Some Effects of Exposure Misclassification on Epidemiological Studies

By

Jun Zou, MSc

A dissertation submitted to the

School of Public Health and the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Public Health

Written under the direction of

George Rhoads, MD, MPH

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2017

ABSTRACT OF THE DISSERTATION

Some Effects of Exposure Misclassification on Epidemiological Studies

By JUN ZOU

Dissertation Director: George Rhoads

In many epidemiological studies the risk factor or exposure of interest is measured with significant error. In well-designed studies this error is non-differential with respect to the outcome, but it nevertheless makes it more difficult to detect associations and it biases estimates of effect toward the null. Less well recognized is that it increases the probability that a significant result, when found, will be a false positive. This is obvious if one considers the extreme example where the observed measure bears little association with the true value and is essentially random, in which case any significant result would have to be an alpha error.

The traditional error model is not realistic in the presence of substantial error because with a fixed observed variance and a large error variance, the parameter variance is constrained. We propose a bivariate normal model,

which makes fewer assumptions than the traditional model and does not constrain the underlying "true" variance. The model implies the need for larger sample sizes to assure that an effect associated with a misclassified variable is sufficiently unlikely to have occurred by chance that it implies the underlying true variable also shows the effect.

A minimal estimate of misclassification can be obtained from the correlation between repeated measurements. When this correlation is low it implies a low correlation of the measurement with the true value and the need for large sample size increases that may make the use of such variables impractical. We use data from the Honolulu Heart Program, a large prospective study of cardiovascular disease to show that risk factors for heart attacks that have stood the test of time mostly are repeatable across a two-years time span with correlations exceeding 0.7. Other risk factors such as diet and physical activity that are believed to cause heart attacks but have been difficult to demonstrate within homogeneous populations have substantially lower repeatability correlations. These considerations emphasize the importance of good measurement of exposure in epidemiological studies.

Acknowledgements

First sincere appreciation to Dr. Rhoads for his patience and his guidance, which inspires and motivates me always.

I would like to acknowledge the valuable input from Dr. Lin, his knowledge and expertise are proved to be essential for the completion of the content of chapter 3.

I am also grateful to Dr. Masaki for her generous support even though I might not get a chance to meet her in person.

I would like express my gratitude to Dr. Demissie, the Chair with picture on SPH website, not just as my committee member, but as mentor who taught me almost everything during my epidemiology journey.

My deepest appreciation belongs to my wife Xiaofei and my son Calvin. Without their understanding and support, this would not have been possible.

Table of Contents

List of tables

Chapter 1

Introduction

Many epidemiological and clinical studies seek to relate a risk factor or treatment to a health outcome. In these kinds of studies either the risk factor (exposure) or the health outcome (disease) may be measured with error. In this thesis I will focus on error in exposure measurement, which can usefully be divided into "differential misclassification" or "non-differential misclassification" with respect to the outcome of interest. It is obvious that differential misclassification, where persons sustaining the outcome are assessed differently from persons who do not sustain the outcome could either exaggerate the effect of the exposure or obscure its effect depending on the direction and magnitude of the differential measurements. This kind of bias is difficult to measure and cannot usually be adjusted by data analysis. Rather, it results from flawed study methods and is best addressed by designing and implementing studies in such a way that the exposure of interest is measured with identical procedures and techniques in all subjects. In this thesis, differential misclassification will not be covered.

While differential misclassification can usually be remedied by excellent study methods, nearly all observational studies will nevertheless include some non-differential measurement error. This reflects the imperfect measurement methods that are available for most clinical and epidemiological variables. Attributes such as height can be measured very well; others such as blood pressure are obviously subject to error but nevertheless are good enough to be related to important disease outcomes; still other variables, including many nutritional and environmental exposures in free living people are measured with still greater error. It is likely that the continuing controversy surrounding the effects of many nutritional and environmental

exposures on chronic diseases is partly due to inability to accurately assess the relevant exposures in free-living people.

In most circumstances non-differential misclassification of exposure is thought to bias the results of the study toward the null. It is easy to appreciate that a true risk difference between exposed and non-exposed individuals will be blurred if some of the exposed individuals are put in the non-exposed group and vice versa. As a consequence, when a statistically significant result is reported for a poorly measured variable, the claim is sometimes made that the true effect is likely to be larger than the observed effect. In one sense these claims are logical but from another perspective they are paradoxical since they imply that when a statistically significant result is found (and such results are found in most published studies), it is potentially more important if the measurement is poor!

In Chapter 2, we will recapitulate and enlarge on a qualitative argument that Rhoads published previously [1] making the case that a) the paradox is explained because significant results (e.g. $p<0.05$) on poor measurements have a high probability of being false positives; b) that the traditional formulation for misclassification on a continuous variable, which is based on an additive, normally distributed error term, is not suitable for badly measured variables because it leaves little room for variation of the true parameter; and c) that a minimal estimate of the extent of misclassification can be developed by repeating the study measurements in a subsample of subjects.

In Chapter 3, we will consider misclassification of a continuous variable in a bivariate normal model that can accommodate extensive error without constraining the variance of the true parameter. We will develop estimates of stricter requirements for observed significance levels

that are needed to reduce the chance of a false positive result to acceptable levels and will describe related power and sample size implications.

In Chapter 4, we test our conclusion that badly measured variables fail to produce scientifically useful results. We examine a large number of variables measured at the Honolulu Heart Study and show that the traditional cardiovascular risk factors which have been demonstrated in study after study are reasonably repeatable within individuals over a two-year period whereas many other variables that might be expected to be predictive of cardiovascular disease, but are not as well measured, have not held up. This analysis also shows that for nearly all variables examined that Pearson and Spearman correlation coefficients were very similar, and it provides substantial information about the extent to which these correlations decline as the time between measurements increases.

In Chapter 5, we summarize an overall conclusion for the thesis.

Chapter 2

Effects of Non-Differential Exposure Misclassification

Many epidemiologic studies are compromised by unreliable measurements. The presence of measurement errors can cause biased and inconsistent parameter estimation and leads to unreliable and erroneous conclusions. Epidemiologists often find themselves assessing exposures which are difficult to measure but are nevertheless believed to have potentially important causal influences on disease. Examples are physical activity, nutritional variables, and individual exposures to some air pollutants. Since the true value of risk factors is usually unknown, most measures of risk factors are simply “approximations” or “surrogates” for some true underlying risk factor. The extent of these measurement errors is difficult to assess because there is usually no “gold standard” to which the measurements can be compared.

The erroneous classification of exposure will result in misclassification bias. Non-differential misclassification, defined as when all classes, groups, or categories of a variable (whether exposure, outcome, or covariate) have the same error rate or probability of being misclassified for all study subjects, will almost always result in underestimation of the strength of an association with the underlying true exposure, when one is present. Thus, it will bias the measure of effect

(e.g. relative risk, risk difference, odds ratio) toward the null and may result in the loss of statistical significance resulting in a Type II error. [1, 2, 3,4,5]

2.1 Non-differential measurement error as a cause of false positive conclusions

As a consequence of this well-known bias toward the null, when a statistically significant result is reported for a poorly measured variable, the claim is sometimes made that the true effect is likely to be larger than the observed effect. As noted in the Introduction, such claims, while supported by this known bias, are paradoxical since they imply that when a statistically significant result is found, it is potentially more important if the measurement is poor.

While it is true that non-differential misclassification biases toward the null, it is also true that as misclassification becomes extreme the observed values have less and less meaning, so that a significant difference, if found, is more and more likely to be a false positive result. In the extreme case, where the variable is a random number, a statistically significant difference between groups with and without a particular health condition would by definition have an alpha error probability of 1.0.

If a significant difference found for a random variable is certain to be an alpha error, whereas a significant difference for a perfectly measured variable can be interpreted at its nominal significance level, then there must be an escalation in the risk of false positive conclusions as one introduces increasing random misclassification into a measurement. This is the root of the paradox mentioned above, which arises because this escalation (and the misclassification that underlies it) is ignored in the usual power calculations and in usual frequentist statistical inference. A result that $p \leq 0.05$ is treated the same whether it is based on a well measured variable or a poorly measured variable, when, in fact a significant result on a very poorly measured variable has much greater than 5% chance of being a false positive. The difference between the standard

frequentist formulation and the argument put forward here arises because prior information about the extent of misclassification can often be obtained and should be taken into account.

2.2 Two Gaussian models of non-differential exposure misclassification.

The usual model that has been used to study the effects of non-differential exposure misclassification of a normally distributed variable has been to assume that the observed value (X) is equal to the sum of the normally distributed true value (T) plus an error term (e) that is uncorrelated with T and is normally distributed with a mean of 0. This results in

$$\mathrm{Var}(X) = \mathrm{Var}(T) + \mathrm{Var}(e).$$

In this model the mean (X) is an unbiased estimate of mean (T) because, on average, the mean error is 0. The model also implies that the variance of the true distribution is equal to the difference between the observed variance and error variance. If the error is large it implies, inappropriately, that the variance of the true variable is small. Moreover, in the presence of substantial error the assumption that it is unbiased will in many cases be untrue. Given the likelihood of publication bias in many areas of epidemiology, it is probable that the error term among published studies of poorly measured exposures is biased away from the null.

2.3 Non-differential measurement error and loss of statistical power

The power of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true (i.e. the probability of not committing a Type II error). The power is in general a function of the possible distributions, often determined by a parameter, under the alternative hypothesis. As the power increases, the chances of a Type II error occurring decrease. The probability of a Type II error occurring is referred to as the false negative rate (β). Therefore, power is equal to $1-\beta$, which is also known as the sensitivity.

Statistical power depends on the following:

- The statistical significance criterion used in the test
- The underlying variation in the population
- The magnitude of the effect of interest in the population
- The sample size used to detect the effect

A significance criterion is a statement of how unlikely a positive result would be, if the null hypothesis of no effect is true. The most commonly used criterion is a probability of 0.05. Under this criterion, the probability of the data implying an effect at least as large as the observed effect when the null hypothesis is true must be less than 0.05, for the null hypothesis of no effect to be rejected. One easy way to increase the power of a test is to carry out a less conservative test by using a larger significance criterion, for example 0.10 instead of 0.05. This increases the chance of rejecting the null hypothesis (i.e. obtaining a statistically significant result) when the null hypothesis is false, that is, reduces the risk of a Type II error (false negative regarding whether an effect exists). But it also increases the risk of obtaining a statistically significant result (i.e. rejecting the null hypothesis) when the null hypothesis is not false; that is, it increases the risk of a Type I error (false positive).[6]

For continuous variables the underlying variation in the population is usually assessed as the variance. The magnitude of the effect of interest in the population can be quantified in terms of an effect size, where there is greater power to detect larger effects. An effect size can be a direct estimate of the quantity of interest, or it can be a standardized measure that also accounts for the variability in the population. For example, in an analysis comparing outcomes in a treated and control population, the difference of outcome means A − B would be a direct measure of the

effect size, whereas (A − B)/σ where σ is the common standard deviation of the outcomes in the treated and control groups, would be a standardized effect size. If constructed appropriately, a standardized effect size, along with the sample size, and required significance level will completely determine the power.

