LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Institut für Statistik

Measurement Error and Study Design in Air Pollution Epidemiology:
Impacts and Recommendations

Bachelor's Thesis

Author: Elisabeth Pangratz

Supervisor: Prof. Dr. Helmut Küchenhoff, Institut für Statistik, LMU
Dr. Veronika Deffner, Institut für Statistik, LMU

Submission date: 14.12.2017

Abstract

Exposure measurement error in epidemiologic air pollution studies can have a severe impact on a study's results as it may introduce bias in health effect estimates and lowers the statistical power. Measurement error is common in air pollution epidemiology as many air pollution studies use ambient fixed site measurements instead of personal air pollution exposure measurements because of lower costs and easier access. While there are sophisticated methods to correct for measurement error during the analysis of a study, this thesis concentrates on how to take measurement error into account during the design stage of a study. Furthermore, it aims at providing a comprehensive review of relevant literature. It regards various aspects concerning the impact of measurement error on the study design and gives recommendations for adjusting the study design to the presence of measurement error.

When designing a study, investigators have to account for a trade-off between using accurate, but costly, measurement methods and a sufficiently large sample size in order to achieve precise health effect estimates and adequate statistical power. As studies have to keep cost constraints, the allocation of resources has to be carefully considered to attain cost-efficient designs. Since taking personal exposure measurements can be many times more expensive and more labour intensive than using ambient monitor measurements, the only use of personal measurements can lead to very small sample sizes when keeping cost constraints. A good trade-off can often be achieved by conducting a validation study additionally to the main study. The primary method for adjusting the study design to measurement error is to increase the sample size. Adjusted sample size calculations for air pollution studies have to be based on realistic assumptions and require correctly specified measurement error models. Further investigations are needed to provide flexible calculation methods and easy access of methods by software implementations.

Contents

1 Introduction

Over the past decades, various studies indicate that exposure to elevated concentrations of air pollution may pose a significant risk on human health (Dockery et al. 1993, p. 1753). Studies in air pollution epidemiology are prone to exposure measurement error. The National Research Council Committee on Research Priorities for Airborne Particulate Matter defines measurement error as follows: "The difference between actual exposures and measured ambient-air concentrations is termed measurement error. Measurement error can occur when measures of ambient air pollution are used as an index for personal exposure." (National Research Council 2001, p. 125).
As it is well known that measurement error in explanatory variables can distort the statistical analysis of the data, the impact and handling of measurement error is part of the most relevant research topics regarding air pollution studies. This thesis is related to the STRATOS initiative (http://stratos-initiative.org/) that has the objective to provide accessible and precise guidance in the design and analysis of observational studies. A group of the initiative (Freedman, Kipnis, Carroll, Deffner, Dodd, Gustafson, Keogh, Küchenhoff, Shaw, Tooze) addresses the topic of measurement error and misclassification.
Measurement error has to be carefully considered in every stage of a study. Generally, one can take steps to limit the degree of measurement error by a study's design, including the conduct of a validation study to determine the structure and extent of the error, and by correcting for measurement error in the statistical analysis. This work concentrates on how to take exposure measurement error into account at the design stage of an epidemiologic air pollution study.
Measurement error is a fundamental challenge in air pollution epidemiology. Available ambient air pollution data and an individual's air pollution exposure can differ substantially. Personal air pollution exposure is rarely if ever measured because it is costly, labour intensive and intrusive for the study subjects. Therefore, the presence of measurement error can be expected in the majority of air pollution studies (Dockery 1993, p. 187). Having inaccurate exposure measurements instead of true exposure data has the potential to bias the inference by attenuating effect coefficients and lessens statistical power to detect truly present associations. If measurement error has a critical magnitude, it can be severe enough to totally negate a study's ability of draw valid conclusions about the exposure effect (Armstrong 1998, p. 654). The power loss by

measurement error leads to the requirement of increasing the sample size in order to achieve equal power as one would have obtained with accurate data. It is therefore of crucial importance to consider measurement error already in the design of a study (Dockery 1993, p. 187). In fact, despite of the wide occurrence and the severe impacts of measurement error, only few epidemiologic investigators take measurement error into account (Shaw et al. 2017).

This work aims at giving a comprehensive review of literature that addresses the study design with regard to measurement error. It regards various aspects concerning the impact of measurement error on the study design and gives recommendations for designing a study in the presence of measurement error. General results for epidemiologic studies are discussed in the context of air pollution studies.

Chapter 2 provides relevant background information concerning air pollution studies, their study designs and the methods of air pollution measurement. Information about measurement error in the context of air pollution studies is given in Chapter 3. The impact of measurement error on effective sample size is described in chapter 4 by regarding the formula of the asymptotic relative efficiency (ARE) for different measurement error models (section 4.1) and by pointing out relevant results for air pollution studies (section 4.2). Chapter 5 provides information on how to take measurement error into account during the design of a study. It consists of three sections. The first section refers to the topic of data collection including considerations about how to choose between different measurement methods. Section 5.2 deals with validation studies which are essential for gaining information about the measurement error in order to adjust the study design or to correct for measurement error by statistical methods. This section contains three sub-sections that discuss the general topic of validation studies as well as the validation study design and the application of validation studies in air pollution epidemiology. Section 5.3 presents different methods for sample size adjustment in the presence of measurement error and reviews the relevance of different approaches for air pollution studies. The thesis concludes with a summary of the considered topics and a discussion that refers to further aspects and literature.

2 Air Pollution Epidemiology

Epidemiology is defined as the ”distribution and determinants of disease frequency in man.” (MacMahon and Pugh 1970). Regarding the field of air pollution epidemiology, scientists investigate the effect of air pollution exposure on human diseases. This chapter provides basic information about air pollution studies by first pointing out general characteristics of air pollution studies with a following explanation of the most common study designs and methods of air pollution measurement.

2.1 Characteristics of Air Pollution Studies

This chapter gives a general introduction into basic characteristics of studies in air pollution epidemiology.

The outcomes of air pollution studies are health variables that can be for example changes in human health determinants, such as vital capacity, lung growth and symptom severity, or events such as the onset of symptoms or death. The exposure of interest, that is assumed to have an impact on the health outcome, can be one or several air pollutants. Some studies are interested in the effect of short-term exposure, whereas others intend to examine the effect of long-term exposure. Acute health effects are temporary and caused by time-varying exposure. On the contrary, chronic health effects are likely to occur due to cumulative exposure or due to complex functions of exposure in a lifetime. The scale of the outcome and the nature of the health effect of interest affect the choice of an appropriate study design (Dominici et al. 2003, p. 245). Modern air pollution studies use advanced statistical models for the analysis of complex air pollution datasets, for example generalized linear models (GLMs), generalized additive models (GAMs), Cox proportional hazard models and hierarchical models (Dominici et al. 2003, p. 244).

There are many factors and difficulties that have to be considered in air pollution epidemiology. First of all, ambient air pollution varies over space and time. In addition, personal exposure to air pollution is not only dependent on ambient pollution but also on individual activity-patterns and individual environments. Furthermore, several confounding variables and the correlation among covariates have to be taken into account. For many diseases the exposure to air pollution is only one of many factors (Dockery 1993, p. 190). Also, when being interested in the effect of a single pollutant, one has

to be aware of the high inter-correlation between the pollutant of interest and other pollutants as ambient air consists of a complex composition of various air pollutants (Dominici et al. 2003, p. 245). Exposure to air pollution is omnipresent and there is no unexposed population which makes it more difficult to evaluate health effects. Moreover, the health effect of air pollution exposure is relatively weak which demands large resources to detect the effects. Additionally, the precise measurement of an individual's exposure to air pollution can be infeasible as measuring personal exposure is very costly. Therefore, many studies make use of surrogates such as ambient exposure data coming from fixed monitors. Thanks to the many existing monitoring networks, investigators have easy access to rich data with a high resolution in time. However, in many applications, monitor data cannot adequately represent personal exposure which leads to measurement error in many air pollution studies (Sheppard et al. 2012, p. 208).

2.2 Study Designs

This section gives an overview and brief description of frequently used study designs in air pollution studies. The broad majority of air pollution studies are observational studies. Many studies are also "opportunistic" studies since they emerge out of the possibility of the use of data that was collected for other reasons. Thus, they can be often limited by the structure of the available data. The most common study designs in air pollution epidemiology are time series, case-crossover, cohort and panel studies (Dominici et al. 2003, p. 244 f.). Case-control studies air pollution epidemiology are rather rare (Dockery 1993, p. 189).

In a time series study, one expects the day-to-day variability of the exposure to be associated with the day-to-day variability of the outcome rate (Nieuwenhuijsen 2015, p. 5). A time series study in air pollution epidemiology typically examines the relationship between aggregated outcomes, typically population-averaged disease outcomes, and spatial aggregated exposure data, typically a daily average ambient concentration of a pollutant. Accordingly, time series studies can be assigned to the category of ecologic studies (Rothman et al. 2008, p. 511 ff.). Ecologic studies use aggregated data leading to "ecological inferences" when conclusions about individuals are drawn from data on an aggregated level. When using aggregated data, one has to be aware of the so-called "ecological fallacy" which happens when inferences on the group-level don't hold for the individual-level (Freedman 1999).

Case-crossover studies can be used to estimate the risk of a rare health event caused by short-term exposure to air pollution. Each case represents his own control as one compares the exposure at the time just prior to the event, the "case time", with the exposure

of several reference times, the “control time”. By this, the constant characteristics of an individual are matched automatically which reduces confounding (Dominici 2003 et al., p. 247). This study design is most useful for intermittent exposures that cause an immediate and transient risk on a rare outcome (Maclure and Mittleman 2000, p. 193).

In contrast to time series, that often concentrate on the day-to-day variation in air pollution and the health outcome with regard to a single location, cohort studies compare the relationship between air pollution and outcome across several geographical locations (Dominici 2003, p. 270). In a cohort study the investigator defines a group of subjects that are heterogeneous in their exposure and follows them over time to record if they develop a disease or health outcome of interest. Finally, risk estimates can be gained by comparing their incidence rates (Nieuwenhuijsen 2015, p. 5). Cohort studies in air pollution epidemiology can be used to estimate the effect of long-term air pollution exposure on the disease outcome. Frequently, cumulative air pollution exposure is used as the exposure variable.

A panel study follows a cohort of study subjects over a certain period to explore changes in the health outcome of interest by taking repeated measurements at different points of time. This design is appropriate if the outcome for a subject varies over time. Furthermore, this design is practical for the investigation of short-term effects of air pollution for a susceptible subpopulation (Dominici et al. 2003, p. 249).

The design of a study always depends on the specific research target of the study and the structure of the data. In summary, time series, panel and case-crossover studies are useful to study the acute effects of short-term exposures, whereas the cohort study design studies a combination of acute and chronic effects of longer exposure times (Dominici et al. 2003, p. 525). The outcomes in cohort and panel studies can be either event counts or continuous measurements , while the outcomes of time series and case-crossover studies should be rare events. The time series study design evaluates aggregated exposures and event counts over a large population, whereas cohort, panel and case-crossover studies include the cases individually in the analysis and often orientate their study on well-defined subgroups (Dominici et al. 2003, p. 252 f.). In general, it can be found that many air pollution studies are semi-ecologic (Dockery 1993, p. 188). Truly ecologic studies observe both exposure and health outcome on an aggregated level. The semi-ecologic, also called semi-individual, study design observes health outcomes on an individual-level and exposure on an aggregated level, measured by a fixed site monitor. In contrasts to ecologic studies that draw conclusions by ecological inference and are therefore prone to ecological fallacy, semi-ecologic studies share their inferential properties with typical individual-level study designs. Just like

other study designs, the semi-ecologic design has to deal with measurement error in the exposure variable (Künzli and Tager 1997).

2.3 Air Pollution Measurement

Ambient fixed site monitors are the most common source of data in air pollution studies. Ambient air pollution concentrations measured at fixed sites are therefore often used as surrogate measurements for personal air pollution exposure. As they are only proxies for the personal exposure, they cannot represent it perfectly which causes a certain level of measurement error. The main reasons for this are the low spatial resolution of fixed site measurements as well as their inability of accounting for the differences in the exposure among the study subjects (Nieuwenhuijsen 2015, p. 252). Nevertheless, ambient monitors have the advantage of providing easy access to extensive exposure data with high temporal resolution, for example daily or hourly measurements (Sheppard et al. 2012, p. 209).

Besides ambient fixed site monitor measurements, other surrogates for personal air pollution exposure are used in air pollution studies, like distance to major roads or traffic intensity on the residential road. Predictions from models that predict exposure for individuals are also used as surrogates measurement for personal exposure (Jerrett et al. 2005, p. 186 ff.). Dionisio et al. (2016) provide a review of such alternative exposure metrics including GIS-based metrics, satellite data, air quality modeling using emissions and meteorological data and human exposure models containing human activity data.

Personal exposure to air pollution can be measured by attaching an exposure monitor or sensor to the person. It is clear that this method of data collection is more directly representative for an individual and thus more informative compared to fixed site measurements or other surrogates (Nieuwenhuijsen 2015, p. 87). However, the universal use of this measurement method is limited. Personal measurements are usually much more expensive than fixed site measurements while many studies have to meet fixed cost constraints. Furthermore, they are more labour intensive. Thus, it is mostly not possible to conduct personal monitoring for a large number of persons while a sufficiently high sample size is essential for a study's quality. Therefore, studies using only personal exposure measurements are rare and most investigators utilize this costly measurement method only for a subpopulation of the overall study population to validate their main measurement method. This issue will be further discussed in section 5.2.

The exposure to air pollution of an individual is complicated to capture by fixed site ambient monitors as it is driven by many factors like movement and activity patterns.

De Nazelle investigates nitrogen dioxide (NO_2) data for different activity spaces by linking space-time air pollution mapping with information about physical activity and geographic location obtained by smartphones for 36 adults in Barcelona (De Nazelle et al. 2013, p. 96).

