age, smoking duration and smoking rate. In both cases, a study effect was added to the model.

4.6.1 Regression Calibration

The mixed models presented in the previous sections (Models 4.3, 4.9, 4.14, 4.18) were used to impute values (when needed) and to obtain parameter estimates for Equation 3.4. After obtaining the corrected measurements, a single measurement was obtained for each home that was lived by an individual. The 6-30 year window was used to decide which homes were going to be used to compute the ETW measurement for the radon exposure. Remember that this window was not necessary for the Iowa study since the individuals all lived in the same home. Due to the various ways of collecting the data, the process mentioned was done for each study separately. Once the ETW was calculated, the data were pooled and the excess odds models fitted to all studies combined. The resulting parameter estimates and confidence intervals can be found in Table 4.4.

4.6.2 SIMEX

In a similar way to the regression calibration, the extra error added as part of the SIMEX simulation step was done study-by-study. That is, in a same process, the error was added to each observation using the information available for each study, and the ETW was computed for each subject within the same study as well. Then the ETW was pooled and the risk model fitted to all studies combined. Note that the risk model was fitted adjusted by study only or by study, sex, and categories of age, smoking rate and smoking duration. This process was repeated for each of the



(a) model with no adjustment (b) model with adjustment

Figure 4.17: SIMEX plots for the combined models

20 contaminated datasets simulated for each value of λ, as explained earlier in this chapter. A simple linear regression of the values of λ and $\hat{\eta}(\lambda)$ was fitted and then used to extrapolate to the case of $\lambda = -1$ (see Figures 4.17a and 4.17b).

4.6.3 Bayesian

The available data for each study were used in order to jointly model the error and risk models, as proposed by the Bayesian method. That is, all the parameters were modeled simultaneously while employing the means of the distributions for radon measurements, the correction equations and the risk models presented in the previous sections. The first part of this process is the error model, which has the

following structure:

$$
\begin{aligned}
\ln \mathrm{Bq}_{ijk}^{l} &\sim N\left(\mu_{ijk}^{l}, \frac{1}{\tau}\right) \\
\mu_{ijk}^{l} &= \begin{cases}
\beta_0 + \beta_1 \mathrm{LEVEL1}_{ijk} + \beta_2 \mathrm{LEVEL2}_{ijk} + & \\
\quad \beta_3 \mathrm{STATUS}_i + \gamma_{0i} + \gamma_{1j(i)}, & \text{if } l = \text{Iowa} \\
\beta_0 + \beta_1 \mathrm{BA}_{ijk} + \beta_2 \mathrm{BR}_{ijk} + \beta_3 \mathrm{OT}_{ijk} + & \\
\quad \beta_4 \mathrm{STATUS}_i + \gamma_{1j(i)}, & \text{if } l = \text{Missouri} \\
\beta_0 + \beta_1 \mathrm{ROOM}_{ijk} + \beta_2 \mathrm{STATUS}_i + \gamma_{1j(i)}, & \text{if } l = \text{Winnipeg} \\
\beta_0 + \beta_1 \mathrm{LEVEL1}_{ijk} + \beta_2 \mathrm{LEVEL2}_{ijk} + & \\
\quad \beta_3 \mathrm{STATUS}_i + \gamma_{1j(i)}, & \text{if } l = \text{Connecticut,} \\
 & \text{Utah/South Idaho}
\end{cases} \\
\beta_p &\sim N(0, 0.0001) \text{ for } p = 0, 1, 2, 3 \\
\gamma_{0i} &\sim N\left(0, \frac{1}{\tau_0}\right) \\
\gamma_{1j(i)}^{l} &\sim N\left(0, \frac{1}{\tau_1}\right) \\
\tau &\sim Gamma(0.001, 0.001) \\
\tau_0 &\sim Gamma(0.001, 0.001) \\
\tau_1 &\sim Gamma(0.001, 0.001)
\end{aligned}
\tag{4.21}
$$