The sample size determines the amount of sampling error inherent in a test result. Other things being equal, effects are harder to detect in smaller samples. Increasing sample size is often the most practical way to boost the statistical power of a test.[6,7]

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size under different assumptions with respect to the desired significance level and effect size to be detected.[7] Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size. In addition, the concept of power is used to make comparisons between different statistical testing procedures: for example, between a parametric and a nonparametric test of the same hypothesis.[8]

The above classical formulation takes no account of the extent of misclassification of the observed variable on the ability to detect differences in the true variable. Power to detect true underlying differences also depends on the magnitude of the correlation between the observed data and the true data (*T*), as well as the degree of collinearity with other variables in the model.[8,9,10]

2.4 Statistical challenge with badly measured variables

Badly measured variables will increase the probabilities of type 2 error. When significant results are found for such variables, misclassification also increases the chance that the result is a false positive. The traditional formulation for misclassification on a continuous variable, which is based on an additive, normally distributed error term, is not suitable for badly measured variables because, given an observed total variance, it leaves little room for variation of the true parameter.

Power analysis is appropriate when the objective is with the correct rejection, or not, of a null hypothesis. In many epidemiology studies, the concern is less about determining if there is or is not a difference but rather with getting a more refined estimate of the population effect size. For example, if there is the correlation of around .50, a sample size of 200 will give us approximately 80% power (alpha = .05, two-sided) to reject the null hypothesis of zero correlation. However, in doing this study we are probably more interested in knowing whether the correlation is .10 or .50 or .90. In this context we would need a much larger sample size in order to reduce the confidence interval of our estimate to a range that is acceptable for our objectives.[10,11,12]

Any statistical analysis involving multiple hypotheses is subject to inflation of the type I error rate if appropriate measures are not taken. Such measures typically involve applying a higher threshold of stringency to reject a hypothesis in order to compensate for the multiple comparisons being made (e.g. as in the Bonferroni method). In this situation, the power analysis should reflect the multiple testing approaches to be used. Thus, for example, a given study may be well powered to detect a certain effect size when only one test is to be made, but the power to detect the same effect size may be much lower if adjustment is made for several tests that are to be performed. Still another consideration is that when we control for the effect of correlated covariates, we will have more power for a fixed sample size, because the error variance (the

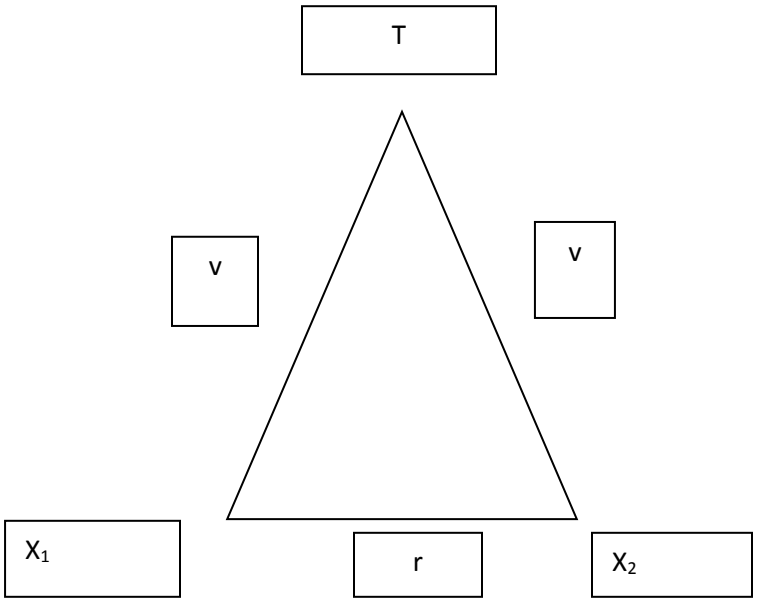
variance of the dependent variable scores after being adjusted for the covariate) will be reduced. The larger the correlation between the covariate and the dependent variable (or, with multiple covariates, the multiple R between covariates and the dependent variable), the greater the reduction of error variance will be.[11,12,13,14]

The investigator who is planning a study for which misclassification of exposure is known to be a problem may wish to assure readers or grant reviewers that the measurement error has been taken into account in planning the investigation. To do so the investigator will need to increase the sample size so that when the study is reported there will be adequate power (e.g. >=80%) to detect the attenuated odds ratio. This adjustment is small when misclassification is modest and relative risks of interest are 2.0 and greater. Unfortunately, when the odds ratio of interest is modest, say less than 1.5, and misclassification is substantial, the increase in sample size required may be prohibitive. In Chapter 3 we will explore these issues in more detail using a bivariate normal model.

2.5 Using Repeat Measurements to *Assess* Exposure Misclassification

In most epidemiological studies based on individual subjects it is assumed that there is an underlying parameter that best captures the effect of the risk factor that we are trying to assess on the outcome. For the intake of a nutrient, for instance, this would probably not be the amount eaten on any particular day but rather some average intake over a period of time that is relevant to the pathogenesis of outcome of interest. We call this the *true* value, which is represented in the figure below as *T.*[1] One or more observations, X_1 (X_2), are correlated with T, but are rarely exactly correct.

Multiple measurements can be made at different times (repeat measurements) as shown by X_1 and X_2. If a "gold standard" were available, it would measure *T* directly and the correlation, *v,* between *T* and X_1 (X_2) could be calculated. This correlation is of considerable interest because if it were perfect we could make an uncompromised assessment of the risk factor status of *T* by simply measuring X_1 (X_2). Unfortunately, a real gold standard generally is unavailable although



clues to the validity of a measurement may be provided indirectly by correlations with serum levels, other physiological parameters, or external information.

An additional approach to assessing the relationship between observed variables and the "true" parameter of interest is to use the consistency of repeated measurements. A convenient summary statistic for this repeatability is the correlation coefficient, *r,* between repeat assessments.

Of course, repeatability does not guarantee validity; but a method that does not yield reproducible results cannot be valid in the sense of providing trustworthy information. If the variables in question are normally distributed and if the errors of repeat assessment are independent of each other (the optimal situation), then *v* can be estimated as *square root of r.* Stated another way, under ideal conditions *r* provides a direct estimate of v^2, the proportion of the variance of the true value that is measured. However, some of the error associated with most methods of diet and other risk factor assessments is likely to repeat systematically each time the

method is applied to a given subject. In other words, the correlation between repeat values, *r,* is likely to be greater than v^2. Hence *r* can be taken as an upper limit of *v square* provided it is measured on a reasonably large sample of subjects.

The notion of *upper* limit should be emphasized. Correlated measurement errors probably are common and may be especially likely to occur when reproducibility becomes an objective in devising the method. If consistency of response is emphasized at the expense making each observation as accurate as possible, repeatability estimates are likely to exaggerate the true quality of the data. In devising measurement methods, validity should be the primary concern because if each measurement is a good estimate of the truth, then repeat estimates will be correlated mainly through their association with the "truth".[1]

2.6 Estimating the Minimal Extent of Misclassification

The usual absence of a "gold" standard has probably contributed to the tendency of investigators to ignore the effects of misclassification on the interpretation of statistical inference. However, since repeatability of a measurement can be assessed in many epidemiological studies as well as in surveys and other data collections, it would appear that much more effort should be devoted to checking the quality of the data in this way, and in making the results of these assessments available to users of the information. Since the extent of misclassification provided by this strategy can be considered a minimal estimate of the problem, it would behoove investigators to account at least for this much measurement error in planning and interpreting their studies. It is likely that the true situation is even worse than the repeat data correlations suggest.

One significant ambiguity in applying this line of thinking is that the correlation between repeat measurements usually falls as time between measurements increases. We will provide examples of this phenomenon in Chapter 4. Several factors are likely to contribute to this decline. These include changes in methodology or circumstances of the measurement such as instrument changes, personnel changes, differing seasons or times of day, fading memory, and many other issues. In addition, exposures related to diet, environment social circumstances, and others truly change over time and are likely to change more over extended time periods than brief ones. That repeatability declines with time does not obviate the need to address the problem for variables that are poorly repeatable within time periods that are short relative to the time frame under which the putative risk factor is believed to affect the outcome. We are not aware of any pragmatic investigation of this issue but believe that for chronic disease outcomes, the time period selected should be short relative to the presumed time required for the risk factor to exert a measurable effect, but long enough to allow variation between different technicians, instrument standardization materials, and week-to-week or seasonal changes in true exposure. For outcomes that are believed to develop over many years a separation of repeat measurements of 3 months to 2 years may seem reasonable. If theory suggests that some average exposure is likely the best predictor of risk, the repeat interval probably should be long enough to allow short term typical true variation to occur. Laboratory error is not the only reason that a single measurement may fail to capture average exposure.

2.7 Repeating Measurements to Improve Accuracy

While repeating measurements to assess error has not been widely advocated, the use of repeated assessments to reduce variability and improve accuracy has been widely used. For instance, many studies have taken multiple blood pressure readings during an initial assessment to achieve a more stable baseline, and dietary recall or diary methods often will try to capture several days of data. This use of repeat measurements is an important tool for reducing intra-individual variation and generally requires that the repeat measurements be done in the same way for all persons in the study. If the measurements were unbiased one could theoretically achieve any desired level of certainty about the exposure in each individual simply by repeating them enough times; but the more subjective and variable the measurements, the less certainty there would seem to be that they are unbiased. No one would argue that any number of nutritional assessments in free living people would result in data that are as valid as one or two careful measurements of height.

To use repeat measurements to reduce measurement error within individuals, the extra measures must generally be collected in all study subjects. In practice, 2 repeating measurements with reasonable correlation ($\rho>0.5$) improve accuracy per HHP data, the details can be found in chapter 4.

2.8 Conclusion

The quality of exposure measurement has important implications for the probability of a statistically significant result being a false positive. This probability ranges from the nominal significance level in the absence of measurement error, often 0.05, to 1.0 for a variable that is known to be random. This implies that barely significant findings for substantially poor measured variables should be viewed with a skepticism that is not always observed in the literature. Moreover, the classic assumptions that measurement error can be modeled as an unbiased additive error term is unlikely to hold when measurement error is extensive, because, with a fixed observed variance, the large error variance can imply unrealistically constrained variation in the true value. The correlations between repeated measurements can provide a useful minimal estimate of the extent of misclassification that is likely to be present, and in Chapter 3 we will develop an alternate, less constraining error model and will use it to look quantitatively at ways in which investigators might reduce false positive claims.