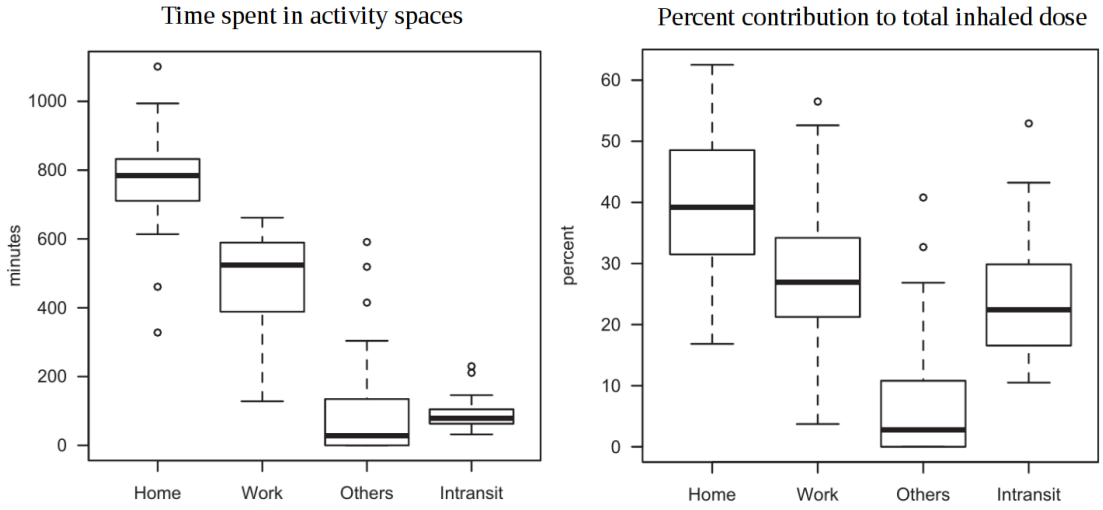

Figure 1: Time spent in activity spaces and percent contribution to total daily inhaled dose of NO_2 from different activity spaces (cf. De Nazelle et al. 2013, p. 96).

Figure 1 shows that the study subjects face the highest NO_2 concentration when they are in transit. Time spent at home which takes 51% of people's daily time contributes to 40% of inhaled NO_2, whereas time at work which occupies 33% of the day contributes to 28% of the inhaled dose. In contrast, time in transit which represents only 6% of the people's daily time accounts for 24% of daily inhaled NO_2 (De Nazelle et al. 2013, p. 95). Therefore, it seems reasonable that fixed site monitoring's ability to represent an individual's air pollution exposure is considerably limited and that individual measurements can provide valuable information in air pollution epidemiology. The discrepancy between personal and fixed site measurements can be expressed by their correlation. Several studies investigate this relationship. Avery et al. (2010) provides a review of studies that examine the within-participant ambient-personal correlation of $PM_{2.5}$. Their work includes 18 studies, published between 1999 and 2008 in 17 cities in Europe and North America, that are found by searching seven electronic reference databases. The studies contain in total 619 non-smoking participants with the age of 6 to 93 years (Avery et al. 2010, p. 218). Avery et al. 2010 give a comprehensive comparison of the study results. The pairs per participant of ambient-personal $PM_{2.5}$ measurements range from 5 to 20 with a median of 8 and were collected over 27 to 547 days. The studies calculate the correlations with either the coefficient by Pearson or

the coefficient by Spearman. The following figure 2 shows the correlation coefficients for the different studies. Some studies have more than one correlation coefficient as they include sub-studies. The within-participant correlation between ambient $PM_{2.5}$ and personal $PM_{2.5}$ is denoted by r (Avery et al. 2010, p. 216).

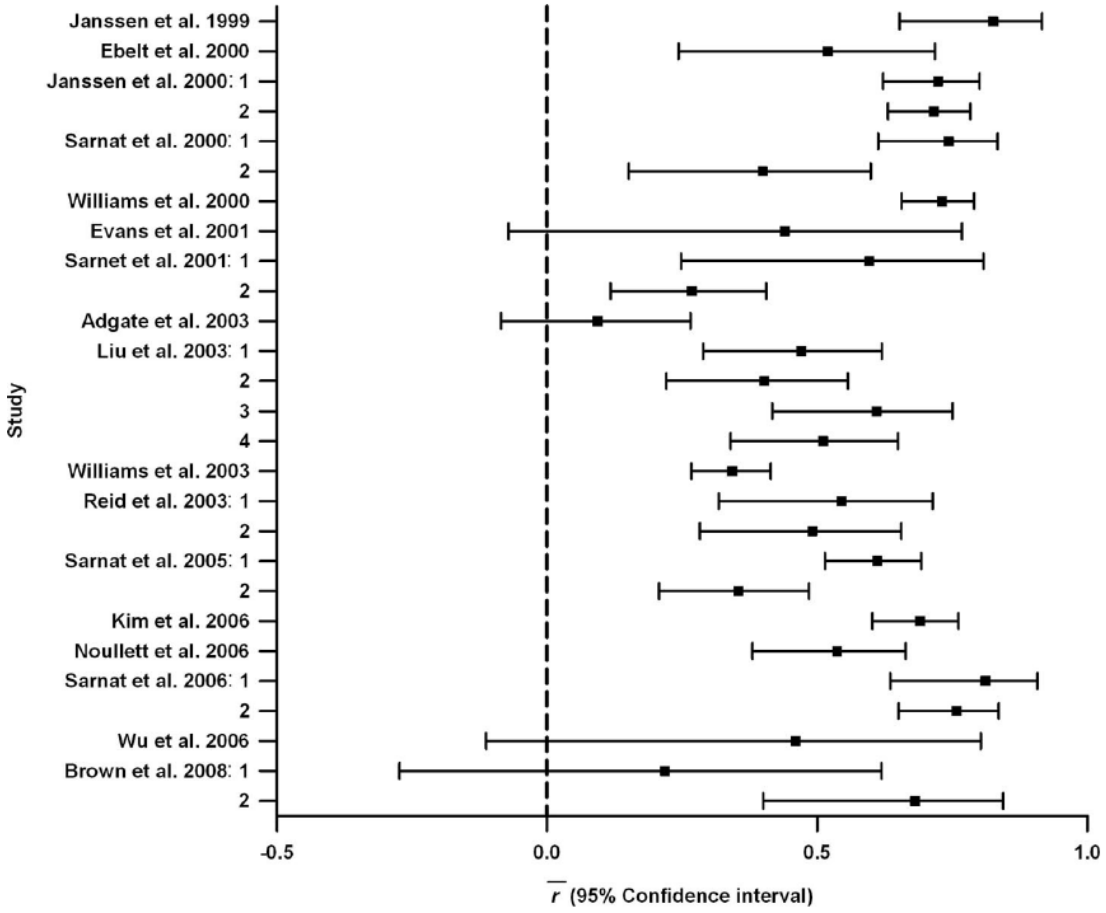

Figure 2: Estimates of $\bar{r}$ (95% confidence interval) of the within-participant correlation between ambient and personal $PM_{2.5}$ for 18 studies (Avery et al. 2010, p. 220).

The correlation coefficients in the 18 studies vary with significant heterogeneity from 0.09 to 0.83 with a median of 0.54. Moreover, Avery finds that correlations are generally higher for higher ambient $PM_{2.5}$ concentrations, higher relative humidity, studies in Europe and with eastern longitudes in North America and lower between-participant variation (Avery et al. 2010, p. 2018 f.). One should note that Avery only reviews $PM_{2.5}$-related studies without regard to other pollutants. In summary, the observed correlation coefficients of different studies vary a lot with sometimes critically low values of the correlation between personal and ambient air pollution. This leads to the conclusion that studies using ambient fixed site measurements instead of personal measurements may suffer from a considerable amount of measurement error.

The fact that fixed site monitoring is the more frequently used method of data collection leads to the necessity of investigating the impact of measurement error in air pollution epidemiology and how to deal with it.

3 Measurement Error

This chapter deals with basic concepts of measurement error with regard to air pollution studies. The first section (3.1) introduces common measurement error types. In the second section (3.2), relevant factors related to measurement error in air pollution studies are explained. Furthermore, as the impact and structure of the measurement error depends on the study design, the different types of error appearing in air pollution studies will be considered separately for aggregated-level studies and individual-level studies in section (3.3). Aggregated level studies examine the association between spatial aggregated outcomes and spatial aggregated exposure data, whereas individual level studies regard the individual exposure of the subjects along with their individual outcome.

3.1 Measurement Error Types

This chapter gives a brief explanation of the basic measurement error models as these terms are fundamentally for the following issues that will be discussed in this thesis.
Generally, measurement error can arise in the outcome variable or in the covariates. While simple regression models naturally handle outcome error since it is represented in the error term of the regression, they assume all explanatory variables are known and measured without error (Sheppard et al. 2012, p. 209). Thus, literature mostly focuses on error in explanatory variables. Those can be differentiated between the variable of interest, which is in epidemiology the exposure that is presumed to have an impact on health, the potential confounders or the potential interaction-variables (Armstrong 1998, p. 651). This thesis concentrates on error in the exposure variable. Frequently used terms for errors in exposure variables are “measurement error” and “misclassification”. The latter term applies to discrete exposure variables (Carroll et al. 2006, p. 32). Since exposure measurements in air pollution epidemiology are continuous, misclassification will not longer be discussed. Measurement error is described by measurement error models. The two most common models are the classical measurement error model and the Berkson error model. They are statistically defined

as (Carroll 2006, p. 3 ff.):

$$1.\ \text{The classical model:}\quad X^* = X + U \tag{1}$$

$$2.\ \text{The Berkson model:}\quad X = X^* + U, \tag{2}$$

where X^* is the error-prone and X the true value of the exposure. U is a random variable which is for the classical model independent of X and $E(U \mid X) = 0$ and for the Berkson model independent of X^* and $E(U \mid X^*) = 0$. Thus, in this case, X^* is a unbiased measure of X. Measurement error is likely to be classical if an error-prone variable is measured uniquely for an individual, whereas it is likely to be of Berkson type if individuals are assigned to a group and given the same value of the error-prone measurement. For classical errors it holds that $Var(X^*) > Var(X)$, whereas for Berkson error it holds that $Var(X) > Var(X^*)$ (Carroll 2006, p. 27 f.).
The error in the mismeasured exposure can either be random, such that some exposure values are underestimated and others overestimated with the true value as their mean, like in the equations above, or it can differ systematically from the true value which is called systematic error (Armstrong 1998, p. 652). For measurement error that is not random but linear, one can extend the classical measurement error model to a linear measurement error model such that the error-prone variable depends linearly on the true value (Buonaccorsi 2010, p. 151):

$$X^* = \alpha_0 + \alpha_X X + U, \tag{3}$$

where U is again the component that represents the random error with mean zero. There are several further extensions for measurement error models, for example specific assumptions for the random component U, that will not be mentioned in this context.
Another important characteristic of the measurement error is whether it is differential or non-differential. Measurement error is called non-differential if the distribution of the outcome Y given (X, Z, X^*) depends only on (X, Z) with Z denoting other error-free covariates. Consequently, variables measured with non-differential error X^* contain no information about the outcome Y other than what is already included in X and Z. Otherwise, measurement error is differential (Carroll et al. 2006, p. 36). A typical example for differential error is the so called "recall bias" in case-control studies where the exposure of the cases is observed with an error that is different from the controls. In contrast, in studies in which exposure is measured previous to the outcome, the error is typically non-differential (Nieuwenhuijsen 2015, p. 202).
For the further understanding, it is important to consider the differences between clas-

sical and Berkson error of continuous explanatory variables regarding their effect on study results. By using X^* as a surrogate of X, assuming single-covariate linear regression and classical, non-differential measurement error, the estimate of the regression coefficient β_{X^*} is a biased estimate of the true coefficient β_X. This result is also called "attenuation" as $|\beta_{X^*}| \leq |\beta_X|$. Unbiased estimates, $|\beta_{X^*}| = |\beta_X|$, occur only for $\beta_X = 0$. That is, when testing the null hypothesis $\beta_{X^*} = 0$, classical measurement error doesn't lead to spurious significant when there is truly no association. Significance tests for testing the null hypothesis $\beta_{X^*} = 0$ remain therefore valid tests of the hypothesis that $\beta_X = 0$ (Zeger et al. 2000, p. 420 f.). In contrast, for non-differential Berkson error and linear regression, $\hat{\beta}_{X^*}$ is an unbiased estimate of β_X. Both types of random non-differential error, classical and Berkson error, reduce a study's statistical power to detect whether β_X is different from zero. That is, the chance to detect truly present significant associations is lower (Armstrong 1998, p. 653 f.). For differential error assuming linear regression and a classical or linear measurement error model, the relation $|\beta_{X^*}| \leq |\beta_X|$ doesn't longer hold as the effect estimates can be biased either downwards or upwards. Significant association can be found even when there is truly no significant association, leading to invalid significant tests for $H_0 : \beta_X = 0$ (Freedman et al. 2017, p. 8).

3.2 Measurement Error in Air Pollution Studies

This chapter focuses on measurement error in air pollution exposure. As described in chapter 2.3, the correlation between personal exposure and ambient fixed monitor measurements is in many cases inadequately low. Thus, when using ambient measurements instead of personal measurements, assuming that individual exposure is the exposure of interest and ambient concentration the surrogate measure, a certain extent of measurement error emerges in the exposure variable.

The measurement error is a result of several factors. An individual's total exposure does not only come from ambient sources, but also from non-ambient sources. Total personal exposure for a subject i at a time point t, $X_{it}^{personal}$, can therefore be described by (Dominici et al. 2003, p. 260):

$$X_{it}^{personal} = X_{it}^{non-ambient} + \alpha_{it} X_{it}^{ambient}. \tag{4}$$

In this equation, $X_{it}^{non-ambient}$ stands for the non-ambient source exposure of the individual i at time t and $X_{it}^{ambient}$ for the ambient concentration at the individual i's spatial location at time t. The parameter α_{it} is an infiltration parameter that stands for the fraction of ambient air pollution concentration that the subject i is exposed to at time t.

It depends on several factors such as outdoor air penetration to indoor environments, the subject's individual activity patterns and pollutant-specific features (Dominici et al. 2003, p. 260).
Furthermore, one needs to consider that ambient data of air pollution concentrations can be prone to instrumental error that is caused by inaccurate measurements of fixed site monitors. There are different types of fixed site monitors that differ in their accuracy and time resolution (Dominici et al. 2003, p. 268).
Figure (3) shows a schematic for the relationship between measured ambient air pollution concentration (X_t^*) and personal exposure (X_{it}) that are connected by the true ambient pollution (X_t') and the indoor exposure (W_{it}) (cf. Zeger et al. 2000 p. 423).

