METHODS TO ACCOUNT FOR OUTCOME MISCLASSIFICATION IN EPIDEMIOLOGY

Jessie K. Edwards

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology.

Chapel Hill
2013

Approved by:

Stephen R, Cole

Amy H. Herring

Andy F. Olshan

David B. Richardson

Melissa A. Troester

Abstract

JESSIE K. EDWARDS: Methods to account for outcome misclassification in epidemiology
(Under the direction of Stephen R. Cole)

Outcome misclassification occurs when the endpoint of an epidemiologic study is measured with error. Outcome misclassification is common in epidemiology but is frequently ignored in the analysis of exposure-outcome relationships. We focus on two common types of outcomes in epidemiology that are subject to mismeasurement: participant-reported outcomes and cause-specific mortality. In this work, we leverage information on the misclassification probabilities obtained from internal validation studies, external validation studies, and expert opinion to account for outcome misclassification in various epidemiologic settings.

This work describes the use of multiple imputation to reduce bias when validation data are available for a subgroup of study participants. This approach worked well to account for bias due to outcome misclassification in the odds ratio and risk ratio comparing herpes simplex virus recurrence between participants randomized to receive acyclovir or placebo in the Herpetic Eye Disease Study. In simulations, multiple imputation had greater statistical power than analysis restricted to the validation subgroup, yet both provided unbiased estimates of the odds ratio.

Modified maximum likelihood and Bayesian methods are used to explore the effects of outcome misclassification in situations with no validation subgroup. In a cohort of textile workers exposed to asbestos in South Carolina, we perform sensitivity analysis using modified maximum likelihood to estimate the rate ratio of lung cancer death per 100 fiber-years/mL asbestos exposure under varying assumptions about sensitivity and specificity. When specificity of outcome classification was nearly perfect, the modified maximum likelihood approach produced estimates that were similar to analyses that ignore outcome misclassification.

Uncertainty in the misclassification parameters is expressed by placing informative prior distributions on sensitivity and specificity in Bayesian analysis. Because, in our examples, lung cancer death is unlikely to be misclassified, posterior estimates are similar to standard estimates. However, modified maximum likelihood and Bayesian methods are needed to verify the robustness of standard estimates, and these approaches will provide unbiased estimates in settings with more misclassification.

This work has highlighted the potential for bias due to outcome misclassification and described three flexible tools to account for misclassification. Use of such techniques will improve inference from epidemiologic studies.

Acknowledgements

I'm grateful for the support of the many people who have been at my side throughout my time at UNC.

Steve Cole has been an exceptional mentor who has encouraged, and challenged, me to grow academically and professionally.

As my thesis committee, Amy Herring, Bob Millikan, Andy Olshan, David Richardson, and Melissa Troester provided invaluable advice about epidemiology, statistics, and the process of writing papers.

My friends in the epidemiology department have pushed me, and continue to inspire me, to stretch beyond my comfort zone and learn new things.

My parents and extended family have enthusiastically charted my progress and had confidence in my abilities even when I did not.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CI	Confidence interval
HEDS	Herpetic Eye Disease Study
HSV	Herpes simplex virus
MI	Multiple imputation
ML	Maximum likelihood
OR	Odds ratio
PI	Posterior interval
RR	Risk ratio

1 Background

The goal of many epidemiologic studies is to obtain an accurate estimate of the effect of an exposure on the occurrence of an event of interest. Measurement bias, selection bias, and confounding are three types of systematic errors that threaten the validity of results from epidemiologic studies. Selection bias and confounding are often considered in analysis of epidemiologic data, but the possible biases arising from measurement error are more routinely ignored.

Two problems occur in studies that disregard the potential for systematic bias: point estimates may be biased and uncertainty about the degree of systematic bias is not quantified. Most epidemiologic studies ignore the potential bias in point estimates and present some quantification of random error, such as the confidence interval, as the sole indicator of uncertainty in their results, ignoring the additional uncertainty that arises from systematic error. This systematic error is often more substantial than random error in large epidemiologic studies that are increasingly common (1,2).

All variables used in epidemiology are subject to mismeasurement. Error in the measurement of exposure, covariates, or outcomes can produce bias in effect estimates from epidemiologic studies, known as information bias. Most work addressing information bias focuses on exposure measurement error (3–8), but error in covariates or outcomes can also produce biased effect estimates (9).

This work begins by describing outcome misclassification, illustrating its potential to cause bias in two distinct types of epidemiologic studies, and describing existing approaches to account for outcome misclassification. Chapter 2 outlines the specific aims of the proposed work, while chapter 3 details two proposed methods to account for outcome misclassification. Chapters 4 and 5 present results from the implementation of these methods in two settings, and chapter 6 summarizes the findings and offers discussion of the results.

1.1 Overview of outcome misclassification

Consider the two-by-two table of binary observed and true outcome measurements, W and D, respectively:

	$W=1$	$W=0$	
$D=1$	a	b	$a+b$
$D=0$	c	d	$c+d$
	$a+c$	$b+d$	

Using this notation, sensitivity, or the probability of being observed to be a case given that a participant is a true case, is defined as $P(W = 1|D = 1)$, and can be calculated from the table above as $a/(a+b)$. Specificity, the probability of being observed to be a non-case given that a participant is actually not a case, is similarly defined as $P(W = 0|D = 0)$, and can be calculated from the table above as $d/(c+d)$.

Sensitivity and specificity allow an investigator to hypothesize about the distribution of the observed data assuming he or she has knowledge of the true distribution

of the data and a guess about the values of sensitivity and specificity. For example, one can calculate the expected probability of a participant's being observed to experience the outcome, $P(W = 1)$, in a study with specific values of sensitivity (se) and specificity (sp) and a known probability of experiencing the outcome, $P(D = 1)$, using the equation:

$$P(W = 1) = P(D = 1) \times se + P(D = 0) \times (1 - sp).$$

In practical applications, however, investigators often have knowledge only about the observed data (i.e., $P(W = 1)$ and $P(W = 0)$, or the margins $a + c$ and $b + d$ in the table above) and wish to make inferences about the true distribution of the outcome variable, $P(D = 1)$ and $P(D = 0)$ (the margins $a + b$ and $c + d$ in the table above).

To do this, investigators need information about the positive predictive value (PPV) and negative predictive value (NPV) of the observed outcome variable. The positive predictive value is the probability that true outcome occurred, given than investigators observed the outcome to occur, $P(D = 1|W = 1)$. Assuming full knowledge of the two-by-two table above, this probability can also be calculated as $a/(a + c)$. Similarly, the negative predictive value is the probability that a participant was not a case, given that he or she was not observed to experience the outcome, or $P(D = 0|W = 0)$. With full knowledge of the two-by-two table above, the negative predictive value could be calculated as $d/(b + d)$.

An investigator can estimate the true probability of the outcome $P(D = 1)$ using the observed probability of the outcome $P(W = 1)$ and the positive and negative predictive values (PPV and NPV) through the formula

$$P(D = 1) = P(W = 1) \times PPV + P(W = 0) \times (1 - NPV).$$

However, positive predictive value and negative predictive value are usually not available unless an investigator conducts an internal validation study nested within the

main study with the possibly misclassified outcome. Sensitivity and specificity are more widely-reported measures of outcome validity because these parameters are functions only of the outcome misclassification process. Positive and negative predictive values depend on the sensitivity and specificity of the observed outcome measure, but also on the prevalence of the outcome in the study population. To see this, consider how sensitivity and specificity can be related to positive and negative predictive value using Bayes' Theorem:

$$P(D = 1|W = 1) = \frac{P(W = 1|Y = 1)P(D = 1)}{P(W = 1|D = 1)P(D = 1) + P(W = 1|D = 0)P(D = 0)},$$

which can be simplified to

$$P(D = 1|W = 1) = \frac{se \times \pi}{se \times \pi + (1 - sp) \times (1 - \pi)},$$

where π represents the $P(D = 1)$ or the disease prevalence.

1.1.1 Differential and nondifferential outcome misclassification

Sensitivity and specificity of the outcome measure can be uniform over the entire study population or can differ by levels of exposure or other covariates. Consider the 2 by 2 table presented above stratified by exposure level to produce a 2 by 2 by k table, where k is the number of exposure levels.

X=1

	W=1	*W=0*	
D=1	a	b	$a + b$
D=0	c	d	$c + d$
	$a + c$	$b + d$	

X=0

	W=1	*W=0*	
D=1	a	b	$a + b$
D=0	c	d	$c + d$
	$a + c$	$b + d$	

Sensitivity and specificity are said to be nondifferential with respect to exposure X if $P(W = 1|D = 1, X = x)$ and $P(W = 0|D = 0, X = x)$ are the same for all possible values of X. Outcome misclassification is said to be differential with respect to exposure status when values of sensitivity and specificity differ across levels of exposure X. Because the positive and negative predictive values are functions of the prevalence of the outcome, they can differ within levels of exposure even with outcome misclassification is nondifferential with respect to exposure status (10).

1.1.2 Types of bias caused by outcome misclassification

Outcome misclassification can cause several distinct problems in epidemiology. First, outcome misclassification can cause errors in the overall estimation of outcome incidence or prevalence. If sensitivity of the observed outcome measure is less than 1, then false negatives can occur, in which a participant truly experiencing the outcome of interest is recorded not to have had the outcome. In specificity of the observed outcome measure is less than 1, false positives can occur, in which a participant who does not experience the outcome of interest is recorded to have the outcome. If more false negatives occur than false positives, the overall probability of the outcome will be underestimated. If more false positives occur, the probability of the outcome will be overestimated. Overall, imperfect sensitivity and specificity will cause error in the estimated incidence or prevalence of disease unless the number of false positives is exactly equal to the number of false negatives. Error in the marginal probability of the outcome causes bias not only in estimates of the overall burden of disease, but also in estimates of disease trends over time.

Error in disease trends is compounded with the misclassification probabilities, sensitivity and specificity, also change over time (11).

Outcome misclassification can also cause bias in estimates of the effect of an exposure variable on an outcome. When outcome misclassification is nondifferential with respect to exposure and the outcome is binary, bias in estimates of the effect of the exposure on the outcome is usually expected to be towards the null. However, this rule does not hold when the outcome has more than two levels or errors in exposure and outcome are not independent of each other (10,12–15). When outcome misclassification is differential with respect to exposure status, bias in estimates of the effect of the exposure on the outcome could be in either direction.

1.1.3 Methods for assessing the probability of misclassification

If a gold-standard measure of the outcome exists, investigators wishing to assess the amount of misclassification in a study may choose to conduct an internal validation study. In this type of study, the gold-standard outcome measurement is taken on a (possibly stratified) random subset of study participants. The gold-standard outcome measure is compared to the fallible outcome observed in the original study to calculate the sensitivity and specificity of the observed outcome measure.

Studies unable to conduct an internal validation study could rely on an external validation study to assess the amount of misclassification likely to have occurred. Sensitivity and specificity can be gleaned from an external validation study if a gold-standard measure exists and such a validation study is available.

Sensitivity and specificity from both internal and external validation studies will only accurately reflect the amount of misclassification in the main study if the relationship between gold-standard outcome and observed outcome is transportable between the validation study and the main study. Transportability implies that misclassification parameters are the same in the validation study and the main study, and would be expected for an internal validation study consisting of a random subgroup of the main study (16). Transportability is not assured for external validation studies or internal validation studies that are not a random subgroup of participants. In both situations, the validation study could represent a group of participants with characteristics different from participants in the main study. External validation studies carry the additional risk of nontransportability if the observed outcome measurement in the validation study was conducted differently from the observed outcome measurement in the main study.

1.2 Error in participant-reported outcomes

The outcome could be recorded with error for a variety of reasons, and the opportunity for outcome misclassification arises throughout the study. This work focuses on two common types of outcome misclassification in epidemiology: error in participant-reported outcomes and misattribution of cause of death on death certificates, leading to misclassification of cause-specific mortality outcomes.

Many epidemiologic studies rely on participants to report disease symptoms and events of interest. Participant-reported outcomes are especially prevalent in studies of recurring nonfatal diseases, such as dermatological conditions (17), allergy, cold, and flu symptoms, and signs gastrointestinal illness (18,19). Participant-reported outcomes are

also known as patient-reported outcomes, a term which has evolved to include "any endpoint derived from patient reports, whether collected in the clinic, in a diary, or by other means, including single-item outcome measures, event logs, symptom reports, formal instruments to measure health-related quality of life, health status, adherence, and satisfaction with treatment." (20) These types of patient-reported outcomes have been used extensively in drug effectiveness research since the 1990s. Willke (20) reports that 30% of drug product labels approved between 1997 and 2002 used patient-reported outcomes as effectiveness study endpoints.

Recording outcomes described by participants is especially useful when time and cost constraints make frequent contact with investigators or physicians difficult or impractical. In such situations, investigators may have contact with the participant at study enrollment and either at the end of the study period or at the end of pre-defined intervals of time, at which point the participant reports any events of interest occurring during the study period or interval. In these settings, participants may be instructed to keep a diary of outcome events on a daily or weekly basis. In other settings, a participant may be contacted by the investigator at only one point in time, at which time the investigator will ask the participant to recall events of interest in his or her past.

Opportunities for bias arise in all study designs using participant-reported outcomes. A prospective study, in which the participant is aware of being under observation and is instructed to keep a diary of outcomes over the study period, may be subject to bias if the participant over-reports symptoms or fails to be diligent about recording events of interest. Likewise, a retrospective study, in which a participant is recruited to the study and then asked to recall symptoms or events of interest in the past, is

also subject to bias if participants fail to remember past events or inflate their number of past events.

Participants may misreport their outcomes due to errors in recall or due to social pressure to report one outcome over another. This social desirability bias arises most often when the outcome of interest is something within the direct control of the participant, such as a behavior, or when the outcome is taboo or embarrassing.

If participants in a prospective study do not know their exposure status, such as in a masked randomized trial, the outcome misclassification from under- or over-reporting symptoms or events of interest is likely to be nondifferential with respect to exposure. This means that the probability that a participant reports an event that occurred during the study period and the probability that a participant falsely reports an event that did not occur during the study period are not different for participants with different exposure values. In a masked randomized trial in which the participant does not know his or her exposure status, exposure is not likely to affect the participant's reporting of the outcome. Because the probability of misclassification is the same for exposed and unexposed groups, effect estimates for binary outcomes subject to nondifferential misclassification are usually biased towards the null.

However, if participants in the prospective study do know their exposure status, outcome misclassification may be differential with respect to exposure. This means that exposed participants may be more or less likely to report an event that actually occurred or falsely report an event that did not occur than their unexposed counterparts. When outcome misclassification is differential with respect to exposure, bias could be in either direction.

Differential outcome misclassification is even more likely in retrospective studies using participant-reported outcomes. When participants are asked to recall prior instances of events or symptoms, it is possible that participants who were exposed are more likely to remember events of disease than participants who were not exposed. For example, a mother living in a city blanketed in smog may be more likely to remember respiratory illnesses in her children than a mother living in an area with no smog, even if the children had similar respiratory histories. Differential outcome misclassification in retrospective studies is similar to recall bias in case control studies, in which cases are more likely to remember exposures than controls even with no association exists between exposure and outcome (21).

1.2.1 Herpetic Eye Disease Study

As an illustrative example, this work addresses outcome misclassification due to errors in participant-reported information in the Herpetic Eye Disease Study, a randomized trial of acyclovir for preventing ocular herpes simplex virus (HSV) recurrence. Ocular HSV infection can cause corneal opacities and vision loss (22), and at least 500,000 people in the United States are infected (23). Treatment of HSV is estimated to cost approximately $17.7 million annually to treat about 59,000 new and recurrent cases (24).

Recurrent infections are a major contributor to vision loss from ocular HSV. After an initial (usually asymptomatic) infection, HSV establishes a latent infection in the trigeminal or other sensory ganglia. From these locations, recurrent viral shedding can lead to infection in the eye. This infection can manifest as blepharitis, characterized by swelling or inflammation of the eyelids, conjunctivitis, characterized by swelling or inflammation of the

membrane lining the eyelids, or dendritic or epithelial keratitis, characterized by a linear branching corneal ulcer. Infection can also lead to stromal keratitis, causing inflammation of the cornea, or iritis, causing inflammation of the anterior uvea, both of which can lead to permanent scarring and decreased vision.

The primary objective of the Herpetic Eye Disease Study was to determine whether treatment with oral acyclovir for one year would prevent ocular recurrences in participants who had had an episode of ocular HSV during the preceding year (22). The outcome of interest in the trial was ocular HSV recurrence diagnosed by an experienced ophthalmologist using slit lamp biomicroscopy. HSV recurrence was assessed after 1, 3, 6, 9, and12 months of treatment; during the post-treatment observation period, after months 13, 15, and 18; and any time new ocular symptoms developed.

However, a companion cohort study was also performed, nested within the randomized trial of acyclovir, to assess psychological stress and other triggers of recurrent HSV infection. In this companion study, HSV recurrence was assess both by an experienced ophthalmologist and through participant self-report in weekly diaries (25). Participants were instructed to record the date of the onset of symptoms of HSV recurrence. Such symptoms included redness and swelling of the eye, blurred vision, sensitivity to light, inflammation of the eyelids, or the sensation of a foreign object in the eye.

Prospective weekly diaries have been used frequently in epidemiology to collect participant-reported data (26–29). Diaries are used most commonly for participants to report disease symptoms (29), specific behaviors (27), and adverse events (30) to minimize errors in recall at the end of the study period. Despite improvements over traditional long-term recall techniques, participant-reported outcomes recorded using

weekly diaries remain subject to outcome misclassification due to factors such as social desireability and lack of medical training to identify events of interest.

This work uses the Herpetic Eye Disease Study and its companion study of triggers of recurrence to explore methods to account for outcome misclassification in a setting with both gold standard (physician diagnosis) and error-prone (weekly diary) outcome measure available.

1.3 Misattribution of cause of death

In addition to accounting for outcome misclassification of participant-reported outcomes, this work focuses on accounting for misclassification of cause-specific mortality outcomes caused by misattribution of cause of death.

Cause-specific mortality is used as an outcome measure in place of disease incidence in many epidemiologic settings, particularly in studies of rapidly fatal diseases, such as certain cancers. Mortality is often chosen as an outcome measure over disease incidence or survival because it is accurately reported, and it is an important indicator of disease burden in its own right (31). In practical terms, a study that performs minimal follow-up on study participants but wishes to assess the relationship between exposure and disease over a long time period may choose disease mortality as an outcome measure because mortality may be the only outcome expected to be reported in a standardized manner. Even in studies following participants closely, mortality may be the preferred outcome measure because it is not as dependent on screening trends or diagnostic techniques as disease incidence. Similarly cause-specific mortality is often chosen over survival as an outcome measure because any inflation in incidence due to diagnosis of disease in patients with mild

or nonmalignant disease causes a spurious increase in survival of those diagnosed as cases (31).

Before the introduction of the national death index in the United States in 1978, vital status was determined in epidemiologic studies by examining sources such the US Social Security Administration, Internal Revenue Service, state vital statistics office, drivers' license files, and US postal service change of address forms. After 1978, vital status was recorded centrally in the National Death Index, a computerized index of death record information on file in state vital statistics offices. For all deaths identified using any of these sources, cause of death is typically abstracted from death certificates, and these causes of death are used to assign outcome statuses to participants in epidemiologic studies.

Cause of death information on the death certificate is typically completed by a physician, medical examiner, or coroner. Figure 1.1 presents the cause of death section on a United States death certificate. In part 1, the physician, medical examiner, or coroner is instructed to report the chain of events leading to directly death, with the immediate cause of death listed first and the underlying cause of death listed last. Part 2 captures all other significant diseases, conditions, or injuries that contributed to death but did not result in the underlying or immediate cause of death. The cause of death reflects the best medical opinion of the person filling out the death certificate and does not need to be supported by a definitive diagnosis in a medical setting.

To translate cause of death information on death certificates into outcome variables in epidemiologic studies, the cause of death on the death certificate is usually translated into international classification of diseases (ICD) codes by a nosologist. In general, investigators chose specific ICD codes to represent the event of interest in a study, and

participants with ICD codes from the cause of death on death certificates matching these ICD codes are designated to have the event of interest. Studies of the reliability and accuracy of cause of death information reported by physicians have revealed that the same patient reviewed by different physicians is likely to be assigned different causes of death (32).