In order to obtain the corrected measurements, the parameters from the error model were used in Equations 4.6 (Iowa), 4.11 (Missouri), 4.16 (Winnipeg) and 4.20 (Connecticut and Utah/South Idaho). These corrected measurements were then averaged in the same way as with the individual analyses. The last part of the modeling process was the specified excess odds model. Risk models were specified for each study individually, where each study had its own intercept to assess the study effects but

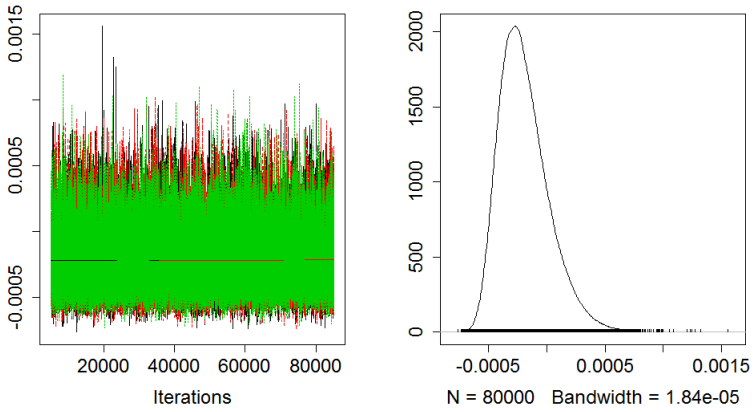
with a shared parameter for the radon effect. That is,

$$
\begin{aligned}
\text{STATUS}_i &\sim Bernoulli(\pi_i) \\
\pi_i &= \frac{\text{odds}_i}{1+\text{odds}_i} \\
\text{odds}_i &= \exp(\alpha_0^l + \alpha_1^l I_{[50\leq age_i\leq 54]} + \alpha_2^l I_{[55\leq age_i\leq 59]} + \\
&\quad \alpha_3^l I_{[60\leq age_i\leq 64]} + \alpha_4^l I_{[65\leq age_i\leq 69]} + \alpha_5^l I_{[70\leq age_i\leq 74]} + \\
&\quad \alpha_6^l I_{[age_i\geq 75]} + \alpha_7^l I_{[9.5<rate_i\leq 19.5]} + \alpha_8^l I_{[19.5<rate_i\leq 29.5]} + \\
&\quad \alpha_9^l I_{[rate_i>29.5]} + \alpha_{10}^l I_{[1\leq dur_i\leq 24]} + \alpha_{11}^l I_{[35\leq dur_i\leq 44]} + \\
&\quad \alpha_{12}^l I_{[dur_i\geq 45]} + \alpha_{13}^l \text{SEX}_i) \times (1+\eta \text{Bq}_i^c) \\
\alpha_g^l &\sim N(0, 0.0001) \quad \text{for } g = 0, 1, ..., 13 \\
\eta &\sim N(0, 0.0001) \prod_{i=1}^{n} I_{[(1+\eta \text{Bq}_i^c)>0]} \\
l &= \text{Iowa, Missouri, Winnipeg, Connecticut, Utah/South Idaho}
\end{aligned}
\tag{4.22}
$$

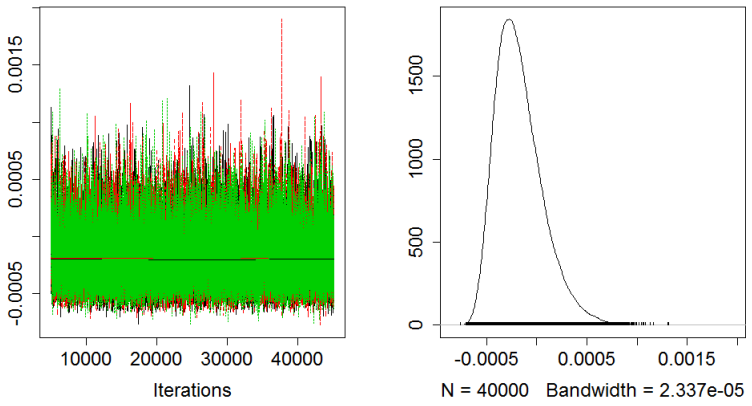
Trace plots for the η node for the risk models with and without adjustment by covariates are presented in Figures 4.18a and 4.18b. The resulting posterior means and credible intervals for that node are shown in Table 4.4.