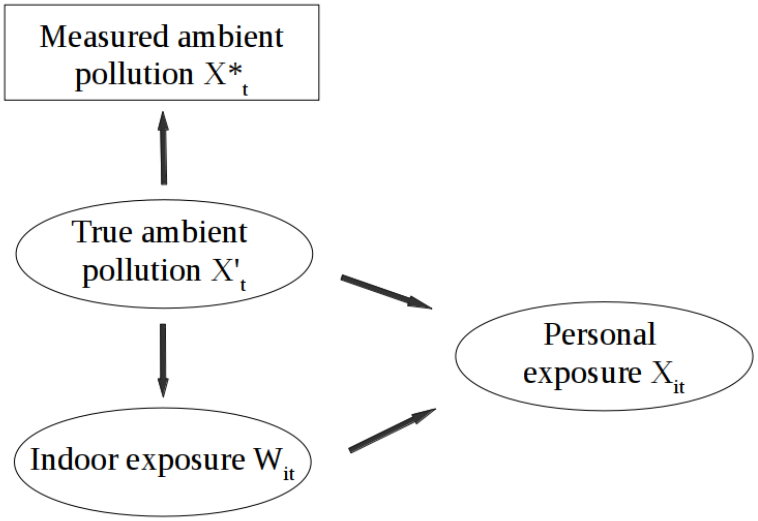

Figure 3: Schematic relating ambient measured pollution level (X_t^*) to personal exposure (X_{it}) by true ambient pollution (X_t') and indoor exposure (W_{it}) (cf. Zeger et al. 2000, p. 423).

Measurement error in air pollution studies is influenced by several study-specific factors. Especially when personal exposure varies a lot among the study subjects, although study subject live in the same area, stationary monitor measurements cannot easily represent the personal concentrations. Such large variation in personal exposures can be caused by outdoor motor vehicle concentrations that are typically very heterogeneous in space. For example, Suh and Zanobetti (2010) examine the adequacy of using ambient measurements for analysing the relationship between traffic-related air pollution and heart rate variability in a study of 30 people living in Atlanta. They conclude that measurement error caused by the usage of ambient measurements instead of personal measurements masks the effect of traffic-related pollution on the outcome. While ambient elemental carbon (EC) and ambient nitrogen dioxide (NO_2) have no

significant effect on the heart rate variability, their respective personal measurements show a significant negative association (Sou and Zanobetti 2010, p. 687 ff.). This indicates that for traffic pollutants ambient concentrations might be a poor surrogate for personal exposure and thus lead to crucial bias.

Another factor that introduces variability in personal exposures is home ventilation. Sarnat et al. (2000) describe a study in Baltimore with 15 non-smoking elderly subjects that wear a multi-pollutant sampler for 12 days during the summer of 1998 and the winter of 1999. They conclude that indoor ventilation is a substantial determinant of personal exposure and of the personal-ambient associations. Whereas this association is strong for individuals that spend a lot of their time in well-ventilated environments, it decreases for people that spend most their time in poorly ventilated environments (Sarnat et al. 2000, p. 1195 f.). This leads to the conclusion that ambient measurements are better proxies for personal exposure when the degree of ventilation for the subjects is high compared to when ventilation levels are low.

Generally, measurement error varies according to space and time. There exist crucial pollutant-specific differences concerning measurement error when using ambient monitor data instead of personal measurements. Some pollutants, like total $PM_{2.5}$ and Ozone (O_3), are homogeneous over space. Other pollutants, especially those caused by traffic, such as carbon monoxide (CO) and nitrogen dioxide (NO_2), are characterized by spatio-temporal variability (Sarnat et al. 2010, p. 135). Sarnat figures out that the effect of monitor site location and distance between sites and study population on estimated relative risks depends on whether the pollutant of interest is homogeneous or heterogeneous over space. For homogeneous pollutants similar relative risk estimates are found regardless of monitor site and distance between site and study population, whereas different relative risk estimates are found for heterogeneous pollutants (Sarnat et al. 2010, p. 144).

While much literature concentrates on the difference between ambient and individual exposure, it is also important to consider the problem of spatial misalignment. Spatial misalignment is caused by data measured at different spatial resolutions. It is common in air pollution studies because exposure and outcome data are often collected independently. In cohort studies, spatial misalignment error occurs because exposure data measured by fixed site monitors has a much lower spatial resolution than the data of the health outcome which is often measured at the individual level. In time series designs, assuming region-wide counts of events like mortality or hospital admissions as the outcome variable and region-wide average pollutant levels as the explanatory variable, spatial misalignment error is caused by the difference between the measurements of the fixed site monitor, which is meant to represent the study population area, and the

true ambient average concentration in this area. The severity of spatial misalignment depends also on the heterogeneity or homogeneity of the pollutant of interest (Peng and Bell 2010, p. 720 f.).
In summary, there are many factors that play a role in the structure and the resulting impact of measurement error in air pollution studies.
Measurement error consists generally of a mixture of different error types which depends on the study design. This topic will be further discussed in the following chapter.

3.3 Measurement Error and Particular Study Designs

For different study designs, exposure measurement error comes in different dominant error-types. The impact of the measurement error is therefore also dependent of design-specific characteristic. This chapter points out the measurement error related differences between aggregated level study designs and individual level study designs. Aggregated level studies examine the association between spatial aggregated outcomes, typically daily event counts in a certain area, and spatial aggregated exposure data, typically a daily average ambient concentration of a pollutant in the same area. These studies can be used for investigating the short-term effects of air pollution on health. As mentioned in section 2.2, a typical study design for such studies is the time-series design. In the following, the particular types of measurement error appearing in such studies will be explained.
Zeger et al. (2000) present a conceptual framework for the evaluation of measurement error in air pollution time-series studies based on a log-linear regression model for risk. This model, which aims to explain an individual's risk of death on a given day by an individual's exposure, is aggregated to form a model whose outcome is the total deaths in a population. Such models, that are based on aggregated data, are used in many time-series studies. Presuming that the personal exposure X_{it} of an individual i at time point t is the true measure and the ambient concentration measured by an inexact monitor X_t^* at time point t is the surrogate measure, the difference between those two measures can be partitioned in three components (Zeger et al. 2000, p. 422):

$$X_{it} - X_t^* = (X_{it} - \bar{X}_t) + (\bar{X}_t - X_t') + (X_t' - X_t^*). \tag{5}$$

The true ambient air pollution concentration at time t is denoted by X_t'. The first part $(X_{it} - \bar{X}_t)$ represents the error caused by the use of averaged instead of individual exposure data. The second part $(\bar{X}_t - X_t')$ corresponds to the difference between averaged personal exposure and the true ambient air pollution concentration. The discrepancy between the true and the measured ambient levels is shown by the third part $(X_t' - X_t^*)$.

It includes instrumental error and spatial variation of ambient concentration levels.
The first term, by having the averaged instead of the individual personal exposure, behaves like Berkson error, whereas the second term $(\bar{X}_t - X'_t)$ is not Berksonian and thought to be a crucial source of bias. The third term, when using the average of available monitors in an area, is likely to be mainly of Berkson-type (Zeger 2000, p. 422). The difference between true ambient levels and average personal exposure $(\bar{X}_t - X'_t)$ is influenced by many factors as personal exposure contains also exposure from non-ambient sources (Zeger et al. 2000, p. 423).
Sheppard et al. (2005) make further investigations for air pollution time series studies regarding explicitly the personal exposure model that was described in section 3.2 as $X_{it}^{personal} = X_{it}^{non-ambient} + \alpha_{it} X_{it}^{ambient}$. They examine the role of the infiltration parameter α_{it}, the contributions of ambient and non-ambient sources exposures, and the spatial heterogeneity in ambient concentrations using a large data set of $PM_{2.5}$ in Seattle including measurements of personal, home-outdoor and fixed site monitors. They make several conclusions for sources of measurement error in time series studies. In summary, by assuming that ambient and non-ambient source exposures are independent of each other, they recommend to use average ambient concentrations of several monitors and conclude that ambient concentration measurements work well in time series studies as they are able to adequately summarize the time-varying population average exposure. By this, good estimates of the health-effect coefficient can obtained even if ambient measurements can only explain a small part of the total variation in personal exposure. This conclusion is, however, restricted to $PM_{2.5}$ which varies little over space (Sheppard et al. 2005, p. 375).
Similarly, Dominici et al. (2003) argue that ambient PM varies generally more over time than over space and that ignoring non-ambient sources of exposure in time series studies may not be problematic since ambient and non-ambient sources of PM are likely to be independent. They also conclude that ambient measurements are an appropriate surrogate for the average exposure of the study population in time series studies (Dominici et al. 2003, p. 268).
The case-crossover study design also uses temporal contrasts and investigates acute effects. Dionisio et al. (2016) concludes that for time-series and case-crossover designs fixed site measurements are adequate as they capture temporal variation. This holds especially for spatially homogeneous pollutants (Dionisio et al. 2016, p. 496 f.).
In summary, these results apply for study designs that driven by temporal contrasts and investigate short-period variations, for example daily variations of exposure and outcome on an aggregated level.
Individual level studies, in contrast to aggregated level studies, regard the individual

exposure of the subjects along with their individual outcome. This is often the case in cohort studies. As already explained in section 2.2, they can be used to estimate the effect of long-term air pollution exposure. Especially for long-term exposure measurements, collecting data with personal monitoring methods is infeasible (Van Roosbroeck et al. 2008, p. 409). Unless the personal monitor is worn the whole period in which exposure can affect disease, it's use is limited by the extent to which the period of personal monitoring is representative of the individual's true exposure (Nieuwenhuijsen 2015, p. 87). Generally, individual level studies attempt to contain sufficient variation in the exposure. Therefore such studies usually compare exposure-outcome associations across several geographical locations. This happens by including study subjects from different communities or different areas within a community (Sheppard et al. 2012, p. 204). The overall aim of those studies is to analyse the effect of individual exposure on individual disease outcomes.

Measurement error in exposure data representing individual exposure can be caused by little spatial resolution of the available air pollution exposure data when assigning a certain value of exposure to an individual. This is typically the case when the investigator has only access to data from a limited number of fixed site monitors. Cohort studies often use ambient community-level measurements of cumulative exposure from fixed site monitors. This approach works good if the variance of the exposure measurements between the communities is large compared to the variance of the exposure within the communities (Dominici et al. 2003, p. 270). Thus, in practice, it has to be checked if this requirement is fulfilled if one intends to use ambient community-level measurements as a surrogate for personal measurements.

As long-term personal measurements are often infeasible and as fixed monitor measurement are limited in their spatial resolution, cohort studies use more and more spatial prediction approaches. These models are used to predict exposures at locations without monitors, for example by land use regression models or spatio-temporal models (Sheppard et al. 2012, p. 208). However, the measurements resulting of the prediction models are still just a surrogate of the true exposure and incorporate a certain degree of measurement error. Szpiro et al. (2011) suggest that by using predicted exposure values, the measurement error generally results in two components. A Berkson-like component arises because of smoothing the exposure surface in the prediction model which lessens the exposure variability, and a classical-like component is caused by the uncertainty of the model estimates. These errors cannot be defined as the standard classical and Berkson error as they are not spatially independent, but are thought to have similar impacts. The Berkson-like error reduces statistical power and can be controlled by increasing the sample size, whereas the classical-like component

is independent of increasing sample sizes and has the potential to induce bias in the effect estimates (Szpiro et al. 2011, p. 612 ff.).

4 Impact of Measurement Error on Effective Sample Size

When designing a study, it is important to make sure that the study has enough statistical power to detect the effect of the exposure variable. One can design a study and determine the sample size such that a prespecified effect will be found with a certain power. Measurement error, however, has the impact to reduce a study's power and effective sample size which diminishes the chance that a truly present significant effect can be detected. Ignoring measurement error can therefore result in large discrepancies between computed and actual study power and significant exposure effects can be missed (Armstrong 1998, p. 654). This chapter deals at first with the general impact of measurement error on the effective sample size by means of the asymptotic relative efficiency. Secondly, the results will be discussed in the context of air pollution studies.

4.1 Asymptotic Relative Efficiency

Lagakos (1988) explores the impact of measurement error on commonly-used statistical tests and the resulting loss of effective sample size. The efficiency of a test statistic T_1 relative to another test statistic T_2 is defined as the ratio of sample sizes N_{T_1}/N_{T_2} required to achieve the same performance (Pratt and Gibbons 1981,p. 345). In this context, the performance of a test statistic is expressed by its statistical power. Lagakos calculates the loss in efficiency by measurement error in terms of asymptotic results requiring large sample sizes. The underlying asymptotic theory is based upon the assumption of local alternatives. The alternatives are the values of the coefficient β in the alternative hypotheses H_1. Under this local alternative assumption a sequence of local alternative hypotheses $\beta^{(n)}$ in the neighborhood of the null hypothesis are tested, whereas tests with fixed alternatives test hypotheses of the form $H_0 = \beta_0$ vs. $H_1 = \beta_1$ (Van der Vaart 1998, chapter 15). The assumption of local alternatives is an approximation that is appropriate for small values β_X of the alternative hypothesis (Tosteson et al. 2003, p. 1070). Lagakos compares the efficiency of a test statistic using X^* with the efficiency of a test statistic using X for local alternatives to the null hypothesis $\beta_X = 0$ by the asymptotic relative efficiency that is denoted by $ARE(X^* : X)$ (Lagakos

1988, p. 258 f.). He relates his calculations to linear regression, logistic regression and proportional hazard models.

For continuous explanatory variables, Lagakos examines the *ARE* for different measurement error models. For the classical non-differential measurement error model with $E(X^* \mid X) = X$ and $Var(X^* \mid X) = \sigma_U^2$ it holds that (Lagakos 1988, p. 268):

$$ARE(X^* : X) = \frac{\sigma_X^2}{\sigma_{X^*}^2} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} = \rho_{XX^*}^2, \tag{6}$$

and for non-differential Berkson errors with $E(X \mid X^*) = X^*$ and $Var(X \mid X^*) = \sigma_U^2$ it holds that (Lagakos 1988, p. 269):

$$ARE(X^* : X) = \frac{\sigma_{X^*}^2}{\sigma_{X^*}^2 + \sigma_U^2} = \rho_{XX^*}^2. \tag{7}$$

The square of the correlation between X and X^*, denoted by $\rho_{XX^*}^2$, is also called the coefficient of validity. The coefficient of validity is a common measure in the topic of measurement error as it is for linear regression and classical measurement error equal to the factor that attenuates the effect estimate (Armstrong 1998, p. 652 f.). Furthermore, the adequacy of a measurement method is often expressed by the correlation between true and measured exposure as it was done in section 2.3.