Because coroners responsible for certifying the underlying cause of death may receive limited medical training as well as the uncertainty inherent in ascribing cause of death for some conditions, underlying cause of death reported on death certificates is error-prone. Misattribution of underlying cause of death has plagued epidemiologic studies of cause-specific mortality (33,34). Studies of etiologic relationships between exposures and cause-specific mortality as well as studies assessing secular trends of cause-specific mortality are subject to bias due to outcome misclassification caused by misattribution of underlying cause of death.

Because misattribution of underlying cause of death can lead to outcome misclassification, cause-specific mortality outcomes abstracted from death certificates have imperfect sensitivity and specificity. Recall that sensitivity is the probability of a true case being classified as such; and specificity is the probability that a true non-case being classified as such. Using autopsy data as a gold standard, sensitivity and specificity of cause of death information from death certificates is imperfect even for well-studied and relatively common endpoints like death due to cancer or coronary heart disease. For example,

Table 1.1 shows estimates ranging from 0.7 to 0.9 and 0.5 to 1.0 for sensitivity and specificity of commonly study causes of death (33–37). In addition to creating bias in estimates of effect in etiologic studies, error in cause of death reporting can also produce bias in secular trends of disease. Cancer epidemiologists have noted the impact of potential misattribution of underlying cause of death on trends of site-specific cancer mortality rates in the US and around the world (36).

Table 1.1: Sensitivity and specificity estimates of cause-specific mortality from the literature.

Study (year)	Outcome	Sensitivity	Specificity
Doria-Rose (2008)	Lung cancer mortality	89%	99%
Selikoff (1992)	Lung cancer mortality	83%	99%
Modelmog (1992)	Any cancer mortality	80%	62%
Modelmog (1992)	All-cause mortality	70%	53%
Lloyd-Jones (1998)	Coronary heart disease mortality	84%	84%

1.3.1 Misattribution of underlying cause of death in occupational cohort studies

Many cohort studies of occupational exposures have used cause-specific mortality as an outcome measure. Cause-specific mortality is often chosen over disease incidence as an outcome measure for practical reasons. Workers are usually enrolled at some point during their employment, at which time their exposure history is recorded for the duration of their employment, along with other relevant covariates. However, the diseases of interest affecting these workers typically occur later in life, at which time few are likely to be working. Because date and cause of death can be identified using publically available information from the National Death Index and death certificates, investigators can identify cause-specific mortality outcomes without performing extensive follow-up on each worker.

Like all studies using cause of death abstracted from death certificates, occupational cohort studies of cause-specific mortality are subject to bias due to outcome misclassification caused by misattribution of underlying cause of death.

1.3.2 Misattribution of cause of death in the South Carolina textile workers study

This work assesses the impact of misattribution of cause of death in the occupational setting using data from a study of workers exposed to asbestos at a South Carolina textile factory. “Asbestos” is the generic name given to a group of naturally occurring silicate minerals with fibrous structure that became commonly used as an insulator for both electrical wires and buildings due to its heat and flame resistant properties after the industrial revolution. Industrial production of asbestos began in the 1850s, and, due to its attractive fire-resistant properties, asbestos was eventually incorporated into many building materials, such as bricks, concrete, pipes, ceiling insulation, drywall, flooring, and roofing materials. However, by the middle of the 20th century, asbestos exposure had been shown to increase the risk of both malignant and non-malignant lung diseases (38).

Despite the subsequent reduction in asbestos used in manufacturing in the United States, asbestos remains a public health concern. More than 30 million tons of asbestos have been mined, processed, and used in the United States since the early 1900s (38), and mining continues in Canada for export to the developing world. In the United States alone, 27 million workers were exposed to asbestos between 1940 and 1979 (39). Although asbestos is no longer mined in the United States, approximately 1000 tons of asbestos is imported into the US each year for use in construction materials, brake linings, and other products (40). Moreover, a substantial amount of asbestos remains in US infrastructure and eventually will be removed, either during remediation or renovations or demolition. Significant production and use is also ongoing in middle-income industrial countries,

including Brazil, India, China and Russia. Therefore asbestos continues to pose important occupational hazards in the US and worldwide (41).

Epidemiologic studies have shown a relationship between exposure to asbestos and lung cancer mortality, though the carcinogenic mechanism is not fully established. The two major types of asbestos fibers are chrysotile fibers and amphiboles, including actinolite, amosite, anthrophyllite, crocidolite, and tremolite (42). Amphibole fibers are more carcinogenic than chrysotile in part because they are more biopersistent in the lung (amphiboles have an estimated half-life in the lungs of decades, while chrysotile fibers have a half-life of months (38)), accumulate in the distal lung parenchyma, and are not cleared as easily. However, both fiber types can induce DNA damage, gene transcription, and protein expression important to modulate cell proliferation and cell death in bronchial and alveolar epithelial cells (43). When asbestos fibers reach the lungs, alveolar epithelial cells and alveolar macrophages internalize the fibers, resulting in oxidative stress and the subsequent generation of reactive oxygen species and reactive nitrogen species, which can cause DNA damage.

Factors determining the probability and severity of disease include the cumulative dose of exposure, time following initial exposure, and the physical-chemical properties of the asbestos fibers. Some researchers have argued for the amphibole hypothesis: that amphiboles are carcinogenic while pure chrysotile may not cause disease because fiber structural characteristics are the primary determinant of toxicity. This hypothesis has been historically difficult to study because tremolite amphibole fibers are frequently mixed with chrysotile fibers in industrial applications. Prior to the study of South Carolina textile workers, few studies analyzing lung cancer specific mortality among factory workers

exposed only to the chrysotile form of asbestos had been conducted. However, like all studies relying on cause of death information from death certificates, estimates from the study of South Carolina textiles workers are subject to bias due to outcome misclassification from misattribution of cause of death.

Unlike the Herpetic Eye Disease Study, no internal gold standard outcome is available for participants in the South Carolina textile workers cohort. In other cohorts, limited validation studies have been performed comparing lung cancer and coronary heart disease mortality reported on death certificates to lung cancer deaths and deaths due to coronary heart disease identified through autopsies and physician diagnosis.

The Life Span Study, which performed autopsies on selected participants who died following the atomic bombings of Hiroshima and Nagasaki, found that death certificates detected that a death was due to lung cancer in only 62% of the cases where the autopsy indicated that lung cancer was the cause of death (36). A more recent study from the Mayo lung clinic in the United States reported that death certificates identified lung cancer as the cause of death in 89% (210/237) of autopsy-confirmed lung cancer cases (34), while specificity was 99%. Sensitivity from other validation studies fell between the estimates from the Life Span Study and the Mayo lung clinic. A study of 4951 deaths occurring among 17,800 workers exposed to asbestos reported that death certificates identified lung cancer as the cause of death in 86% of the deaths designated as lung cancer deaths by autopsy and other medical evidence (37).

This work uses the estimates and of sensitivity and specificity from these validations studies to inform methods to account for outcome misclassification of lung cancer mortality in the South Carolina cohort.

1.4 Existing methods to account for outcome misclassification

1.4.1 Algebraic approaches

Approaches to account for bias in crude effect estimates due to use of a misclassified binary outcome variable have existed for more than half a century (44). These approaches use simple bias correction formulas to account for misclassification in two-by-two tables. In these approaches, "true" counts of outcomes in each strata of exposure are predicted from the observed number of outcomes and given values of sensitivity and specificity.

Algebraic approaches to account for outcome misclassification can be deterministic (44) or probabilistic (10,45). Both types of algebraic methods to account for misclassification can be used as part of a sensitivity analysis in which the investigator evaluates the changes in the point estimate of the effect of the exposure on the outcome due to different hypothesized values for sensitivity and specificity. The probabilistic analysis offers an advantage in that it also provides a means to assess the uncertainty in the final point estimates due to outcome misclassification (10).

1.4.2 The EM algorithm

More recently, investigators have developed maximum likelihood approaches for logistic regression to produce effect estimates accounting for outcome misclassification while adjusting for relevant confounders (46,47).

A first existing approach (46) uses an expectation maximization algorithm (48) to estimate parameters corrected for misclassification in logistic regression. To do this,

investigators perform standard logistic regression considering each individual as both diseased and non-diseased with weights determined by the probability that the study subject is truly diseased given the data. Specifically, for individuals designated as cases by the error prone outcome indicator (W_i=1), the probability that the i^{th} individual is truly diseased is the predicted value of a positive test for that individual calculated from the covariates, regression coefficients, sensitivity, and specificity using Bayes's Theorem. For individuals designated as non-cases by the error prone outcome variable (W_i=0), the probability that the i^{th} individual is truly diseased is the predicted positive value of a negative test. Because the probabilities depended on regression parameters, they are recalculated after the logistic regression parameters are estimated, which leads to new probabilities and thus, new regression parameters. The processes of estimating the probabilities and the logistic regression parameters are repeated alternately until the parameter estimates converge.

This method can account for differential outcome misclassification (with respect to exposure or covariates) by assigning different values of sensitivity and specificity to individuals with different sets of values for exposure or covariates. This method can also incorporate internal validation data to estimate sensitivity and specificity.

1.4.2 Direct maximum likelihood

A second existing approach accounts for misclassification in logistic regression by directly specifying the likelihood function to include adjustments for sensitivity and specificity. This approach allows incorporation of assumed values of sensitivity and specificity, external validation data, or internal validation data. In the case of studies with

external or internal validation data, the sensitivity and specificity are estimated from the data based on specified covariates. It can also be extended to the case control setting when internal validation data are available (47).

The direct maximum likelihood approach for a main study with a validation subgroup specifies the likelihood for the logistic regression model relating exposure to outcome as the product of the likelihood for the main study and the likelihood for the validation subgroup. In both likelihood terms, sensitivity and specificity are based on associations between observed outcome, gold standard outcome, and exposure defined using a logistic model

$$\eta_d = \text{logit}[\Pr(W = 1|D = d, X = x)] = \theta_0 + \theta_1 d + \theta_2 X + \boldsymbol{\theta_3 Z}, \text{for } d = 0,1.$$

Sensitivity and specificity are calculated as

$$\text{SE}_i = \Pr(W = 1|D = 1, X = x_i, \boldsymbol{Z} = \boldsymbol{z_i}) = \frac{\exp(\eta_{i1})}{1 + \exp(\eta_{i1})},$$

and

$$\text{SP}_i = \Pr(W = 0|D = 0, X = x_i, \boldsymbol{Z} = \boldsymbol{z_i}) = \frac{\exp(\eta_{i0})}{1 + \exp(\eta_{i0})}.$$

The likelihood for the main study is then modified by the estimated values of sensitivity and specificity for each observation and multiplied by the likelihood for the validation study. Lyles (47) presents approaches to account for outcome misclassification in situations with internal validation data, external validation data, and assumed values of sensitivity and specificity.

The direct maximum likelihood approach applied to account for outcome misclassification in logistic or log binomial regression varies according to the link function. In the logistic model,

$$\Pr(D = 1|X = x, \boldsymbol{Z} = \boldsymbol{z})] = \frac{\exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z})}{1 + \exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z})} ,$$

And in the log binomial model,

$$\Pr(D = 1|X = x, \boldsymbol{Z} = \boldsymbol{z})] = \exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z}) .$$

1.5 Summary

Outcome misclassification is a neglected problem in epidemiologic research despite common use of study endpoints subject to measurement error. Several methods exist to account for misclassification of binary outcomes, but these methods are not straightforward to program using standard statistical software or to extend to the time-to-event setting.

Figures

CAUSE OF DEATH (See instructions and examples)

32. **PART I.** Enter the chain of events—diseases, injuries, or complications—that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

Approximate interval: Onset to death

IMMEDIATE CAUSE (Final disease or condition resulting in death) → a. Rupture of myocardium — Minutes

Due to (or as a consequence of):

Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**

b. Acute myocardial infarction — 6 days

Due to (or as a consequence of):

c. Coronary artery thrombosis — 5 years

Due to (or as a consequence of):

d. Atherosclerotic coronary artery disease — 7 years

PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I.

Diabetes, Chronic obstructive pulmonary disease, smoking

33. WAS AN AUTOPSY PERFORMED? ■ Yes ☐ No

34. WERE AUTOPSY FINDINGS AVAILABLE TO COMPLETE THE CAUSE OF DEATH? ■ Yes ☐ No

35. DID TOBACCO USE CONTRIBUTE TO DEATH?
■ Yes ☐ Probably
☐ No ☐ Unknown

36. IF FEMALE:
■ Not pregnant within past year
☐ Pregnant at time of death
☐ Not pregnant, but pregnant within 42 days of death
☐ Not pregnant, but pregnant 43 days to 1 year before death
☐ Unknown if pregnant within the past year

37. MANNER OF DEATH
■ Natural ☐ Homicide
☐ Accident ☐ Pending Investigation
☐ Suicide ☐ Could not be determined

Figure 1.1. Example of a completed cause of death section on a US death certificate

2 Specific Aims

Misclassification of outcome variables is a threat to the accuracy of epidemiologic studies. Outcome misclassification occurs when investigators observe an error-corrupted version of the event of interest instead of the true outcome status of study participants. Both binary event indicators and continuous outcomes are subject to misclassification, but standard approaches to analyzing cohort data typically assume such biases are absent.

Outcome misclassification can occur in many settings, but this work focuses on two specific types of error in outcome measurements: misclassification due to incorrect information reported by study participants and misattribution of cause of death on death certificates leading to outcome misclassification in mortality studies. This work develops methods to account for outcome misclassification in diverse situations both with and without validation data and in both binary regression models and time-to-event analyses.

Specifically, I aim to account for outcome misclassification in binary regression models in situations with internal validation data using multiple imputation. I hypothesize that accounting for outcome misclassification in logistic and binomial regression using multiple imputation will produce estimates of the odds ratio and risk ratio that are not biased by outcome misclassification. Applying this approach to data on self-reported ocular herpes recurrence from the Herpetic Eye Disease study will produce estimates of the effect

of acyclovir on herpes recurrence that are not biased by incorrect outcome information supplied by participants.

I also aim to account for outcome misclassification in the time-to-event setting with no validation data using maximum likelihood and Bayesian methods. I hypothesize that accounting for outcome misclassification using a modified likelihood function in Poisson models will provide estimates of the rate ratio not biased by outcome misclassification when the values of sensitivity and specificity indicated by the investigator are correct. Specifying prior distributions for sensitivity and specificity will capture the uncertainty in sensitivity and specificity. Applying this approach to data from the South Carolina textile workers' cohort will produce estimates of the effect of asbestos exposure on lung cancer death that are not biased by misattribution of cause of death on death certificates.

3 Methods

This thesis focuses on three methods to account for outcome misclassification in a range of epidemiologic settings: multiple imputation, maximum likelihood, and Bayesian analysis. The principle features of the methods are described below, and specific applications of the methods are detailed in chapter 4 (multiple imputation) and chapter 5 (maximum likelihood and Bayesian analysis).

3.1 Multiple imputation to account for outcome misclassification

This work begins by describing the use of multiple imputation to address outcome misclassification in studies with internal validation data. Multiple imputation is a standard technique for handling missing data (50,51). We use multiple imputation to account for outcome misclassification by viewing outcome misclassification as a missing data problem.

Briefly, analyses using multiple imputation to impute a single missing variable (X) do so by examining the relationships between the value of X and other covariates $\boldsymbol{Z}$ for observations in which X is not missing. These relationships are then used to impute the value of X for the observations in which X is missing. This process is repeated K times, and point estimates are averaged over the resulting K cohorts. The multiple imputation process produces estimates that are consistent and asymptotically normal (and asymptotically

efficient as $K \rightarrow \infty$) if *X* is missing at random within strata of $\boldsymbol{Z}$. Missing at random implies that missingness of *X* may depend on $\boldsymbol{Z}$ but not on *X* itself (after controlling for $\boldsymbol{Z}$) (52).

In a study with an internal validation subgroup, the possibly misclassified outcome (*W*) is available for all participants, but the gold-standard outcome (*D*) is available only for participants in the validation subgroup. If participants are selected into the validation subgroup randomly within strata of exposure (*X*) and covariates ($\boldsymbol{Z}$), information on the gold-standard outcome can be said to be missing at random in the full cohort. The missing at random assumption allows us to exploit the relationships between *D, W, X*, and $\boldsymbol{Z}$ among participants in the validation subgroup to impute values for *D* for all other participants.

To account for outcome misclassification using multiple imputation, we use the logistic method for monotone missing data (53). As a first step, the gold-standard outcome *D* is regressed on the possibly misclassified outcome *W*, the exposure *X*, and other relevant covariates *Z* in the validation subgroup using the logistic regression model shown in Equation 3.1.

$$\mathrm{P}(D = 1 \mid W, X, \boldsymbol{Z}) = \frac{\exp(\alpha_0 + \alpha_1 W + \alpha_2 X + \alpha_3 WX + \boldsymbol{\alpha_4 Z})}{1 + \exp(\alpha_0 + \alpha_1 W + \alpha_2 X + \alpha_3 WX + \boldsymbol{\alpha_4 Z})} \quad 3.1$$

Regression parameters are assumed to follow a multivariate Gaussian distribution with mean vector $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\boldsymbol{\alpha}}_{\boldsymbol{4}})$ and covariance matrix $(\widehat{\boldsymbol{\Sigma}}_{wxz})$ estimated from the logistic regression model above. Regression parameters are drawn for each of *K* imputations from the posterior predictive distribution of parameters. Drawing regression coefficients for each imputation allows uncertainty about the relationship between *W, X,* and *D* to propagate through the analysis (50).

A new variable, D'_k, is created to represent the imputed outcome, where k indexes the number of imputations. By definition, $D'_k = D$ for participants in the validation subgroup. For participants not in the validation subgroup, D'_k is imputed based on regression coefficients drawn for each imputation. For each imputation, D'_k is assigned by a random draw from a Bernoulli distribution with probability p_k, where

3.2

$$p_k = \frac{\exp\left(\hat{\alpha}_0^k + \hat{\alpha}_1^k W + \hat{\alpha}_2^k X + \hat{\alpha}_3^k WX + \hat{\boldsymbol{\alpha}}_{\boldsymbol{4}}^{\boldsymbol{k}} \boldsymbol{Z}\right)}{1 + \exp\left(\hat{\alpha}_0^k + \hat{\alpha}_1^k W + \hat{\alpha}_2^k X + \hat{\alpha}_3^k WX + \hat{\boldsymbol{\alpha}}_{\boldsymbol{4}}^{\boldsymbol{k}} \boldsymbol{Z}\right)}.$$

At this point in the analysis, K datasets exist with imputed outcomes D'_k for all participants. The relationship between exposure and outcome in these datasets can be analyzed with any type of analysis model desired. For example, to estimate the risk ratio for the effect of exposure X on true outcome D,

$$\frac{P(D = 1|X = 1, \boldsymbol{Z} = \boldsymbol{z})}{P(D = 1|X = 0, \boldsymbol{Z} = \boldsymbol{z})},$$

we can use binomial regression to estimate the effect of exposure X on the imputed outcome in each of the K datasets. The binomial model for the imputed outcome given the exposure and relevant covariates for $k = 1, 2, \ldots, K$ is

3.3

$$\mathrm{P}(D'_k = 1|\, X, \boldsymbol{Z}) = \exp\left(\beta_0^k + \beta_1^k X + \boldsymbol{\beta}_{\boldsymbol{2}}^{\boldsymbol{k}} \boldsymbol{Z}\right).$$

The estimated risk ratio from this model is $\exp\left(\bar{\beta}_1\right) = \exp\left(K^{-1} \sum_{k=1}^{K} \hat{\beta}_1^k\right)$, where $\hat{\beta}_1^k$ is the natural log of the estimated risk ratio from the k^{th} imputed dataset. The variance for $\bar{\beta}_1$ is given by

3.4

$$V(\bar{\beta}_1) = \frac{1}{K}\sum_{k=1}^{K}\hat{V}(\hat{\beta}_1^k) + \left(1+\frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}(\hat{\beta}_1^k - \bar{\beta}_1)^2.$$

The choice of effect measure is flexible. For example, if one wished to estimate the odds ratio in place of the risk ratio

$$\frac{P(D=1|X=1,\boldsymbol{Z}=\boldsymbol{z})/P(D=0|X=1,\boldsymbol{Z}=\boldsymbol{z})}{P(D=1|X=0,\boldsymbol{Z}=\boldsymbol{z})/P(D=0|X=1,\boldsymbol{Z}=\boldsymbol{z})},$$

logistic regression could be used as the analysis model in place of log binomial regression. The logistic models for the imputed outcome given treatment group and relevant covariates for k=1 to K are

3.5

$$\mathrm{P}(D'_k = 1|\, X, \boldsymbol{Z}) = \frac{\exp(\beta_0^k + \beta_1^k X + \boldsymbol{\beta_2^k Z})}{1+\exp(\beta_0^k + \beta_1^k X + \boldsymbol{\beta_2^k Z})}$$

and the odds ratio is given by is $\exp(\bar{\beta}_1) = \exp(K^{-1}\sum_{k=1}^{K}\hat{\beta}_1^k)$, where $\hat{\beta}_1^k$ is the natural log of the estimated odds ratio from the k^{th} imputed dataset. Variance is again computed using equation 3.4.