Chapter 3

Modeling Non-Differential Exposure Misclassification Using a Bivariate Normal Distribution

3.1 Background

Non-differential misclassification of exposures is a common limitation of observational epidemiological studies, and we have argued in Chapter 2 that repeated independent measurements provide a useful approach for estimating a minimal extent of this problem for specific study variables. This idea was developed in a preliminary way by Rhoads in 1987 with dietary variables.[1] Some of the most important questions that remain unresolved in human nutrition concern the relation of diet to the development of chronic disease. While coronary heart disease has been a major focus of work in this area, it has estimated that 30% or more of cancers also are attributable to dietary habits, and thousands of papers have been written exploring the nutritional causes of various tumors. Conditions as diverse as osteoporosis, varicose veins, and diverticulitis, are likely to have nutritional causes. [15]

Unfortunately, findings in this field have often been inconsistent or have failed to reproduce associations with physiological attributes that can be demonstrated in controlled experimental situations. For example, many experimental studies have shown that saturated animal fats will raise serum cholesterol, that international differences in serum cholesterol can be explained by differing fat intake, and that countries with low fat intake and low serum cholesterol (which tend to go together) have low rates of coronary heart disease.[16] But within single geographically defined cohorts it has been difficult to show relationships between dietary fat and either serum cholesterol or coronary heart disease.[16]

3.2 Using repeatability to provide a minimal estimate of measurement error

There is assumed the true value of intake of a nutrient that we are trying to assess-- not the intake on any particular day but rather some average intake over a period of time-- that is relevant to the pathogenesis of chronic disease. This *true* value is represented in the figure as *T.*

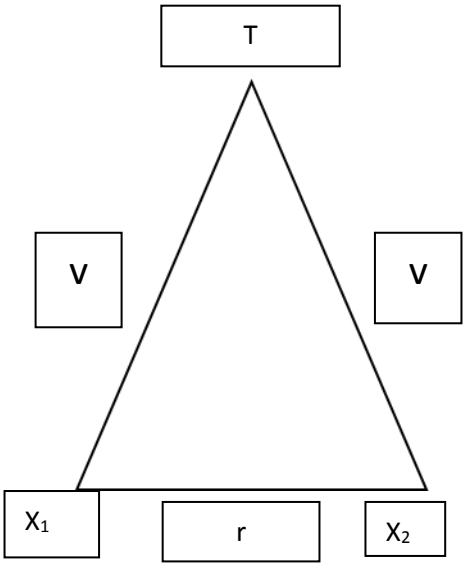


Fig. 1 Reliability, repeatability and validity (The true value is represented in the figure as T)

Using some dietary method, such as a 24 hours' dietary recall, an attempt may be made to assess *T* yielding observation(s), $X_1(X_2)$, which are correlated with T, but are rarely exactly correct. Multiple measurements can be made at different times (repeat measurements) as shown by X_1 and X_2. If a *gold* standard were available, it would measure T without error and the correlation, *v,* between *T* and $X_1(X_2)$ could be calculated. If this correlation were perfect, we could make an uncompromised assessment of the risk factor status of *T* by measuring X_1. *U*nfortunately, a real gold standard generally is unavailable, especially because the *truth* sought

is rarely the intake on one day or week but rather some average intake over a sustained period of time. Thus, clues to the validity of a measurement may have to be provided indirectly by correlations with serum levels, with other physiological parameters, or by examining the consistency of repeated measurements. A convenient summary statistic for this repeatability is the correlation coefficient, *r,* between repeat assessments separated by a time period that is appropriate for the disease in question. For most chronic diseases, an interim period of 6 months or more would seem appropriate. Of course, repeatability does not guarantee validity; but a method that does not yield reproducible results cannot be valid. If the variables in question are normally distributed and if the errors of repeat assessment are independent of each other (the optimal situation), then *v* can be estimated as *square root of r.* Stated another way, under ideal conditions *r* provides a direct estimate of *v square,* the proportion of the variance of the true value that is measured. However, some of the error associated with most methods of diet assessment is likely to repeat systematically each time the method is applied to a given subject. In other words, the correlation between repeat values, *r,* is likely to be greater than *v square.* Hence *r* can be taken as an upper limit of *v square* provided it is measured on a reasonably large sample of subjects. The notion of *upper* limit should be emphasized. Correlated errors probably are common and seem especially likely to occur when reproducibility becomes an objective in devising the method. For instance, at Framingham, the apparent reproducibility of the Burke interview was high. However, some subjects were excluded when a regular eating pattern could not be established from the interview. The method emphasizes consistency of response and in so doing may yield repeatability estimates that exaggerate the apparent quality of the data. In devising dietary methods, validity should be the primary concern. [1]

3.3 Mathematics proof

In Rhoads paper, he made the statement that under some assumptions *v* can be estimated as *square root of r.* Stated another way, under ideal conditions *r* provides a direct estimate of *v square, assuming* the variables in question are normally distributed and that the errors of repeat assessments are independent of each other (the optimal situation). for analyzing this we introduce not only the traditional approaches but also some new theory development, which is absent on any paper so far on this topic. The following mathematics proof provides the further details of this work.

3.4 Mathematical Model of Repeated Independent Observations Under Normal Theory

X_1, X_2, T are three random variables. T is random variable for true value, X_1 and X_2 are two repeated measurements for X random variable as the sample value to estimate T value.

Given error term ε is as Var (ε_1) = Var (ε_2), which implies E(ε_1) = E(ε_2) = 0

Cov (T, ε_1) = Cov (T , ε_2) = Cov(ε_1, ε_2) = 0 (X, ε_1 and ε_2 are iid, independent identically distribution)

$X_1 = T + \varepsilon_1$

$X_2 = T + \varepsilon_2$

Corr (X_1, T) = $\frac{Cov(X_1,T)}{\sqrt{Var(T)}\sqrt{Var(X_1)}}$

Corr (X_2, T) = $\frac{Cov(X_2,T)}{\sqrt{Var(T)}\sqrt{Var(X_2)}}$

Corr (X_1, X_2) = $\frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)}\sqrt{Var(X_2)}}$

Cov (X_1, X_2) = Cov (($T+ \varepsilon_1$) ($T_+ \varepsilon_2$)) = Var (T) + Cov (T, ε_1) + Cov (T, ε_2) + Cov(ε_1, ε_2) = Var (T)

Given Var (ε_1) = Var (ε_2)

$$\sqrt{\text{var}(X_1)} = \sqrt{\text{var}(T + \varepsilon_1)} = \sqrt{\text{var}(T) + \text{var}(\varepsilon_1)}$$

$$\sqrt{\text{var}(X_2)} = \sqrt{\text{var}(T + \varepsilon_2)} = \sqrt{\text{var}(T) + \text{var}(\varepsilon_2)}$$

Corr (X_1, X_2) = $\frac{\text{var}(T)}{\sqrt{\text{var}(T) + \text{var}(\varepsilon_2)}\sqrt{\text{var}(T) + \text{var}(\varepsilon_1)}}$

Corr (T, X_1) = $\frac{Cov(T, T + \varepsilon_1)}{\sqrt{\text{var}(T)(\text{var}(T) + \text{var}(\varepsilon_1))}} = \frac{\sqrt{\text{var}(T)}}{\sqrt{\text{var}(T) + \text{var}(\varepsilon_1)}}$

Corr (T, X_2) = $\frac{Cov(T, T + \varepsilon_2)}{\sqrt{\text{var}(T)(\text{var}(T) + \text{var}(\varepsilon_2))}} = \frac{\sqrt{\text{var}(T)}}{\sqrt{\text{var}(T) + \text{var}(\varepsilon_2)}}$

Corr (T, X_1) × Corr (T, X_2) = $\frac{\text{var}(t)}{\sqrt{\text{var}(t) + \text{var}(\varepsilon_1)}\sqrt{\text{var}(t) + \text{var}(\varepsilon_2)}}$ = Corr (X_1, X_2)

Corr (X_1, X_2) = Corr (T, X_1) × Corr (T, X_2) Given Var (ε_1) = Var (ε_2), equal sign holds.

if the errors of repeat assessment are independent of each other (the optimal condition), then v can be estimated as $r^{½}$. It means under ideal conditions r provides a direct estimate of v^2, the proportion of the variance of the true value that is measured. However, some of the error associated with the most methods of diet assessment is likely to repeat systematically each time the method is applied to a given subject. In other words, the correlation between repeat value,

r, is likely to be greater than v square. Hence r can be taken as an upper limit of v^2 provided it is measured on a reasonably large sample of subjects. The notion of upper limit should be emphasized. Correlated errors probably are common and seem especially likely to occur when reproducibility becomes an objective in devising a measurement method. Note that in developing this model we ignore the possibility that repeat measurements might be negatively correlated as we have not been able to imagine a useful measurement where that would be the case.

As noted in Chapter 2, the classical model of measurement error, is unlikely to hold when the error is substantial. As the variance of the error term increases, the usual model implies that the observed variance grossly overestimates the true variance which then is unrealistically small. Moreover, the assumption that the measurements with so much error are unbiased, i.e. that the mean of the observed distribution accurately centers on the true mean, becomes a matter of faith with little objective support. Researchers often assume that the variance of the observed data roughly estimates the variance of the true data and a model is needed that allows that possibility even in the presence of substantial error. We note, however, that epidemiologic inferences can be made in the presence of biased measurement as long as the bias is unrelated to the outcome and a linear relationship between the observed and true values is maintained.

The usual model is a special case of a bivariate normal distribution. By relaxing the special case assumptions, the problems noted above can be avoided. Thus, to make sense of measurements with substantial error we posit a bivariate normal distribution between the observed and true distributions without making further assumptions about the distribution of the errors. Of course, a bivariate normal distribution is a special case of the multivariate normal distribution.

3.5 The Multivariate Normal Distribution

The multivariate normal distribution is an extension of the one-dimensional univariate normal distribution to higher dimensions. A random vector (rv) is defined to be *k*-variate normally distributed if and only if every linear combination of its *k* components is univariate normally distributed. Furthermore, its importance derives from the multivariate central limit theorem (CLT). That is when the sample size is large, the average of any random vectors is approximately multivariate normally distributed regardless of the components of the random vector are correlated or not.

> The multivariate normal distribution is to be "non-degenerate" when the symmetric covariance matrix $\boldsymbol{\Sigma}$ is positive definite. The distribution has density as
>
> $$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$
>
> where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. Note the equation above can reduce to that of the univariate normal distribution if $\boldsymbol{\Sigma}$ is a 1×1 matrix (i.e. a real number, or *k*=1).[16]

3.6 The Bivariate Normal Distribution

The bivariate normal distribution is the form of the multivariate normal distribution that has most applicability to exposure measurement error.