Lagakos also gives an extension of the Berkson model that can be used when exposure is measured at different geographical areas and the same exposure measurement is assigned to all subjects living in a certain geographical area. This is often the case in air pollution studies with various fixed site monitors that give approximated exposure measurements for the nearby located individuals. In this model the true exposure for a region j satisfies $E(X \mid X^* = X_j^*) = X_j^*$ with $Var(X \mid X^* = X_j^*) = \sigma_{U_j}^2$ for $j = 1, ..., k$ and the *ARE* becomes (Lagakos 1988, p. 269):

$$ARE(X^* : X) = \frac{\sum_j \pi_j (X_j^* - \mu)^2}{\sum_j \pi_j \sigma_{U_j}^2 + \sum_j \pi_j (X_j^* - \mu)^2}, \tag{8}$$

where $\mu = E(X) = E(X^*)$ and π_j is the proportion of the population in region j. Although Lagakos considers only least-squares tests for linear models and likelihood tests for logistic and proportional hazards models, the overall result $ARE = \rho_{XX^*}^2$ holds for all generalized linear models for one parameter exponential families (Lagakos 1988, p. 270).

4.2 Application to Air Pollution Epidemiology

For studies in air pollution epidemiology, the measurement error comes in complex forms with a mixture of different error types. Therefore, the formulas described in the previous chapter cannot be simply transfered to applications in air pollution epidemiology. One should rather use formulas that make it possible to extend the measurement error model to individual demands.

A simple approach, that may be helpful in this context, is given in a paper by Tosteson and Tsiatis (1988). They provide an extension of the *ARE*-formula for measurement error consisting of a mixture of non-differential classical and Berkson error. The measurement error model including both error types can be described by means of an intervening variable S (Tosteson and Tsiatis 1988, p. 512):

$$X = S + U_B, \quad X^* = S + U_C, \tag{9}$$

where S is a random variable with mean μ_S and variance σ_S^2. U_B and U_C are independent random variables with means zero and variances σ_B^2 and σ_C^2. The variable S corresponds to the true mean exposure, U_B to the Berkson error and U_C to the classical error. Assuming a score test of a generalized linear model leads to the formula of the asymptotic relative efficiency (Tosteson and Tsiatis 1988, p. 512):

$$ARE(X^* : X) = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_C^2} \cdot \frac{\sigma_S^2}{\sigma_S^2 + \sigma_B^2}. \tag{10}$$

As the $ARE(X^* : X) < 1$, the formula shows again a reduction of the power of the score test by the combination of classical error and Berkson error. The mixture of these error types can also be considered during the analysis stage of a study. Mallick et al. (2002) present Bayesian methods for semiparametric regression modeling with predictors measured with both classical and Berkson error (Mallick et al. 2002).

In conclusion, measurement error of classical type, Berkson type or a mixture of both types reduces the power and the effective sample size of a study. However, the tests and p-values remain valid for non-differential measurement error which is likely in air pollution studies. This means that mistaken "significant" results, when there is actually no effect, aren't more likely with measurement error than without measurement error. For differential error, which is rather rare in air pollution studies, these results don't hold. In some cases, one may obtain "significant results" for effect estimates even though they is in fact no significant association (Armstrong 1998, p. 654 f.). In summary, one requires a larger number of study subjects for detecting a certain effect slope with a given power when having erroneous exposure values instead of true val-

ues (Devine 2003, p. 331 f.). Formulas for calculating the measurement error adjusted sample size will be shown in the following chapter.

5 Study Design in the Presence of Measurement Error

The correction of biased coefficients due to measurement error is generally possible by using statistical methods during the analysis. Nonetheless, those methods cannot bring back lost power (Armstrong 1998, p. 655). It is therefore necessary to consider and minimize measurement error already in the design stage of a study. In fact, the big majority of air pollution studies doesn't take measurement error into account at all. The topic group 4 of the STRATOS initiative for measurement error and misclassification evaluates the current practice for addressing measurement error in observational epidemiology by a literature survey. Only 42% of the reviewed articles of air pollution cohort studies, that have been searched using general search terms related to the research area, mention measurement error as a potential problem. Only 6% use a method to adjust for measurement error (Shaw et al. 2017).
This chapter comprises comprehensive information concerning the three topics data collection, validation studies and sample size adjustment in the presence of measurement error.

5.1 Data Collection

As the impact of exposure measurement error on a study's result can be serious, one has to try to collect the data such that measurement error is minimized. A study has limited resources because of its budget and the more accurate a measurement method, the higher are the costs. One has to reasonably organize a study's budget and decide how much resources should be spent on data collection. If one designs a study while accounting for measurement error, one has to consider that the use of inaccurate and cheap measurement methods can lead to unfeasibly large sample sizes. Then, is it necessary to reduce the measurement error and more accurate measurement methods may be needed. An intuitive way to choose the best measurement method is to calculate the required sample sizes for each measurement instrument under consideration. Less precise measurement methods will demand higher sample sizes than more accurate measurement methods. After all, one can determine the best measurement method

by choosing the method with the lowest cost for the respective required sample size (McKeown-Eyssen and Tibshirani 1994, p. 419).

Armstrong (1996) gives a more formal criterion for allocating resources in improving accuracy of the measurement method. As more accurate measurements demand costs that could be spent in other ways like increasing the sample size, Armstrong gives a formal method regarding the trade-off between improving accuracy and increasing sample size. His criterion maximizes study power and compares study designs by their *ARE*. He extends the usual definition of the *ARE*, presented in chapter 4, by including the costs per subject. The commonly defined *ARE* compares two study designs by the ratio of the sample sizes required to attain equal statistical power of detecting an association. The extended *ARE* is defined as the ratio of the total study costs to achieve equal power (Armstrong 1996, p. 192). Assuming that the total costs of a study are proportional to the sample size, the basic cost of including a subject in the study is denoted as C_I, the cost of measuring the true exposure X as C_X and the cost of measuring the surrogate X^* as C_{X^*}. The total costs of two studies S_X and S_{X^*}, one study observing the perfect measure X with sample size N_X and the other study the surrogate measure X^* with sample size N_{X^*}, can be expressed by $N_X(C_I+C_X)$ and $N_{X^*}(C_I+C_{X*})$, respectively. The *ARE* of study S_{X^*} relative to study S_X is defined as (Armstrong 1996, p. 194):

$$ARE(X^*:X) = \frac{N_X(C_I+C_X)}{N_{X^*}(C_I+C_{X*})} = \frac{N_X}{N_{X^*}} \cdot \frac{C_I+C_X}{C_I+C_{X^*}}. \tag{11}$$

As $\frac{N_X}{N_{X^*}}$ equals the validity coefficient $\rho^2_{XX^*}$, it follows that (Armstrong 1996, p. 194):

$$ARE(X^*:X) = \rho^2_{XX^*} \cdot \frac{C_I+C_X}{C_I+C_{X^*}}. \tag{12}$$

Study S_X is preferable compared to study S_{X^*} if $ARE(X^*:X) < 1$ which is true when:

$$\rho^2_{XX^*} < \frac{C_I+C_{X^*}}{C_I+C_X}. \tag{13}$$

This formula can also be used for comparing an approximate measure X_1^* with a more precise, but still approximate measure X_2^*. The application of the formula above can be carried out analogously with replacing $\rho^2_{XX^*}$ by the ratio $\rho^2_{XX_2^*}/\rho^2_{XX_1^*}$. Armstrong provides also an extension of the formula that allows to include overhead costs of a study that are not proportional to the number of study subjects. In summary, the conclusion of this approach is that spending resources on improving accuracy of measurement is worthy up to the point at which the proportional increase in total costs per individ-

ual is greater that the proportional increase of $\rho^2_{XX^*}$ (Armstrong 1996, p. 194). One has to notice that this criterion leads to a design with maximum power, but doesn't necessarily lead to a design with minimum bias. Furthermore, this approach assumes non-differential errors. Non-differential errors, however, are reasonable to assume for many air pollution studies. Also, confounding variables Z aren't considered. For this, one can replace the coefficient of validity $\rho^2_{XX^*}$ by the square of the partial correlation $\rho^2_{XX^*|Z}$, assuming that the confounder Z is measured without error (Armstrong 1996, p. 196).

Regarding air pollution epidemiology, one may assume that the measurements from personal monitors can be treated as the true values X and measurements from fixed site monitors as the surrogate values X^*. For example, if the true and the surrogate measures are correlated at ρ_{XX^*}=0.5, then $\rho^2_{XX^*}$=0.25, which implies that a study based on personal monitors will only be worthwhile if the total cost per subject for personal monitors $(C_I + C_X)$ is less than four times the total cost per subject for fixed site measurements $(C_I + C_{X^*})$ (Armstrong 1996, p. 194). As the costs for personal monitor measurements are many times higher than the costs for fixed site measurements, this will not hold for air pollution studies leading to the conclusion that using only the gold standard measure is less effective than using approximates measurements with more study subjects. For correlation coefficients that are larger than 0.5, the only use of the accurate measure will be even less effective. However, this is only a simplified theoretical representation of the reality. In fact, a mixture of the observation of accurate and surrogate measurements can be useful. This will be discussed in the following chapter regarding validation studies including the allocation of resources to the main study and the validation study. General methods for determining main study sample sizes will be presented in section 5.3.

Regarding the collection of the data, White et al. (1994) deal with the effect of exposure variance in the presence of measurement error. Selecting a population with larger exposure variance in contrast to one with smaller exposure variance reduces under certain assumptions the required sample size with an even more pronounced effect for exposures measured with error (White et al. 1994, p. 1994 ff.). This however, cannot generally be transfered to air pollution studies. A population with larger exposure variance in a certain area, that is represented by a single ambient monitor, leads to increased measurement error in the ambient measurements that try to represent the personal exposures. The result of White (1994) does only hold when the true exposure variance is captured by the measurements, for example when exposure is measured by personal monitors or when the study includes different locations with different monitors for each location. In this case, the overall exposure variance has to be explained

by the exposure variance between the different locations, not by the exposure variance within the locations.
Regarding the air pollution monitor placement in a study area, studies in air pollution epidemiology often locate the ambient monitors in an informal fashion, preferring to place monitors near locations that are thought to be interesting for land use, traffic or population characteristics (Kanaroglou et al. 2005, p. 2400). Kanaroglou et al. present a formal methodology for locating a fixed number of monitors in urban areas that was applied in Toronto, Canada. They suggest to determine a continuous surface over the study area whereby higher values of the surface indicate an increased need for monitors. The surface is based on two criteria. The first indicates that an increased number of monitors has to be placed for areas with higher spatial variability of the pollutant. The second criteria demands an increased number of monitors in areas with a high density of the at-risk populations, for example specified by socio-demographic characteristics. The optimal locations for the predefined number of monitors are found by an algorithm for location-allocation problems (Kanaroglou et al. 2005, p. 2400). When collecting data by ambient monitors, such criteria can be useful in the design stage of the study in order to ensure maximal benefit of the collected data.

5.2 Validation Studies

Many epidemiologic air pollution studies don't account for exposure measurement error because of insufficient knowledge about the error (Sheppard et al. 2012, p. 204). With regard to this, validation studies are of great importance as they are necessary for gaining information about the measurement error in a specific study.

5.2.1 Characteristics of Validation Studies

A validation study aims at collecting information about the structure and extent of the measurement error in order to be able to work out an appropriate measurement error model and its parameters. This information can be used for adjustment of the study design, such as an increase of sample size, or for adjustment of the later statistical analysis with measurement error correction methods.
A validation study requires measurements of the true value of the variable of interest as well as its imperfect measurements that are used in the main study. The true measurement is also called the "gold standard". These accurate measurements are usually collected using sophisticated technology and are usually much more expensive than the inaccurate measurements. This is also the main reason why in most studies these accurate measurements can't be collected for the whole study population and the vali-

dation study sample size n is mostly much smaller than the main study sample size N (Spiegelman 1994, p. 192). After obtaining the gold standard measurements of the validation study, they have to be analysed to get information about the error by comparing them with the crude measurements of the main study. First of all, the error structure should be figured out, indicating whether the error is systematic, for example linearly dependent on the true measurements, or random. This is necessary to work out an appropriate measurement error model. This can be for example simply done by plotting both measures against each other. Furthermore, the magnitude of the measurement error in the exposure variable is an important indicator for the impact and severity of the error (Nieuwenhuijsen 2015, p. 204). By means of validation studies, the extent of the error can be quantified. For continuous exposure variables, as used in air pollution studies, a simple approach of quantifying the error is to calculate the validation coefficient, the square of the correlation between the gold standard measurements X obtained from the validation study and the crude measurements X^* (Nieuwenhuijsen 2015, p. 206):

$$\rho^2_{XX^*} = \frac{\sigma^2_X}{\sigma^2_{X^*}} = \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_U}. \tag{14}$$

The validation coefficient has already been used in chapter 4 to express the loss of effective sample size. Alternatively, one can express the relationship between the inaccurate measurements and the true exposure in terms of a regression equation (Holford and Stack 1995, p. 344):

$$X = \lambda_0 + \lambda_{X^*} X^* + \varepsilon. \tag{15}$$

The slope of this regression λ_{X^*} is also called the calibration coefficient. In linear regression with classical error it is equal to $\rho^2_{XX^*}$. Furthermore, it is called the attenuation factor as the attenuation of the effect coefficient can in linear regression with classical error be described as $\beta X^* = \lambda_{X^*}\beta X$ (Nieuwenhuijsen 2015, p. 209). It is therefore important for the correction of biased coefficients. The ability of quantifying the measurement error requires the gold standard measurements and makes validation studies inevitable. This leads to the necessity of considering them in the study's design stage. In an internal validation study a subset of the main study population is selected to validate the data. Thus, data for the validation study subjects include the imperfect measurements of exposure and the health outcome, which are collected for all main study participants, and additionally the gold standard measurements. To achieve unbiased estimates of the measurement error distribution, it is essential for the validation study to select a representative subset of the main study population. Random sampling

is the most common way to realize this (Holford and Stack 1995, p. 348).
In an external validation study the study subjects are not a subset of the main study population. The data could be obtained by another investigator. Therefore the data of the n study subjects gained by an external validation study include no information about the subject's outcome status (Holford and Stack 1995, p. 354). Figure (4) shows a schematic describing internal vs. external validation studies (cf. Holford and Stack 1995, p. 349).