Regardless of the choice of effect estimate, the analysis model need not match the imputation model. Covariates likely to influence the outcome misclassification process may be used in the imputation model (**3.1**) but excluded from the analysis model if they do not meet the criteria for covariate inclusion in the analysis model. Addition of covariates to the imputation model rarely reduces the precision of the final estimate, and any decline in precision is generally offset with a reduction in bias (51).

The imputation model shown in Equation **3.1** can be used to account for outcome misclassification that is differential or nondifferential with respect to exposure. Investigators often assume outcome misclassification is nondifferential if the person who

assesses the outcome does not have knowledge of the participants' exposure status. The assumption of nondifferential misclassification implies that $\alpha_3 = 0$ in the imputation model shown in Equation **3.1**. In models where α_3 is allowed to be different from 0, separation of data points can occur if the size of the validation subgroup is small. Firth's correction (54) can be applied in these settings to prevent separation of data points (55). Firth's correction uses a modified score function to obtain maximum likelihood estimates when response variables can be perfectly predicted by a linear combination of risk factors (55), a situation known as separation (56) or monotone likelihood (57). Firth's correction may be viewed as a multivariable extension of a continuity correction.

3.2 Maximum likelihood to account for outcome misclassification

The following section discusses the use of modified maximum likelihood to account for outcome misclassification. In this section, we assume no validation data are available and perform sensitivity analyses by setting values of sensitivity and specificity.

The maximum likelihood approach to account for outcome misclassification in logistic regression was outlined by Lyles (47). Briefly, for a logistic regression model comparing the odds of outcome Y between exposure groups X controlling for covariates $\boldsymbol{Z}$,

$$logit[P(Y = 1|X, \boldsymbol{Z})] = \beta_0 + \beta_1 X + \sum_{m=2}^{L+1} \beta_m Z_{j(m-1)},$$

each independent record i contributes the following likelihood term

$$L = P(Y = 1|X = x, \boldsymbol{Z} = \boldsymbol{z})^{y_i} P(Y = 0|X = x, \boldsymbol{Z} = \boldsymbol{z})^{(1-y_i)}. \quad (3.6)$$

In this model, the odds ratio of interest is given by $\exp(\beta_1)$.

If a misclassified version of the outcome variable, *W,* is observed in place of the gold standard outcome *Y,* the estimated odds ratio is subject to bias. To account for outcome misclassification in the logistic model, the likelihood is rewritten in terms of *W,* sensitivity, and specificity.

3.7

$$L = \{[(1 - sp) \times P(Y = 0|X = x, \boldsymbol{Z} = \boldsymbol{z}) + se \times P(Y = 1|X = x, \boldsymbol{Z} - \boldsymbol{z})]^{w_i} \times [sp \times P(Y = 0| = x, \boldsymbol{Z} - \boldsymbol{z}) + (1 - se) \times \{(Y = 1| = x, \boldsymbol{Z} - \boldsymbol{z})]^{(1-w_i)}$$

Here, we extend this approach to account for outcome misclassification due to misattribution of cause of death. When outcome misclassification is thought to be attributable only to misspecification of the cause of death, several issues emerge. First, the dates of death are assumed to be correct. Second, participants alive at the end of the study are not subject to outcome misclassification. Similarly, the misclassification probabilities apply only to deaths observed to occur during the study.

We choose to study the relationship between cause-specific mortality and an exposure of interest using Poisson regression. When estimating the rate ratio of death due to cause *A* per unit increase in exposure *X* using Poisson regression, the parameter estimating the desired rate ratio is $\exp(\beta_1)$ in the Poisson model below,

3.8

$$\lambda_j = \exp\left(\beta_0 + \beta_1 X_j + \sum_{m=2}^{L+1} \beta_m Z_{j(m-1)}\right),$$

where λ_j represents the rate of death due to cause *A* in strata *j*, X_j is the exposure, and $\boldsymbol{Z}$ is a $J \times L$ matrix with columns for each of L covariates in the model.

To fit this model, person-time contributed by study participants is grouped into strata of distinct covariate patterns. Strata are indexed by $j = 1, 2, \ldots, J$. In studies of

continuous exposures or covariates, these variables must be categorized. Each strata contains a count of the number of person years contributed to that strata (n_j) and the number of deaths (d_j). In each strata, w_j deaths are attributed to the cause of interest (cause A), though the true number of deaths due to cause A (y_j) is unobserved. The true number of person-years and deaths remains n_j and d_j, respectively, under the assumption that the dates of death are correct.

Under the ideal model specified above, the likelihood expression would be

$$L = \prod_{j=1}^{J} \lambda_j^{y_j} \exp(-\lambda_j n_j). \tag{3.9}$$

The first term $\lambda_j^{y_j}$, captures the number of events occurring in strata j, and the second term, $\exp(-\lambda_j n_j)$, takes the number of person-years contributed in strata j into account.

However, because we observe w_j possibly misclassified deaths due to cause A in place of y_j true deaths due to cause A, the model above cannot be fit directly. Instead, standard analyses typically fit the model

$$\lambda_j' = \exp\left(\gamma_0 + \gamma_1 X_j + \sum_{m=2}^{L+1} \gamma_m Z_{j(m-1)}\right),$$

where λ_j' represents the rate of a possibly misclassified version of the outcome variable, w_j and $\exp(\gamma_1)$ represents an estimate of the rate ratio possibly biased by outcome misclassification.

To account for misclassification using a modified likelihood function, I begin by specifying the Poisson likelihood for a situation with two causes of death and no outcome misclassification.

3.10

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{J} \lambda_j^{y_j} \, \mu_j^{(d_j - y_j)} \exp\{-(\lambda_j + \mu_j) n_j\}$$

where λ_j is described above, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_4)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ \ldots, \beta_4)$, and μ_j is the estimated rate of deaths due to causes other than cause *A* for strata *j*,

$$\mu_j = \exp\left(\alpha_0 + \alpha_1 X_j + \sum_{m=2}^{L+1} \alpha_m \, Z_{j(m-1)}\right),$$

where $\boldsymbol{Z}$ is a $J \times L$ matrix with columns for the L covariates included in the analysis. As in the likelihood function for one cause of death, deaths due to cause *A* contributed to the first term, $\lambda_j^{y_j}$, deaths due to other causes contributed to the second term, $\mu_j^{(d_j - y_j)}$, and person-time is taken into account in the third term, $\exp\{-(\lambda_j + \mu_j) n_j\}$.

Because the true number of deaths due to cause *A* is unavailable, the likelihood is modified to use the count of potentially misclassified deaths due to cause *A* for each stratum, w_j , and the misclassification probabilities (i.e., sensitivity and specificity) to restructure the likelihood as:

3.11

$$L_{\text{modified}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{J} \left\{\lambda_j(se) + \mu_j(1 - sp)\right\}^{w_j} \left\{\lambda_j(1 - se) + \mu_j(sp)\right\}^{(d_j - w_j)}$$

$$\times \exp\{-\left[\lambda_j(se) + \mu_j(1 - sp) + \lambda_j(1 - se) + \mu_j(sp)\right] n_j\}.$$

Under the assumption that sensitivity and specificity are correct, the modified likelihood function will be maximized at the same estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as the likelihood function 3.10, though the modified likelihood function will be more diffuse, resulting in wider intervals. Using the modified likelihood function above, a sensitivity analysis can be

performed by setting sensitivity and specificity to plausible values identified using the existing literature or through expert opinion.

3.3 Bayesian analysis to account for outcome misclassification

The sensitivity analysis approach outlined above requires the investigator to set the value of sensitivity and specificity. When an investigator suspects outcome misclassification is present based on external validation studies or expert opinion, Bayesian methods offer an appealing alternative to sensitivity analysis.

Uncertainty about the amount of misclassification in the data can be acknowledged explicitly by placing informative prior distributions on sensitivity and specificity that reflect beliefs about the amount of misclassification in the data and certainty about those beliefs. Bayes's theorem offers a method to combine these prior probability distributions for sensitivity and specificity with the observed data characterized by likelihood function 3.6 to obtain a posterior distribution of the parameter(s) of interest. In this case, the parameter of interest is β_1 from model 3.9. Non-informative, null-centered priors are placed on the regression parameters.

The posterior distribution can rarely be expressed in closed form. However, a large number of samples can be drawn from the posterior distribution using Markov chain Monte Carlo. We more closely approximate the posterior distribution as more samples are drawn, allowing calculation of statistics of interest, such as the mean, median, and 95% highest posterior density intervals. If posterior draws have a normal distribution, 95% posterior intervals will approximate the 95% highest posterior density intervals.

3.4 Application to the Herpetic Eye Disease Study

We implement the multiple imputation approach to account for outcome misclassification due to error in participant reported outcomes in the Herpetic Eye Disease Study. The Herpetic Eye Disease Study is a randomized trial of acyclovir for preventing ocular herpes simplex virus (HSV) recurrence at 58 university and community-based sites in the United States (22). Participants were 12 years of age and older and had an episode of ocular HSV in the 12 months before the study, but their disease had been inactive during the 30 days preceding the study. During the study, the 703 participants were randomized to receive either oral acyclovir or placebo. The acyclovir group received 400 mg of acyclovir twice daily for 12 months, and the placebo group received oral placebos with the same frequency. The goal of the study was to compare the 12-month incidence of ocular HSV recurrence between the group randomized to receive acyclovir and the group randomized to receive placebo. Information was collected on age, race, gender, and number of ocular recurrences prior to randomization. Participants returned for five follow-up visits during the one-year treatment period and an additional three follow-up visits during the six months immediately following the treatment period. HSV recurrences that were diagnosed during the trial were treated with topical corticosteroids and antivirals, though oral acyclovir or placebo was continued for the duration of the one-year treatment period.

Nested within the Herpetic Eye Disease Study, the Ocular HSV Recurrence Factor Study was designed to evaluate the psychological, environmental, and biological triggers of ocular HSV recurrence between 1992 and 1998. Patients in the Herpetic Eye Disease study were eligible to participate in the recurrence factors study if they were at least 18 years of age. Participants in the recurrence factor study completed weekly diaries to track acute and

chronic stressors, including illnesses, injuries, menstrual periods, sun exposure, and emotional and financial stresses. This analysis was limited to the 308 Herpetic Eye disease study patients who also enrolled in the recurrence factors study.

The outcome of interest was a binary indicator of HSV recurrence over the 12-month study period. Among participants in the recurrence factors study, HSV recurrence was assessed in two ways. First, participants recorded symptoms of HSV recurrence in their weekly diaries. In addition, study-certified ophthalmologists examined participants using microscopy when symptoms were apparent or at planned study visits in months 1, 3, 6, 9, and 12. This analysis will consider physician-diagnosed ocular recurrence to be the gold-standard outcome measure, represented by D, and participant-reported symptoms to be the error-prone outcome measure, represented by W.

The effect of interest was the odds ratio comparing the incidence of HSV recurrence among participants randomized to receive acyclovir to incidence of HSV recurrence among participants randomized to placebo. The first analysis was conducted using physician-diagnosed HSV recurrence as the outcome measure on all participants. The odds ratio was estimated as $\exp(\beta_1)$ in the logistic regression model

$$\mathrm{P}(D = 1 | X, \boldsymbol{Z}) = \frac{\exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z})}{1 + \exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z})},$$

where D represents physician-diagnosed recurrence, X is treatment assignment to acyclovir ($X = 1$) or placebo ($X = 0$), and $\boldsymbol{Z}$ is a vector of covariates including age, sex, and number of previous recurrences.

To assess the performance of multiple imputation to account for outcome misclassification, a hypothetical validation subgroup of 91 participants was selected from

the 308 participants in the recurrence factors study. Participant-reported symptoms of HSV recurrence (*W*) was assumed to be available for all participants, but physician-diagnosed HSV recurrence was assumed to be available only for participants randomly selected to be in the validation subgroup. Thus, the hypothetical validation subgroup mimicked an internal validation subgroup randomly sampled from the main study.

We compared results of an ideal analysis on the full cohort of 308 participants using the physician diagnosis as the outcome variable with results from four methods for handling outcome misclassification: 1) the naïve analysis, in which participant-reported outcome (*W*) represented the outcome status for all 308 participants; 2) the validation subgroup, in which the physician-diagnosed outcomes (*D*) were compared between those receiving acyclovir and those receiving placebo in the validation subgroup of 91 participants; 3) a direct maximum likelihood approach (47) to account for outcome misclassification and; 4) multiple imputation to account for outcome misclassification. Direct maximum likelihood and multiple imputation approaches were evaluated under the assumptions of both differential and nondifferential misclassification of the outcome with respect to treatment group.

3.5 Application to study of South Carolina textile workers study

Both maximum likelihood and Bayesian approaches were applied to account for outcome misclassification due to misattribution of cause of death in the South Carolina textile workers' study. The study population comprised workers at a textile production plant in South Carolina that produced asbestos beginning in 1896 and asbestos textile products beginning in 1909 (58). The study enrolled 3072 men and women who were

employed at the plant for at least one month between 1 January 1940 and 31 December 1965. Employment records were used to obtain information on date of birth, year of study entry, race (Caucasian or non-Caucasian), sex, and employment status in each year.

Detailed work histories, including plant department, job held by the participant, and start and end dates, were available for each participant in the cohort. Cumulative exposure to the chrysotile form of asbestos was estimated using a job exposure matrix to link work history to industrial hygiene sampling measurements taken between 1930 and 1975, as previously described (59). Industrial hygiene data were collected from the company insurance carrier, the State Board of Health, the US Public Health Service, and the company sampling program (58). Chrysotile exposure concentrations, expressed as fibers longer than 5 micrometers per milliliter of air (fibers/mL), were estimated for each day of each participant's work history. Yearly exposure values were calculated as the product of the proportion of the year worked and the average daily exposure concentration and reported as fiber-years per milliliter (fiber-y/mL). To capture the appropriate exposure window for the effect of asbestos on lung cancer mortality, exposure values were lagged 10 years, meaning that the risk of lung cancer in each year was not affected by asbestos exposure in the prior 10 years. Chrysotile was the only type of asbestos ever processed at the plant as a raw fiber, indicating that confounding exposure by other types of asbestos figures, such as crocidolite, is unlikely.

Participants were followed for cause of death through January 1, 2001. Between 1940 and 1978, vital status was determined by the US Social Security Administration, Internal Revenue Service, Veterans Affairs, state drivers' license files, vital statistics offices, and postal mail correction services. Participants not found in one of these sources were

traced using local telephone listings, property records, voter records, records of funeral homes, and other local sources (42).

Between 1979 and 2001, vital status was determined using the National Death Index. Those confirmed alive in 1979 and not found in the National Death Index were assumed to be alive at the end of the study. For those who died, cause of death was determined through examination of death certificates and coded by a qualified nosologist into the revision of the International Classification of Diseases (ICD) in effect at the date of each death.

In this cohort, we estimated the effects of asbestos exposure on the rate of death due to lung cancer. A death was classified as a death due to lung cancer using ICD-7, ICD-8, and ICD-9 code 162. Cause of death reported on the death certificate was the error-prone version of the outcome variable, but, unlike the earlier example from the Herpetic Eye Disease Study, no internal validation subgroup was available to assess the possible outcome misclassification. To account for outcome misclassification, we performed sensitivity analysis using modified maximum likelihood and the Bayesian analysis placing informative prior distributions on sensitivity and specificity as described above. The effect of interest was the rate ratio of lung cancer death per 100 f-y/mL of asbestos exposure.

3.6 Simulations

We used simulations to explore the finite sample properties of the methods to account for outcome misclassification proposed in this section. Each simulation experiment was repeated for scenarios involving varying parameters of interest, including total sample

size, effect size, misclassification parameters, and size of the validation subgroup, if applicable.

In each simulation experiment, the distribution of exposure was designed to mimic a real-data example. For example, in the simulations to evaluate multiple imputation to account for outcome misclassification, the exposure distribution reflected the distribution of acyclovir exposure in the Herpetic Eye Disease Study data. In the simulation experiments to assess the performance of modified maximum likelihood to account for outcome misclassification, the exposure distribution was designed to mimic the distribution of asbestos exposures among participants in the South Carolina textile workers study.

True outcome variables were generated based on the exposure and effect size of interest. The effect size was varied across scenarios. In the simulations designed to assess the performance of multiple imputation, the true binary outcome indicator was generated directly from the exposure and effect size. In the simulations for maximum likelihood, the true time to death due to cause *A* was generated along with a true time to death due to all other causes. If the time to death due cause *A* was less than the time to death due to other causes, the death was a death due to cause *A.* Otherwise, the simulated participant was said to have died from other causes.

In each simulation, a misclassified version of the outcome variable was generated based on the true outcome variable and the values of the misclassification parameters. Values of sensitivity and specificity were altered for each simulated scenario.

For each scenario, 10,000 cohorts were simulated based on the characteristics and parameters described above. The analysis of interest was conducted in each simulated cohort, and outcome misclassification was address by either multiple imputation (for the

cohorts mimicking the Herpetic Eye Disease study data) or modified maximum likelihood (for the cohorts mimicking the South Carolina textile workers study).

We assessed the performance of each method to account for outcome misclassification by comparing bias, 95% confidence interval coverage, statistical power, and mean-squared error between the standard analysis with the analysis using multiple imputation or modified maximum likelihood to account for outcome misclassification. Bias was defined as 100 times the difference between the average estimated effect size and true effect size, and confidence interval coverage was calculated as the proportion of simulations in which the estimated Wald-type confidence limits included the true value. Statistical power was calculated as the percentage of simulations in which the Wald-type confidence interval excluded the null value. The bias-precision tradeoff was considered through examination of the mean-squared error, which was the sum of the square of the bias and the square of the standard deviation of the bias.

4 Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data

Outcome misclassification is widespread in epidemiology, but methods to account for it are rarely used. We describe the use of multiple imputation to reduce bias when validation data are available for a subgroup of study participants. This approach is illustrated using data from 308 participants in the multicenter Herpetic Eye Disease Study between 1992 and 1998 (48% female, 85% Caucasian, median age 49 years). The odds ratio (OR) comparing acyclovir and placebo groups on the gold-standard outcome (physician-diagnosed herpes simplex virus recurrence) was 0.62 (95% confidence interval (CI): 0.35, 1.09). We discarded physician diagnosis except a 30% validation subgroup to compare methods. Multiple imputation (OR=0.60; 95% CI: 0.24, 1.51) was compared to naïve analysis using self-reported outcomes (OR=0.90; 95% CI: 0.47, 1.73), analysis restricted to the validation subgroup (OR=0.57; 95% CI: 0.20, 1.59), and direct maximum likelihood (OR=0.62; 95% CI: 0.26, 1.53). In simulations, multiple imputation and direct maximum likelihood had greater statistical power than analysis restricted to the validation subgroup, yet all three provided unbiased estimates of the OR. The multiple imputation approach was extended to estimate risk ratios using log-binomial regression. Multiple imputation has advantages regarding flexibility and ease of implementation for epidemiologists familiar with missing data methods.

4.1 Introduction

Misclassification of outcome variables is common in epidemiology and threatens the validity of inferences from epidemiologic studies (1,2). However, standard approaches to epidemiologic data analysis typically assume outcome misclassification is absent. Although approaches to account for bias in crude effect estimates due to use of a misclassified binary outcome have existed for more than half a century (46), these methods are rarely used because epidemiologists commonly wish to present results adjusted for several confounding variables. More recently, investigators have developed maximum likelihood approaches (2,3) to produce odds ratio estimates that account for outcome misclassification while adjusting for relevant confounders using logistic regression, but these methods have not been widely applied in the epidemiologic literature. Here we describe an alternative approach to account for outcome misclassification using missing data methods that are familiar to epidemiologists.