In the 2-dimensional nonsingular case (*k* = rank(Σ) = 2), the joint density function f (X_1, X_2) of two jointly normally distributed variables X_1 and X_2 can be written as

f (X_1, X_2) = $1/[2\pi\sigma_{x1}\,\sigma_{x2}\sqrt{1-\rho^2}]$ exp{(-1/[2(1-ρ^2)])([(X_1- μ_{x1})/σ_{x1}]^2 - 2ρ[(X_1- μ_{x1})/σ_{x1}][(X_2- μ_{x2})/σ_{x2}] +[(X_2- μ_{x2})/σ_{x2}]^2 }

where ρ is the correlation between X_1 and X_2 and $\sigma_{x1} > 0$ and $\sigma_{x2}. > 0$. In the matrix notation of Section 3.5, we have

$$\mu = \begin{pmatrix} \mu_{x1} \\ \mu_{x2} \end{pmatrix}, \ \Sigma = \begin{bmatrix} \sigma_{x1}^2 & \rho\sigma_{x1}\sigma_{x2} \\ \rho\sigma_{x1}\sigma_{x2} & \sigma_{x2}^2 \end{bmatrix}$$

It can be shown that the marginal mean and variance of X_i are μ_{xi} and σ_{xi}^2 respectively, i=1, 2. In the following, we will use μ_x and σ_x^2 for the mean and variance of random variable X respectively. In the bivariate case, the equivalent condition for bivariate normality is to verify that any distinct linear combinations of X_1 and X_2 are normal in order to conclude that the vector $[X_1, X_2]$ is bivariate normal. [16]

3.7 The Range of ρ

The value range of ρ under the given assumptions is the first step to be explored. In this thesis, we assume at n dimensional space, such as n random variables (rv), the correlation (**ρ)** will be the same for any two of them, it is often referred to as compound symmetry correlation structure.

Let

$X \sim N(\mu, \Sigma)$

Σ = Variance- Covariance Matrix; and Σ is defined as non-positive definite Matrix, hence we have

$|\Sigma| = \det(\Sigma) \geq 0$. We will show that $-\frac{1}{N-1} \leq \rho \leq 1$ if the correlation (**ρ)** is the same for any two of components of the normally distributed random vector.

With *equal correlation* assumption, we have

$$\Sigma = \begin{bmatrix} \sigma_{x1}^2 & \rho\sigma_{x1}\sigma_{x2} & \cdots & \rho\sigma_{x1}\sigma_{xN} \\ \rho\sigma_{x2}\sigma_{x1} & \sigma_{x2}^2 & \cdots & \rho\sigma_{x2}\sigma_{xN} \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma_{xN}\sigma_{x1} & \rho\sigma_{xN}\sigma_{x2} & \cdots & \sigma_{xN}^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{x1} & 0 & \cdots & 0 \\ 0 & \sigma_{x2} & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{xN} \end{bmatrix} \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_{x1} & 0 & \cdots & 0 \\ 0 & \sigma_{x2} & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{xN} \end{bmatrix} = \text{DRD}$$

where $D = \begin{bmatrix} \sigma_{x1} & 0 & \cdots & 0 \\ 0 & \sigma_{x2} & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{xN} \end{bmatrix}$ and $R = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$.

Since Det $(D) = \sigma_{x1}\sigma_{x2} \ldots \sigma_{xN} > 0$, we have Det(R) ≥ 0. We will use two methods to show that $-\frac{1}{N-1} \leq \rho \leq 1$.

Method 1. Direct calculation of Det(R):

$$\text{Det (R)} = \begin{vmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 1+(N-1)\rho & 1+(N-1)\rho & \cdots & 1+(N-1)\rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{vmatrix}$$ add all rows to the 1st row

$$= (1 + (N\text{-}1)\,\rho) \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{vmatrix}$$

$$= (1 + (N\text{-}1)\,\rho) \begin{vmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1-\rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1-\rho \end{vmatrix}$$ minus all other rows by ρ times the first row

= (1 + (N-1) ρ) $(1-\rho)^{N-1}$ by the property of upper triangular matrix.

Since det (R) ≥ 0 and $|\rho| \leq 1$, we have (1 + (N-1) ρ) $(1-\rho)^{N-1}) \geq 0$. This implies that 1 + (N-1) $\rho \geq 0$ for $\rho \neq 1$, therefore $\rho \geq -\frac{1}{N-1}$

Method 2. Since R can be written as

R = $(1-\rho)$**I** + ρ **11**$^{\mathrm{T}}$

Where **I** is N x N identity matrix and **1** = $(1, 1, \cdots, 1)^{\mathrm{T}}$ is N by 1 matrix with elements of 1 for all entries. Since Σ *is non-negative definite* and D is diagonal matrix with standard deviation on the main diagonal*, by definition, we have for any x ϵ R* ^{N}, x$^{\mathrm{T}}\Sigma$x ≥ 0 and hence x$^{\mathrm{T}}R$x ≥ 0. By Cauchy inequality, we have

$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} 1 \times x_i \leq \sqrt{\sum_{i=1}^{N} 1^2} \sqrt{\sum_{i=1}^{N} x_i^2} = \sqrt{\mathrm{N} \sum_{i=1}^{N} x_i^2}$

Therefore

0 ≤ x$^{\mathrm{T}}R$x = x$^{\mathrm{T}}$($(1-\rho)$**I** + ρ **1**$^{\mathrm{T}}$**1**)x

= $(1-\rho)$x$^{\mathrm{T}}$x + ρ (x$^{\mathrm{T}}$**1**) (**1**$^{\mathrm{T}}$x)

= $(1-\rho)$x$^{\mathrm{T}}$x + ρ $(\sum_{i=1}^{N} x_i)^2$

= $(1-\rho)$ $\sum_{i=1}^{N} x_i^2$ + ρ $(\sum_{i=1}^{N} x_i)^2$

≤ $(1-\rho)$ $\sum_{i=1}^{N} x_i^2$ + ρ $N \sum_{i=1}^{N} x_i^2$ by Cauchy inequality

= $((1-\rho)$ + ρ $N)$ $\sum_{i=1}^{N} x_i^2$

This implies that $(1-\rho)$ + ρ $N \geq 0$ and $\rho \geq -\frac{1}{N-1}$

3.8 Testing mean difference under measurement error

First we will show that If two random variables are independent, then they are uncorrelated.

Proof: If X and T are two independent random variables, then

E[XT] = ∫ ∫ X T $p_{X,T}$(x, y) $d_x\, d_T$

= ∫ ∫ XT $p_X(x)\ \ p_T(T) d_x\, d_T$

= (∫ X $p_X(x)\ d_x$) (∫ T $p_T(T) d_T$)

= E[X]E[T]

Therefore, Cov (X, T) = EXT − EXET = EXET − EXET = 0. This implies that X and T are uncorrelated.

In general, if two random variables are uncorrelated, they can still be dependent. However, in normal distribution special case, it can be shown that if two random variables are uncorrelated, then they are independent as well. This implies that in the multivariate normal situation, lack of correlation implies independence.

Assume that X is observed and has measurement error: $X = T + \varepsilon$, where T is the true exposure and not measured and ε is the measurement error and independent of T. It is clear that X and T are correlated. We have

$$\sigma_X^2 = var(X) = var(T + \varepsilon) = var(T) + var(\varepsilon) = \sigma_T^2 + \sigma_\varepsilon^2 \geq \sigma_T^2.$$

This implies that the variance of X is larger than the variance of T. We also have

$$Corr(X,T) = \frac{Cov(X,T)}{\sqrt{var(X)var(T)}} = \frac{Cov(T+\varepsilon,T)}{\sqrt{var(X)var(T)}} = \frac{var(T)}{\sqrt{var(X)var(T)}} = \sqrt{\frac{\sigma_T^2}{\sigma_T^2+\sigma_\varepsilon^2}} = \rho.$$

This implies that $\sigma_X^2 = \rho^2 \sigma_T^2$.

Under normality assumption, $\binom{X}{T}$ will follow a bivariate normal distribution:

$$\binom{X}{T} \sim N\left(\binom{\mu_X}{\mu_T}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_T \\ \rho\sigma_X\sigma_T & \sigma_T^2 \end{pmatrix}\right)$$

Now assume that X_i^D and X_i^N are for observed disease and control groups with iid normal distribution as follows:

$X_1^D, X_2^D, \ldots, X_m^D \overset{iid}{\sim} N(\mu_X^D, \sigma_X^2)$ and $X_1^N, X_2^N, \ldots, X_n^N \overset{iid}{\sim} N(\mu_X^N, \sigma_X^2)$,

Similarly, T_i^D and T_i^N are for unobserved true values of disease and control groups with iid normal distribution as follows:

$T_1^D, T_2^D, \ldots, T_m^D \overset{iid}{\sim} N(\mu_T^D, \sigma_T^2)$ and $T_1^N, T_2^N, \ldots, T_n^N \overset{iid}{\sim} N(\mu_T^N, \sigma_T^2)$,

then

$$\begin{pmatrix} X_i^j \\ T_i^j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X^j \\ \mu_T^j \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_T \\ \rho\sigma_X\sigma_T & \sigma_T^2 \end{pmatrix}\right)$$

Where $j = D \; or \; N$.

The test statistics for testing the difference in means between disease and non-disease groups are:

$$Z_X = \frac{\overline{X}_D - \overline{X}_N}{\sqrt{\sigma_X^2\left(\frac{1}{m}+\frac{1}{n}\right)}} \sim \text{N}\left(\frac{\mu_X^D - \mu_X^N}{\sqrt{\sigma_X^2\left(\frac{1}{m}+\frac{1}{n}\right)}}, 1\right) \qquad (1)$$

$$Z_T = \frac{\overline{T}_D - \overline{T}_N}{\sqrt{\sigma_T^2\left(\frac{1}{m}+\frac{1}{n}\right)}} \sim \text{N}\left(\frac{\mu_T^D - \mu_T^N}{\sqrt{\sigma_T^2\left(\frac{1}{m}+\frac{1}{n}\right)}}, 1\right) \qquad (2)$$

The covariance between Z_X and Z_T is:

$$\text{Cov}\,(Z_X\,,Z_T) = Cov\left(\frac{\overline{X}_D-\overline{X}_N}{\sqrt{\sigma_X^2\,(\frac{1}{m}+\frac{1}{n})}}\,,\frac{\overline{T}_D-\overline{T}_N}{\sqrt{\sigma_T^2\,(\frac{1}{m}+\frac{1}{n})}}\right) = \frac{Cov(\overline{X}_D-\overline{X}_N,\overline{T}_D-\overline{T}_N)}{\sqrt{\sigma_X^2\,(\frac{1}{m}+\frac{1}{n})}\sqrt{\sigma_T^2\,(\frac{1}{m}+\frac{1}{n})}}$$

Per the independent assumptions, $\overline{X}_D, \overline{X}_N, \overline{T}_D$, and $\overline{T}_N$ are mutually independent, we have

$$Cov\big(\overline{X}_D-\overline{X}_N,\overline{T}_D-\overline{T}_N\big) = Cov\left(\overline{X}_D,\overline{T}_D\right) + Cov\,(\overline{X}_N,\overline{T}_N) = \rho\frac{1}{m}\sigma_X\sigma_T + \rho\frac{1}{n}\sigma_X\sigma_T = (\frac{1}{m}+\frac{1}{n})\rho\sigma_X\sigma_T$$

and

$$\text{Cov}\,(Z_X\,,Z_T) = \frac{(\frac{1}{m}+\frac{1}{n})\rho\sigma_X\sigma_T}{\sqrt{\sigma_X^2\,(\frac{1}{m}+\frac{1}{n})}\sqrt{\sigma_T^2\,(\frac{1}{m}+\frac{1}{n})}} = \rho \qquad (3)$$