4.6.4 Comparison

Even after combining the data, there was no radon effect found regardless of the correction method. This was also the case for the model adjusted by covariates. In fact, it seems like the adjustment by covariates was not necessary after pooling the data.



(a) model with no adjustment



(b) model with adjustment

Figure 4.18: Trace and density plots for the η node in the combined models

CHAPTER 5
DISCUSSION

5.1 Summary

The goal of this dissertation was to provide a Bayesian alternative to the correction of measurement error with an application to radon and lung cancer data. We accomplished that goal by using a Bayesian hierarchical approach to jointly model the error model and the risk model. The performance of this method was compared to the performance of two existing frequentist methods used for the correction of measurement error: regression calibration and Simulation Extrapolation (SIMEX). We started by providing a background of measurement error, the risk model and the data that motivated this work. Then in Chapter 2, we presented the regression calibration method, followed by the SIMEX method. That chapter ended with a background of Bayesian statistics that was the foundation for the method presented at the end of the chapter. Chapter 3 compared the three methods by performing a simulation study with different scenarios. Finally, the data presented in Chapter 1 was analyzed by using the three correction methods.

The simulation study showed that when we assume a small effect, the SIMEX method performs better in terms of the point estimate, bias and lower MSE. The regression calibration method is overestimating the results and has the largest bias, MSE and parameter estimate. A similar situation is found in the case where the effect is increased to 0.14 per 100 Bq/m^3, and when we have two studies. If we reduce the sample size to half but keep the higher effect, then all methods perform well in terms of coverage. However, the bias of the regression calibration and Bayesian methods increases. The results from the Bayesian method were consistent throughout all scenarios. We agree with Allodji et al. (2015) in that it will be useful to apply more than one correction method when analyzing data measured with error.

Finally, the three methods were applied to datasets from five North America case-control studies to assess the possible effect of indoor radon exposure on the risk of developing lung cancer. The analyses were performed individually for each study and for the pooled data. If we focus on the point estimate results, the effect was positive only for the Iowa study, where individuals were required to live in the same home for 20 consecutive years and the error model was fitted using the home level information. In fact, the only time in which the risk model showed a significant effect was when we used the radon exposure as quantified by Krewski et al. (2006) in the Iowa study. The effect was increased for the model fitted with adjustment by categories of age, smoking duration and smoking rate. From the correction methods, the Bayesian correction was the one that had a largest effect of indoor radon exposure. The Connecticut and Utah/South Idaho study also had alpha-track detectors in different home levels but the participants had lived in more than one home. Individuals from the Connecticut study had a negative risk estimate, even after correcting the indoor radon exposure measurements and fitting the adjusted risk model. In the case of the individuals from the Utah/South Idaho study, the Bayesian risk estimate was positive as well as the estimates from the adjusted risk model for the uncorrected and corrected data. In neither case a significant effect was found. For subjects that had indoor radon exposure measured in the home rooms (kitchen and bedroom for Missouri; bedroom and bathroom for Winnipeg), all point estimates were negative, except the one obtained for the unadjusted model using the Bayesian correction and the Missouri data. Once again, no statistical effect of radon exposure in lung cancer risk was found. This negative estimate and no effect trend was similar for the pooled data analysis. Based on the results from the c-index, most of the models had a better discrimination ability when we adjusted by covariates.

5.2 Contributions, Advantages and Disadvantages

This dissertation provides some contributions to the epidemiology field. To the best of our knowledge, this pooled data set has not being been analyzed using methods that correct for measurement error. We were able to analyze to correct for measurement error by using regression calibration, SIMEX and the proposed Bayesian method. Moreover, this proposed method fits a joint model so all parameters are modeled simultaneously. The Bayesian model also implements a new method for analyzing missing values in this pooled data by assigning random effects to those missing measurements. We also contribute to the comparison of the performance of regression calibration and SIMEX as has been done before (Allodji et al., 2012). Those contributions are extended by adding the comparison of regression calibration and SIMEX to the Bayesian model.