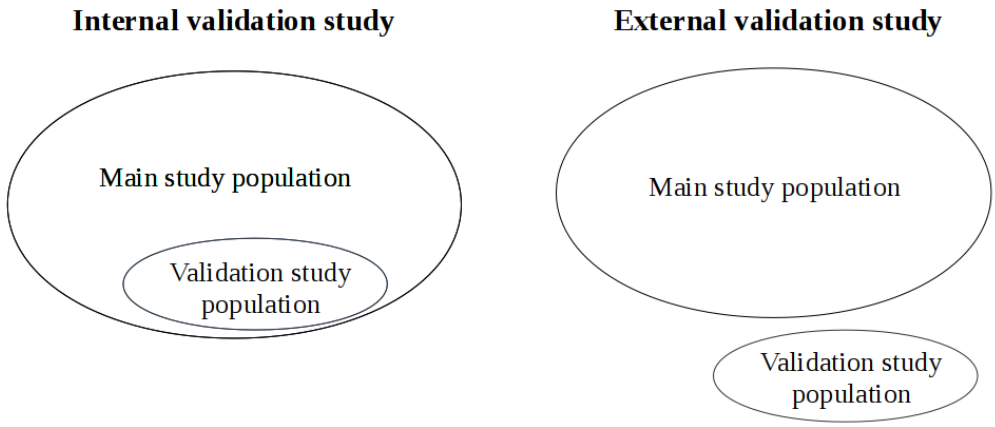

Figure 4: Schematic diagram of internal vs. external validation studies (cf. Holford and Stack 1995, p. 349).

When designing a validation study, one has to make sure that the information gained by the validation study, for example a validation coefficient, can be transfered to the main study. As explained in the previous section, the validation coefficient depends upon the error distribution as well as upon the distribution of the true exposures. Consequently, if the distributions of the true exposures differ in the main and validation study, the coefficient is not portable. The same holds for the calibration coefficient (Nieuwenhuijsen 2015, p. 219). For an external validation study is it necessary to particularly assure the portability of coefficients, whereas for internal studies the portability is fulfilled by random sampling of the validation study population out of the main study population (Spiegelman 1994, p. 197). Furthermore, the internal validation study has the advantage that additional information is collected for the subjects in the main study that are also part of the validation study (Nieuwenhuijsen 2015, p. 218). In summary, from validity and efficiency considerations an internal validation study design is preferable over an external one.
Besides internal and external validation studies, one can use repeated measurement studies to validate the error-prone data. When measurement error is random, then tak-

ing repeated observations with the inaccurate measuring tool provides the possibility to estimate the variance of the random error (Holford and Stack 1995, p. 341). This method of taking repeated measurements can be used to assess instrumental measurement error, for example for inaccurate monitors.

5.2.2 Validation Study Sample Size

The parameters obtained by validation studies, like validation or calibration coefficients, should be as precise as possible. An adequate and intuitive approach for identifying the optimal validation study sample size is to figure out how the variance of the estimated parameter of interest depends upon the sample size. In addition, for determining the sample size, it is required to find a good trade-off between precise estimates and the resulting costs (Nieuwenhuijsen 2015, p. 220). In the following, more concrete examples for such approaches are presented.
We assume linear regression with a single error-prone explanatory variable X^* whose error is non-differential and follows the linear measurement error model $X^* = \alpha_0 + \alpha_X X + U$. Further, we assume that the validation study is designed to provide an estimate of the calibration factor λ_{X^*} and that λ_{X^*} is estimated by the slope of the regression of X on X^*. Additionally, we assume that the relation $\beta_{X^*} = \lambda_{X^*}\beta_X$ holds. The biased coefficient β_{X^*} can therefore be corrected by dividing it by the estimate of λ_{X^*} obtained from the validation study. Consequently, the variance of the adjusted estimate of β_X depends on the uncertainty of the estimated λ_{X^*}. Rosner provides a formula for the variance of the adjusted $\hat{\beta}_X$ (Rosner et al. 1989, p. 1054):

$$Var(\hat{\beta}_X) \approx \frac{Var(\hat{\beta}_{X^*})}{\lambda^2} + \frac{\hat{\beta}^2_{X^*} Var(\hat{\lambda})}{\lambda^4}. \tag{16}$$

The second part of the right hand side of formula (16) represents the additional uncertainty in the adjusted $\hat{\beta}_X$ caused by the uncertainty of $\hat{\lambda}$. To minimize this additional uncertainty, one should determine the validation study sample size such that second part of the right hand side of (16) remains a small fraction f of the first part. A formula for the validation study sample size can be determined by assuming that the main study has 50% power to detect $\hat{\beta}_{X^*}$ as statistically significant at the 5% level, that is $Var(\hat{\beta}_{X^*}) \approx \beta^2_{X^*}/4$. This yields $Var(\hat{\lambda}) \approx \mathrm{f}\lambda^2/4$. Assuming that the measurements of the validation study equal the true values X, the formula for validation study sample size n can be derived as (Freedman et al. 2017, p. 19 f.)

$$n \approx \frac{4(1-\rho^2_{XX^*})}{\mathrm{f}\rho^2_{XX^*}}. \tag{17}$$

For this formula, one has to determine $\rho^2_{XX^*}$ in advance. Furthermore, the fraction f has to be decided. For example, a value of the fraction $f = 0.1$ provides that the variance of the adjusted $\hat{\beta}_X$ will not increase by more than 10% because of the uncertainty in the estimate of λ.

There are other approaches regarding the validation study sample size. Generally, there are two ways to determine the validation study sample size n. The first is to determine the values of N and n that will optimize the study design by setting a fixed cost constraint. The other is to specify the size of the effect that one wants to detect with a given power, and then find the values of N and n that minimize the costs (Spiegelman and Gray 1991, p. 855 f.).

Greenland (1988a) provides a method to determine the optimal proportion of subjects $P = n/(N+n)$ allocated to the validation study by minimizing the variance of the estimated association between true exposure and disease while fixing the overall cost. He only addresses this method to the case of binary exposures. Since air pollution exposure measurements are continuous, we will focus on validation study design methods related to continuous exposure variables. Spiegelman and Gray (1991) present approaches to identify the optimal allocation of study sample sizes N and n of the main and the validation study for relative risk modeling in the case of a single continuous exposure. The approach of Spiegelman and Gray is presented in the following.

Spiegelman and Gray's approach is based on minimizing the proposed budget given adequate statistical power to the test of the hypothesis of interest. Their proposed methods for efficient study designs hold for studies in which the outcome is binary and random in the design and exposure is observed at the baseline (Spiegelman and Gray 1991, p. 853). One should note that at least this is usually not true for case-control studies. The model on which their calculations are based is the logistic regression model (Spiegelman and Gray 1991, p. 852):

$$f(Y \mid X) = \frac{exp\{(\alpha + \beta X)Y\}}{1 + exp\{\alpha + \beta X\}}, \tag{18}$$

where Y is the observed binary disease outcome and X is the true exposure. The parameter α equals the logit of the sample disease prevalence at $X = 0$ and β, the parameter of interest, equals the logit of $Pr(Y = 1 \mid X = x + 1)$ minus the logit of $Pr(Y = 1 \mid X = x)$. They describe the relationship between the true and the imperfect exposure with a generalized Gaussian measurement error model (Spiegelman and Gray 1991, p. 853):

$$f(X \mid X^*) = \frac{exp\{-\frac{1}{2}(X - \alpha - \gamma X^*)^2/\sigma^2\}}{\sqrt{2\pi\sigma^2}}. \tag{19}$$

When X follows a Gaussian distribution, the generalized Gaussian measurement error model (19) can also be transformed to the classical error model $X^* = X + U$ with $U \sim N(0, \sigma_U^2)$ by setting suitable values for α, γ and σ^2. This structural model, assuming that X has a well-defined conditional distribution given X^*, is incorporated in the main model linking the outcome and the observed values X^*. In addition they make the conditional independence assumption $f(Y \mid X, X^*) = f(Y \mid X)$ which is equivalent to the assumption of non-differential errors (Spiegelman and Gray 1991, p. 853). They provide different design optimality approaches for internal and external validation studies. The design optimization uses log-likelihood functions which differ for internal and external validation studies. For an internal validation study, the log-likelihood function can be separated in three parts (Spiegelman and Gray 1991, p. 854):

$$L = \sum_{i=1}^{N} l_{1i}(\alpha, \beta, \alpha', \gamma, \sigma^2) + \sum_{i=1}^{n} l_{2i}(\alpha', \gamma, \sigma^2) + \sum_{i=1}^{n} l_{3i}(\alpha, \beta). \tag{20}$$

The first part represents the likelihood contribution of the N main study subjects, where l_1 is the log of the model for the observed data $f(Y \mid X^*)$ that consists of the logistic model with true exposure $f(Y \mid X)$ of formula (18) and the measurement error model $f(X \mid X^*)$ of formula (19). The second part corresponds to the likelihood contribution of the validation study subjects n where l_2 is the log of (19). The third part is the log of the logistic model of formula (18).
When the validation study is external, the corresponding log-likelihood function consists only of the first and the second part since the third part cannot be determined as there is no information available about the outcome status of the validation study subjects (Spiegelman and Gray 1991, p. 854):

$$L = \sum_{i=1}^{N} l_{1i}(\alpha, \beta, \alpha', \gamma, \sigma^2) + \sum_{i=1}^{n} l_{2i}(\alpha', \gamma, \sigma^2). \tag{21}$$

By means of this full likelihood approach including the information of the main and the validation study, one can calculate maximum likelihood estimates of β and approximate the variance of this estimator for power calculations. The total cost function, which has to be specified, depends on the unit costs r_X, r_{X^*} and r_Y for observing X, X^* and Y, and on the sample sizes N and n. The total cost function of main and validation study for internal validation studies is given by (Spiegelman 1994, p. 194):

$$C(N, n) = min\{(r_Y + r_{X^*})N + (r_X + r_{X^*} + r_Y), (r_Y + r_X)n\}, \tag{22}$$

where $(r_Y + r_X)n$ is the total cost of a fully validated design with accurate measurements for all subjects. For external validation studies the total cost function becomes (Spiegelman 1994, p. 194):

$$C(N,n) = (r_Y + r_{X^*})N + (r_X + r_{X^*})n. \tag{23}$$

To summarize, the optimization criteria of Spiegelman and Gray minimizes the cost function with respect to the values N and n while keeping the power constraints. Those constraints assure that the study will be able to discriminate between two specified effect levels of the outcome-exposure relationship with a fixed power (Spiegelman and Gray 1991, p. 856). For power calculations they choose the discriminatory power criterion which is proposed by Greenland (1988b). The sample size which fulfills the discriminatory power criterion is given by the maximum of two sample sizes that are calculated by the common specification of power and sample size, where the role of two hypothesized values β_L and β_U, between which the study is designed to discriminate, are alternated. This means that at first β_L plays the role of the null hypothesis and β_U the role of the alternative hypothesis and then the roles are reversed. The presented approach for the main study and validation study sample size calculations requires numerical methods. (Spiegelman and Gray 1991, p. 855 ff.). Several quantities have to be identified in advance which are the costs of observing X, X^* and Y, the prevalence of disease in the study population, the two hypothesized values of β_X between which the study is designed to discriminate, the mean and variance of X^*, the parameters of the distribution of X given X^* and the desired confidence level and power for the hypothesis test (Spiegelman 1994, p. 198).

The conclusions found by Spiegelman and Gray after applying the method to different scenarios can be summarized as follows. The relative validation study sample size is highly dependent on the reliability of the error-prone measure and the cost of the gold standard measure X relative to the error prone measure X^* used in the main study. With decreasing reliability of the crude measure and with decreasing relative validation study unit cost, the optimal relative validation study size increases. Different scenarios included also different values for the disease probability and hypothesized relative risk values, where slightly different results are obtained for internal and external validation studies (Spiegelman and Gray 1991, p. 858 ff.). In most cases one can conclude that the internal validation study is considerably more cost-efficient than the external one (Spiegelman and Gray 1991, p. 863).

It is important to keep in mind that the approach of Spiegelman and Gray is only valid for the classical measurement error model (Spiegelman and Gray 1991, p. 866) while the classical measurement error assumption is not adequate in many practical appli-

cations in epidemiology. For instance, in air pollution studies usually a part of the measurement error is of Berkson-type. Another issue that is worth considering is the normality assumption on the conditional distribution on X given X^*. This assumption may not be fulfilled as the exposure distribution in epidemiology is often highly non-normal like the distribution of exposure to air pollutants (Spiegelman and Gray 1991, p. 867). Transformations of the data may be needed. However, the work of Spiegelman and Gray gives a profound insight into a way of realizing validation study and main study sample size calculations under the important aspect of cost-efficiency. Since the gold standard measurements are in many cases extremely expensive and validation studies are the only way of gaining useful information about the measurement error, it is highly recommended to carefully consider the proportion of overall study resources allocated to the main study and the validation study.

There is an important aspect that should be considered when using validation study results for the correction of attenuated coefficients. In practice, in many epidemiologic applications the putative gold standard is measured with at least a small extent of error. Measurements of such imperfect reference instruments are called "alloyed gold standards". Wacholder et al. (1993) examine the influence of measurement error in alloyed gold standards when a validation study is used to correct effect estimates. Their results apply to corrections in proportional hazard and logistic regression models (Wachholder et al. 1993, p. 1253). They conclude that in general all common methods that are used to correct for attenuation yield biased estimates for the corrected coefficients when the validation data has an alloyed gold standard (Wachholder et al. 1993, p. 1256). The relation between the biased corrected estimate and the true slope depends on the error of the alloyed gold standard measurements and on the correlation of the error in the crude measure of the main study with the error in the alloyed gold standard of the validation study. Positive correlation between the errors is likely when the two measures are expected to make similar mistakes, while independent errors can be assumed when the two measures are taken with different and independent processes. Negative correlations are rare but can occur when the measure used in the validation study is too specific and the measure used in the main study is too sensitive, or the other way round (Wachholder et al. 1993, p. 1257). If the errors are independent, alloyed gold standards imply overcorrection of the effect estimates. Positive correlation of the errors diminishes the overcorrection, whereas negative correlation accentuates the overcorrection (Wachholder et al. 1993, p. 1254). Since less than perfect gold standards are common in epidemiology, particularly for environmental, occupational and nutritional exposures, it is recommended to be cautious with careless adoption of error corrections based on validation studies (Wachholder et al. 1993, p. 1256). For the study design,

regarding validation studies, it is important to keep in mind the problem of alloyed gold standards. One can take steps to improve the accuracy of the validation study measurements or, if alloyed gold standard measurements are inevitable, to estimate the error in the alloyed gold standard measurements.