Thus, from normal theory, $(Z_X\,,Z_T)$ is bivariate normal distributed with

$$\begin{pmatrix}Z_X\\Z_T\end{pmatrix} \sim \text{N}\left(\begin{pmatrix}\frac{\mu_{X-}^D\,\mu_X^N}{\sqrt{\sigma_x^2\,(\frac{1}{m}+\frac{1}{n})}}\\\frac{\mu_{T-}^D\,\mu_T^N}{\sqrt{\sigma_T^2\,(\frac{1}{m}+\frac{1}{n})}}\end{pmatrix},\quad\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}\right) \qquad (4)\ \text{Under Ha}$$

Hence the conditional distribution of Z_T given Z_X is

$$Z_T|Z_X \sim N\left(\frac{\mu_{T-}^D\,\mu_T^N}{\sqrt{\sigma_T^2\,(\frac{1}{m}+\frac{1}{n})}} + \rho\left(Z_X - \frac{\mu_{X-}^D\,\mu_X^N}{\sqrt{\sigma_X^2\,(\frac{1}{m}+\frac{1}{n})}}\right),\quad 1-\rho^2\right). \qquad (5)$$

Under null hypothesis H_0: $\mu_T^D = \mu_T^N$, the joint distribution of Z_T and Z_X is

$$\begin{pmatrix}Z_X\\Z_T\end{pmatrix} \sim \text{N}\left(\begin{pmatrix}\frac{\mu_{X-}^D\,\mu_X^N}{\sqrt{\sigma_X^2\,(\frac{1}{m}+\frac{1}{n})}}\\0\end{pmatrix},\quad\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}\right), \qquad (6)$$

the conditional distribution is

$$Z_T|Z_X \sim N\left(\rho\left(Z_X - \frac{\mu_{X-}^{D} \mu_X^N}{\sqrt{\sigma_X^2\ (\frac{1}{m}+\frac{1}{n})}}\right),\quad 1-\ \rho^2\right) \qquad (7)$$

Let $Z_X^a = Z_X\ -\ \frac{\mu_{X-}^{D} \mu_X^N}{\sqrt{\sigma_X^2\ (\frac{1}{m}+\frac{1}{n})}}$, then per equation (7), we get the following equation (8) and (9)

E $(Z_T|Z_X) =\ \rho Z_X^a$ (8) Under null for true value

P $(Z_T \geq C|Z_X)$ = P $\left(\frac{Z_T-\rho Z_X^a}{\sqrt{1-\rho^2}} \geq \frac{C-\rho Z_X^a}{\sqrt{1-\rho^2}}\ |Z_X\right)$ = 1 - ϕ $\left(\frac{C-\rho Z_X^a}{\sqrt{1-\rho^2}}\right)$ (9) Under null for true value

Where ϕ is the CDF of standard normal distribution and C is critical value adjusted with ρ.

Under the measurement error model, we have $\mu_X^N = \mu_T^N$ and $\mu_X^D = \mu_T^D$, hence under null hypothesis H_0: $\mu_T^D = \mu_T^N$, we have $\mu_X^D = \mu_X^N$ as well, therefore

$Z_X^a = Z_X\ -\ \frac{\mu_{X-}^{D} \mu_X^N}{\sqrt{\sigma_X^2\ (\frac{1}{m}+\frac{1}{n})}} =\ Z_X$ (10) Under null for observed distribution

Given one side test for Z_X = 1.64 (α = 0.05), or two sided test Z_X = 1.96 (α = 0.025). Per equation (8), the conditional Expected value (Z_X is not a random variable (rv) here, Z_X is fixed.)

E $(Z_T|Z_X) = \rho Z_X$ = 1.64 ρ (one-side) and = 1.96 ρ (two-sides).

We can see that the expected conditional true test statistic value is smaller than the observed one, this will lead to the increase of true Type I error if we use the conventional cut-off value for the test. We will discuss in more details in the following section.

3.9 The Conditional Expected Value Under Null

Using equation 8 under the null hypothesis we can calculate **expected** Z_T given a measured value, Z_X, as E $(Z_T|Z_X) = \rho Z_X$. The observed Z value in the study sample only provides

evidence against the null hypothesis if it implies that the true Z value in the sample is away from the null. Given the observed deviation and an independent estimate of the correlation between the observed and true distributions, one can find the expected value of Z_T. As shown in Table 3.1 below, observed Z values that just meet conventional levels of statistical significance do not imply critical Z values for the true distribution. Therefore, usual interpretation of these observed values will be associated with an excessive number of false positive conclusions. The lower the correlation estimate, the less relevance Z_X has for Z_T.

Equation 8 can also be used to calculate how extreme an observed value needs to be to imply that the expected true value will meet a conventional (or other specified) level of significance when the correlation is believed to be less than 1.0. For example, to achieve an expected true deviation (Z_T) of 1.96 with a correlation between repeated measures of 0.8 one would need to observe Z_X of 2.45. as shown in the right hand column of Table. If the correlation between repeat values was as low as 0.3 one would need to observe Z_X of 6.53. Since the repeat value correlation is believed to provide a minimal estimate of the extent of misclassification, these estimates of how extreme Z_X is needed could be too small.

3.10 The Conditional Critical Value for Test Statistics

In this section, we will also first discuss, if the conventional cut-off value for the observed test statistic Z_X is used, what is the expected value of the true test statistics? On this other hand, we will also discuss that given critical values for true test statistic Z_T such as 1.64 or 1.96 (i.e., true significance level of 10 or 5%), what is the observed Z_X should be to achieve the claim significance level?

First, if conventional cut-off value such as 1.64 or 1.96 is used for the observed value of Z_X , then E ($Z_T|Z_X$) $= \rho Z_X$ = ρ1.64 or ρ1.96. This implies, for example, for ρ = 0.5, E ($Z_T|Z_X$) = 0.82 or 0.98 as shown in Table 3.1. Hence the true type I error would be inflated.

On the other hand, in order that E ($Z_T|Z_X$) ≥ 1.64 (α = 0.05), or E ($Z_T|Z_X$) ≥ 1.96 (α = 0.025), Per equation (8), E ($Z_T|Z_X$) $= \rho Z_X$ the observed Z_X should be greater than 1.64/ρ (α = 0.05), or 1.96/ρ (α = 0.025). As an example for ρ = 0.5, Z_X ≥ 3.28 or 3.92 as shown in Table 3.1.

Table 3.1 shows the relationships between the observed critical values and expected ones under several different scenarios of ρ and the significance level of 5 and 10%.

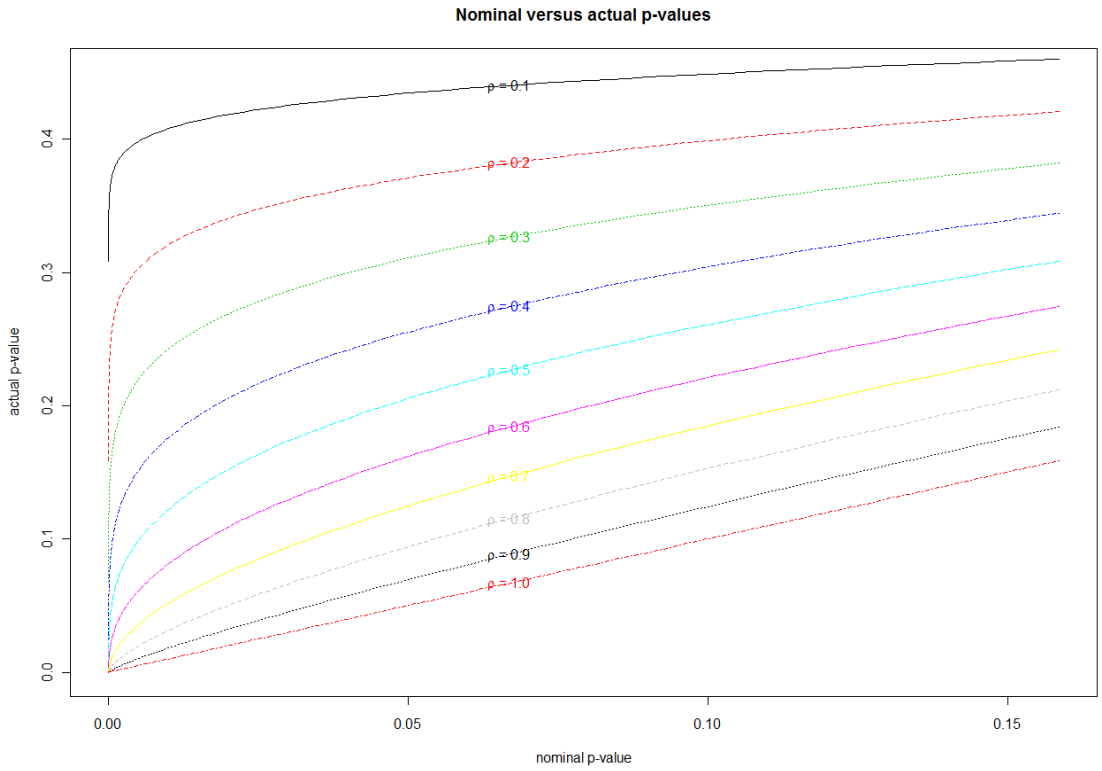
Table 3.1: The Conditional Expected Value (Z_T and Z_X) under Different ρ Scenarios and α Levels

Expected Value Under null for observed distribution $\mu_X^D = \mu_X^N$	ρ	E ($Z_T \mid Z_X$) if One Side (Z_X =1.64)	Observed Z_X required when E($Z_T \mid Z_X$) ≥ 1.64 (α = 0.05) $Z_X \geq 1.64/\rho$	E ($Z_T \mid Z_X$) if One-sides (Z_X = 1.96)	Observed Z_X required when E($Z_T \mid Z_X$) ≥ 1.96 (α = 0.025) $Z_X \geq 1.96/\rho$
E($Z_T \mid Z_X$)	0.3	0.492	5.47	0.588	6.53
E($Z_T \mid Z_X$)	0.4	0.656	4.1	0.784	4.9
E ($Z_T \mid Z_X$)	0.5	0.82	3.28	0.98	3.92
E ($Z_T \mid Z_X$)	0.6	0.984	2.73	1.176	3.27
E ($Z_T \mid Z_X$)	0.7	1.148	2.34	1.372	2.8
E ($Z_T \mid Z_X$)	0.8	1.312	2.05	1.568	2.45
E ($Z_T \mid Z_X$)	0.9	1.476	1.82	1.764	2.18
E ($Z_T \mid Z_X$)	1	1.64	1.64	1.96	1.96

These relationships are shown graphically in Figure 1. Using expected Z_T given Z_X (E ($Z_T \mid Z_X$)) for the determination of P-value (considered as actual P-value), the figure shows the graph for

nominal (if *X* measured without measurement error) versus actual p-values (assuming ρ is the correlation between true and observed values). The numbers on the graph are ρ values. When ρ is 1.0 the P-values are identical resulting in the straight line at the bottom of the figure, but if ρ is 0.7, corresponding to a correlation between repeat observations of 0.49, the observed one-sided P-value would need to be 0.01 to correspond to an actual P-value 0.05.