Methods to account for misclassification rely on information relating the observed outcome to the gold standard outcome measure. This relationship can be estimated by comparing the observed outcome to the gold standard outcome in a validation subgroup that is a random subset of the main study or in external data. In this paper, we focus on the former case where internal validation data are available for a subgroup of the population under study. We treat outcome misclassification as a missing data problem where the true outcome status is known only for participants in the validation subgroup and missing for all other participants (4). This perspective allows misclassification bias to be addressed by applying well-established methods for handling missing data (16,50,53). In the sections that follow, we describe an approach to account for outcome misclassification using

multiple imputation to estimate odds ratios and risk ratios, provide examples using cohort data (25), and explore some finite sample properties of the proposed method by Monte Carlo simulation.

4.2 Methods

Study population

We illustrate the use of multiple imputation to account for outcome misclassification using data from the Herpetic Eye Disease Study, a randomized trial of acyclovir for preventing ocular herpes simplex virus (HSV) recurrence at 58 university and community-based sites in the United States (22). Participants were 12 years of age and older and had an episode of ocular HSV in the 12 months before the study, but their disease had been inactive during the 30 days preceding the study. During the study, the 703 participants received either oral acyclovir or placebo for 12 months. The goal of the study was to compare the 12-month incidence of ocular HSV recurrence between the group randomized to receive acyclovir and the group randomized to receive placebo. Information was also collected on age, race, gender, and number of ocular recurrences prior to randomization. Here, we restrict analyses to the 308 of 703 participants who co-enrolled in a study that collected weekly diaries about ocular HSV symptoms and possible triggers between 1992 and 1998(60).

Outcome Ascertainment and Validation

The outcome of interest was a binary indicator of HSV recurrence over the 12-month study period (any recurrence versus none) assessed in two ways. Study-certified

ophthalmologists examined participants using microscopy when symptoms were apparent or at planned study visits in months 1, 3, 6, 9, and 12. In addition, participant-reported HSV recurrence was obtained from a weekly diary. We consider participant-reported HSV recurrence to be the observed, and possibly mismeasured, version of the outcome variable ($W = 1$ if the participant reported any recurrence, $W = 0$ otherwise), and physician-diagnosed HSV recurrence to be the gold standard (D = 1 if the ophthalmologist diagnosed a recurrence, D = 0 otherwise). We randomly sampled 30% (n = 91) of the 308 participants to treat as a validation subgroup. In this analysis, we assume that W was available for all participants and D was observed only for those selected to be in this hypothetical validation subgroup.

Statistical methods

We used logistic regression to estimate the odds ratio comparing ocular HSV recurrence between participants randomly assigned to acyclovir and those assigned to placebo. We compared results of an ideal analysis on the full cohort of 308 participants using the physician diagnosis as the outcome variable with results from four methods for handling outcome misclassification: 1) the naïve analysis, in which W represented the outcome status for all 308 participants; 2) the validation subgroup, in which the physician-diagnosed outcomes (D) were compared between those receiving acyclovir and those receiving placebo in the validation subgroup of 91 participants; 3) a direct maximum likelihood approach (47) to account for outcome misclassification and; 4) multiple imputation to account for outcome misclassification. Direct maximum likelihood and multiple imputation approaches were evaluated under the assumptions of both differential

and nondifferential misclassification of the outcome with respect to treatment group. We further extended the direct maximum likelihood and multiple imputation approaches to estimate risk ratios using log-binomial regression.

The direct maximum likelihood approach accounted for outcome misclassification using the method described by Lyles et al (47). This approach included data from all participants, with those in the validation subgroup providing data on the correctly classified outcome and those not in the validation subgroup providing data on the misclassified outcome. In contrast, the naïve analysis included data from all participants, but used only the misclassified outcome, and the validation analysis included data from participants in the validation subgroup only, but used the correctly classified outcome. To account for nondifferential misclassification in the direct maximum likelihood approach, we estimated the sensitivity and specificity from the records in the validation subgroup. These values were used to compute the likelihood to be maximized, which was a product of the main study likelihood and the validation sample likelihood, as detailed in Appendix 1. To relax the assumption of nondifferential misclassification, we added treatment group to the model for sensitivity and specificity.

Multiple imputation is a standard technique for handling missing data (50,51). We use multiple imputation to account for outcome misclassification by exploiting the relationships between *D, W,* treatment group (X), and other covariates ($\boldsymbol{Z}$) among participants in the validation subgroup to impute values for *D* for all other participants.

The first step is to model the relationship between physician-diagnosed HSV recurrence and participant-reported HSV recurrence in the validation subgroup. In this example, we use the logistic regression method for monotone missing data (50). To do this,

we regress physician-diagnosed HSV recurrence (*D)* on participant-reported HSV recurrence (*W*), treatment group (*X*), and other covariates ($\boldsymbol{Z}$) using a logistic regression model

$$\mathrm{P}(D = 1|\, W, X, \boldsymbol{Z}) = \frac{\exp(\alpha_0 + \alpha_1 W + \alpha_2 X + \alpha_3 WX + \ \boldsymbol{\alpha_4 Z})}{1 + \exp(\alpha_0 + \alpha_1 W + \alpha_2 X + \alpha_3 WX + \ \boldsymbol{\alpha_4 Z})}. \quad 4.1$$

We then draw a set of regression coefficients for each of *K* imputations from the posterior predictive distribution of the parameters. We set *K* = 40 in this analysis. Assume parameters follow a multivariate Gaussian distribution with mean vector $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \widehat{\boldsymbol{\alpha}_4})$ and covariance matrix ($\widehat{\boldsymbol{\Sigma}}_{wxz}$) estimated from the logistic regression model above. Drawing regression coefficients for each imputation allows uncertainty about the relationship between *W, X,* and *D* to propagate through the analysis (50).

A new variable, D'_k, is created to represent the imputed outcome. For participants in the validation subgroup, $D'_k = D$, where *k* indexes the number of imputations. For participants not in the validation study, values for D'_k are imputed based on the regression coefficients drawn for that imputation. For each imputation, D'_k is assigned by a random draw from a Bernoulli distribution with probability p_k, where

$$p_k = \frac{\exp\left(\hat{\alpha}_0^k + \hat{\alpha}_1^k W + \hat{\alpha}_2^k X + \ \hat{\alpha}_3^k WX + \widehat{\boldsymbol{\alpha}_4^k}\, \boldsymbol{Z}\right)}{1 + \exp\left(\hat{\alpha}_0^k + \hat{\alpha}_1^k W + \hat{\alpha}_2^k X + \ \hat{\alpha}_3^k WX + \widehat{\boldsymbol{\alpha}_4^k}\, \boldsymbol{Z}\right)}\ . \quad 4.2$$

The analysis model is then used to compare imputed outcomes between treatment and placebo groups conditional on other covariates ($\boldsymbol{Z}$). In the example, we first use a logistic regression model to estimate the odds ratio comparing imputed HSV recurrence for participants assigned to acyclovir and those assigned to placebo in each imputation and

combine results using standard multiple imputation techniques (51). The logistic models for the imputed outcome given treatment group and relevant covariates for k=1 to 40 are

$$\mathrm{P}(D'_k = 1 | X, \boldsymbol{Z}) = \frac{\exp(\beta_0^k + \beta_1^k X + \boldsymbol{\beta_2^k Z})}{1 + \exp(\beta_0^k + \beta_1^k X + \boldsymbol{\beta_2^k Z})}. \quad (4.3)$$

The estimated odds ratio is

$$\exp(\bar{\beta}_1) = \exp\left(K^{-1} \sum_{k=1}^{K} \hat{\beta}_1^k\right), \quad (4.4)$$

where $\hat{\beta}_1^k$ is the natural log of the estimated odds ratio from the k^{th} imputed dataset. The variance for $\bar{\beta}_1$ is given by

$$V(\bar{\beta}_1) = \frac{1}{K} \sum_{k=1}^{K} \hat{V}(\hat{\beta}_1^k) + \left(1 + \frac{1}{K}\right)\left(\frac{1}{K-1}\right) \sum_{k=1}^{K} (\hat{\beta}_1^k - \bar{\beta}_1)^2. \quad (4.5)$$

In a closed cohort, it may be preferable to estimate the risk ratio instead of the odds ratio (61–63). To illustrate the ability of the proposed multiple imputation approach to estimate different parameters of interest, we also use a log-binomial regression model to estimate the risk ratio comparing imputed HSV recurrence for participants assigned to acyclovir and those assigned to placebo in each imputation. To estimate a risk ratio, the binomial model for the imputed outcome given treatment group and relevant covariates for k = 1 to 40 is used in place of the logistic model shown in Equation 4.3.

$$\mathrm{P}(D'_k = 1 | X, \boldsymbol{Z}) = \exp(\beta_0^k + \beta_1^k X + \boldsymbol{\beta_2^k Z}). \quad (4.6)$$

Multiple imputation can be used to account for misclassification of the outcome that is differential or nondifferential with respect to treatment group. The assumption of nondifferential misclassification implies that $\alpha_3 = 0$ in the imputation model (Equation 4.2). In models where α_3 was allowed to be different from 0, because the validation subgroup was relatively small, we used Firth's correction (54) to prevent separation of data points (55). Firth's correction uses a modified score function to obtain maximum likelihood estimates when response variables can be perfectly predicted by a linear combination of risk factors (55), a situation known as separation (56) or monotone likelihood (57). Firth's correction may be viewed as a multivariable extension of a continuity correction. Appendix 2 provides the SAS code for multiple imputation to account for outcome misclassification. Alternatively, one could use standard programs for multiple imputation included in many statistical software packages, such as SAS's PROC MI (SAS Institute, Cary, NC) or IVEware (University of Michigan, Ann Arbor, MI).

Although the cohort originated as part of a randomized trial, selection into the cohort for analysis was dependent on the participant keeping a weekly diary, which could have been influenced by several covariates. To estimate measures of association that were not biased by this selection, we adjusted for age, sex, and number of previous HSV occurrences by including these covariates in the $\boldsymbol{Z}$ vector in all analyses.

Simulation study

The bias, 95% confidence interval coverage, mean squared error, and statistical power for each method were evaluated under 15 simulation scenarios (Appendix 3). Each scenario represented different values of key parameters: sensitivity, specificity, size of the

validation subgroup, and total sample size. One set of simulations was designed to mimic the example; that is, for each trial, 300 participants were generated with values for treatment group, true disease status, reported disease status, and whether or not that individual was in the validation subgroup.

Another set of scenarios used the same parameter values but simulated a study of 1000 participants. For each simulation, 10% or 30% of participants were randomly selected for the validation subgroup. In each scenario, the odds ratio for the effect of acyclovir on ocular HSV recurrence was estimated using each of the four methods described above and summarized over the 10,000 simulations.

4.2 Results

Study participants were 48% female, 85% white, and had a median age of 49 years. Table 1 presents the data on self-reported recurrence (W) and physician diagnosed recurrence (D) from the Herpetic Eye Disease Study. Of the 308 study participants with both outcome measures available, 91 were randomly selected for the hypothetical validation subgroup. Of the 14 participants in the validation subgroup who reported HSV recurrences, 8 were diagnosed with HSV recurrence by a physician; of the 77 participants who did not report HSV recurrence, 65 had no physician diagnosed recurrence. Specificity of self-reported HSV recurrence was 0.9 (95% CI: 0.8, 1.0) and did not differ by treatment group. Sensitivity appeared to be higher for participants assigned to acyclovir (sensitivity = 0.5; 95% CI: 0.3, 0.6) than for participants assigned to placebo (sensitivity = 0.3; 95% CI: 0.2, 0.5), though the difference was imprecise. The sensitivity and specificity of self-

reported HSV recurrence in the validation subgroup were similar to the sensitivity and specificity in the full cohort.

Table 2 presents estimates of the odds ratio from each method to account for outcome misclassification. In the complete data, the odds ratio comparing the gold standard outcome measure, physician-diagnosed HSV recurrence, between treatment groups was 0.62 (95% CI: 0.35, 1.09; standard error (SE) = 0.29). The odds ratio comparing self-reported HSV recurrence between participants assigned to acyclovir and those assigned to placebo was 0.90 (95% CI: 0.47, 1.73; SE = 0.33). Restricting the analysis to the 91 participants in the validation subgroup yielded an estimate of the odds ratio of 0.57 (95% CI: 0.20, 1.59; SE = 0.52). While this result was similar to the estimate from the complete data, it was less precise, as expected based on the smaller sample size. Assuming outcome misclassification was nondifferential with respect to treatment group, the direct maximum likelihood approach estimated an odds ratio of 0.62 (95% CI: 0.26, 1.53; SE = 0.46). Assuming differential misclassification, the estimated odds ratio from direct maximum likelihood was 0.59 (95% CI: 0.22, 1.55; SE = 0.49). Accounting for outcome misclassification through multiple imputation produced estimated odds ratios of 0.60 (95% CI: 0.24, 1.51; SE = 0.47) and 0.62 (95% CI: 0.24, 1.61; SE = 0.49) assuming nondifferential and differential misclassification, respectively. Estimates from the direct maximum likelihood and multiple imputation approaches were similar in magnitude to estimates from the validation subgroup alone, and marginally more precise.

Table 3 presents results from several analyses of the risk ratio. Direct maximum likelihood and multiple imputation produced estimates of the risk ratio that were similar to the estimate of the risk ratio from the complete data using physician-diagnosed recurrence

as the outcome measure (RR=0.68; 95% CI: 0.44, 1.07; SE=0.23). Accounting for outcome misclassification using direct maximum likelihood produced an estimated risk ratio of 0.68 (95% CI=0.34, 1.38; SE=0.36) assuming nondifferential misclassification and 0.65 (95% CI=0.31, 1.40; SE=0.39) assuming differential misclassification. The estimated risk ratios from the multiple imputation approach were 0.69 (95% CI: 0.35, 1.36; SE = 0.35) and 0.69 (95% CI: 0.34, 1.41; SE=0.36) assuming nondifferential and differential misclassification, respectively. Estimates of the risk ratio from both direct maximum likelihood and multiple imputation were similar in magnitude to estimates from analysis limited to the validation subgroup (RR=0.61; 95% CI: 0.27, 1.35, SE=0.41), and slightly more precise.

Simulation Results

Results from the simulations indicated that multiple imputation removed bias due to outcome misclassification under all combinations of sensitivity, specificity, and validation subgroup sizes explored. Naïve estimates were biased dramatically towards the null in scenarios with both nondifferential and differential misclassification, with bias increasing as sensitivity decreased (Tables 4 and 5). In contrast, the multiple imputation approach yielded estimates of the odds ratio with less bias than the naïve analysis in all scenarios examined. Bias in odds ratios estimated by multiple imputation was similar in magnitude to bias in estimates from analyses limited to the validation subgroup and bias in estimates obtained using direct maximum likelihood. Bias decreased as the proportion of participants in the validation subgroup increased, but all three correction methods succumbed to finite sample bias when the total number of subjects in the validation subgroup was small.

Confidence intervals from the naïve analysis showed poor coverage, which varied as a function of sensitivity and sample size. Confidence intervals from multiple imputation maintained appropriate coverage, as did those from the validation subgroup and direct maximum likelihood.

Multiple imputation and direct maximum likelihood generally had smaller mean squared error than analysis limited to the validation subgroup. However, all three methods to account for outcome misclassification typically had larger mean squared error than the naïve analysis because the added imprecision of the correction methods offset the corresponding reduction in bias.

In simulations under a true odds ratio of 1, both direct maximum likelihood and multiple imputation preserved the type-1 error rate of 5%. Results from simulations under a true odds ratio of 0.5 indicated that both direct maximum likelihood and multiple imputation had higher statistical power than limiting analysis to the validation subgroup at levels of sensitivity commonly seen in the literature (0.9 and 0.6), but that all three non-naïve methods had similar statistical power at low values of sensitivity (0.3), as seen in the example (Figure). Analyses accounting for misclassification using multiple imputation were slightly less powerful than those using direct maximum likelihood. As expected, statistical power for the methods to account for outcome misclassification increased as the sensitivity of the observed outcome measure increased. Despite a pronounced null bias, the naïve analysis had high statistical power when sensitivity was large due to its high precision. However, when sensitivity decreased, bias in the naïve analysis caused power to fall well below that of the other methods.

4.3 Discussion

Multiple imputation performed well to account for bias due to outcome misclassification in the Herpetic Eye Disease Study example and the scenarios explored through simulation. Estimates from multiple imputation were similar in magnitude to estimates from the complete data using the gold standard outcome and were marginally more precise than estimates from analysis limited to the validation subgroup. Multiple imputation produced estimates that were similar in magnitude and precision to estimates obtained using direct maximum likelihood to account for outcome misclassification. These results were supported in Monte Carlo simulations, where multiple imputation yielded estimates with little bias in all scenarios and was sometimes more statistically powerful than analyses limited to the validation subgroup.

Both multiple imputation and direct maximum likelihood have been used to handle traditional missing data situations (50,64,65) and exposure measurement error (4,6). Both approaches have been shown to provide consistent and asymptotically normal estimates. The direct maximum likelihood approach produces estimates that are asymptotically efficient, while multiple imputation produces estimates that approach asymptotic efficiency as the number of imputations increases (52). While multiple imputation employs a two-stage estimation procedure, it can be implemented with standard missing data methods. In contrast, though direct maximum likelihood methods perform estimation in a single step, these methods must be programmed explicitly using a procedure that is able to obtain maximum likelihood estimates given a general likelihood expression, such as the SAS procedure NLMIXED. We could have addressed outcome misclassification with other

techniques to handle missing data, such as inverse probability weights or the expectation maximization (EM) algorithm. We chose to use multiple imputation because the standard inverse probability weighted estimator is inefficient (66) and the EM algorithm is more difficult to implement in standard software.

In the example, we demonstrated that the multiple imputation approach can be easily adapted to estimate risk ratios using log-binomial regression. Had the binomial model not converged, we could have applied the multiple imputation approach with any standard method to estimate the risk ratio, including the "copy method" applied in the binomial model (67,68), modified Poisson regression (69), or Bayesian techniques (70). More importantly, flexibility in the choice of analysis models enables the multiple imputation techniques illustrated here to be further extended to account for misclassification of non-binary outcomes by altering the imputation and analysis models. For continuous outcomes measured with error, the observed outcome measure and covariates could be regressed on the gold-standard outcome measure in the validation subgroup using linear regression. Coefficients from this model could be used to impute outcomes for study participants not in the validation subgroup, and the complete dataset could then be analyzed using the appropriate analysis model.

Another advantage of the multiple imputation approach is it easily allows researchers to include different sets of variables in the imputation model and the analysis model. Performing the imputation and analysis using different models avoids the problem of conditioning on variables influencing only the relationship between the observed and gold standard outcome in the final analysis model. Likewise, the imputation model could be

altered to employ more flexible prediction functions in place of the linear-logistic model used to impute outcomes in this example (71).

Although the present work focuses on estimation of effect measures in a closed cohort, flexibility in choice of analysis model allows the multiple imputation approach to be extended to account for outcome misclassification in analysis of time-to-event outcomes in situations where the event type is subject to error but the event date is assumed to be known. In this scenario, the event indicator could be imputed using the monotone logistic method, and the hazard ratio or rate ratio would be estimated in each imputation and summarized using Equation 4.4.

Measurement error methods typically assume that the relationship between the true outcome and the observed outcome variable is monotonic, which implies that the observed outcome measure either increases, plateaus, or decreases with increasing levels of the gold standard measure, but does not decrease following an increase or vice versa (15). Monotonicity is ensured for binary outcome variables (as in the example), but must be considered for non-binary outcomes.

In the example, accounting for misclassification with multiple imputation and direct maximum likelihood offered only slight gains in precision over analysis limited to the validation subgroup. We expect estimates from multiple imputation and direct maximum likelihood to be more precise than estimates from the validation subgroup because these methods use information from all participants in the study to estimate the effect size, while analysis limited to the validation subgroup discards all information on participants missing the gold standard outcome measure. Because, in our example, the observed outcome was a poor proxy for the gold standard outcome, the imputation model contained a high degree of

uncertainty that propagated through to the variance of the final effect estimate. Larger gains in precision would be expected if sensitivity in the example data were higher or the proportion of participants in the validation substudy were smaller. However, in the example, when the proportion of participants in the validation substudy was further reduced, the absolute numbers in the validation substudy became so small that results became unstable.

In simulations, we used mean squared error to assess the tradeoff between bias and precision. Despite its large bias, the naïve analysis had smaller mean squared error than methods to account for outcome misclassification in most of the scenarios explored through simulation. Because mean squared error places equal weight on bias and variance, the precision of the naïve analysis offset its bias. In large sample sizes, where mean-squared error is dominated by bias instead of random error, the non-naïve methods will be superior to the naïve analysis. The simulation results can be interpreted only under the assumption that the underlying data generating mechanism matches the parametric models used to simulate the data. It is unclear how multiple imputation and direct maximum likelihood would have performed under a misspecified analysis model.