5.2.3 Validation Studies in Air Pollution Epidemiology

In air pollution studies, the inaccurate measure is usually a measure of ambient levels whereas the gold standard measurements of the validation study come from personal monitors that are attached to the individuals. The use of personal monitors is very costly as it requires additional technical input. Because of the high costs and labour intensity, it is mostly not possible to take personal measurements for a large number of subjects. Furthermore, the duration of measuring personal exposures is limited as it causes inconvenience for study subjects (Nieuwenhuijsen 2015, p. 87). Because of these factors, the sample size of validation studies in air pollution epidemiology is limited. However, when personal exposure is the exposure of interest of the study, the measurements of personal monitors from a validation study can provide useful information (Nieuwenhuijsen 2015, p. 87). As the individual measurements are directly representative for the true exposure of each person in the measuring period, they could be seen as either a gold standard or something approximating such a standard. In the case that the measurements from personal monitors still contain error, for example instrumental error of monitors, the measurements meet only an alloyed gold standard. Regarding the problem of alloyed gold standards, for air pollution studies it is mostly adequate to assume that the errors in the crude measure and in the gold standard measure are uncorrelated. This is reasonable as the measurement methods are objective (Van Roosbroeck 2008, p. 411), in contrast to other epidemiologic studies that obtain their data for example by interviewing subjects.

For example in air pollution epidemiology, Van Roosbroeck et al. (2008) analyses validation studies for NO_2 and soot . The conducted main study is interested in the association between traffic-related pollutants and respiratory symptoms in children. The collected data includes several disease outcomes and outdoor measurements of soot and NO_2 between 1997 and 1998 for 2083 children from different dutch schools located near freeways. The validation study for NO_2 is an internal validation study with 67 validation study subjects out of 2083 main study subjects. The surrogate measure is the home outdoor concentration, whereas the gold standard measurements come from personal monitors attached to the subject. Average personal and average outdoor concentrations are available for one to four 1-week periods in each of the four seasons. The validation study for soot is external and conducted 8 years later on school children

in Utrecht with a validation study size of 45. Personal measurements from this validation study are available for four 48-hour periods (Van Roosbroeck 2008, p. 410). They assume a linear measurement error model and estimate the model parameters by a linear regression equation $X = \lambda_0 + \lambda_{X^*}X^* + \lambda_Z'Z' + \varepsilon$, including confounders Z'. Furthermore, they calculate correlation coefficients between surrogate and "true" measurements ρ_{XX^*} and obtain the values 0.53 for soot and 0.35 for NO_2. After adjustment for measurement error by the regression calibration method, the estimated effect coefficients become two to three times larger than the original estimates. For example, the adjusted prevalence ratio for the outcome "current phlegm" is 5.3 (95% confidence interval: 1.2 – 22.6), whereas the original result is 2.2 (95% confidence interval: 1.3-3.9). The much larger confidence interval for the adjusted value results from the small validation study sample size and the increased estimate of uncertainty by taking measurement error into account. Van Roosbroeck points out that the validation study sample size is too small to give reliable corrected effect estimates. In general, validation studies with adequate power will remain difficult to realize because of the infeasibility of measuring long-term personal exposure (Van Roosbroeck 2008, p. 412 ff.).

The increase in the effect estimates after adjusting for measurement error in the described validation study shows the crucial importance of accounting for measurement error and underlines the usefulness of the conduct of validation studies in air pollution epidemiology. As the correlation between true and surrogate exposure is in many air pollution studies using ambient monitor measurements inadequately low, one can expect heavy changes in the effect coefficients by measurement error adjustment for other air pollution studies. Even though a part of measurement error in air pollution epidemiology is of Berkson type, the complex mixture of measurement error types including classical error components results in biased effect coefficients. It remains difficult to give general results for air pollution studies. Kioumourtzoglou et al. (2014) provides another article related to the topic of validation studies in air pollution epidemiology. Articles for studies, that take measurement error into account during the design stage of the study by conducting validation studies in advance, cannot be found indicating the importance of improving the current practice with regard to this topic.

5.3 Sample Size Calculation

The determination of the sample size is a crucial part of designing a study. The sample size in an epidemiologic study can typically be calculated by using a criterion that expresses the desired statistical power or the desired precision of effect estimates. There is a broad variety of methods and software in order to calculate sample sizes, but those assume implicitly that the data is free of measurement error (Devine 2003, p. 316). As presented in chapter 4, measurement error has the impact to reduce the effective sample size of a study which induces the requirement of increasing the sample size in order to meet the criteria.

Devine (2003) suggests an approach of sample size calculation assuming a linear regression model which bases on a power criterion by Dupont and Plummer (1998). One has to specify a value for the regression coefficient β_X that one wishes to detect with a certain power at a certain significance level α and tests the null hypothesis $H_0 : \beta_X = 0$. When assuming that N is large, the sample size estimator for error-free data is given by (Devine 2003, p. 326):

$$N = \frac{(z_{1-\alpha} + z_{1-\delta})^2 \sigma^2_{\varepsilon,X}}{\beta_X^2 \sigma_X^2}. \tag{24}$$

In this formula, $z_{1-\alpha}$ and $z_{1-\delta}$ are the $(1-\alpha)$ and the $(1-\delta)$ quantiles of the standard normal distribution where δ equals the type 2 error and thus $(1-\delta)$ equals the power. One has to insert the value for β_X, that has to be specified before. The values $\sigma^2_{\varepsilon,X}$ and σ_X^2 denote the residual variance of the regression of X on the outcome and the variance of the true explanatory variable X. Both can be estimated by for example using information from previous studies (Devine 2003, p. 326 f.).

Further on, Devine assumes that the data is not error-free but measured with classical measurement error $X^* = X + U$ such that the values U_i are independent of X and normally distributed with mean zero and variance σ_U^2. Again a linear regression is fitted, now with the values of X^*. For again obtaining the required number of study subjects to achieve the desired power to test $H_0 : \beta_X = 0$, Devine suggests to use the previous formula and simply substitute the relevant parameters related to X^*(Devine 2003, p. 332):

$$N^* = \frac{(z_{1-\alpha} + z_{1-\delta})^2 \sigma^2_{\varepsilon,X^*}}{(\rho^2_{XX^*} \beta_{X^*})^2 \sigma^2_{X^*}}, \tag{25}$$

where $\sigma^2_{\varepsilon,X^*}$ is the residual variance of the regression of X^* on the outcome and $\sigma^2_{X^*}$ the variance of X^*. Again, both values can be estimated. Furthermore, the coefficient

β_X from formula (24) is substituted by $\rho^2_{XX^*}\beta_{X^*}$ according to the result that classical measurement error under linear regression attenuates the true β_X by the factor $\rho^2_{XX^*}$. As expected, the formula gives high sample sizes N* for low correlations between surrogate and true measure.

Devine evaluates the adjusted sample size estimator by a simulation approach for a linear regression of daily exercise time on the body mass index. He repeats the simulation experiment for different values of the correlation ρ_{XX^*}. The resulting empirical powers for the adjusted values N^* using the suggested formula are nearly identical to the target power of 0.9 (Devine 2003, p. 329, 333).

This approach requires knowledge about $\rho^2_{XX^*} = \sigma^2_X/\sigma^2_{X^*}$ which can be obtained by a validation study with the gold standard measure X. One has to notice that the suggested approach applies only to the classical measurement error. Devine's considerations are based on the fact that for classical measurement error $\sigma^2_X < \sigma^2_{X^*}$ and that the coefficient β_{X^*} is attenuated towards the null while the true β_X is the parameter that is wanted to be tested (Devine 2003, p. 330 ff.). Clearly this is not the case for many practical applications where a considerable part of the error is of Berkson-type like in many air pollution studies.

McKeown-Eyssen and Tibshirani (1994) give another example for sample size adjustment in the presence of measurement error for logistic regression. They assume a linear measurement error model $X^* = \alpha_0 + \alpha_X X + U$, where U is assumed to be normally distributed and independent of X. Moreover, they refer their work to case-control studies. The calculations are based on a sample size formula for error-free variables by McKeown-Eyssen and Thomas (1985) that is applicable for normally distributed exposures X and exponential relations between exposure X and the risk of disease $r(X)$ such that $r(X) = exp(a+bX)$ (McKeown-Eyssen and Tibshirani 1994, p. 415 f.). Similarly as in the paper from Devine (2003), the formula for the adjusted sample size N^* is derived after inserting the respective values for X^* for the attenuated coefficient β_{X^*} and the increased exposure variance $\sigma^2_{X^*}$ into the formula. After simplifying, one obtains (McKeown-Eyssen and Tibshirani 1994, p. 416):

$$N^* = \frac{N}{\rho^2_{XX^*}}. \tag{26}$$

While this approach is similar to that presented by Devine, the difference between formula (26) and the formula (25) by Devine is that (26) restricts itself to logistic regression with exponential risks whereas (25) is limited to linear regression. Furthermore, the approach by McKeown-Eyssen and Tibshirani is targeted on case-control studies.

For a practical application, it is important to regard the adequacy of the assumptions

that were made for this derivation. The exposure X is assumed to be normally distributed (McKeown-Eyssen and Tibshirani 1994, p. 416 f.). As air pollution exposure data is skewed, this assumption does not automatically hold and data transformations may be necessary. Another assumption is that the random error U is normally distributed and independent of the level of the exposure (McKeown-Eyssen and Tibshirani 1994, p. 417). In practice, this assumption has to be ensured by applying for example a logarithmic transformation to the data. However, assumptions have to be proofed individually for different studies since general results may be difficult as studies in air pollution epidemiology vary a lot in their designs and available data. The approach by Devine which is based on the classical measurement error assumes normal distribution and independence of the error as well. The formulas by McKeown-Eyssen and Tibshirani and Devine both assume a single error-prone covariate and don't take confounders into account. In fact, it is essential for models used in air pollution epidemiology to adjust for confounders. The confounders itself may be measured with or without error. Further research is needed for sample size determination under measurement error with confounding variables (McKeown-Eyssen and Tibshirani (1994), p. 417 f.). Finally, the main concern for the approaches by Devine and McKeown-Eyssen and Tibshirani is that they only refer to linear regression and logistic regression while studies in air pollution epidemiology also use more sophisticated models like generalized linear models (GLMs).

Tosteson et al. (2003) give methods for sample size calculations that can be applied to a broad variety of GLMs. Their asymptotic theory is built upon the properties of the score test modified for measurement error without the simplifying assumption of small relative risks or other modeling restrictions. Previous calculations for the asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates employ the local alternative condition under a sequence of alternatives $\beta^{(n)}$ such that $\sqrt{n}\beta^{(n)} \to \tau$, where $\|\tau\| > 0$ (Tosteson and Tsiatis 1988, p. 510; Tosteson et al. 2003, p. 1072). The approach of Tosteson et al. does not rely on the local alternative assumption and considers asymptotic theory appropriate for fixed alternatives (Tosteson et al 2003, p. 1070). Furthermore, this approach is applicable for various measurement error models (Tosteson et al. 2003, p. 1078 f.). Additionally, the calculations by Tosteson et al. take into account a confounder Z, measured without error. They use a general measurement error model that is able to express the relationship between X and X^* in a flexible way. All they assume is the usual conditional independence assumption (Tosteson et al. 2003, p. 1071):

$$f_{Y,X^*|Z,X}(Y_i, X_i^* \mid Z_i, X_i) = f_{Y|Z,X}(Y_i \mid Z_i, X_i) f_{X^*|Z,X}(X_i^* \mid Z_i, X_i), \tag{27}$$

where f is a conditional probability density or mass function. The assumption requires that the surrogate X^* is independent of the outcome Y given the true exposure X and the confounder Z, corresponding to non-differential error (Tosteson et al. 2003, p. 1071). This assumption is likely to be satisfied in many practical applications of air pollution epidemiology as the measurements by fixed site monitors are objective and unlikely to contain information about the outcome status Y given X. They derive the asymptotic distribution of the score test statistic under measurement error for the null hypothesis $H_0 : \beta_X = 0$ and the asymptotic power function of the test for fixed alternatives.

In contrast to the general derivation, Tosteson et al. find that when making the more restricted assumption of local alternatives, suitable for small alternatives β_X, the calculations simplify to the already known relation $ARE = \rho^2_{XX^*}$ yielding sample size adjustment by $N^* = N/(\rho^2_{XX^*})$. This holds for simple regression models and linear measurement error and is consistent with the findings of other authors (Tosteson et al. 2003, p. 1073). The approach by Tosteson et al., however, does not rely on these assumptions. They underline this advantage by showing that the simplified adjusted sample size assuming local alternatives seriously underestimates the correct sample size for larger alternatives of β_X (Tosteson 2003, p. 1078). In practice, when conducting sample size adjustments, it is important to consider this issue, depending on whether one tests for local or large alternatives to $H_0 : \beta_X = 0$.

The calculation of measurement error adjusted sample sizes N^*, proposed by Tosteson et al., require specification of the joint distribution of (Z, X, X^*). Additionally, they involve multi-dimensional integral evaluations. A web-based demonstration program for sample size calculations has been implemented for continuous covariates and a scalar Z in logistic regression with joint normality of (Z, X, X^*). It is applicable for classical, Berkson, and general measurement error models. One has to specify the correlations ρ_{XX^*} and ρ_{ZX}. For classical error, X^* and Z are conditionally independent given X and $\rho_{ZX^*} = \rho_{ZX}\rho_{XX^*}$. For Berkson error, X and Z are conditionally independent given X^* and $\rho_{ZX^*} = \rho_{ZX}/\rho_{XX^*}$. For the general measurement error model, no assumptions are set for the correlations among X, X^* and Z, and ρ_{XX^*} and ρ_{ZX^*} must be specified independently (Tosteson et al. 2003, p. 1073).