Figure 3.1: One-sided P values corresponding to z scores expected in the "true" distribution under the null hypothesis compared to observed P values for different estimates of ρ.



3.11 Bayes implication with observed Z_X to predict statistical inference

Given observed Z_X and based on the conditional probability per equation (9) and (10), in order to be (1-γ) sure that $Z_T \geq Z_{1-\alpha}$, we have

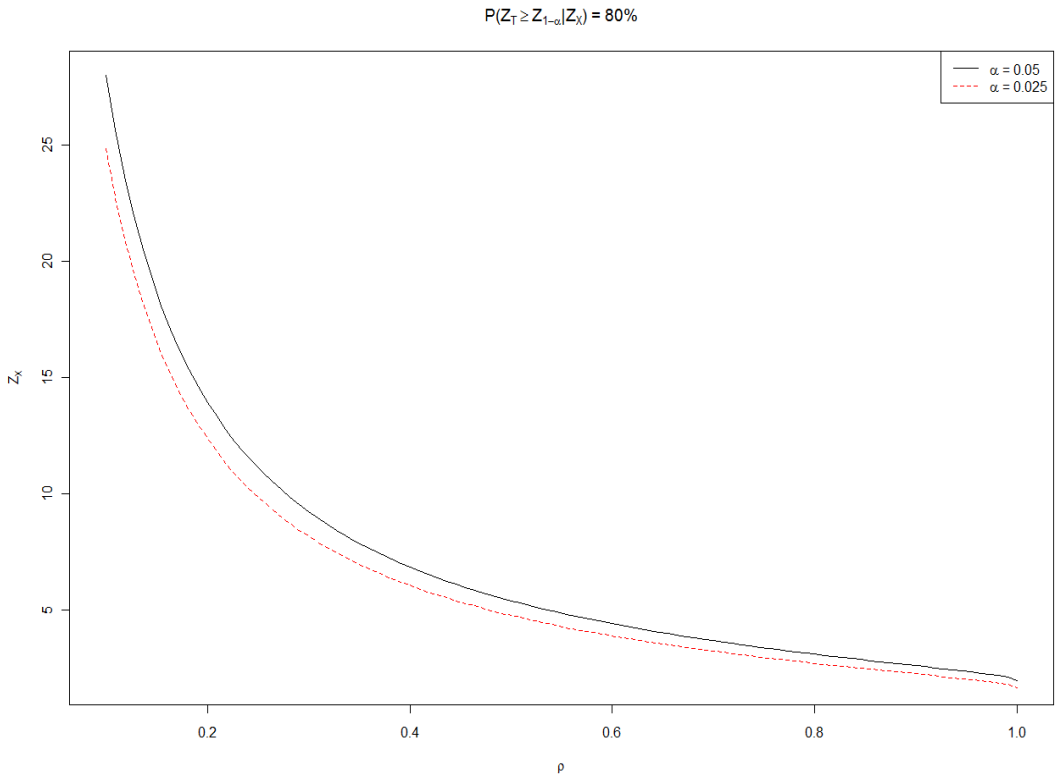
$(Z_{1-\alpha} - \rho\, Z_X)/(\sqrt{1-\rho^2}) = Z_\gamma \Rightarrow (Z_{1-\alpha} - \rho\, Z_X) = Z_\gamma \sqrt{1-\rho^2} \Rightarrow \rho\, Z_X = Z_{1-\alpha} - Z_\gamma \sqrt{1-\rho^2}$

So we get $Z_X = (Z_{1-\alpha} - Z_\gamma\sqrt{1-\rho^2}) / \rho$ (10a) Under null for observed distribution

(α is the type 1 error, γ is the probability).

Figure 2 shows the plots of ρ versus Z_X for P ($Z_T \geq Z_{1-\alpha} | Z_x$) = 80% and 90% probability respectively. From the figure, for example, for α = 0.05 and $Z_{1-\alpha}$ = 1.64, in order to be 80% sure that $Z_T \geq 1.64$, with ρ = 0.6, then Z_X should be 3.86 or greater, if ρ = 0.3, Z_X should be 8.16 or greater. The detail Z_x with ρ range from (0, 1) is showed in table 2. This is the alternative way to show the same message of (table 3.1 and figure 3.1).

Figure 3.2: ρ versus Z_X for P ($Z_T \geq Z_{1-\alpha} | Z_x$) = 80% and 90%



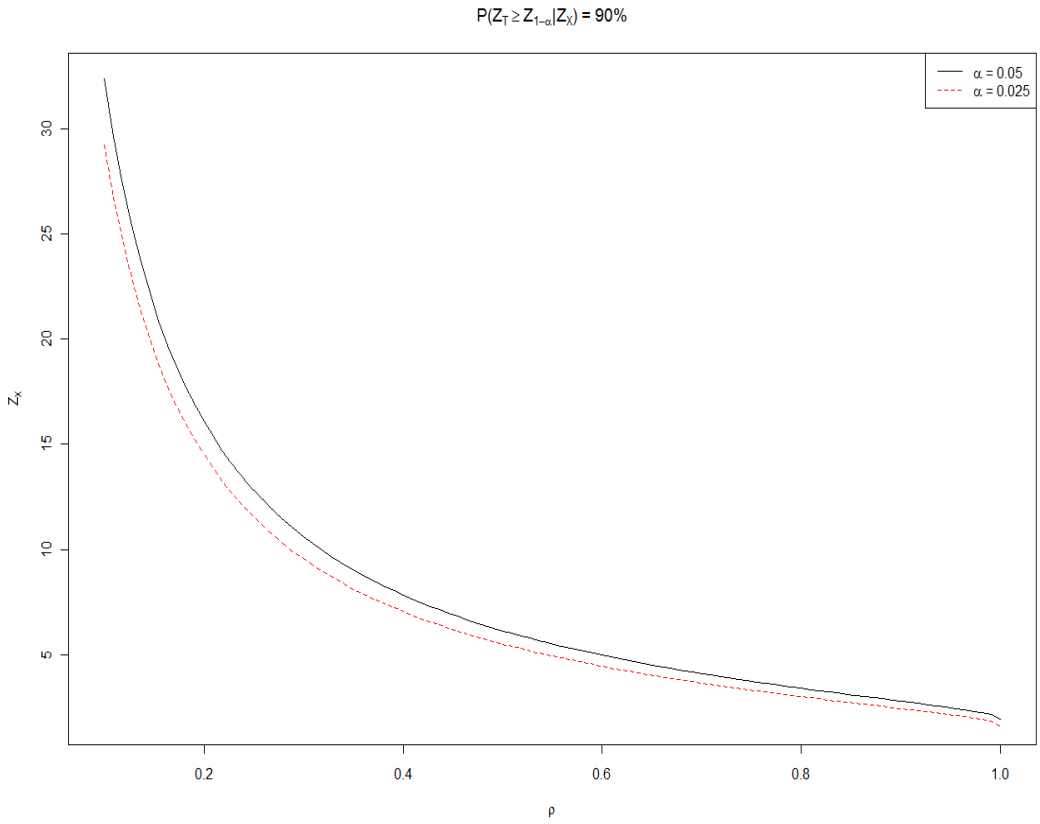


Table 3.2 shows values from the above figures for selected values of ρ. Note that ρ is the square root of the correlation between repeat measurements, so that ρ=0.3 corresponds to a correlation between repeat measurements of 0.09, ρ = 0.5 corresponds to a correlation between measures of 0.25, ρ = 0.7 corresponds to an observed correlation of 0.49, and ρ = 0.9 corresponds to an observed correlation of 0.81.

Table 3.2 The Critical Observed Z_X Value under Different ρ Scenarios α and β Errors

Z_X Value Under null for observed distribution $\mu_X^D = \mu_X^N$	ρ	ρ^2	Required minimal observed Z_X when $Z_T \geq Z_{1-\alpha}$ ($Z_{1-\alpha}$=1.64) 80% Sure	Required minimal observed Z_X when $Z_T \geq Z_{1-\alpha}$ ($Z_{1-\alpha}$=1.64) 90% Sure	Required minimal observed Z_X when $Z_T \geq Z_{1-\alpha}$ ($Z_{1-\alpha}$ = 1.96) 80% Sure	Required minimal observed Z_X when $Z_T \geq Z_{1-\alpha}$ ($Z_{1-\alpha}$ = 1.96) 90% Sure
Z_X	0.3	0.09	8.16	9.56	9.21	10.61
Z_X	0.4	0.16	6.04	7.05	6.83	7.84
Z_X	0.5	0.25	4.75	5.51	5.38	6.14
Z_X	0.6	0.36	3.86	4.45	4.39	4.98
Z_X	0.7	0.49	3.21	3.66	3.66	4.11
Z_X	0.8	0.64	2.69	3.02	3.08	3.41
Z_X	0.9	0.81	2.24	2.45	2.59	2.80
Z_X	1	1	1.64	1.64	1.96	1.96

If ρ =0, then we get P ($Z_T \geq Z_{1-\alpha} | Z_x$) = ϕ ($Z_{1-\alpha}$) = 1- (1- α) = α, meaning that X does not provide any information on *T*. On the other hand,

If ρ =1,

P ($Z_T \geq Z_{1-\alpha} | Z_X$) = 1 if $Z_{1-\alpha} < Z_X$ (10b)

or

P ($Z_T \geq Z_{1-\alpha} | Z_X$) = 0 if $Z_{1-\alpha} > Z_X$ (10c)

This is logical since *X* would have the same impact as *T* because of zero measurement error.