We have shown that multiple imputation works well to account for both nondifferential and differential outcome misclassification. When the degree of misclassification varied across levels of exposure, we often saw separation of data points in the imputation model. Separation is likely to occur when the positive predictive value of the observed outcome is high. In this analysis, we applied Firth's correction to obtain point estimates in these models. Alternatively, Bayesian methods could be used to address the

problem of separation by incorporating prior information to stabilize regression coefficients.

A limitation of both multiple imputation and direct maximum likelihood approaches is that they depend on correct specification of the model relating the observed outcome to the gold standard outcome measure. Estimates of the association between exposure and outcome could be biased if the relationship between observed and gold standard measurements is not transportable, implying that it is not consistent between the validation subgroup and the complete data. Obtaining a representative validation subgroup is vital to any method using a validation study to account for misclassification, as these methods typically assume that information on the gold standard outcome measure is missing at random. Because inclusion in the validation subgroup determines if the gold standard outcome is missing for a participant, the probability of being included in the validation study must be independent of that participant's gold standard outcome, given the observed outcome and the covariates. When information on the gold standard outcome measure is not missing at random, the transportability assumption may not be met.

We must also consider the possibility that the gold-standard measurement is itself misclassified. A fundamental limitation of all validation studies is that they assume that the gold standard outcome measure represents the true outcome. In the example, physician diagnosis may have been misclassified if a participant experienced a recurrence of HSV that resolved before the opportunity for physician diagnosis or if errors occurred during chart abstraction. In situations in which the gold standard measurement is itself subject to non-negligible error, using methods that rely on validation data to account for outcome misclassification may yield biased and falsely precise estimates (72).

Under the assumptions mentioned above, applying multiple imputation to account for outcome misclassification removes bias in effect estimates from logistic and log-binomial regression. This technique uses well-established missing data methods that can be implemented using standard statistical software and provides an opportunity for data analysts to account for outcome misclassification in wide range of statistical models.

4.4 Tables and Figures

Table 4.1. Characteristics of full cohort and validation subgroup [a] classified by self-reported recurrence and physician-diagnosed recurrence of ocular herpes simplex virus during 12 months of follow-up, 308 participants in the multicenter Herpetic Eye Disease Study followed for 12 months between 1992 – 1998

	Full Cohort (N=308)		Validation subgroup (N=91)	
	Placebo (No.)	Acyclovir (No.)	Placebo (No.)	Acyclovir (No.)
No self-reported recurrence [b]				
No diagnosed recurrence	104	119	30	35
Diagnosed recurrence	26	14	8	4
Self-reported recurrence				
No diagnosed recurrence	11	10	3	3
Diagnosed recurrence	12	12	4	4
Sensitivity [c]	0.32	0.46	0.33	0.50
Specificity [d]	0.90	0.92	0.91	0.92

[a] Self-reported outcomes and physician records were available for all 308 participants. We sampled a synthetic validation subgroup of 91 participants for the purposes of illustration

[b] Participants reported ocular HSV recurrences through a weekly diary and were seen by an ophthalmologist every 3 months. Self-reported recurrences refers to data obtained from patient diaries and diagnosed recurrences refer to results of examination by the study ophthalmologist

[c] Sensitivity was the proportion of patients with a physician-diagnosed recurrence who also self-reported a recurrence

[d] Specificity was the proportion of participants without a physician-diagnosed recurrence who did not self-report a recurrence

Table 4.2. Estimates of the odds ratio comparing recurrence of ocular herpes simplex virus between participants randomized to acyclovir or placebo from various models, 308 participants in the multicenter Herpetic Eye Disease Study followed for 12 months between 1992 – 1998

Model	No. Outcomes	No. at Risk	Adjusted OR [a]	95% CI	SE for ln(OR)
Complete data, physician diagnosed recurrence					
Acyclovir group	26	155	0.62	0.35, 1.09	0.29
Placebo group	38	153	1		
Total	64	308			
Naïve analysis					
Acyclovir group	22	155	0.90	0.47, 1.73	0.33
Placebo group	23	153	1		
Total	45	308			
Validation subgroup [b]					
Acyclovir group	8	46	0.57	0.20, 1.59	0.52
Placebo group	12	45	1		
Total	20	91			
Direct maximum likelihood (nondifferential)			0.62	0.26, 1.53	0.46
Direct maximum likelihood (differential)			0.59	0.22, 1.55	0.49
Multiple imputation (nondifferential)			0.60	0.24, 1.51	0.47
Multiple imputation (differential)			0.62	0.24, 1.61	0.49

* OR, odds ratio; CI, confidence interval; No., number; SE, standard error; ln(OR), natural log of the odds ratio
[a] All models were adjusted for race, sex, age, and number of previous recurrences [b] Validation subgroup includes 91 participants

Table 4.3. Estimates of the risk ratio comparing recurrence of ocular herpes simplex virus between participants randomized to acyclovir or placebo from various models, 308 participants in the multicenter Herpetic Eye Disease Study followed for 12 months between 1992 – 1998

Model	No. Outcomes	No. at Risk	Adjusted RR [a]	95% CI	SE for ln(RR)
Complete data, physician diagnosed recurrence					
Acyclovir group	26	155	0.68	0.44, 1.07	0.23
Placebo group	38	153	1		
Total	64	308			
Naïve analysis					
Acyclovir group	22	155	0.93	0.55, 1.59	0.27
Placebo group	23	153	1		
Total	45	308			
Validation subgroup [b]					
Acyclovir group	8	46	0.61	0.27, 1.35	0.41
Placebo group	12	45	1		
Total	20	91			
Direct maximum likelihood (nondifferential)			0.68	0.34, 1.38	0.36
Direct maximum likelihood (differential)			0.65	0.31, 1.40	0.39
Multiple imputation (nondifferential)			0.69	0.35, 1.36	0.35
Multiple imputation (differential)			0.69	0.34, 1.41	0.36

* RR, risk ratio; CI, confidence interval; No., number; SE, standard error; ln(RR), natural log of the risk ratio

[a] All models were adjusted for race, sex, age, and number of previous recurrences

[b] Validation subgroup includes 91 participants

Table 4.4. Bias and 95% confidence interval coverage for simulation studies [a] under 9 scenarios for nondifferential misclassification

				Naive			Validation			Direct ML [b]			MI [c]		
Sensitivity	Specificity	N	Percent [d]	Bias [e]	Cover [f]	MSE [g]	Bias	Cover	MSE	Bias	Cover	MSE	Bias	Cover	MSE
0.9	0.9	1000	10	24	62	8	-5	96	47	-4	96	33	2	97	27
		1000	30	24	62	8	-1	95	10	-1	95	6	-1	95	6
		300	30	24	85	13	-5	96	47	-4	96	42	2	97	27
0.6	0.9	1000	10	35	40	15	-5	96	47	-5	96	48	-3	96	29
		1000	30	35	40	15	-1	95	10	-1	95	8	-1	95	8
		300	30	35	76	21	-5	96	47	-4	96	48	-3	96	33
0.3	0.9	1000	10	51	20	30	-5	96	47	-5	96	54	-4	95	35
		1000	30	51	20	30	-1	95	10	-1	95	9	-1	95	9
		300	30	51	66	38	-5	96	47	-5	96	58	-3	96	37

*MI, multiple imputation; ML, maximum likelihood; MSE, mean squared error;

[a] Results are summarized over 10,000 simulations

[b] Direct maximum likelihood

[c] Multiple imputation

[d] Percent of all participants included in the validation subgroup

[e] Bias was defined as 100 times the difference between the average estimated log odds ratio and the true log odds ratio

[f] Confidence interval coverage was calculated as the percentage of simulations in which the estimated 95% Wald-type confidence limits included the true value

[g] MSE was calculated as the sum of the bias squared and the variance

Table 4.5. Bias and 95% confidence interval coverage for simulation studies [a] under 6 scenarios for differential misclassification.

				Naive			Validation			Direct ML [b]			MI [c]		
Sensitivity [d]	Specificity	N	Percent [e]	Bias [f]	Cover [g]	MSE [h]	Bias	Cover	MSE	Bias	Cover	MSE	Bias	Cover	MSE
(0.95, 0.85)	0.9	1000	10	34	36	14	-5	96	47	-5	96	41	4	100	12
		1000	30	34	36	14	-1	95	10	-1	95	6	1	97	6
		300	30	34	75	19	-5	96	47	-4	96	42	3	99	17
(0.70, 0.50)	0.9	1000	10	60	4	39	-5	96	47	-5	96	46	3	98	17
		1000	30	60	4	39	-1	95	10	-1	95	8	0	96	7
		300	30	60	47	45	-5	96	47	-5	96	49	2	98	23

* MI, multiple imputation; ML, maximum likelihood; MSE, mean squared error;

[a] Results are summarized over 10,000 simulations

[b] Direct maximum likelihood

[c] Multiple imputation

[d] Sensitivity differs by exposure group; presented as (sensitivity for $X = 1$, sensitivity for $X = 0$)

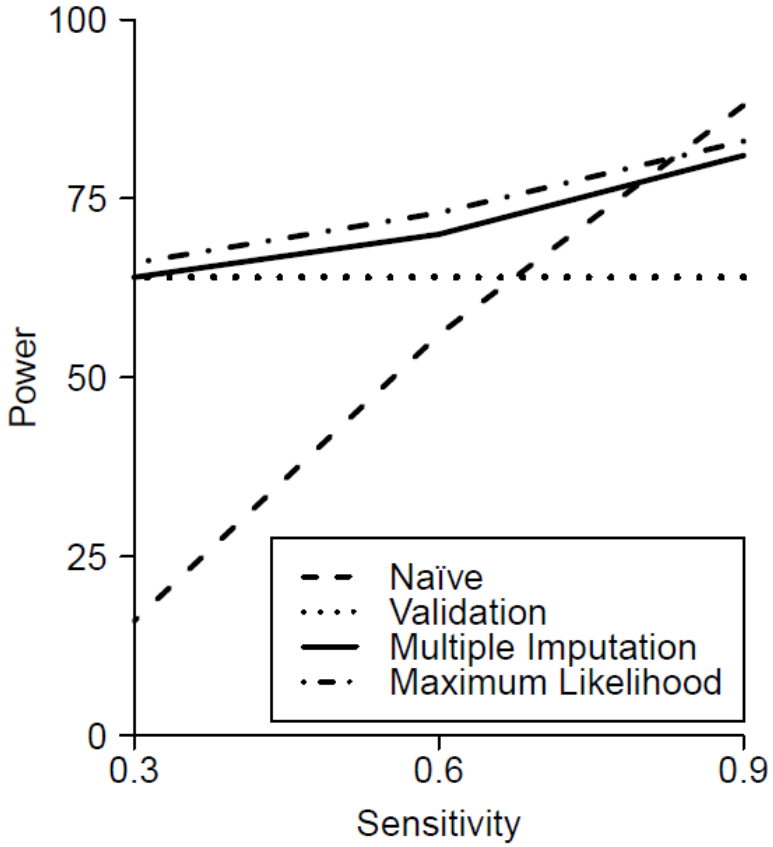
[e] Percent of all participants included in the validation subgroup

[f] Bias was defined as 100 times the difference between the average estimated log odds ratio and the true log odds ratio

[g] Confidence interval coverage was calculated as the percentage of simulations in which the estimated 95% Wald-type confidence limits included the true value

[h] MSE was calculated as the sum of the bias squared and the variance

Figure 4.1. Relationship between statistical power and sensitivity of the observed outcome measure in simulations with a 30% validation subgroup and a total sample size of 1000 for the naïve analysis, analysis limited to the validation subgroup, the direct maximum likelihood method, and the multiple imputation method to account for outcome misclassification



5 Accounting for outcome misclassification in the effect of occupational asbestos exposure on lung cancer mortality

Asbestos is a known cause of lung cancer, but the outcome typically used to quantify the relationship between asbestos exposure and lung cancer, lung cancer death, is subject to misclassification. We used modified maximum likelihood and Bayesian methods to explore the effects of outcome misclassification on the rate ratio of lung cancer death per 100 fiber-y/mL of asbestos exposure. The standard covariate-adjusted estimate of the rate ratio was 1.94 (95% confidence interval [CI]: 1.55, 2.44), and modified maximum likelihood produced similar results when we assumed that the specificity of outcome classification was 0.98, regardless of sensitivity. With sensitivity and specificity assumed to be 0.90, estimated rate ratios were farther from the null, and less precise (rate ratio = 2.97; 95% CI: 1.34, 6.56). With specificity constrained above 0.95 in the Bayesian analysis, the posterior estimate was similar to the standard estimate, but when specificity was centered at 0.95 (and sensitivity centered at 0.85), the rate ratio rose to 2.04 (95% posterior interval [PI]: 1.61, 2.56). In the present context, standard estimates for the effect of asbestos on lung cancer death were similar to estimates accounting for the relatively minor misclassification. However, modified maximum likelihood and Bayesian methods were needed to verify the robustness of standard estimates, and these approaches will provide unbiased estimates in settings with more misclassification.

5.1 Introduction

The relationship between occupational exposure to asbestos and lung cancer mortality has been examined for over half a century, and epidemiological studies have provided strong evidence that asbestos is a lung carcinogen (42,58,59). Although asbestos is no longer mined in the United States, approximately 1000 tons of asbestos is imported into the US each year for use in construction materials, brake linings, and other products (40). Moreover, a substantial amount of asbestos remains in US infrastructure and eventually will be removed, either during remediation, renovations, or demolition. Significant production and use is also ongoing in middle-income industrial countries, including Brazil, India, China and Russia. Therefore asbestos continues to pose important occupational hazards in the US and worldwide (41).

Most analyses of asbestos exposure in occupational settings have estimated the effect of asbestos on lung cancer mortality in place of lung cancer incidence for practical reasons. Because many countries have comprehensive databases containing standardized information about deaths, investigators can identify the observed deaths that are due to lung cancer, while the dates of lung cancer incidence are typically less well-known. The number of lung cancer deaths approximates the number of incident lung cancer cases because the time between lung cancer incidence and death is relatively short and few effective treatments have existed.

Outcome ascertainment in studies of lung cancer mortality involves determining both the date and cause(s) of death. In the current paper we focus on the scenario in which the underlying cause of death is used to classify each decedent with respect to the outcome. In most cases, particularly in developed countries, the date of death is recorded with

typically negligible error. However, misattribution of cause of death remains more likely. If such misattribution results in a death due to lung cancer being classified as a death due to other causes, or vice versa, the outcome is misclassified. Despite evidence of imperfect sensitivity and specificity for cause of death abstracted from death certificates (32,33,35–37,49,73,74), most studies of occupational asbestos exposure assume no misclassification.

To present estimates of the effect of asbestos exposure on lung cancer mortality that account for outcome misclassification, we propose an approach that uses modified maximum likelihood to estimate rate ratios under chosen values of sensitivity and specificity, as in a sensitivity analysis. We expand this approach to account for uncertainty in the values of sensitivity and specificity by placing informative prior distributions on sensitivity and specificity. Until the discussion, we assume that the date of death is measured without error but that the cause of death is subject to misclassification. We illustrate these approaches using data from a cohort of textile workers in South Carolina, United States assembled to assess the relationship between the chrysotile form of asbestos and lung cancer mortality.

5.2 Methods

Study population

The study population comprised workers at a textile production plant in South Carolina that produced asbestos beginning in 1896 and asbestos textile products beginning in 1909 (58). The study enrolled 3072 men and women who were employed at the plant for at least one month between 1 January 1940 and 31 December 1965. Employment

records were used to obtain information on date of birth, year of study entry, race (Caucasian or non-Caucasian), sex, and employment status in each year.

Exposure assessment

Detailed work histories were available for each participant in the cohort. Cumulative exposure to the chrysotile form of asbestos was estimated using a job exposure matrix to link work history to industrial hygiene sampling measurements taken between 1930 and 1975, as previously described (59). Chrysotile exposure concentrations, expressed as fibers longer than 5 micrometers per milliliter of air (fibers/mL), were estimated for each day of each participant's work history. Yearly exposure values were calculated as the product of the proportion of the year worked and the average daily exposure concentration and reported as fiber-years per milliliter (fiber-y/mL). To capture the appropriate exposure window for the effect of asbestos on lung cancer mortality, exposure values were lagged 10 years, meaning that the risk of lung cancer in each year was not affected by asbestos exposure in the prior 10 years.

Mortality ascertainment

Participants were followed for lung cancer death through January 1, 2001. Between 1940 and 1978, vital status was determined by the US Social Security Administration, Internal Revenue Service, Veterans Affairs, state drivers' license files, vital statistics offices, and postal mail correction services. Between 1979 and 2001, vital status was determined using the National Death Index. Those confirmed alive in 1979 and not found in the National Death Index were assumed to be alive at the end of the study. For those who died,

cause of death was determined through examination of death certificates and coded by a qualified nosologist into the revision of the International Classification of Diseases (ICD) in effect at the date of each death. The underlying cause of death was used to define the outcome. A death was considered to be a case if it was classified as a death due to lung cancer (defined as ICD-8 and ICD-9 code 162, and ICD-10 codes C33 – C44) (42,58,59).

Statistical methods

The 121,010 person years contributed by 3072 participants were grouped into 3059 populated strata, indexed as $j = 1, 2, \ldots, J$. These strata are defined by sex, age (5 year intervals from 15 to 90), and year at study entry (1 year intervals from 1940 to 1965), and cumulative asbestos exposure. Cumulative asbestos exposure was categorized into 1 fiber - y/mL intervals for values 10 fiber-y/mL and under, 5 fiber-y/mL intervals for values from 10 to 50 fiber-y/mL, 10 fiber-y/mL intervals for values from 50 to 100 fiber-y/mL, and 25 fiber-y/mL intervals for values above 100 fiber-y/mL, and the category score was set to the mean value of asbestos exposure for each interval.

In strata j, we have n_j person-years with d_j deaths. We observe w_j possibly misclassified lung cancer cases, but the true unobserved number of lung cancer cases is y_j. The true number of person-years and deaths remains n_j and d_j, respectively under the assumption that the dates of death are correct.

We would like to estimate the effect of occupational asbestos exposure on lung cancer mortality by estimating the rate ratio of lung cancer death per 100 fiber-y/mL of asbestos exposure. The parameter estimating the desired rate ratio is $\exp(\beta_1)$ in the Poisson model

$$\lambda_j = \exp\left(\beta_0 + \beta_1 X_j + \sum_{m=2}^{4} \beta_m Z_{j(m-1)}\right),$$

where λ_j represents the rate of true lung cancer deaths in strata j, X_j is the cumulative asbestos exposure, and $\boldsymbol{Z}$ is a $J \times 3$ matrix with columns for sex, log(age), and year of study entry.

However, because we observe w_j possibly misclassified lung cancer cases in place of y_j true lung cancer cases, the model above cannot be fit directly. Instead, standard analyses typically fit the model

$$\lambda_j' = \exp\left(\gamma_0 + \gamma_1 X_j + \sum_{m=2}^{4} \gamma_m Z_{j(m-1)}\right),$$

where λ_j' represents the rate of a possibly misclassified version of the outcome variable, w_j. We fit this second (naïve) model to the data from the South Carolina textile plant cohort, where w_j is the number of deaths due to lung cancer recorded on death certificates in strata j.

We account for misclassification of the outcome using values of sensitivity and specificity of lung cancer classification obtained from existing literature. Sensitivity is defined as the probability that a participant is correctly classified as a lung cancer case, given that the participant died of lung cancer. Specificity is the probability that a participant is classified as a non-lung cancer death, given than the participant died of a cause other than lung cancer. Because validation studies report varying estimates of the accuracy of cause-of-death information obtained from death certificates, we perform a sensitivity analyses in which we set sensitivity and specificity to each of several plausible

values. We then adopt a Bayesian approach by placing informative prior distributions on parameters to incorporate uncertainty in the sensitivity and specificity.

Sensitivity analysis

We first demonstrate how to modify the Poisson likelihood to account for outcome misclassification by setting the values of sensitivity and specificity, as one would in a sensitivity analysis. We begin by specifying the Poisson likelihood for a situation with two causes of death and no misclassification,

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{J} \lambda_j^{y_j} \, \mu_j^{(d_j - y_j)} \exp\{-(\lambda_j + \mu_j) n_j\}$$

where λ_j is described above, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_4)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ \ldots, \beta_4)$, and μ_j is the estimated rate of other types of death for strata *j*,

$$\mu_j = \exp\left(\alpha_0 + \alpha_1 X_j + \sum_{m=2}^{4} \alpha_m \, Z_{j(m-1)}\right),$$

where $\boldsymbol{Z}$ is a $J \times 3$ matrix with columns for sex, age, and year of study entry.