There are a few advantages of the proposed method to the existing ones. First, we can incorporate information from previous studies in the analysis at hand rather than treat the problem as an isolated one. Such previous information could come from the detector measurement error, European studies, a conference, meetings, descriptive studies, or even previous frequentist inferences. As with any other Bayesian analysis, we are also able to make inferences based on the observed data and make probability statements about quantities of interest.

While this new method is providing new contributions to he epidemiological and statistical field, it also has some disadvantages. The implementation of the method could be computationally intensive. Furthermore, like any Bayesian analysis, it could be criticized for being subjective when choosing the priors.

5.3 Future Work

Although this dissertation provides contributions to the epidemiology and statistical fields, more work can be done. In the case of the application, bootstrap could be used to adjust the standard errors in order to account for the estimation of the parameters in the calibration equation. This adjustment is suggested by Carroll et al. (2006) but in reality is not implemented by most of the applied problems that correct for measurement error using regression calibration. Moreover, a model selection can be done to determine if there are other covariates that should be included in the model. In terms of the statistical approach, a sensitivity analysis can be done to explore how much the prior choices affect the results, if they do. We could also expand the simulation scenarios to include different numbers of cases and controls for each study and even simulating the 6-30 exposure time window used in the application presented before.

REFERENCES

Alavanja, M. C., R. C. Brownson, J. H. Lubin, E. Berger, J. Chang, and J. D. Boice (1994). Residential radon exposure and lung cancer among nonsmoking women. *Journal of the National Cancer Institute 86*(24), 1829–1837.

Allodji, R. S., B. Schwartz, I. Diallo, C. Agbovon, D. Laurier, and F. de Vathaire (2015). Simulation–extrapolation method to address errors in atomic bomb survivor dosimetry on solid cancer and leukaemia mortality risk estimates, 1950–2003. *Radiation and environmental biophysics 54*(3), 273–283.

Allodji, R. S., A. Thiébaut, K. Leuraud, E. Rage, S. Henry, D. Laurier, and J. Bénichou (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*(30), 4428–4443.

Bäverstam, U. and G.-A. Swedjemark (1991). Where are the errors when we estimate radon exposure in retrospect? *Radiation Protection Dosimetry 36*(2-4), 107–112.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics 7*(4), 434–455.

Canty, A. and B. D. Ripley (2016). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-18.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Chu, H., L. Nie, and S. R. Cole (2011). Estimating the relative excess risk due to interaction: a bayesian approach. *Epidemiology 22*(2), 242–248.

Cook, J. R. and L. A. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association 89*(428), 1314–1328.

Council, N. R. (1988). *Health risks of radon and other internally deposited alpha-emitters: BEIR IV*. National Academies Press.

Cowles, M. K. (2013). *Applied Bayesian statistics: with R and OpenBUGS examples*, Volume 98. Springer Science & Business Media.

Darby, S., D. Hill, H. Deo, A. Auvinen, J. M. Barros-Dios, H. Baysson, F. Bochicchio, R. Falk, S. Farchi, A. Figueiras, et al. (2006). Residential radon and lung cancerdetailed results of a collaborative analysis of individual data on 7148 persons with lung cancer and 14 208 persons without lung cancer from 13 epidemiologic studies in europe. *Scandinavian journal of work, environment & health*, 1–84.

Darby, S., E. Whitley, P. Silcocks, B. Thakrar, M. Green, P. Lomas, J. Miles, G. Reeves, T. Fearn, and R. Doll (1998). Risk of lung cancer associated with residential radon exposure in south-west england: a case-control study. *British Journal of Cancer 78*(3), 394.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press. ISBN 0-521-57391-2.

Fearn, T., D. Hill, and S. Darby (2008). Measurement error in the explanatory variable of a binary regression: regression calibration and integrated conditional likelihood in studies of residential radon and lung cancer. *Statistics in medicine 27*(12), 2159–2176.

Field, R. W., D. J. Steck, B. J. Smith, C. P. Brus, E. L. Fisher, J. S. Neuberger, C. E. Platz, R. A. Robinson, R. F. Woolson, and C. F. Lynch (2000). Residential radon gas exposure and lung cancer the iowa radon lung cancer study. *American Journal of Epidemiology 151*(11), 1091–1102.