3.12 The Conditional Power and Sample Size

Since

$$\begin{pmatrix} Z_X \\ Z_T \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{\mu_{X-}^{D} \mu_X^N}{\sqrt{\sigma_x^2 \left(\frac{1}{m}+\frac{1}{n}\right)}} \\ \frac{\mu_{T-}^{D} \mu_T^N}{\sqrt{\sigma_T^2 \left(\frac{1}{m}+\frac{1}{n}\right)}} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \quad \text{(4) Under Ha}$$

the conditional distribution is

$$Z_T | Z_X \sim N\left(\frac{\mu_{T-}^{D} \mu_T^N}{\sqrt{\sigma_T^2 \left(\frac{1}{m}+\frac{1}{n}\right)}} + \rho \left(Z_X - \frac{\mu_{X-}^{D} \mu_X^N}{\sqrt{\sigma_X^2 \left(\frac{1}{m}+\frac{1}{n}\right)}} \right), \quad 1 - \rho^2 \right). \quad (5)$$

Therefore

$$E\left(Z_T | Z_X\right) = \frac{\mu_{T-}^{D} \mu_T^N}{\sqrt{\sigma_T^2 \left(\frac{1}{m}+\frac{1}{n}\right)}} + \rho \left(Z_X - \frac{\mu_{X-}^{D} \mu_X^N}{\sqrt{\sigma_X^2 \left(\frac{1}{m}+\frac{1}{n}\right)}} \right) = \mu_{(Z_T | Z_X)}, \quad (6)$$

and the power function is

$$\text{P}\ (Z_T \geq C = Z_{1-\alpha}|Z_X) = \text{P}\left(\frac{Z_T - \mu(Z_T|Z_X)}{\sqrt{1-\rho^2}} \geq \frac{Z_{1-\alpha} - \mu(Z_T|Z_X)}{\sqrt{1-\rho^2}}\ |Z_X\right)$$

$$= 1 - \phi\left(\frac{Z_{1-\alpha} - \mu(Z_T|Z_X)}{\sqrt{1-\rho^2}}\right) = 1 - \beta \qquad \text{(11) Power Function}$$

where ϕ is CDF of standard normal distribution.

Under Ha: $\mu_{T\neq}^{D}\ \mu_T^N$, we assuming that effect sizes for *T* and *X* are

$d_T = \frac{\mu_{T-}^{D}\ \mu_T^N}{\sigma_T}$ and $d_X = \frac{\mu_{X-}^{D}\ \mu_X^N}{\sigma_X}$.

Since $\sigma_X^2 = \rho^2\sigma_T^2$ and $\mu_{T-}^{D}\ \mu_T^N = \mu_{X-}^{D}\ \mu_X^N$,

$d_X = \frac{\mu_{X-}^{D}\ \mu_X^N}{\sigma_X} = \rho\frac{\mu_{T-}^{D}\ \mu_T^N}{\sigma_T} = \rho d_T$.

Let n' be the harmonic mean of n and m, that is

$$\frac{1}{n'} = \frac{1}{m} + \frac{1}{n}$$

$$\mu_{(Z_T|Z_X)} = \frac{\mu_{T-}^{D}\ \mu_T^N}{\sqrt{\sigma_T^2\ (\frac{1}{m}+\frac{1}{n})}} + \rho\left(Z_X - \frac{\mu_{X-}^{D}\ \mu_X^N}{\sqrt{\sigma_X^2\ (\frac{1}{m}+\frac{1}{n})}}\right) \qquad (6)$$

$$\mu_{(Z_T|Z_X)} = \sqrt{n'}\ d_T + \rho\ (Z_X - \sqrt{n'}d_X)$$

$$= \sqrt{n'}\ (d_T - \rho\ d_X) + \rho Z_X \qquad (12)$$

Therefore the conditional power of $1 - \beta$ given observed Z_X for the test is

$$\text{P}\ (Z_T \geq C = Z_{1-\alpha}|Z_X, H_a) = \text{P}\left(\frac{Z_T - \mu(Z_T|Z_X)}{\sqrt{1-\rho^2}} \geq \frac{Z_{1-\alpha} - \mu(Z_T|Z_X)}{\sqrt{1-\rho^2}}\ |Z_X\right)$$

$= 1 - \Phi \left(\frac{Z_{1-\alpha} - \sqrt{n'}\,(d_T - \rho\, d_X) + \rho Z_X}{\sqrt{1-\rho^2}} \right) = 1 - \beta$ (13)

Hence we have

$\frac{Z_{1-\alpha} - \sqrt{n'}\,(d_T - \rho\, d_X) + \rho Z_X}{\sqrt{1-\rho^2}} = Z_\beta$ Since $Z_\beta = -Z_{(1-\beta)}$ (14)

$Z_{(1-\alpha)} - (\sqrt{n'}\,(d_T - \rho\, d_X) + \rho Z_X) = Z_\beta \sqrt{1-\rho^2}$

$\sqrt{n'} = \frac{Z_{1-\alpha} + Z_{1-\beta}\sqrt{1-\rho^2} - \rho Z_X}{(d_T - \rho\, d_X)}$

Since $\sigma_X^2 = \rho^2 \sigma_T^2$, and $d_X = \rho d_T$.

$n' = \left(\frac{Z_{1-\alpha} + Z_{1-\beta}\sqrt{1-\rho^2} - \rho Z_X}{(d_T - \rho\, d_X)} \right)^2 = \left(\frac{Z_{1-\alpha} + Z_{1-\beta}\sqrt{1-\rho^2} - \rho Z_X}{\left(\frac{d_X}{\rho} - \rho\, d_X\right)} \right)^2 = \left(\frac{Z_{1-\alpha} + Z_{1-\beta}\sqrt{1-\rho^2} - \rho Z_X}{d_X\,(1-\rho^2)/\rho} \right)^2$ (15)

as the Sample size function in term of d_X.

Or in terms of true underlying effect size d_T,

$n' = \left(\frac{Z_{1-\alpha} + Z_{1-\beta}\sqrt{1-\rho^2} - \rho Z_X}{d_T\,(1-\rho^2)} \right)^2$ (16) Sample size function

Table 3.3 shows sample size increases that would likely be needed to achieve the same confidence in results with poorly measured variables that apply to well measured variables. At low values for ρ and ρ^2 very large increases in sample size are needed to achieve the high observed Z values that are required to make inferences about the underlying weakly correlated true data. These estimates of sample size increases do not incorporate all the considerations that are relevant to calculating statistical power under the bivariate normal model. Indeed, when measurements are poor it may be difficult to know how to estimate what effect size in the

true data one should be looking for since practical experience is nearly always limited to observed data.

A striking feature shown in the tables is the increase in observed Z score that is required in the bivariate normal model for even quite well measured variables. For instance, with ρ =0.9 the required observed Z-score is 2.80 to provide confidence that the true Z score is 1.96. Mathematically, this results from uncoupling the mean of the true distribution from the observed distribution. The evidence against the use of the traditional error model, which assumes the observed mean is unbiased, is quite strong for poorly measured variables since it implies restricted variation in the true data. But little restriction is implied for well measured variables, so the bivariate normal distribution may allow too much drift of the observed mean and thus over-estimate the effect of modest measurement error.

Table 3.3. Increase in Number of Observations Needed to Expect to Achieve 80% Confidence that True Mean in the Study Sample Exceeds the Critical Value ($Z_T \geq 1.96$)

Correlation of Observed with True Value (ρ)	Correlation Between Repeat Observations (ρ^2)	Required Minimal Observed Zx to Provide 80% Confidence $Z_T \geq 1.96$	Percent Increase in Sample Size Required for Increase in Needed Zx*
0.3	0.09	9.21	948.2
0.4	0.16	6.83	526.5
0.5	0.25	5.38	322.5
0.6	0.36	4.39	206.3
0.7	0.49	3.66	132.5
0.8	0.64	3.08	81.1
0.9	0.81	2.59	42.7
1.0	1.00	1.96	0.0

*Calculated using Z_X from table 3.2 and $Z_\theta = 1.28$

3.13 Discussion and Conclusion

We have explored the use of a bivariate normal distribution model to address two deficiencies of the classical error model that is used for continuous exposure variables in epidemiological (and many other) studies. These deficiencies are evident mainly when measurement error is substantial. In this setting a significant finding may paradoxically be taken to be of increased importance because the variable was poorly measured, and the point estimate is assumed to biased toward the null. In addition to this problem, the decomposition of the observed variance into true and error variances implies shrinking true variance as error variance increases. In the presence of poor measurement this can yield unrealistically low estimates of true variability.

The bivariate normal model provides a way of quantitating the effect of misclassification on statistical inference in epidemiology that has heretofore been largely ignored. Moreover, it implies no restriction on the variance of the true parameter and it accommodates the reality that poorly measured variables may provide very little information about the underlying truth, thus increasing the chance that any significant findings may be false positives. While it is likely true that some significant results on poorly measured variables underestimate true effects, it is also true that poor measurement will increase the probability of false positive conclusions.

A probable shortcoming of the bivariate normal model is that it appears likely to be too conservative for variables that are quite well measured.

Chapter 4

Repeatability of Epidemiological Data Collected at the Honolulu Heart Study and Implications for Data Interpretation.

4.1. Objectives

In this Chapter we investigate the repeatability of a wide variety of epidemiological exposure data and explore the implications of the bivariate normal model that is developed in Chapter 3 for planning sample sizes and developing conclusions from variables that are highly repeatable or not so repeatable. The data selected for this purpose are from the Honolulu Heart Program (HHP), which was initiated in 1965 by the National Heart Lung and Blood Institute (NHLBI), NIH, as a prospective cohort study of the antecedents and causes of cardiovascular disease among Japanese Americans living in Hawaii. Using the extensive data collected in this cohort we explore the following three issues: a) Given the inherent variation in the distributions of various exposure risk factors, would it make a substantial difference if one used a rank order statistic to assess repeatability rather than the traditional Pearson correlation coefficient which assumes normal distributions? b) How large is the effect of varying the time interval between repeat measurements of exposure risk factors on the correlations that are obtained, and what advice can be given as to a reasonable time interval to use? c) Among variables that might reasonably have been expected to be related to coronary heart disease in this cohort (and more generally in the many other cohort studies of CHD that have been done) are those that have turned out to

be significant, generally accepted predictors better measured than other variables that would be expected to be related to CHD but for which no consistent relationships have been found?

4.2 Overview of the Honolulu Heart Program

The Honolulu Heart Program was initiated because the Hawaii Japanese population was known to have low incidence of coronary heart disease compared to whites, but a higher incidence than the indigenous Japanese. The study provided opportunities to investigate relationships among disease frequencies, pathologic findings, and disease predictors in the cohort and to compare the findings in this population with those in other populations, especially cohorts of Japanese men resident in Japan and the continental U.S. [1]

The HHP study population was comprised of American men of Japanese ancestry born in 1900-1919 and living on the island of Oahu in 1965. Of approximately 14,000 such men believed from intercensal estimates to be available, 8006 were located and participated in a first examination that was conducted in the years 1965-68. The age range at baseline was 45-68 years.[1,17]

The first examination was followed by a second examination two years later (1967-70) in which 7498 men participated, comprising about 95% of the survivors.[18] A third examination was funded by the National Cancer Institute six years after baseline (1971-74) and enrolled 6860 men;[19] and a fourth examination was conducted 25 years after baseline (1991-93) when the 3,741 participating survivors were aged 71-93 years.[20] The National Institute on Aging conducted a fifth examination of approximately 2,705 (1994-1996) survivors focusing on dementia and its precursors which was completed in 1996. In addition to the above examinations aimed at the entire cohort, three examinations of a sub-sample of participants were conducted between 1970 and 1982 to collect more detailed lipid measurements (Lipoprotein Exams I, II and III).[21] All of these efforts were implemented through contracts with

the Kuakini Medical Center in Honolulu, which has played a central role in these landmark studies of Japanese migrants to Hawaii.