However, because the true number of lung cancer deaths is unavailable, we must modify the likelihood to use the count of potentially misclassified lung cancer deaths for each stratum, w_j and the misclassification probabilities (i.e., sensitivity and specificity) to restructure the likelihood as:

$$L_{\text{modified}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{J} \{\lambda_j(se) + \mu_j(1 - sp)\}^{w_j} \{\lambda_j(1 - se) + \mu_j(sp)\}^{(d_j - w_j)}$$

$$\times \exp\{-[\lambda_j(se) + \mu_j(1 - sp) + \lambda_j(1 - se) + \mu_j(sp)] n_j\}.$$

Under the assumption that sensitivity and specificity are correct, the modified likelihood function will be maximized at the same estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as the likelihood function above, though the modified likelihood function will be more diffuse.

To identify plausible values of the misclassification probabilities, we turned to existing literature on the accuracy of cause of death information reported on death certificates. The Life Span Study, which performed autopsies on selected participants who died following the atomic bombings of Hiroshima and Nagasaki, found that death certificates listed lung cancer as the underlying cause of death in only 62% of the cases where the autopsy indicated that lung cancer was the cause of death (36). A more recent study from the Mayo lung clinic in the United States reported that death certificates identified lung cancer as the underlying cause of death in 89% (210/237) of autopsy-confirmed lung cancer cases (34), while specificity was 99%. A validation study conducted in the Third National Cancer Survey found that lung cancer was recorded as the underlying cause of death on the death certificate in 95% (9568/10059) of lung cancer cases diagnosed by hospital physicians (73). Sensitivity from other validation studies fell between the estimates from the Life Span Study and the Mayo lung clinic. For example, a study of 4951 deaths occurring among 17,800 workers exposed to asbestos reported that death certificates identified lung cancer as the cause of death in 86% of the deaths designated as lung cancer deaths by autopsy and other medical evidence (37).

We allowed sensitivity to range from 0.6, as seen in the Life Span Study, to 0.9 as seen in the Mayo Lung Clinic study. Because few validation studies provided the specificity of death certificates to identify lung cancer deaths, we investigated three plausible values for specificity: 0.98, 0.95, and 0.90. We estimated the rate ratio of lung cancer death per

100 fiber-years/mL of asbestos exposure for the following scenarios: 1) Assuming no misclassification of cause of death, which corresponds to the standard naïve analysis of these data; 2) setting specificity to 0.98 and sensitivity to 0.9, 0.8, and 0.6; 3) setting specificity to 0.95 and sensitivity to 0.9, 0.8, and 0.6; and 4) setting specificity to 0.90 and sensitivity to 0.9, 0.8, and 0.6. In all scenarios, we assumed that outcome misclassification was nondifferential with respect to cumulative asbestos exposure. To compare more directly with the Bayesian analysis, we also examined one scenario in which specificity was set to 0.95 and sensitivity was set to 0.85. The sensitivity analysis was performed using SAS procedure NLMIXED (V 9.3, SAS Institute, Cary, NC), and code to perform this analysis is available in appendix 4.

We evaluated the performance of this method to account for outcome misclassification through Monte Carlo simulations. Bias, 95% confidence interval coverage, and mean squared error were compared between standard naïve methods and the analysis using modified maximum likelihood to set values of sensitivity and specificity for 5 scenarios with varying degrees of outcome misclassification. The design and results of the simulation study are detailed in appendix 5.

Prior distributions for sensitivity and specificity

We are rarely certain about the amount of misclassification present in the data. In this analysis, we acknowledge this uncertainty by placing informative prior distributions on sensitivity and specificity that reflect beliefs about the amount of misclassification in the data and certainty about those beliefs. Based on the existing literature, we assumed sensitivity was uniformly distributed between 0.75 and 0.95. Specificity was assumed to be

uniformly distributed between 0.9 and 1.0. We explored the effects of tightening the prior distributions and shifting the prior distribution for specificity. Posterior estimates of the rate ratio for lung cancer mortality per 100 f-y/mL of asbestos exposure were obtained using Markov chain Monte Carlo simulation by the SAS procedure MCMC. We report posterior intervals, which were equivalent to the highest posterior density intervals. Models were run for 500,000 iterations with a burn-in of 50,000 iterations. Every third value after burn-in was retained in the calculation of the posterior estimates and 95% posterior intervals (PI) to minimize autocorrelation. Convergence was assessed through examination of trace and autocorrelation plots. SAS code is provided in appendix 4.

5.3 Results

The study enrolled 3072 textile workers between 1940 and 1965. The cohort was predominantly male and Caucasian and enrolled at a median age of 23 years (table 1). The median occupational exposure to asbestos at study entry was 0.2 fiber-y/mL and the median cumulative occupational exposure to asbestos at the end of follow-up was 5.0 fiber-y/mL. One hundred ninety-eight lung cancer deaths and 1763 other deaths were recorded between 1940 and 2001, and 265 participants were lost to follow-up (9%) (Table 1).

Table 2 provides the estimated ratio ratios for lung cancer death per 100 fiber-y/mL cumulative asbestos exposure under several assumptions about sensitivity and specificity. Assuming perfect sensitivity and specificity of cause of death information, as in standard analyses, the rate of lung cancer deaths increased by a factor of 1.94 per 100 fiber-y/mL (95% CI: 1.72, 2.62) after adjustment for sex, race, age, and year of study entry.

The rate ratios under scenarios assuming varying degrees of nondifferential outcome misclassification were further from the null than the rate ratio from the standard analysis assuming perfect sensitivity and specificity. The change in the rate ratio was determined primarily by the higher specificity. With specificity set to 0.98, the rate ratio was relatively unchanged as 2.03 (95% CI: 1.57, 2.61), 2.02 (95% CI: 1.57, 2.60), and 2.00 (95% CI: 1.56, 2.56) when sensitivity was varied as 0.9, 0.8, and 0.6, respectively. When specificity was reduced to 0.95, the estimated rate ratios were 2.19 (95% CI: 1.60, 3.00), 2.17 (95% CI: 1.59, 2.98), and 2.12 (95% CI: 1.56, 2.89) for sensitivity set to 0.90, 0.80, and 0.60, respectively. When specificity was further reduced to 0.90, estimates of the rate ratio were even farther from the null but much less precise. With specificity at 0.9 and sensitivity set to 0.9, 0.8, and 0.6, the estimated rate ratios were 2.97 (95% CI: 1.34, 6.56), 3.03 (95% CI: 1.32, 6.94), and 3.07 (95% CI: 1.54, 6.10).

In simulations, revised estimates using modified maximum likelihood showed little bias in all scenarios examined, even when sensitivity and specificity were quite low. Similarly, confidence limits from the revised estimates showed appropriate coverage, and mean squared error was improved for the revised estimates when compared to the standard estimates under all combinations of sensitivity and specificity. Detailed numeric results from the simulations are provided in appendix 2.

Bayesian analysis

Results from the Bayesian analysis are presented in table 3. Rate ratios were attenuated when compared to results using modified maximum likelihood, but remained further from the null than results from the standard analysis. Under our consensus prior

(sensitivity uniformly distributed around 0.85 and specificity uniformly distributed around 0.95), the rate ratio was 2.04 (95% PI: 1.61, 2.56). Tightening the priors for sensitivity and specificity produced an estimated rate ratio of 2.14 (95% PI: 1.69, 2.78), and shifting the prior for specificity to be uniformly distributed between 0.95 and 1.0, moved the rate ratio to 1.96 (95% PI: 1.53, 2.51), similar to the value given by standard methods. Shifting the prior for sensitivity to include all of the values of sensitivity estimated from the previous validation studies referenced had little effect on the posterior estimate of the rate ratio or posterior interval.

5.4 Discussion

Misclassification of cause of death has been a concern in analysis of cancer trends and etiologic research in cancer epidemiology for decades (36,37,73–75). We accounted for misclassification of lung cancer death in a cohort of textile factory workers exposed to asbestos using a modified maximum likelihood approach. The covariate-adjusted rate ratio of lung cancer death per 100 fiber-y/mL of asbestos exposure of 1.94 obtained using standard methods rose to over 3.00 when sensitivity and specificity were assumed to be poor, though rose only to 2.18 under likely values of sensitivity and specificity.

Estimates of the rate ratio from a sensitivity analysis assuming imperfect sensitivity and specificity were always further from the null than the standard analysis assuming perfect outcome classification, though less precise. When informative prior distributions were placed on the parameters determining outcome misclassification rather than setting these values to constants, estimates of the rate ratio fell between estimates from standard analysis and estimates from the sensitivity analysis. In simulations with imperfect

sensitivity and specificity, using modified maximum likelihood to account for outcome misclassification removed bias in all scenarios examined and resulted in smaller mean squared error than the standard analysis.

Sensitivity analysis showed that estimates of the rate ratio were relatively insensitive to changes in hypothetical values of sensitivity, but changed substantially when specificity was altered. The sensitivity of the rate ratio to changes in specificity is not surprising; when the event is rare, even small changes in the specificity result in considerable changes in the number of events assumed to have occurred.

A similar pattern emerged in the Bayesian analysis. When specificity was constrained to be between 0.95 and 1 and sensitivity was allowed to range from 0.75 to 0.95, the posterior estimate of the rate ratio was similar to the rate ratio estimated with standard methods. However, when specificity was centered at 0.95 (and sensitivity was centered at 0.85), the posterior rate ratio fell between estimates of the rate ratio from the sensitivity analysis, in which sensitivity and specificity were set to constant values, and standard methods, in which sensitivity and specificity were assumed to be 1. When the prior distributions for sensitivity and specificity were tightened, posterior estimates of the rate ratio moved away from the estimate from standard methods and towards the estimate from the sensitivity analysis and posterior intervals were tighter.

Because the specificity of lung cancer mortality reported on death certificates is high, posterior estimates of the rate ratio per 100 f-y/mL of asbestos exposure were similar to the adjusted rate ratio from standard analysis. We would expect posterior rate ratios to differ more dramatically from the standard estimates of the rate ratio for outcomes subject to more severe misclassification. For example, asbestos has been implicated in the elevated

risk of mortality from coronary heart disease seen in cohorts of miners, mill workers, and shipyard workers (76–80). Unlike lung cancer, the specificity of coronary heart disease reported on death certificates is low (33,81). We expect that future studies of the relationship between asbestos and coronary heart disease mortality that account for outcome misclassification using methods such as those described here would produce estimates of the rate ratio that differ substantially from standard estimates.

We use results from existing validation studies on the accuracy of cause of death information on death certificates to inform the sensitivity analysis and Bayesian prior distributions. Here, we discuss the misclassification probabilities in terms of sensitivity and specificity instead of the detection rates and confirmation rates often presented in such validation studies. Sensitivity and detection rate both refer to the probability that the underlying cause of death recorded on the death certificate is lung cancer, given that a participant died of lung cancer. The confirmation rate often reported in validation studies is the probability that a participant died of lung cancer, given that the death certificate listed lung cancer as the underlying cause of death, and is also known as the positive predictive value. We choose to frame our methods to account for outcome misclassification in terms of sensitivity and specificity instead of detection and confirmation rates because the confirmation rate is sensitive to changes in the prevalence of the outcome.

This work extends existing approaches to account for outcome misclassification to the time-to-event setting. Magder and Hughs (46), Lyles (47), and Edwards (82) have illustrated maximum likelihood-based approaches to account for outcome misclassification in logistic regression. Here, we apply a modified likelihood function to account for misclassification between lung cancer deaths and other types of death in Poisson

regression, similar to work by Sposto et al (83) and Stamey et al (84), who have used a maximum likelihood approach (the EM algorithm) and Bayesian methods, respectively, in this setting. The current work complements the methods set forth in these papers by providing a less-computationally intensive direct maximum likelihood solution to account for the misclassification of outcomes, as well as applying this likelihood function in the Bayesian setting.

In studies without validation data, posterior estimates are largely determined by the prior distributions for sensitivity and specificity. Priors can be elicited through expert opinion or formed from the existing literature. Here, we have constructed the prior distributions based on three previous validation studies: the atomic bomb survivor data, the Mayo lung clinic study, and a validation study of workers exposed to asbestos. If available, internal or external validation data could be formally incorporated into the modified likelihood function in either the maximum likelihood or the Bayesian method described above, following Lyles (47). When validation data are available, the prior distributions for sensitivity and specificity will exert less influence on the posterior estimate of the rate ratio than when the prior contains most of the information about the misclassification parameters.

While sensitivity and specificity arise from a joint distribution, here we have chosen to specify independent prior distributions for sensitivity and specificity. In practice, independent priors on sensitivity and specificity are easier to elicit both from experts and the existing literature, and independent priors simplify computational aspects of both the sensitivity analysis and the Bayesian analysis. However, ignoring the correlation between

sensitivity and specificity may cause the uncertainty in the posterior estimates of the rate ratio to be either understated or overstated (85,86).

Both the sensitivity analysis and the Bayesian method presented above could be extended to account for outcome misclassification that is differential with respect to exposure. In both cases, sensitivity and specificity would be specified as a function of exposure. Differential outcome misclassification may be of interest in studies of self-reported outcomes or other situation in which the person recording the outcome of interest is aware of the participant's exposure status. In this analysis, because the coroner or medical professional assigned to complete the death certificate was likely unaware of the participant's cumulative asbestos exposure, outcome misclassification was assumed to be nondifferential with respect to exposure. Similarly, outcome classification may depend on covariates other than exposure status. If investigators believe that the validity of cause of death information on death certificates improves over time, sensitivity and specificity could be written as a function of calendar time or other relevant covariates.

In this analysis, we have assumed that the date of death was correct and that only the cause of death was subject to error. Under this assumption, the event time, and thus the person-time contributed by each participant, is assumed to be measured correctly, though the event indicator is error-prone. However, if the date of death were recorded incorrectly, a death was never recorded, or a death was falsely recorded, the event times would also be subject to error. Under these conditions, the modified maximum likelihood approach and the Bayesian methods presented here would be insufficient to account for the bias due to outcome mismeasurement. When the outcome is death, event times are usually correct in countries that require standardized reporting of all deaths. However, studies of other

outcomes, such as disease incidence, are more likely to be confronted with mismeasured event times, especially if detection of the disease is difficult.

Here, we have presented maximum likelihood and Bayesian methods to account for misclassification of lung cancer-specific mortality in a cohort of textile workers exposed to asbestos. Results from the sensitivity analysis and Bayesian analysis suggest that, at likely values of sensitivity and specificity, outcome misclassification of lung cancer death is not likely to produce substantial bias in standard estimates of the rate ratio for the effect of asbestos exposure on lung cancer death. However, sensitivity analysis suggests that standard methods to estimate rate ratios for outcomes subject to greater probability of misclassification, particularly those subject to poor specificity, are likely to produce biased estimates. Both the sensitivity analysis and the Bayesian analysis provided approaches to account for outcome misclassification in estimation of the rate ratio under varying beliefs about the misclassification parameters.

5.5 Tables

Table 5.1. Characteristics of 3072 textile workers in South Carolina, United States, 1940 – 2001

Characteristic:	Median (IQR)	Percent (*n*)
Age at study entry (years)	23 (19, 29)	
Calendar year at study entry	1943 (1941, 1946)	
Male		58.9 (1807)
Caucasian		81.4 (2500)
Cumulative asbestos exposure in fiber-y/ml at end of follow-up	4.99 (1.45, 21.38)	
Lung cancer deaths		6.5 (198)
Non lung cancer deaths		57.4 (1763)
Loss to follow-up		8.6 (265)

IQR, Interquartile range

Table 5.2. Rate ratio of lung cancer mortality per 100 fiber-years/mL cumulative asbestos exposure, South Carolina, United States, 1940 – 2001, under several outcome misclassification scenarios

Model	Specificity	Sensitivity	RR	95% CI
Crude	1	1	3.52	2.86, 4.33
Adjusted [a]	1	1	1.94	1.55, 2.44
	0.98	0.90	2.03	1.57, 2.61
	0.98	0.80	2.02	1.57, 2.60
	0.98	0.60	2.00	1.56, 2.56
	0.95	0.90	2.19	1.60, 3.00
	0.95	0.85	2.18	1.59, 2.99
	0.95	0.80	2.17	1.59, 2.98
	0.95	0.60	2.12	1.56, 2.89
	0.90	0.90	2.97	1.34, 6.56
	0.90	0.80	3.03	1.32, 6.94
	0.90	0.60	3.07	1.54, 6.10

RR, Rate Ratio; CI, Confidence Interval

[a] Adjusted for sex, race, age, and year of study entry

Table 5.3 Rate ratio of lung cancer mortality per 100 fiber-years/mL cumulative asbestos exposure, South Carolina, United States, 1940 – 2001, under several independent prior distributions reflecting beliefs about outcome misclassification

Model	Specificity Details	Sensitivity Details	RR	95% CI/PI
Crude	1	1	3.52	2.86, 4.33
Adjusted [a]	1	1	1.94	1.55, 2.44
Consensus prior	Uniform (0.9, 1.0)	Uniform (0.75, 0.95)	2.04	1.61, 2.56 [b]
Alternate variance	Uniform(0.925, 0.975)	Uniform(0.8, 0.9)	2.14	1.69, 2.78 [b]
Alternate priors	Uniform (0.95, 1)	Uniform (0.75, 0.95)	1.96	1.53, 2.51 [b]
	Uniform (0.9, 1.0)	Uniform(0.6, 0.9)	2.05	1.58, 2.56 [b]

RR, Rate Ratio; CI, Confidence Interval; PI, Posterior Interval

[a] Adjusted for sex, race, age, and year of study entry

[b] Posterior intervals

6 Discussion

6.1 Summary

Effect estimates are likely to be biased when the outcome of interest is misclassified, except in special cases. Bias due to outcome misclassification is small when specificity is near 1 or the number of false positive outcomes is low. Many epidemiologists assume either perfect outcome measurement or perfect specificity and routinely disregard outcome misclassification in their analyses. However, we have seen that bias due outcome misclassification may occur in studies of both of the commonly-used types of outcomes explored in the preceding chapters. In this work, we proposed three tools to account for outcome misclassification in epidemiologic analyses: multiple imputation, modified maximum likelihood, and Bayesian methods.

We first examined the use of participant-reported symptoms as an indicator of the outcome of interest. Many studies use participant recall or diaries to track outcomes without invading privacy or requiring frequent contact between the participant and the study team. In the Recurrence Factors sub-study of the Herpetic Eye Disease Study, the sensitivity of participant-reported HSV recurrence was less than 50% and specificity was about 90%, resulting in substantial bias in the naïve odds ratio (or risk ratio).

We next examined the use of death certificates to identify deaths due to a cause of interest. Misattribution of cause of death information on death certificates can cause outcome misclassification. In the example from the South Carolina textile workers cohort, we suspected that lung cancer death was subject to mild outcome misclassification, which resulted in little bias in the naïve rate ratio. In sensitivity analysis, however, we saw that the naïve rate ratio would be substantially biased if specificity had been below 95%, which may occur for other causes of death.

Multiple imputation, modified maximum likelihood, and Bayesian methods provide tools to account for outcome misclassification in a variety of epidemiologic settings. In this work, we demonstrated that multiple imputation can be used to account for outcome misclassification in studies with an internal validation subgroup. Modified maximum likelihood allowed sensitivity analysis of effect estimates under known or hypothesized values of sensitivity and specificity. This modified maximum likelihood approach was extended to incorporate prior distributions that expressed uncertainty about the values of sensitivity and specificity using Bayesian analysis.

Both multiple imputation and the Bayesian analysis using the modified likelihood function accounted for the uncertainty inherent in outcome misclassification. Multiple imputation allowed this uncertainty to propagate through the analysis by redrawing coefficients for the relationship between the gold standard outcome and the error-prone outcome for each imputation. Uncertainty was handled in the Bayesian analysis through direct manipulation of the width of the uniform prior distributions for sensitivity and specificity.

6.1.1 Multiple imputation

We used multiple imputation to account for outcome misclassification in a study with an internal validation subgroup. In simulations, multiple imputation produced estimates of the odds ratio and risk ratio with little bias and appropriate confidence interval coverage under scenarios with varying degrees of differential and nondifferential outcome misclassification.