Figures (5) and (6) show outputs of the program for sample size and power calculations for logistic regression with classical measurement error.

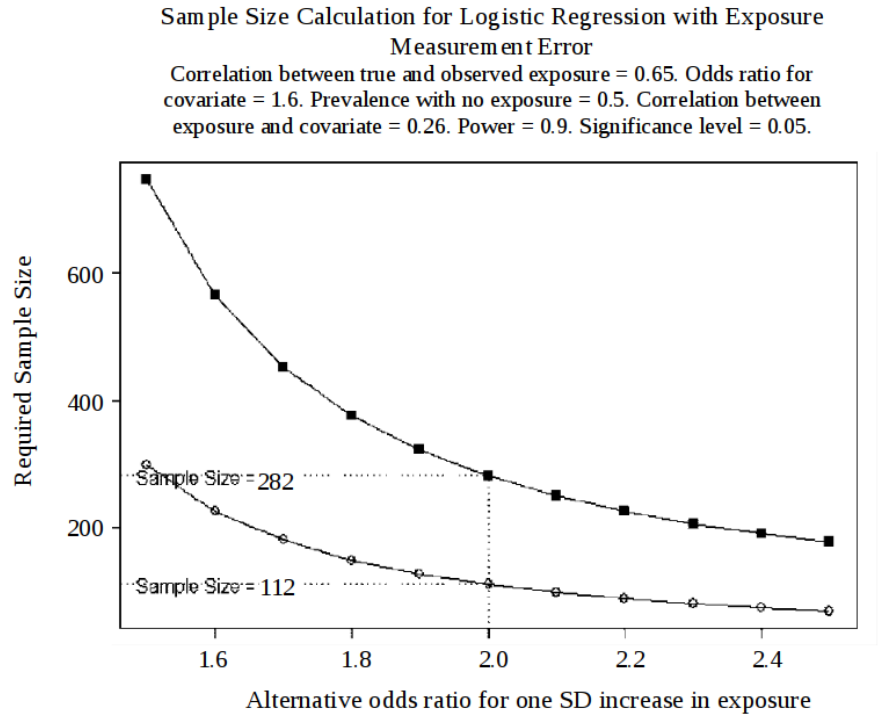

Figure 5: Sample size calculation with measurement error (filled box) and without measurement error (empty circle)(cf. Tosteson et al. 2003, p. 1074).

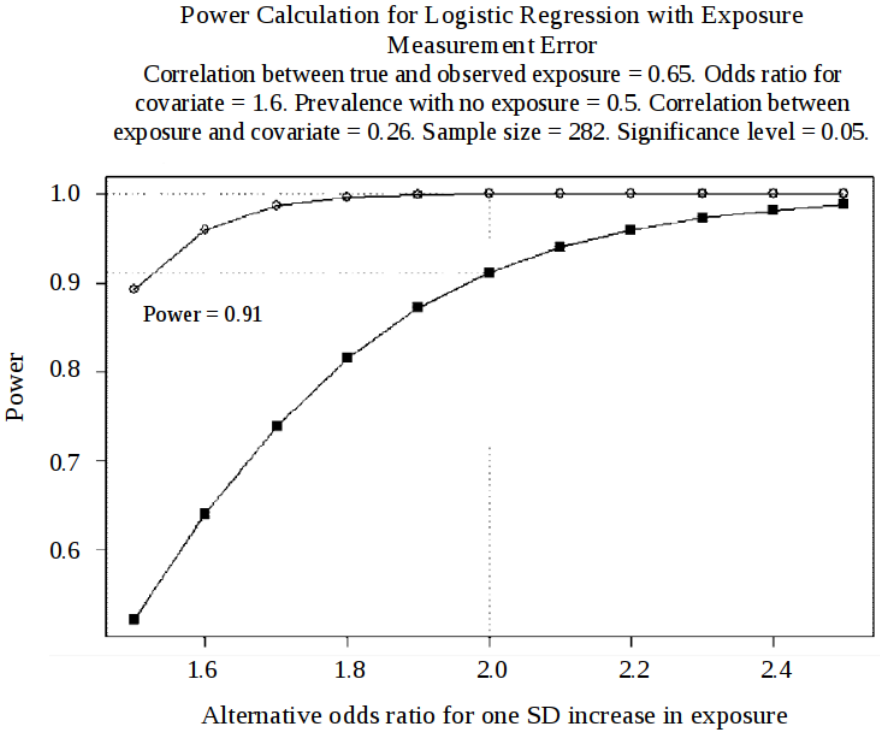

Figure 6: Power calculation with measurement error (filled box) and without measurement error (empty circle)(cf. Tosteson et al. 2003, p. 1074).

Figure (5) shows the required sample size for a fixed power dependent on different values of the alternative odds ratio for one standard deviation increase in exposure. The required sample sizes are given for exposure measured with and without measurement error. They decrease with increasing alternative odds ratios. Larger sample sizes are required in the presence of measurement error, especially for small alternatives. Figure (6) shows the power for a fixed sample size, again dependent on alternative odds ratios for one standard deviation increase in exposure. Power increases with increasing alternative odds ratios. Under measurement error, the power is considerably reduced, particularly for small alternatives. The programm, on which the figures are based, is not accessible anymore.

Besides the joint normality of (Z,X,X^*), it is also possible to use the same methods by specifying other forms of the joint distribution. Regarding the joint normality assumption of (Z,X,X^*) of the implemented program, Tosteson et al. find a moderate to high sensitivity to the introduction of skewed exposure and measurement error distributions (Tosteson et al. 2003, p. 1076 f.). For skewed exposures, as air pollution exposure, it is therefore necessary to adequately specify the joint distribution in order to obtain correct power functions. Furthermore, Tosteson et al. verify the properties of their power function by computer-simulations using a logistic regression model. The simulations show that the power function leads to correct sample sizes for detecting small as well as large alternatives (Tosteson et al. 2003, p. 1076).

In summary, approaches of three papers, Devine (2003), McKeown-Eyssen and Tibshirani (1994) and Tosteson et al. (2003) were described. The approach by Devine is restricted to linear regression under classical measurement error, while McKeown-Eyssen and Tibshirani derive their calculations for logistic regression under linear measurement error. In contrast, Tosteson et al. give sample size calculation methods that are applicable to a broad variety of GLMs and measurement error models with the inclusion of confounders. They base their calculations upon the asymptotic theory of the properties of the score test for fixed alternatives under measurement error. All approaches assume non-differential error and a single error-prone exposure variable. Concerning the relevance of these methods in air pollution epidemiology, the approach by Tosteson et al. seems more important than the previously described methods by Devine and McKeown-Eyssen and Tibshirani. The models used in air pollution epidemiology are more complex than simple linear regression and vary by the study design. Tosteson's approach is very flexible and can be adapted to individual requirements by specifying the measurement error model and the joint distribution of (Z,X,X^*). More research has to be done in order to investigate how such methods can be used in specific practical applications in air pollution epidemiology. The increase

of sample size by the factor $1/(\rho^2_{XX^*})$, which is proposed by Devine for linear regression and by McKeown-Eyssen and Tibshirani for logistic regression, can serve as an approximation of the required sample size for other models.
Sample size adjustment by using the equation $N^* = N/(\rho^2_{XX^*})$ of course involves the calculation of N, the sample size that would be needed for the true measure X. Since N has to be estimated as well, one also has to expect a certain degree of uncertainty in N. To estimate N, one has to use any information available about the distribution of X, in the best case observations of X by an internal validation study, and apply an adequate sample size calculation method depending on the study design and model used for statistical analysis. For example the required sample size N for error-free variables in a case-control study, assuming normally distributed exposure X and exponential risks of disease $r(X) = exp(a + bX)$, can be calculated by (McKeown-Eyssen and Tibshirani 1994, p. 416):

$$N = \frac{2(t_\alpha - t_\delta)^2}{\sigma^2_X \beta^2_X}, \tag{28}$$

where t_α and t_δ are the critical values of the t distribution for the desired significance level α and power $1-\delta$. One has to specify σ^2_X, the true exposure variance in the study population. Furthermore, one has to specify β_X, the slope parameter of the association between exposure and outcome that is regarded as clinically or epidemiologically important (McKeown-Eyssen and Tibshirani 1994, p. 416). Consequently, calculating N involves to already make assumptions or estimates for these parameters.
However, it is clear, that measurement error in the exposure variable can call for an enormous increase of sample size which is dependent on the coefficient of validity $\rho^2_{XX^*}$. Figure (7) shows sample size adjustment factors $(1/\rho^2_{XX^*})$ dependent on the correlation ρ_{XX^*} to illustrate sample size adjustment according to the formula $N^* = N/\rho^2_{XX^*}$.

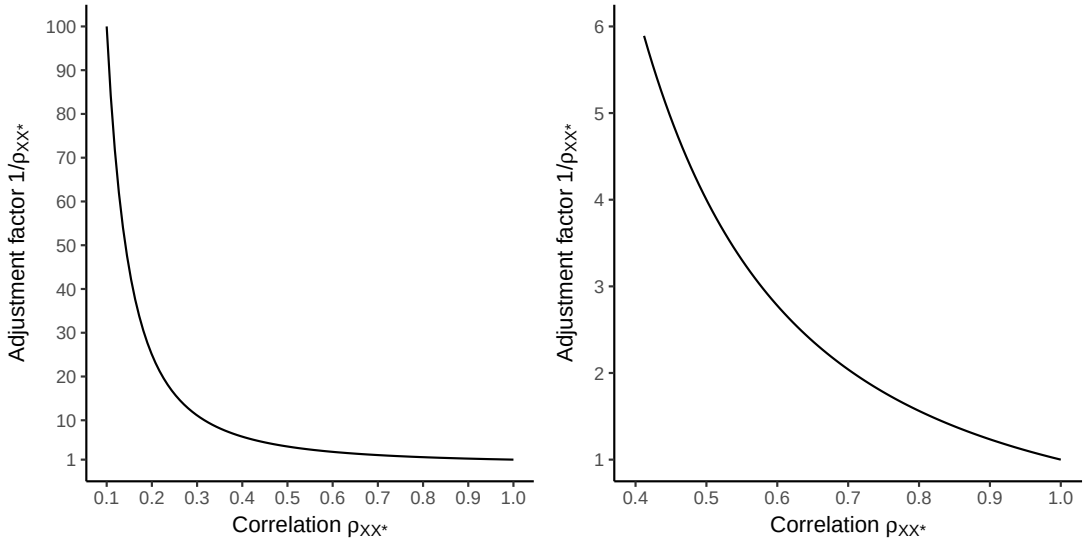

Figure 7: Sample size adjustment factors dependent on the correlation ρ_{XX^*}.

The plot on the left side shows the adjustment factors $1/(\rho^2_{XX^*})$ for a whole range of values for ρ_{XX^*}. The plot on the right side is an enlarged version of the left plot for values of ρ_{XX^*} from 0.4 to 1.0. As can be seen, the impact of measurement error on sample size can be considerable. If the correlation between X and X^* is less than 0.5, then at least a fourfold increase of sample size will be required. Only for correlations larger than 0.7 the increase in required sample size will be less than twofold.

For air pollution studies, supposing that the true measure X is measured by personal monitors and the surrogate X^* is obtained from ambient monitors, the correlation coefficients ρ_{XX^*} vary a lot among different air pollution studies. As described in section 2.3, Avery et al. (2010) provide a review of 18 studies that examine the within-participant ambient-personal correlation of $PM_{2.5}$. The correlation coefficients vary with significant heterogeneity from 0.09 to 0.83 with a median of 0.54 (Avery et al. 2010, p. 2018 f.). This leads to the conclusion that at least for some studies the required increase of sample size for measurement error adjustment can be tremendous. This sample size increase may sometimes be infeasible to realize in practice. Then, one should better try to reduce measurement error by improving measurement methods in order to achieve higher correlation coefficients ρ_{XX^*}. However, the determination of the sample size is an important step of designing a study which is crucial for a study's ability to detect significant associations between outcome and exposure. All of the sample size adjustment methods demand information from previous validation studies which underlines the necessity of validation studies.

6 Summary and Discussion

In summary, it is recommended to design a study such that the need of precise results, like accurate health risk estimates, and the need for a cost-efficient study design are in a good balance. This means in a concrete way that collecting data of personal exposures might indeed lead to precise risk estimates but are on the other hand often very expensive and can thus reduce the sample size because of cost constraints which can then reduce the statistical power of the study. Thus, in some cases exposure measurement from fixed site monitors can be more cost-efficient. As opposed to that, if the surrogate exposure, which is usually the fixed site monitor measurement, is a bad representative of the personal exposure, measurement error will have a severe impact on the study's results. That is, measurement error reduces the effective sample size which reduces the statistical power of a study to detect truly significant results. Also, measurement error can cause in some cases critical bias in the estimated coefficients that describe the association between exposure and outcome. In the case of severe measurement error due to inaccurate surrogate measurements, the only use of the cheap surrogate measurements might be inadequate. In many cases one can achieve a good trade-off with precise results and efficient allocation of resources by conducting a validation study, additionally to the main study. Validation studies have to be well-designed to allocate resources optimally to the main and the validation study. The primary method to adjust the study design for measurement error is to increase the sample size. Appropriate sample size adjustment approaches have to be applied while using correctly specified measurement error models adapted to individual needs.