Fuller, W. A. (2009). *Measurement error models*, Volume 305. John Wiley & Sons.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*(1), 29–36.

Heid, I., H. Küchenhoff, J. Miles, L. Kreienbrock, and H. Wichmann (2004). Two dimensions of measurement error: classical and berkson error in residential radon exposure assessment. *Journal of Exposure Science and Environmental Epidemiology 14*(5), 365–377.

Heid, I., H. Küchenhoff, J. Wellmann, M. Gerken, L. Kreienbrock, and H. Wichmann (2002). On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. *Statistics in medicine 21*(21), 3261–3278.

Hornung, R. W. and T. J. Meinhardt (1987). Quantitative risk assessment of lung cancer in us uranium miners. *Health Physics 52*(4), 417–430.

Krewski, D., J. H. Lubin, J. M. Zielinski, M. Alavanja, V. S. Catalan, R. William Field, J. B. Klotz, E. G. Létourneau, C. F. Lynch, J. L. Lyon, et al. (2006). A combined analysis of north american case-control studies of residential radon and lung cancer. *Journal of Toxicology and Environmental Health, Part A 69*(7-8), 533–597.

Letourneau, E., D. Krewski, N. Choi, M. Goddard, R. McGregor, J. Zielinski, and J. Du (1994). Case-control study of residential radon and lung cancer in winnipeg, manitoba, canada. *American journal of epidemiology 140*(4), 310–322.

Lubin, J., J. Boice Jr, C. Edling, R. Hornung, G. Howe, et al. (1994). Lung cancer following radon exposure among underground miners: a joint analysis of 11 studies. *Washington, DC: US Govt Print Off*.

Lubin, J., Z. Wang, L. Wang, J. Boice Jr, H. Cui, S. Zhang, S. Conrath, Y. Xia, B. Shang, J. Cao, et al. (2005). Adjusting lung cancer risks for temporal and spatial variations in radon concentration in dwellings in gansu province, china. *Radiation research 163*(5), 571–579.

Lubin, J. H., J. D. Boice, C. Edling, R. W. Hornung, G. R. Howe, E. Kunz, R. A. Kusiak, H. I. Morrison, E. P. Radford, J. M. Samet, et al. (1995). Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. *Journal of the National Cancer Institute 87*(11), 817–827.

Lubin, J. H., J. D. Boice Jr, and J. M. Samet (1995). Errors in exposure assessment, statistical power and the interpretation of residential radon studies. *Radiation research 144*(3), 329–341.

Lubin, J. H., J. M. Samet, and C. Weinberg (1990). Design issues in epidemiologic studies of indoor exposure to rn and risk of lung cancer. *Health Physics 59*(6), 807–817.

McGregor, R., P. Vasudev, E. Letourneau, R. McCullough, F. Prantl, and H. Taniguchi (1980). Background concentrations of radon and radon daughters in canadian homes. *Health Physics 39*(2), 285–289.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News 6*(1), 7–11.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radford, E. P. and K. S. C. Renard (1984). Lung cancer in swedish iron miners exposed to low doses of radon daughters. *New England Journal of Medicine 310*(23), 1485–1494.

Sandler, D. (1999). Indoor radon and lung cancer risk: a case-control study in connecticut and utah. In *Radiation Research*, Volume 151, pp. 103–104. RADIATION RESEARCH SOC 2021 SPRING RD, STE 600, OAK BROOK, IL 60521 USA.

SAS Institute Inc (2002-2012). *User's Guide*. NC, USA.

Souza, A. D. and H. S. Migon (2004). Bayesian binary regression model: an application to in-hospital death after ami prediction. *Pesquisa Operacional 24*(2), 253–267.

Spiegelman, D., A. McDermott, and B. Rosner (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American journal of clinical nutrition 65*(4), 1179S–1186S.

Zhang, W., K. Chaloner, M. K. Cowles, Y. Zhang, and J. T. Stapleton (2008). A bayesian analysis of doubly censored data using a hierarchical cox model. *Statistics in medicine 27*(4), 529.