Multiple imputation was also used to account for outcome misclassification in the Herpetic Eye Disease study data. The naïve odds ratio comparing participant-reported HSV recurrence between treatment arms randomized to receive oral acyclovir and placebo was biased towards the null due to poor sensitivity and specificity of participant-reported HSV recurrence. Using multiple imputation to account for outcome misclassification produced an odds ratio with little bias. However, multiple imputation offered only slight gains in precision over analysis limited only to the internal validation subgroup.

Multiple imputation is useful in studies with internal validation data because it leverages information in the complete data to obtain more precise estimates than analysis limited to the validation subgroup. This advantage is seen most clearly when the validation subgroup is a small proportion of the entire study population and outcome misclassification is mild. In the data from the Herpetic Eye Disease Study multiple imputation offered only a slight advantage over limiting analysis to the validation subgroup even though the validation study was small. Because the observed outcome was a poor proxy for the gold standard outcome, the imputation model contained a high degree of uncertainty that propagated through to the variance of the final effect estimate.

Both multiple imputation and modified maximum likelihood can be used to estimate odds ratios and risk ratios when validation data are available. In this analysis, multiple imputation produced results similar to results using modified maximum likelihood to account for outcome misclassification, though results using the modified likelihood function were slightly more precise.

Any method using an internal validation subgroup to account for outcome misclassification must make two critical assumptions about the validation subgroup. The first is that the relationship between the observed, error-prone, outcome and the gold-standard outcome is the same in the validation subgroup and in the complete data. This transportability assumption could be violated if participants are not selected into the validation subgroup at random (or at random within strata of covariates). The second assumption is that the gold-standard outcome is measured without error. If the gold-standard outcome is, itself, subject to error, multiple imputation will produce results with residual bias that are misleadingly precise (72).

Unfortunately, internal validation data are often unavailable. In studies without an internal validation subgroup, sensitivity and specificity can be estimated from external validation subgroups or expert opinions about the misclassification probabilities. However, multiple imputation is not easily extended to account for outcome misclassification when only sensitivity and specificity are known and no information is available about the positive or negative predictive value. Because positive predictive value is a function of the prevalence of the outcome in addition to the sensitivity and specificity, it is rarely transportable from one population to another. In these situations, we turn to modified maximum likelihood and Bayesian methods.

6.1.2 Modified maximum likelihood and Bayesian analysis

Modified maximum likelihood and Bayesian analysis worked well to account for outcome misclassification in studies in which no internal validation subgroup was available. Here, we focused on the scenario in which the event indicator was subject to error but the times were thought to be correct, as motivated by cause of death from the National Death Index. In simulated scenarios with varying degrees of outcome misclassification, modified maximum likelihood produced rate ratios with little bias and appropriate coverage.

In the cohort of South Carolina textile workers exposed to asbestos, modified maximum likelihood was used to conduct a sensitivity analysis in which rate ratios were estimated under varying assumptions about the amount of misclassification of lung cancer mortality. Changes in sensitivity had little effect on the rate ratio, but changes in specificity produced dramatic shifts in the rate ratio.

To allow uncertainty about sensitivity and specificity, we placed informative prior distributions on these parameters. Based on existing validation studies, sensitivity was assumed to follow a uniform distribution bounded by 0.75 and 0.95. Specificity was assumed to be uniformly distributed between 0.9 and 1.0. Using these priors, the posterior estimate of the rate ratio was very similar to the rate ratio from the standard analysis. We would expect posterior rate ratios to differ more dramatically from the standard estimates of the rate ratio for outcomes subject to more severe misclassification.

While placing prior distributions on sensitivity and specificity and using Markov chain Monte Carlo to sample from the posterior distribution of the parameter of interest is

appealing, it has several limitations. First, Markov chain Monte Carlo is computationally intensive. Condensing the data into a dataset with one record per distinct covariate pattern reduced the time required for each run, but the program still required several hours to obtain posterior estimates. Second, sensitivity and specificity are weakly identifiable from the observed data at best (46). For this reason, almost all of the information about sensitivity and specificity is contained in the prior distributions, and convergence to a stable posterior distribution is difficult. In addition, posterior draws of sensitivity and specificity may not have fully explored the uniform distributions specified by the priors, resulting in underestimation of the variance and artificially tight posterior intervals.

6.2 Future directions

In chapters 3, 4, and 5, we described multiple imputation, modified maximum likelihood, and Bayesian methods to account for outcome misclassification. These methods worked well in the examples to confirm earlier findings; after accounting for misclassification acyclovir was shown to be more effective than placebo at preventing ocular HSV recurrences and asbestos was shown to increase the risk of lung cancer. The next step is to use these methods to account for outcome misclassification in emerging research areas. Use of these methods will improve inferences from epidemiologic studies, but may require extensions to be applied to specific types of analyses.

In chapter 5, we saw that outcome misclassification did not produce bias in effect estimates when specificity of the outcome measure was high. In the sensitivity analysis, estimates of the rate ratio for lung cancer death per 100 f-y/mL of asbestos exposure accounting for outcome misclassification were similar to estimates from the standard

model under the most likely value of specificity (98%). However, for outcome measures with lower specificity, bias due to outcome misclassification becomes more pronounced and may have an impact on epidemiologic findings.

For example, asbestos has been implicated in the elevated risk of mortality from coronary heart disease seen in cohorts of miners, mill workers, and shipyard workers (76–80). Unlike lung cancer, the specificity of coronary heart disease reported on death certificates is low (33,81).

In the Framingham heart study, 84% (635/758) of the deaths identified by a physician panel as deaths due to coronary heart disease were classified as a death due to coronary heart disease on the death certificate. Similarly, 84% of deaths designated as non-coronary heart disease deaths by the physician panel were classified as deaths due to causes other than coronary heart disease on the death certificate (33). A validation study conducted in the Atherosclerosis Risk in Communities (ARIC) study produced similar results. In the ARIC study, the sensitivity of death certificate classification of coronary heart disease death was 81%, and specificity was 72% (81).

Due to the high probability of outcome misclassification of coronary heart disease mortality, studies using coronary heart disease mortality reported on death certificates are likely to produce biased effect estimates. Future work could use the methods developed here to account for outcome misclassification in studies of coronary heart disease mortality. As a first step, we used the methods outline in chapter 5 to conduct a sensitivity analysis to examine the effect of possible outcome misclassification on the rate ratio of death due to coronary heart disease per 100 f-y/mL of asbestos exposure in the cohort of textile workers in South Carolina. Results are summarized in the appendix. Many outcomes

used in epidemiology are subject to low specificity and could benefit from similar sensitivity analyses.

In chapter 4, we saw that multiple imputation can be used to account for outcome misclassification in studies with validation data. We focused on accounting for outcome misclassification of a binary outcome in logistic and log-binomial regression, but multiple imputation could be extended to account for outcome misclassification in other types of models as well. Extending the multiple imputation approach to account for misclassification in the time-to-event setting presents several challenges. A time-to-event outcome has two components: a time and an event indicator. Multiple imputation can be extended with minor modification to account for error in either of these components if an internal validation subgroup is available.

For example, consider the situation in which the time from the origin to the event was potentially mismeasured. If an internal validation subgroup related the observed times to a gold-standard measure of the time-to-event and all participants were known to have experienced the event of interest, multiple imputation could use information in the validation subgroup to impute the time-to-event in the complete data. Unlike the binary outcomes discuss above, the relationship between the gold standard measure of time and the observed time would be modeled with a parametric accelerated failure time model, and this model would be used to impute the times-to-event in the complete data. Challenges in the imputation of the time to event arise when participants drop out of the study or experience competing events.

Alternatively, a situation could arise in which the times were known but the event indicator was potentially misclassified, as in chapter 5. If internal validation study relating

the observed outcome indicator to the gold-standard outcome indicator were available, multiple imputation could be used to impute the outcome in the complete data. Using multiple imputation to account for a misclassified event indicator in time-to-event data would differ from the application of multiple imputation in the Herpetic Eye Disease Study data only in the choice of analysis model. In the time-to-event setting, the analysis model would likely be a Cox proportional hazards model or parametric accelerated failure time model instead of a binary regression model.

In chapter 5, we placed informative prior distributions on sensitivity and specificity and used Markov chain Monte Carlo to sample from the posterior distribution of the rate ratio. However, the convergence of this procedure was difficult to achieve and highly dependent on the specification of the model and starting values for parameters. Future work using this approach should assess model convergence carefully before basing inference on results obtained from Markov chain Monte Carlo.

As an alternative, future work could place informative prior distributions on sensitivity and specificity using prior data records, as proposed by Greenland for bias analysis (72). In this approach, prior data records are appended to the dataset and combined with the existing data using missing data methods to compute posterior effect estimates. Encoding beliefs about the misclassification parameters using prior data records avoids the computational challenges of Markov chain Monte Carlo and allows more flexibility in the choice of prior.

6.3 Conclusions

This work has highlighted the potential for bias due to outcome misclassification and described three tools to account for misclassification in a variety of epidemiologic settings. Use of such techniques in concert with validation data or expert knowledge and appropriate methods to account for confounding and selection bias will improve inference from epidemiologic studies.

Appendix 1: Direct maximum likelihood to account for outcome misclassification

We compare the proposed multiple imputation approach to account for outcome misclassification to a direct maximum likelihood approach outlined in Carroll et al (16) and detailed by Lyles et al (47). The direct maximum likelihood approach for a main study with a validation subgroup specifies the likelihood for the logistic regression model relating exposure to outcome as the product of the likelihood for the main study and the likelihood for the validation subgroup. In both likelihood terms, sensitivity and specificity are based on associations between observed outcome, gold standard outcome, and exposure defined using a logistic model

$$\eta_d = \text{logit}[\Pr(W = 1|D = d, X = x)] = \theta_0 + \theta_1 d + \theta_2 X + \boldsymbol{\theta_3 Z}, \text{ for } d = 0,1.$$

Sensitivity and specificity are calculated as

$$\text{SE}_i = \Pr(W = 1|D = 1, X = x_i, \boldsymbol{Z} = \boldsymbol{z_i}) = \frac{\exp(\eta_{i1})}{1 + \exp(\eta_{i1})},$$

and

$$\text{SP}_i = \Pr(W = 0|D = 0, X = x_i, \boldsymbol{Z} = \boldsymbol{z_i}) = \frac{\exp(\eta_{i0})}{1 + \exp(\eta_{i0})}.$$

The likelihood for the main study is

$$L_m = \prod_{i=1}^{n_m} \{[(1 - \text{SP}_{\text{x}_i}) \times \Pr(D = 0|X = x_{i,}\boldsymbol{Z} = \boldsymbol{z_i})$$
$$+ \text{SE}_{\text{x}_i} \times \Pr(D = 1|X = x_{i,}, \boldsymbol{Z} = \boldsymbol{z_i})]^{w_i} \times [\text{SP}_{\text{x}_i} \times \Pr(D = 0|X = x_i, \boldsymbol{Z} = \boldsymbol{z_i})$$
$$+ (1 - \text{SE}_{\text{x}_i}) \times \Pr(D = 1|X = x_i, \boldsymbol{Z} = \boldsymbol{z_i})]^{(1-w_i)}\},$$

and the likelihood for the validation study is

$$L_v = \prod_{j=1}^{n_v} \{[\mathrm{SE}_{x_j} \times \Pr(D = 1 | X = x_j, \boldsymbol{Z} = \boldsymbol{z_j})]^{w_j d_j} \times [(1 - \mathrm{SP}_{x_j})$$

$$\times \Pr(D = 0 | X = x_j, \boldsymbol{Z} = \boldsymbol{z_j})]^{w_j (1 - d_j)} \times (1 - \mathrm{SE}_{x_j})$$

$$\times \Pr(D = 1 | X = x_j, \boldsymbol{Z} = \boldsymbol{z_j})]^{(1 - w_j) d_j} \times [\mathrm{SP}_{x_j}$$

$$\times \Pr(D = 0 | X = x_j, \boldsymbol{Z} = \boldsymbol{z_j})]^{(1 - w_j)(1 - d_j)} \},$$

where i indexes the n_m participants in the main study (but not the validation subgroup) (i=1, … , n_m) and j indexes the n_v participants in the validation subgroup (j=1, … , n_v). As in the main body of the paper, D is an indicator of the gold standard outcome status, W is the observed outcome status, X is the treatment group, and $\boldsymbol{Z}$ is the vector of covariates.

The direct maximum likelihood approach differed between the logistic model and log binomial model only in the choice of link function. In the logistic model,

$$\Pr(D = 1 | X = x, \boldsymbol{Z} = \boldsymbol{z})] = \frac{\exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z})}{1 + \exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z})} ,$$

And in the log binomial model,

$$\Pr(D = 1 | X = x, \boldsymbol{Z} = \boldsymbol{z})] = \exp(\beta_0 + \beta_1 X + \boldsymbol{\beta_2 Z}) .$$

Appendix 2: SAS code for multiple imputation to account for outcome misclassification

We adapted SAS code from Cole (4) to perform multiple imputation to account for outcome misclassification. Here, we present SAS code to account for outcome misclassification that is differential with respect to treatment group. We will illustrate use of Firth's correction to prevent separation of data points when modeling the relationship between gold standard and observed exposure in the validation subgroup in the imputation model. The SAS code below could be adapted to account for nondifferential outcome misclassification by removing the interaction between treatment group and observed outcome in the imputation model.

```
*step 1: Fit logistic regression model relating gold standard outcome to observed outcome
in validation subgroup;
data m; col1=.; col2=.; col3=.; col4=.; col5=.; col6=.;
data a; set a; wx=w*x;
proc logistic data=a descending covout outest=b(keep=_name_ intercept w x wx z1 z2)
        noprint;  where r=1;
        model d=w x wx z1 z2/firth;
data bb; set b; if  _name_="d"; bh0=intercept; bh1=w; bh2=x; bh3=wx; bh4=z1; bh5=z2;
        keep bh0-bh5;
data cov; set b; if _name_^="d"; keep intercept w x wx z1 z2 ;

*step 2: Sample coefficients for each imputation;
```

```
proc iml;
  use cov; read all into cov;          *variance-covariance matrix;
  use bb; read all into mu; mu=mu`;          *means;
  v=nrow(cov);                         *number of variables;
  n=40;                                *number of imputations;
  seed=222;
  l=t(root(cov));                      *cholesky root of cov matrix;
  z=normal(j(v,n,seed));               *generate nvars*samplesize normals;
  d=l*z;                          *premultiply by cholesky root;
  d=repeat(mu,1,n)+d;                  *add in the means;
  td=t(d);
  create m from td;                    *write out sample data to sas dataset;
  append from td;
quit;
data m; set m;  retain _imputation_ 0; _imputation_=_imputation_+1;
   b0=col1; b1=col2; b2=col3; b3=col4; b4=col5; b5=col6;
   keep _imputation_ b0-b5 ;
data aa; merge d bb;  do _imputation_=1 to 40; output; end;
proc sort data=m; by  _imputation_;
proc sort data=aa; by  _imputation_;

*step 3: impute outcome for records not in the validation subgroup;
data c; merge aa m; by  _imputation_; call streaminit(9);
```

```
   if r=1 then d_imp=d;
   else  d_imp=rand("bernoulli",1/(1+exp(-(b0+b1*w+b2*x+b3*wx+b4*z1+b5*z2))));


*step 4a: Fit Logistic analysis model in each imputation;
proc logistic data=c outest=e covout noprint desc; by _imputation_;model d_imp=x z1 z2;
*step 4b: Fit Binomial analysis model in each imputation;
proc genmod data=c desc;
model d_imp=x z1 z2 / link=log dist=bin wald type3;
by _imputation_;
ods output parameterestimates=f;
run;
*step 4c: Summarize Logistic results over all imputations;
proc mianalyze data=e;  modeleffects x;
   title " multiple imputation to account for outcome misclassification";
run;


*step 4d: Summarize Binomial results over all imputations;
data f;
        set f;
        if parameter="x";
proc mianalyze data=f;
        modeleffects estimate;
        stderr stderr;
        ods output parameterestimates=mi3(keep= parm estimate stderr);
run;
```

Appendix 3: Monte Carlo Simulation Methods for Chapter 4

For each of 15 simulated scenarios, records were generated with values for treatment group (x), true outcome status (d), observed outcome status (w), and an indicator of inclusion in the validation subgroup (r). Half of the simulated records were assigned to each treatment group. True outcomes were simulated based on a Bernoulli distribution with the probability of being a case generated from a logistic regression model with the β coefficient for acyclovir equal to -0.693 (a true odds ratio of 0.5). Error-prone outcomes were generated based on hypothetical values for the accuracy of the outcome measure; one set of simulations generated observed outcome data with nondifferential misclassification with sensitivity set to be 0.90, 0.60, or 0.30, and the other assumed that outcome classification was more sensitive in exposed participants (sensitivity=0.95 and 0.70 for the two scenarios, respectively) than unexposed participants (sensitivity=0.85 and 0.50). Specificity was 0.90 in all scenarios.

For each scenario, a designated proportion was chosen to be included in the validation subgroup. For each record, r was sampled from a Bernoulli distribution with probability equal to the proportion included in the validation subgroup. As in the real data example, the true outcome was assumed to be known only for records where $r = 1$.

Each scenario was simulated 10,000 times. Bias was defined as 100 times the difference between the average estimated log odds ratio and the true log odds ratio. Confidence interval coverage was calculated as the percentage of simulations in which the estimated Wald-type confidence limits included the true value. Bias and precision were considered together using mean squared error, which was calculated as the sum of the

square of the bias and the variance. Statistical power was calculated as the percentage of simulations in which the Wald-type confidence interval excluded the null value.

Appendix 4: Computer programs for Chapter 5

A. SAS code to account for outcome misclassification in a Poisson model using direct maximum likelihood and known values of sensitivity and specificity

The SAS code below is applied to the dataset with the 120,010 person-years summarized into 4183 strata of distinct covariate patterns. In the code below, lam = λ_j, mu = μ_j, wj = w_j, dj = d_j, and nj = n_j. In this example, sensitivity is assumed to be 85% and specificity is assumed to be 95%.

```
%let se=0.85; %let sp=0.95;
title "rate ratio accounting for outcome misclassification (se=&se, sp=&sp)";
proc nlmixed data=tabled gconv=1e-15 fconv=1e-15;
        parms b0=-5 b1=.5 a0=-5 a1=0 b2=0 b3=0  b4=0 a2=0 a3=0  a4=0;
        se=&se; sp=&sp;
        lam=exp(b0+b1*asbestos+b2*sex+b3*log(age)+b4*year);
        mu=exp(a0+a1*asbestos+a2*sex+a3*age+a4*year);
        lik=(lam*se+mu*(1-sp))**(wj)*( lam*(1-se)+mu*sp)**(dj-wj)*exp(-(lam*se+mu*(1-
sp)+ lam*(1-se)+mu*sp)*nj);
        model nj~general(log(lik));
run;
```

B. SAS code to account for outcome misclassification in a Poisson model by placing prior distributions on sensitivity and specificity

The SAS code below places informative prior distributions on sensitivity and specificity. In this example, sensitivity is given a uniform prior distribution from 0.75 to 0.95, and specificity is given a uniform prior distribution from 0.9 to 1.

```
proc mcmc data=table ntu=1000 nmc=500000 nbi=50000 nthin=3 seed=215  outpost=p1 ;
        parms b0=-8 b1=1 a0=-3 a1=0 b2=-1 b3=2  b4=0 a2=-1 a3=0  a4=0;
        parms se=.9 sp=.97;
        prior b0 b1 b2 b3 b4 a0 a1 a2 a3 a4~normal(0,var=100);
        prior se~uniform(0.75, 0.95);
        prior sp~uniform(0.9, 1);
        lam=exp(b0+b1*asbestos+b2*sex+b3*log(age)+b4*year);
        mu=exp(a0+a1*asbestos+a2*sex+a3*age+a4*year);
        lik=(lam*se+mu*(1-sp))**(wj)*( lam*(1-se)+mu*sp)**(dj-wj)*exp(-(lam*se+mu*(1-
sp)+ lam*(1-se)+mu*sp)*nj);
        model nj~general(log(lik));
        ods select PostSummaries PostIntervals;
title "priors se=0.75 to 95,  sp=0.9 to 1";
run; quit; run;
```

Appendix 5: Monte Carlo simulations for Chapter 5

A. Design

We used simulation to explore the finite sample properties of using the modified maximum likelihood estimates to account for outcome misclassification. Simulations were performed with sensitivity and specificity assumed to be known. The simulations were intended to mimic the data from the cohort of textile workers exposed to asbestos in South Carolina. Let *i* index simulated participants in each strata of distinct covariate patterns (i =1, … ,n_j), where n_j is the number of participants in strata *j,* and *X* represent exposure ranging from 0 to 500 (mean = 45, standard deviation = 28). The time to death due to lung cancer (R) and time death due to other causes (S) followed exponential distributions with means determined by the exposure value. In expectation, a 100-unit increase in exposure decreased the time to lung cancer (R) by one-half and the time to non-lung cancer death (S) by one-third. The total time (T) contributed by each record was the minimum of *R* and *S.*

Cause of death was represented by δ. If death due to lung cancer occurred before death due to other causes would have occurred ($R < S$) then δ was set to 1. Otherwise, if death due to other causes occurred before death due to lung cancer ($S < R$), then δ was set to 2. Simulated participants were censored after 5 years; for participants with $T > 5$, δ was set to 0.