After determining the study design, it is generally recommended to conduct a simulation-based evaluation of the study design prior to data collection to verify the adequacy of planned sample sizes, assumptions on parameters and other design attributes (Devine 2003, p. 336). A carefully considered study design can be even more valuable than a complex statistical analysis. More investigations are needed to allow for a practically orientated and feasible application of the proposed methods. For this, it is necessary to provide accessible guidelines and software to calculate power and adjusted sample sizes in the presence of measurements error. They have to be applicable for a broad class of models, like logistic regression models, GLMs or GAMs, and it has to be possible to adapt measurement error models to individual requirements. An accurate specification of the measurement error model is important as measurement error in en-

vironmental epidemiology is often a mixture of different error types.
It is important to notice that all the points discussed in this thesis assume that the relationship between the health outcome and the true exposure is of interest. Sometimes this may not be the case. Since regulations affect the ambient air, one may rather be interested in predicting the effect resulting from a change of measured ambient exposure instead of personal exposure (Thomas et al. 1993, p. 90).
Furthermore, one should keep in mind that models in air pollution studies also have to control for confounders although confounders are mostly not considered in measurement error related literature. In general, the impact of measurement error on the effect estimates of interest is thought to be increased by the presence of confounders, either measured with or without error (Armstrong 1998, p. 654). Further research is needed regarding how to include confounders into the relevant formulas.
In addition, most literature concerning measurement error in air pollution studies refers only to a single pollutant. Also, because the US Environmental Protection Agency (EPA) regulates the different pollutants independently, many studies investigate the effect of a particular pollutant on health. In fact, the correct identification of this effect is challenging as it demands an accurate adjustment for confounding caused by a complicated mixture of co-pollutants. Further work regarding this issue will provide a benefit for the whole field of air pollution epidemiology (Dominici et al. 2003, p. 268 f.).
Current air pollution studies should also keep in mind recently developed exposure metrics, like GIS-based metrics or models using meteorological or human activity data. Dionisio et al. review the relevance of these exposure metrics and conclude that they have the potential to reduce measurement error by increasing spatial resolution of the exposure data compared to the only use of fixed site monitor data (Dionisio et al. 2016, p. 496 ff.).
There is extensive literature that deals with the effect of measurement error on effect estimates and that provides sophisticated methods for correcting biased coefficients by statistical fixes. There is also a large number of articles related to measurement error and epidemiology. However, literature that gives concrete and feasible instructions for designing an epidemiology study in the presence of measurement error is limited. Future investigations are needed as this topic can be highly complex, particularly in air pollution epidemiology.
Further literature, that deals with different topics mentioned in this thesis, is useful for a deeper understanding. Related to air pollution measurement, Hsu et al. (2012) investigate factors affecting personal air pollution exposure by conducting studies in Seattle and New York. Setton et al. (2011) examine the impact of ignoring daily mobility patterns for studies of traffic-related air pollution. Furthermore, Szpiro et al. (2010)

propose an exposure prediction approach that considers complex spatio-temporal correlations and accommodates spatio-temporal misaligned data. Baxter et al. (2013) and Özkaynak et al. (2013) present methods for alternative exposure metrics, like different exposure prediction approaches, and give future recommendations. Further, Goldman et al. (2012) refer to geostatistical air pollution modeling and use modeled exposures to examine measurement error in a time-series design. Gryparis et al. (2008) provide a framework for measurement error in exposure predictions related to the topic of spatial misalignment. Furthermore, concerning spatial misalignment and controlling for unmeasured confounding, Lee and Sarran (2015) propose an approach that addresses both challenges and provide software in form of an R-package for application of their proposed model. Regarding sample size and power calculations, Tosteson and Tsiatis (1988) show calculations for the efficient score test statistic for surrogate covariates and the asymptotic relative efficiency. Devine and Smith (1998) provide another paper that deals with the effects of measurement error concerning the required sample size under measurement error. With regard to the difficulty in air pollution epidemiology that humans are simultaneously exposed to a complex mixture of different pollutants, Dominici et al. (2010) discuss the topic of shifting from a single-pollutant to a multi-pollutant approach from different perspectives.

After all, much work remains to be done in the development of efficient designs in air pollution epidemiology that will help to answer current questions about the impact of air pollution on health. Such studies will be of high importance since they will ultimately influence regulatory policies of governments to protect public health.

References

Armstrong, B. G. (1996). Optimizing power in allocating resources to exposure assessment in an epidemiologic study, *American journal of epidemiology* **144**(2): 192–197.

Armstrong, B. G. (1998). Effect of measurement error on epidemiological studies of environmental and occupational exposures., *Occupational and environmental medicine* **55**(10): 651–656.

Avery, C. L., Mills, K. T., Williams, R., McGraw, K. A., Poole, C., Smith, R. L. and Whitsel, E. A. (2010). Estimating error in using ambient $PM_{2.5}$ concentrations as proxies for personal exposures, *Epidemiology (Cambridge, Mass.)* **21**(2): 215–223.

Baxter, L. K., Dionisio, K. L., Burke, J., Sarnat, S. E., Sarnat, J. A., Hodas, N., Rich, D. Q., Turpin, B. J., Jones, R. R., Mannshardt, E., Kumar, N., Beevers, S. D. and Özkaynak, H. (2013). Exposure prediction approaches used in air pollution epidemiology studies: key findings and future recommendations, *Journal of Exposure Science and Environmental Epidemiology* **23**(6): 654–659.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*, CRC Press.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*, 2. edn, CRC Press.

De Nazelle, A., Seto, E., Donaire-Gonzalez, D., Mendez, M., Matamala, J., Nieuwenhuijsen, M. J. and Jerrett, M. (2013). Improving estimates of air pollution exposure through ubiquitous sensing technologies, *Environmental Pollution* **176**: 92–99.

Devine, O. (2003). The impact of ignoring measurement error when estimating sample size for epidemiologic studies, *Evaluation & the health professions* **26**(3): 315–339.

Devine, O. J. and Smith, J. M. (1998). Estimating sample size for epidemiologic studies: the impact of ignoring exposure measurement uncertainty, *Statistics in medicine* **17**(12): 1375–1389.

Dionisio, K. L., Baxter, L. K., Burke, J. and Özkaynak, H. (2016). The importance of the exposure metric in air pollution epidemiology studies: When does it matter, and why?, *Air Quality, Atmosphere & Health* **9**(5): 495–502.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G. J. and Speizer, F. E. (1993). An association between air pollution and mortality in six u.s. cities, *New England Journal of Medicine* **329**(24): 1753–1759.

Dominici, F., Peng, R. D., Barr, C. D. and Bell, M. L. (2010). Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach, *Epidemiology (Cambridge, Mass.)* **21**(2): 187–194.

Dominici, F., Sheppard, L. and Clyde, M. (2003). Health effects of air pollution: A statistical review, *International Statistical Review* **71**(2): 243–276.

Dupont, W. D. and Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression, *Controlled clinical trials* **19**(6): 589–601.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy, *International Encyclopedia of the social & Behavioral sciences* **6**: 4027–4030.

Freedman, L. S., Shaw, P. A., Deffner, V., Dodd, K. W., Gustafson, P., Carroll, R. J., Keogh, R. H., Kipnis, V., Küchenhoff, H. and Tooze, J. A. (2017). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology, *unpublished* .

Goldman, G. T., Mulholland, J. A., Russell, A. G., Gass, K., Strickland, M. J. and Tolbert, P. E. (2012). Characterization of ambient air pollution measurement error in a time-series health study using a geostatistical simulation approach, *Atmospheric environment* **57**: 101–108.

Greenland, S. (1988a). Statistical uncertainty due to misclassification: implications for validation substudies, *Journal of clinical epidemiology* **41**(12): 1167–1174.

Greenland, S. (1988b). On sample-size and power calculations for studies using confidence intervals, *American Journal of Epidemiology* **128**(1): 231–237.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. and Coull, B. A. (2008). Measurement error caused by spatial misalignment in environmental epidemiology, *Biostatistics* **10**(2): 258–274.

Holford, T. R. and Stack, C. (1995). Study design for epidemiologic studies with measurement error, *Statistical Methods in Medical Research* **4**(4): 339–358.

Hsu, S.-I., Ito, K., Kendall, M. and Lippmann, M. (2012). Factors affecting personal exposure to thoracic and fine particles and their components, *Journal of Exposure Science and Environmental Epidemiology* **22**(5): 439–447.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J. and Giovis, C. (2005). A review and evaluation of intraurban air pollution exposure models, *Journal of Exposure Analysis and Environmental Epidemiology* **15**(2): 185–204.

Kanaroglou, P. S., Jerrett, M., Morrison, J., Beckerman, B., Arain, M. A., Gilbert, N. L. and Brook, J. R. (2005). Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach, *Atmospheric Environment* **39**(13): 2399–2409.

Kioumourtzoglou, M.-A., Spiegelman, D., Szpiro, A. A., Sheppard, L., Kaufman, J. D., Yanosky, J. D., Williams, R., Laden, F., Hong, B. and Suh, H. (2014). Exposure measurement error in $pm_{2.5}$ health effects studies: A pooled analysis of eight personal exposure validation studies, *Environmental Health* **13**(1): 2–11.

Künzli, N. and Tager, I. (1997). The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies., *Environmental Health Perspectives.* **105**(10): 1078–1083.

Lagakos, S. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable, *Statistics in medicine* **7**(1-2): 257–274.

Lee, D. and Sarran, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies, *Environmetrics* **26**(7): 477–487.

Maclure, M. and Mittleman, M. (2000). Should we use a case-crossover design?, *Annual review of public health* **21**(1): 193–221.

MacMahon, B. and Pugh, T. F. (1970). *Epidemiology: principles and methods.*, Boston: Little Brown & Co. Published in Great Britain by J. & A. Churchill, London.

Mallick, B., Hoffman, F. O. and Carroll, R. J. (2002). Semiparametric regression modeling with mixtures of berkson and classical error, with application to fallout from the nevada test site, *Biometrics* **58**(1): 13–20.

McKeown-Eyssen, G. E. and Thomas, D. C. (1985). Sample size determination in case-control studies: the influence of the distribution of exposure, *Journal of chronic diseases* **38**(7): 559–568.

McKeown-Eyssen, G. E. and Tibshirani, R. (1994). Implications of measurement error in exposure for the sample sizes of case-control studies, *American journal of epidemiology* **139**(4): 415–421.

National Research Council (2001). *Research Priorities for Airborne Particulate Matter: III. Early Research Progress*, Vol. 3, National Academies Press.

Nieuwenhuijsen, M. J. (2015). *Exposure assessment in environmental epidemiology*, 2. edn, Oxford University Press.

Özkaynak, H., Baxter, L. K., Dionisio, K. L. and Burke, J. (2013). Air pollution exposure prediction approaches used in air pollution epidemiology studies, *Journal of Exposure Science and Environmental Epidemiology* **23**(6): 566–572.

Peng, R. D. and Bell, M. L. (2010). Spatial misalignment in time series studies of air pollution and health data, *Biostatistics* **11**(4): 720–740.

Pratt, J. W. and Gibbons, J. D. (1981). Asymptotic relative efficiency, *Concepts of Nonparametric Theory*, Springer, pp. 345–424.

Rosner, B., Willett, W. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error, *Statistics in medicine* **8**(9): 1051–1069.

Rothman, K. J., Greenland, S. and Lash, T. L. (2008). *Modern epidemiology*, 3. edn, Lippincott Williams & Wilkins.

Sarnat, J. A., Koutrakis, P. and Suh, H. H. (2000). Assessing the relationship between personal particulate and gaseous exposures of senior citizens living in baltimore, md, *Journal of the Air & Waste Management Association* **50**(7): 1184–1198.

Sarnat, S. E., Klein, M., Sarnat, J. A., Flanders, W. D., Waller, L. A., Mulholland, J. A., Russell, A. G. and Tolbert, P. E. (2010). An examination of exposure measurement error from air pollutant spatial variability in time-series studies, *Journal of Exposure Science and Environmental Epidemiology* **20**(2): 135–146.

Setton, E., Marshall, J. D., Brauer, M., Lundquist, K. R., Hystad, P., Keller, P. and Cloutier-Fisher, D. (2011). The impact of daily mobility on exposure to traffic-related air pollution and health effect estimates, *Journal of Exposure Science and Environmental Epidemiology* **21**(1): 42–48.

Shaw, P. A., Deffner, V., Dodd, K. W., Freedman, L. S., Keogh, R. H., Kipnis, V., Küchenhoff, H. and Tooze, J. A. (2017). Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations., *unpublished* .

Sheppard, L., Burnett, R. T., Szpiro, A. A., Kim, S.-Y., Jerrett, M., Pope, C. A. and Brunekreef, B. (2012). Confounding and exposure measurement error in air pollution epidemiology, *Air Quality, Atmosphere & Health* **5**(2): 203–216.

Sheppard, L., Slaughter, J. C., Schildcrout, J., Liu, L. S. and Lumley, T. (2005). Exposure and measurement contributions to estimates of acute air pollution effects, *Journal of Exposure Analysis and Environmental Epidemiology* **15**(4): 366–376.

Spiegelman, D. (1994). Cost-efficient study designs for relative risk modeling with covariate measurement error, *Journal of Statistical Planning and Inference* **42**(1-2): 187–208.

Spiegelman, D. and Gray, R. (1991). Cost-efficient study designs for binary response data with gaussian covariate measurement error, *Biometrics* pp. 851–869.

Suh, H. H. and Zanobetti, A. (2010). Exposure error masks the relationship between traffic-related air pollution and heart rate variability (hrv), *Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine* **52**(7): 685–692.

Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D. and Kaufman, J. D. (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies, *Environmetrics* **21**(6): 606–631.

Szpiro, A. A., Sheppard, L. and Lumley, T. (2011). Efficient measurement error correction with spatially misaligned data, *Biostatistics* **12**(4): 610–623.

Thomas, D., Stram, D. and Dwyer, J. (1993). Exposure measurement error: influence on exposure-disease relationships and methods of correction, *Annual review of public health* **14**(1): 69–93.

Tosteson, T. D., Buzas, J. S., Demidenko, E. and Karagas, M. (2003). Power and sample size calculations for generalized regression models with covariate measurement error, *Statistics in medicine* **22**(7): 1069–1082.

Tosteson, T. D. and Tsiatis, A. A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates, *Biometrika* **75**(3): 507–514.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, Vol. 3, Cambridge university press.

Van Roosbroeck, S., Li, R., Hoek, G., Lebret, E., Brunekreef, B. and Spiegelman, D. (2008). Traffic-related outdoor air pollution and respiratory symptoms in children: the impact of adjustment for exposure measurement error, *Epidemiology* **19**(3): 409–416.

Wacholder, S., Armstrong, B. and Hartge, P. (1993). Validation studies using an alloyed gold standard, *American Journal of Epidemiology* **137**(11): 1251–1258.

White, E., Kushiz, L. H. and Pepe, M. S. (1994). The effect of exposure variance and exposure measurement error on study sample size: Implications for the design of epidemiologic studies, *Journal of clinical epidemiology* **47**(8): 873–880.

Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D. and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: concepts and consequences, *Environmental Health Perspectives* **108**(5): 419–426.

List of Figures

Statutory Declaration

I hereby declare that this thesis is my own original work and that information which has been directly or indirectly taken from other sources has been noted as such.

Munich, December 14, 2017

Elisabeth Pangratz