Error-prone cause of death indicator δ^* was generated based on δ and values of sensitivity and specificity. We simulated five possible scenarios with varying degrees of outcome misclassification: 1) both sensitivity and specificity set to 1; 2) specificity set to 0.95 and sensitivity set to 0.9; 3) specificity set to 0.95 and sensitivity set to 0.6; 4) both

sensitivity and specificity set to 0.9; and 5) specificity set to 0.9 and sensitivity set to 0.6. In all scenarios, outcome misclassification was nondifferential with respect to exposure and other measured covariates. For each scenario, δ^* was sampled from a Bernoulli distribution with probability determined by sensitivity and specificity. Where δ = 1, the probability that δ^*= 1 was equal to the value of sensitivity; where δ = 2, the probability that δ^*= 1 was equal to 1 – specificity. If $\delta^*$was not drawn to be 1 (lung cancer death), $\delta^*$was set to 2 (other death). If T > 5 years, then δ^* was set to 0.

Each scenario was simulated 10,000 times. For each simulated cohort, we summarized the data into J strata of distinct covariate patterns following the same categorization used for the actual data and calculated two counts for each strata: y_j, the sum of all actual lung cancer deaths in each strata, $\sum_{i=1}^{n_j} I(\delta_i = 1)$, and w_j, the sum of all reported lung cancer deaths in each strata, $\sum_{i=1}^{n_j} I(\delta_i^* = 1)$. We used Poisson regression to estimate the rate ratio of the lung cancer death per 100-unit increase in exposure. We estimated the true rate ratio (using y_j as the count of lung cancer deaths) and the naïve rate ratio (using w_j as the count of lung cancer deaths) with standard methods. We then compared these results to results using the method described above using modified maximum likelihood to account for outcome misclassification by setting values of sensitivity and specificity.

We evaluated the performance of this method to account for outcome misclassification by comparing bias and 95% confidence interval coverage between the standard analysis using w_j as the count of lung cancer deaths and the analysis using modified maximum likelihood to set values of sensitivity and specificity. Bias was defined as 100 times the difference between the average estimated log rate ratio and true log rate ratio, and confidence interval coverage was calculated as the proportion of simulations in

which the estimated Wald-type confidence limits included the true value. The bias-precision tradeoff was considered through examination of the mean-squared error, which was the sum of the square of the bias and the square of the standard deviation of the bias.

B. Results

Appendix table 1 compares the performance of the standard method and the modified maximum likelihood estimate to account for misclassification in the rate ratio for 10,000 simulated cohorts under several scenarios of outcome misclassification. As expected, the standard estimates of the rate ratio were biased towards the null when sensitivity and specificity were imperfect and bias increased as the degree of outcome misclassification increased. In contrast, revised estimates accounting for sensitivity and specificity using modified maximum likelihood showed little bias, even when sensitivity and specificity were quite low. The confidence limits from the revised estimates showed appropriate coverage in all scenarios examined.

Mean squared error was improved for the revised estimates when compared to the standard estimates under all combinations of sensitivity and specificity. The difference in mean squared error between the standard and revised estimates was small in scenario 2, where sensitivity was 0.9 and specificity was 0.95, because the inflated standard error of the revised estimate offset the small bias in the standard estimate. However, as sensitivity and specificity decreased, the difference in mean squared error became more pronounced. In the scenario with the most extreme outcome misclassification (sensitivity of 0.6 and specificity of 0.9), the bias in the standard estimate overwhelmed the increase in standard

error of the revised estimate, resulting in a large improvement in mean squared error for the revised estimate when compared to the standard estimate.

Table A.1 Results accounting for outcome misclassification using Poisson regression in 10,000 simulated cohorts [a]

Scenario	Method	Rate ratio	Bias [b]	95% CI Coverage [c]	Mean squared error [d]
1. Specificity = 1, Sensitivity = 1	Truth	2.00	0	95	0.77
2. Specificity = 0.95, Sensitivity = 0.9	Standard	1.89	-5	91	1.15
	Revised	2.00	0	95	0.95
3. Specificity = 0.95, Sensitivity = 0.6	Standard	1.84	-8	89	1.96
	Revised	2.01	0	95	1.30
4. Specificity = 0.9, Sensitivity = 0.9	Standard	1.80	-10	79	1.93
	Revised	2.01	0	95	0.96
5. Specificity = 0.9, Sensitivity = 0.6	Standard	1.72	-15	72	3.55
	Revised	2.01	1	95	1.46

[a] The models accounting for imperfect sensitivity and specificity did not converge in 6, 7, 9, and 5 simulated cohorts for scenarios 2,3, 4, and 5, respectively.

[b] Bias was defined as 100 times the difference between the true ln(rate ratio) and the estimated ln(rate ratio)

[c] 95% confidence interval coverage was the proportion of iterations in which the estimated 95% confidence interval contained the true value

[d] Mean squared error was the sum of the square of the bias and the square of the standard deviation of the bias

Table A.2 Comparison of average standard errors and standard deviations of point estimates from 10,000 simulated cohorts [a]

Scenario	Method	Mean β	Mean standard error	Standard deviation of β
1. Specificity = 1, Sensitivity = 1	Truth	0.695	0.087	0.088
2. Specificity = 0.95, Sensitivity = 0.9	Naïve	0.639	0.091	0.092
	Revised	0.691	0.094	0.095
3. Specificity = 0.95, Sensitivity = 0.6	Naïve	0.612	0.110	0.114
	Revised	0.698	0.110	0.114
4. Specificity = 0.9, Sensitivity = 0.9	Naïve	0.588	0.089	0.091
	Revised	0.696	0.097	0.098
5. Specificity = 0.9, Sensitivity = 0.6	Naïve	0.542	0.108	0.112
	Revised	0.700	0.115	0.121

[a] The models accounting for imperfect sensitivity and specificity did not converge in 6, 7, 9, and 5 simulated cohorts for scenarios 2,3, 4, and 5, respectively.

Appendix 6: Sensitivity analysis of rate ratios of coronary heart disease per 100 f-y/mL asbestos exposure under misclassification scenarios

Table A.3 Rate ratios of mortality due to coronary heart disease per 100 fiber-years/mL cumulative asbestos exposure, South Carolina, United States, 1940 – 2001, under several outcome misclassification scenarios

			Coronary Heart Disease	
Model	Specificity	Sensitivity	RR	95% CI
Crude	1	1	2.60	2.14, 3.16
Adjusted [a]	1	1	1.37	1.10, 1.70
	0.98	0.90	1.38	1.09, 1.75
	0.98	0.80	1.38	1.09, 1.75
	0.98	0.60	1.38	1.09, 1.74
	0.95	0.90	1.41	1.07, 1.85
	0.95	0.85	1.41	1.07, 1.85
	0.95	0.80	1.41	1.07, 1.85
	0.95	0.60	1.40	1.07, 1.84
	0.90	0.90	1.50	1.02, 2.20
	0.90	0.80	1.50	1.02, 2.19
	0.90	0.60	1.48	1.01, 2.17
	0.85	0.90	1.91	1.07, 3.42
	0.85	0.80	1.92	1.08, 3.41
	0.85	0.60	1.91	1.14, 3.22

RR, Rate Ratio; CI, Confidence Interval

[a] Adjusted for sex, race, age, and year of study entry

References

1. Kaplan G a. How big is big enough for epidemiology? *Epidemiology (Cambridge, Mass.)*. 2007;18(1):18–20.

2. Samet JM. Data: to share or not to share? *Epidemiology (Cambridge, Mass.)*. 2009;20(2):172–4.

3. Spiegelman D. Approaches to uncertainty in exposure assessment in environmental epidemiology. *Annual Review of Public Health*. 2010;31:149–63.

4. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*. 2006;35(4):1074–1081.

5. Zeger SL, Thomas D, Dominici F, et al. Exposure Measurement Error in Time-Series Studies of Air Pollution : Concepts and Consequences. *Environmental Health*. 2000;108(5).

6. Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine*. 2008;27(30):6332–6350.

7. Stram DO, Langholz B, Huberman M, et al. Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau Uranium Miners cohort. *Health physics*. 1999;77(3):265–75.

8. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. *Annual Review of Public Health*. 1993;14:69–93.

9. Ogburn EL, VanderWeele TJ. Analytic results on the bias due to nondifferential misclassification of a binary mediator. *American journal of epidemiology*. 2012;176(6):555–61.

10. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International journal of epidemiology*. 2005;34(6):1370–6.

11. Clegg LX, Feuer EJ, Midthune DN, et al. Impact of reporting delay and reporting error on cancer incidence rates and trends. *Journal of the National Cancer Institute*. 2002;94(20):1537–45.

12. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am. J. Epidemiol.* 1990;132(4):746–748.

13. Chavance M, Dellatolas G, Lellouch J. Correlated nondifferential misclassifications of disease and exposure: application to a cross-sectional study of the relation between handedness and immune disorders. *International journal of epidemiology*. 1992;21(3):537–46.

14. Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology (Cambridge, Mass.)*. 1992;3(3):210–5.

15. Weinberg CA, Umbach DM, Greenland S. When Will Nondifferential Misclassification of an Exposure Preserve the Direction of a Trend? *American Journal of Epidemiology*. 1994;140(6):565–571.

16. Carroll RJ, Ruppert D, Stefanski LA, et al. Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition. London, UK: Chapman and Hall/CRC; 2006.

17. Townshend AP, Chen CM, Williams HC. How prominent are patient-reported outcomes in clinical trials of dermatological treatments? *BRITISH JOURNAL OF DERMATOLOGY*. 2008;(5):1152–1159.

18. Denslow S, Edwards J, Horney J, et al. Improvements to water purification and sanitation infrastructure may reduce the diarrheal burden in a marginalized and flood prone population in remote Nicaragua. *BMC International Health and Human Rights*. 2010;10(30).

19. Hunter PR, Chalmers RM, Hughes S, et al. Self-reported diarrhea in a control group: a strong association with reporting of low-pressure events in tap water. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. 2005;40(4):e32–4.

20. Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Controlled clinical trials*. 2004;25(6):535–52.

21. Coughlin SS. Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology*. 1990;43(1):87–91.

22. Herpetic Eye Disease Study Group. Acyclovir for the prevention of recurrent herpes simplex virus eye disease. *The New England Journal of Medicine*. 1998;339(5):300–306.

23. Farooq A V, Shukla D. Herpes simplex epithelial and stromal keratitis: an epidemiologic update. *Survey of ophthalmology*. 2012;57(5):448–62.

24. Lairson DR, Begley CE, Reynolds TF, et al. Prevention of Herpes Simplex Virus Eye Disease. 2013;121:108–112.

25. Herpetic Eye Disease Study Group. Psychological stress and other potential triggers for recurrences of herpes simplex virus eye infections. *Archives of Ophthalmology*. 2000;118(12):1617–1625.

26. Lemmens P, Knibbe R, Tan F. Weekly Recall and Diary Estimates of Alcohol Consumption in a General Population Survey. *Journal of Studies on Alcohol and Drugs*. 1988;49(02):131.

27. Heeb J-L, Gmel G. Measuring alcohol consumption: a comparison of graduated frequency, quantity frequency, and weekly recall diary methods in a general population survey. *Addictive behaviors*. 2005;30(3):403–13.

28. Totterdell P, Wood S, Wall T. An intra-individual test of the demands-control model: A weekly diary study of psychological strain in portfolio workers. *Journal of Occupational and Organizational Psychology*. 2006;79(1):63–84.

29. Fukuoka M, Yano S, Giaccone G, et al. Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial) [corrected]. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2003;21(12):2237–46.

30. Hiller L, Dunn JA, Higgins HB, et al. Optimising patient recall of adverse events over prolonged time periods. *Trials*. 2011;12(Suppl 1):A74.

31. Bailar JC, Gornik HL. Cancer Undefeated. *The New England Journal of Medicine*. 2007;336(22):1569–1574.

32. Messite J. Accuracy of Death Certificate Completion. The need for formalized physician training. *Journal of the American Medical Association*. 1996;275(10):794–796.

33. Lloyd-Jones DM, Martin DO, Larson MG, et al. Accuracy of death certificates for coding coronary heart disease as the cause of death. *Annals of internal medicine*. 1998;129(12):1020–6.

34. Doria-Rose VP, Marcus PM. Death certificates provide an adequate source of cause of death information when evaluating lung cancer mortality: an example from the Mayo Lung Project. *Lung cancer (Amsterdam, Netherlands)*. 2009;63(2):295–300.

35. Modelmog D, Rahlenbeck S, Trichopoulos D. Accuracy of death certificates: a population-based, complete-coverage, one-year autopsy study in East Germany. *Cancer causes & control : CCC*. 1992;3(6):541–6.

36. Hoel DG, Ron E, Carter R, et al. Influence of death certificate errors on cancer mortality trends. *Journal of the National Cancer Institute*. 1993;85(13):1063–8.

37. Selikoff IJ, Seidman H. Use of death certificates in epidemiological studies, including occupational hazards: variations in discordance of different asbestos-associated diseases on best evidence ascertainment. *American journal of industrial medicine*. 1992;22(4):481–92.

38. Kamp DW. Asbestos-induced lung diseases: an update. *Translational research : the journal of laboratory and clinical medicine*. 2009;153(4):143–52.

39. Murphy RLH. The Diagnosis of Nonmalignant Diseases Related to Asbestos. *American Journal of Respiratory and Critical Care Medicine*. 1987;136(6):1516–1517.

40. Bang KM, Mazurek JM, Storey E, et al. Malignant mesothelioma mortality - United States, 1999-2005. *Morbidity and Mortality Weekly Report*. 2009;58(15):393–396.

41. Department of Heath and Human Services, CDC, NIOSH. Current Intelligence Bulletin 62: Asbestos Fibers and Other Elongate Mineral Particles: State of the Science and Roadmap for Research. 2011.

42. Dement JM, Harris RL, Symons MJ, et al. Exposures and mortality among chrysotile asbestos workers. Part II: mortality. *American journal of industrial medicine*. 1983;4(3):421–33.

43. Liu G, Beri R, Mueller A, et al. Molecular mechanisms of asbestos-induced lung epithelial cell apoptosis. *Chemico-biological interactions*. 2010;188(2):309–18.

44. Bross I. Misclassification in 2 X 2 Tables. *Biometrics*. 1954;10(4):478–486.

45. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology (Cambridge, Mass.)*. 2003;14(4):451–8.

46. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*. 1997;146(2):195–203.

47. Lyles RH, Tang L, Superak HM, et al. Validation Data-based Adjustments for Outcome Misclassification in Logistic Regression: An Illustration. *Epidemiology*. 2011;22(4):589–597.

48. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1977;39(1):1–38.

49. Maudsley G, Williams EM. "Inaccuracy' in death certification--where are we now? *Journal of public health medicine*. 1996;18(1):59–66.

50. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: Wiley; 1987.

51. Little RJA, Rubin DB. Statistical Analysis with Missing Data, Second Edition. New York, NY: Wiley-Interscience; 2 edition; 2002.

52. Allison PD. Missing Data (Quantitative Applications in the Social Sciences). Thousand Oaks, CA: Sage Publications, Inc; 2001.

53. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581–592.

54. Firth D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika*. 1993;80(1):27–38.

55. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine*. 2002;21(16):2409–2419.

56. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1984;71(1):1–10.

57. Bryson MC, Johnson ME. The Incidence of Monotone Likelihood in the Cox Model. *Technometrics*. 1981;23(4):381–383.

58. Dement JM, Harris RL, Symons MJ, et al. Exposures and mortality among chrysotile asbestos workers. Part I: exposure estimates. *American journal of industrial medicine*. 1983;4(3):399–419.

59. Hein MJ, Stayner LT, Lehman E, et al. Follow-up study of chrysotile textile workers: cohort mortality and exposure-response. *Occupational and environmental medicine*. 2007;64(9):616–25.

60. Herpetic Eye Disease Study Group. Oral Acyclovir for Herpes Simplex Virus Eye Disease: Effect on Prevention of Epithelial Keratitis and Stromal Keratitis. *Archives of Ophthalmology*. 2000;118(8):1030–1036.

61. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*. 1987;125(5):761–768.

62. Lee J, Chia K. Estimation of prevalence rate ratios for cross sectoinal data: an example in occupational epidemiology. *British Journal of Industrial Medicine*. 1993;50(9):861–864.

63. Axelson O, Fredriksson M, Ekberg K. Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies. *Occupational and Environmental Medicine*. 1994;51(8):574–574.

64. Stuart EA, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American journal of epidemiology*. 2009;169(9):1133–1139.

65. Janssen KJM, Donders ART, Harrell FE, et al. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*. 2010;63(7):721–727.

66. Clayton D, Spiegelhalter D, Dunn G, et al. Analysis of longitudinal binary data from multi- phase sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 1998;60(1):71–87.

67. Deddens JA, Petersen MR. Approaches for estimating prevalence ratios. *Occupational and Environmental Medicine*. 2008;65(7):481, 501–506.

68. Petersen MR, Deddens JA. A revised SAS macro for maximum likelihood estimation of prevalence ratios using the COPY method (Letter). *Occupational and environmental medicine*. 2009;66(9):639.

69. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology*. 2004;159(7):702–706.

70. Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. *Epidemiology*. 2010;21(6):855–862.

71. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). New York, NY: Springer; 2009.

72. Greenland S. Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. *International journal of epidemiology*. 2009;38(6):1662–1673.

73. Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American journal of public health*. 1981;71(3):242–50.

74. Gobbato F, Vecchiet F, Barbierato D, et al. Inaccuracy of death certificate diagnoses in malignancy: an analysis of 1,405 autopsied cases. *Human pathology*. 1982;13(11):1036–8.

75. Cameron HM, Mcgoogan E. A prospective study of 1152 hospital autopsies: Analysis of inaccuracies in clinical diagnoses and their significance. 1981;133(July 1980):285–300.

76. Sjögren B. Mortality among British asbestos workers. *Occupational and environmental medicine*. 2009;66(12):854–5.

77. Enterline PE, Hartley J, Henderson V. Asbestos and cancer: a cohort followed up to death. *British journal of industrial medicine*. 1987;44(6):396–401.

78. McDonald JC, Liddell FD, Dufresne A, et al. The 1891-1920 birth cohort of Quebec chrysotile miners and millers: mortality 1976-88. *British journal of industrial medicine*. 1993;50(12):1073–81.

79. De Klerk NH, Musk AW, Cookson WO, et al. Radiographic abnormalities and mortality in subjects with exposure to crocidolite. *British journal of industrial medicine*. 1993;50(10):902–6.

80. Peto J, Doll R, Hermon C, et al. Relationship of mortality to measures of environmental asbestos pollution in an asbestos textile factory. *The Annals of occupational hygiene*. 1985;29(3):305–55.

81. Coady SA, Sorlie PD, Cooper LS, et al. Validation of death certificate diagnosis for coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study. *Journal of clinical epidemiology*. 2001;54(1):40–50.

82. Edwards JK, Cole SR, Troester MA, et al. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *American Journal of Epidemiology*. (10.1093/aje/kws340 in press).

83. Sposto R, Preston DL, Shimizu Y, et al. The Effect of Diagnostic Misclassification on Non-Cancer and Cancer in A-Bomb Mortality Dose Response Survivors. 1992;48(2):605–617.

84. Stamey JD, Young DM, Jr JWS. A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression. 2008;(October 2007):2440–2452.

85. Chu H, Wang Z, Cole SR, et al. Sensitivity analysis of misclassification: a graphical and a Bayesian approach. *Annals of Epidemiology*. 2006;16(11):834–841.

86. Gustafson P, Greenland S. Curious phenomena in Bayesian adjustment for exposure misclassification. *Statistics in medicine*. 2006;25(1):87–103.