While many disease outcomes occurring in the HHP cohort were ascertained at the examinations, these were supplemented by an active surveillance program that reviewed the death certificates of Japanese men born in 1900-1919 as well as discharges of such men with selected diagnoses from all civilian hospitals on the island of Oahu. Hospital surveillance was done through December 1999, death certificate surveillance through November 2013.[22, 23, 24, 25]

4.3 Source of HHP Data Used for this Thesis

The HHP Public Use Data Set comprises 12 data files: four cohort examinations, three Lipoprotein sub-examinations, three questionnaire data files, and two surveillance files of deaths and morbid events which occurred during 1965-1994. A coding manual for each file provides variable names, the unit of measurement for measured values, and the meanings of codes for categorical variables. An identification number for each cohort member provides linkage between files. Ranges of values for measured quantities are included in the coding manuals. For certain variables, categories have been collapsed to prevent extreme, or unusual (rare) values from being utilized to identify individual cohort members. In all such cases, the corresponding variable is marked with an asterisk (*) in the coding manual and the specific details on how the variable values were modified are described. Age has been grouped to protect confidentiality.

The data sets and data documentation were provided by the National Heart, Lung and Blood Institute, NIH, for this thesis. The University of Medicine and Dentistry of New Jersey Institutional Review Board (IRB) approved this use of the Honolulu Heart Program data on August 17, 2012.

4.4 General Observations about Correlation Coefficients

A correlation coefficient measures the strength of a linear relationship existing between two continuous variables. The most commonly used correlation coefficient is called the Pearson correlation coefficient, but there are other coefficients, such as Spearman, Kendall Tau b and Hoeffding Dependence Coefficients. Most correlation coefficients vary between +1.0 (perfect positive association that allows one to specify the second variable if the first is known) and -1.0 (perfect negative association) with values near zero indicating little or no association. An important assumption concerning a correlation coefficient is that each pair of x and y data points is independent of any other pair. That is, each pair of points has to come from a separate subject.

It is common to have a statistical significance level associated with the coefficient, which gives the probability of obtaining a sample correlation coefficient as large as or larger than the one obtained when, in fact, there is no association in the underlying population. The significance of a correlation coefficient is a function of the magnitude of the correlation and the sample size. With a large number of data points, even a small correlation coefficient can be significant. But statistical significance is not the same as importance or strength. Thus, finding that a correlation coefficient between repeated measurements is statistically significant is not informative with respect to the extent of misclassification present. That depends on the magnitude of the coefficient. When the variables are normally distributed, the observed Pearson correlation coefficient can be squared to provide a sample estimate of the proportion of the variance in one of the variables that can be explained by variation in the other variable.

In evaluating a correlation coefficient, it is useful to review a scatter plot of the data to identify non-linear relationships as well as outliers. It often turns out that one or two extreme data

points can cause the correlation coefficient to be much larger (or smaller) than expected. A keypunching error in data entry can dramatically alter a Pearson correlation coefficient.

4.5 Comparison of Pearson and Spearman Correlation Coefficients as Measures of Repeatability Across a Broad Range of Variables

Although a number of different correlation statistics have been developed, the Pearson correlation coefficient is the most widely used and is a mathematically useful statistic for defining the extent of co-linearity in a bivariate normal distribution. It evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.[2] However, there often is concern that the Pearson correlation might be distorted by skewed distributions, non-linear relationships, or extreme observations in the data.

In order to evaluate whether distributional distortions are a serious problem in measuring repeatability of epidemiological variables, we sought to test whether the Pearson correlation provides results that are similar to those that would be obtained from a non-parametric correlation statistic. The Spearman coefficient is an obvious choice since it is calculated in the same way as the Pearson correlation except for using rank order data that ignore the extent of separation between ranked observations. Thus, the Spearman coefficient is less influenced by extreme observations than is the Pearson coefficient. It should be noted that establishing rank order of exposure is an important accomplishment in the context of epidemiology since it allows comparison of subjects with high exposure with those with low exposure as is done, for instance, when quintiles of an exposure are related to outcomes.

Table 4.1 compares Pearson and Spearman correlations for a wide variety of risk factor measures collected at the Honolulu Heart Study at Exam 1 (completed in 1968), with those collected at Exam 2 (completed in 1970), Exam 3 (completed in 1974), and Exam 4 (completed in 1993). The two measures of correlation are remarkably coherent. For the comparison across 2 years from Exam 1 to Exam 2, the repeatability correlations for 23 attributes were within 0.01 for nineteen, 0.02 for three and 0.05 for one. For the comparison 16 variables measured six years apart (from Exam 1 to Exam 3), the Pearson and Spearman coefficients were within 0.01 for 15 with one that differed by 0.03. For the 25-year interval (Exam 1 to Exam 4) there was a little more discrepancy, but 12 of 15 were within .02 and the maximum difference remained at .05—still very good agreement. The largest discrepancies were for serum triglyceride which has a positive skewed distribution. In 8 of the 10 discrepancies greater than .01, the Spearman coefficient was higher.

It is common practice to transform variables that are not normally distributed before calculating Pearson correlation coefficients. Such transformations do not change the rank order so will not affect the Spearman coefficients, but they usually have the effect of reducing the importance of extreme values and would be expected to bring the Pearson coefficients closer to the Spearman values, i.e. usually to raise the Pearson estimates. Using the Spearman values to estimate repeatability is likely, therefore, to give a more conservative estimate of the extent of misclassification. For studies of a particular exposure it would be desirable to compare the correlations on raw and transformed variables. Given the overall close agreement seen here, the decision as to which coefficient to use will in most instances be trivial.

Table 4.1. Correlation Coefficients Between Repeat Measures of Selected Attributes Measured as Continuous Variables at More Than One Examination. Honolulu Heart study

Attributes	Correlation Coefficient Exam 1 vs Exam 2 2 Year Interval *(Maximum N = 7498)*		Correlation Coefficient Exam 1 vs Exam 3 6 Year Interval *(Maximum N = 6860)*		Correlation Coefficient Exam 1 vs Exam 4 25 Year Interval *(Maximum N = 3741)*	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Systolic BP (Average)	0.75	0.75	0.64	0.65	0.32	0.32
Systolic BP (By nurse)	0.66	0.66	0.57	0.57	NA	NA
Systolic BP (By Physician)	0.70	0.70	0.60	0.61	NA	NA
Diastolic BP (Average)	0.70	0.69	0.60	0.60	0.25	0.24

Diastolic BP (By nurse)	0.60	0.59	0.51	0.51	NA	NA
Diastolic BP (By Physician)	0.67	0.65	0.57	0.56	NA	NA
Serum cholesterol	0.73	0.74	0.62	0.63	0.39	0.40
Triglyceride	0.56	0.61	NA	NA	0.35	0.40
Uric acid	0.71	0.72	0.59	0.62	NA	NA
Hematocrit	0.69	0.68	0.61	0.62	0.39	0.41
Height (Standing)	0.97	0.96	0.96	0.96	0.78	0.83
Weight	0.95	0.95	0.92	0.92	0.62	0.64
BMI	0.93	0.92	0.89	0.88	0.64	0.63
Skinfold Subscapular	0.77	0.79	0.76	0.77	0.39	0.39
Skinfold Triceps	0.67	0.68	0.54	0.55	0.37	0.37
Physical Activity	NA	NA	NA	NA	0.10	0.11
Cigarettes/Day	0.85	0.87	NA	NA	0.47	0.51

Years Smoked	0.85	0.85	NA	NA	0.73	0.71
Total Vital Capacity	0.87	0.87	0.85	0.85	0.58	0.57
FEV1	0.93	0.92	0.91	0.90	0.58	0.59
Arm Girth	0.68	0.66	NA	NA	NA	NA
Chest Depth	0.85	0.85	NA	NA	NA	NA
Height (Sitting)	0.72	0.70	NA	NA	NA	NA
Biacromial Diameter	0.67	0.67	NA	NA	NA	NA

All correlations are statistically significant

4.6 Repeatability correlations and elapsed time between measurements

We have argued in previous chapters that poor repeatability of an exposure measure has serious implications for the usefulness of an assessment. In Chapter 3 we also argued that correlations between repeated measures provide an indication of the extent of misclassification and can be used to adjust required significance levels and sample sizes to reduce the likelihood of false positive conclusions. However, as illustrated in Table 4.1, the correlation between repeat measurements tends to fall as the time elapsed increases.

The extent of this decline varies considerably by type of variable. Anthropometric measures which are largely determined by early adulthood and persist with minor changes thereafter tend to correlate quite well over the full 25-year period. These include height, and vital capacity and presumably would include other measures of body size if they were had been measured at Exam 4. In these data weight and BMI were also well established in most men at baseline since the subjects were all 45 or older at enrollment. Physiologic measures such as systolic blood pressure, diastolic blood pressure, serum cholesterol and uric acid change more as the time between examinations is lengthened. For these variables the correlations from exam1 to exam 4, which were separated by 25 years, are much lower than those seen for the two year and six year intervals. All correlations between Exam1 and Exam 2 for these risk factors are above 0.7 whereas all correlations between Exam1 and Exam 4 are near or below 0.4.

For correlations across the first two years there is presumably only modest change in the underlying usual values in most subjects, so that most of the observed variation is probably due to day-to-day activities, physiologic variation, measurement error and laboratory variation. In contrast the large separation in time between Exams 1 and 4 is accompanied by the shift from middle age into old age for most of these men with substantial changes in physical activity,

physiology and possibly diet, in addition to increased opportunity for changes in laboratory or clinic methods, intercurrent illness, and other influences.

It is interesting that despite the variety of personal attributes studied the extent of the decline in correlation over 25 years is strongly related to the extent of repeatability over the first two years. For instance, among the 15 attributes that were measured at Exams 1,2, and 4, the seven of those had Spearman correlations of 0.8 or better over the first two years maintain an average correlation with the baseline of 0.64 at 25 years whereas those that were correlated at lower levels over the first two years have an average correlation with baseline at 25 years of only 0.36. It is unclear to what extent this pattern can be generalized to other kinds of exposure variables such as environmental exposures or diet which were not represented in the data source used here.

As stated in Chapter 2, the goal of exposure measurement is usually to characterize some average exposure over a significant period of time that is relevant to the pathogenesis of the disease in question. If this is to be done for most subjects with a single measurement, we need to know that the measurement would get similar results if it were done at another equally appropriate time. Repeating the measure for a subsample of subjects provides an assessment of this. The repeat measurement should not be so close to the original that it artificially inflates the extent of agreement, as might be the case if it is done by the same technician (when multiple technicians are involved) or is included in the same lab run. But it should not be so remote in time that substantial secular changes in the characteristic are likely. An appropriate time interval in most instances will be dictated by good judgment concerning the probable time frame in which the exposure contributes to the pathogenesis of the disease as well as the willingness of subjects and the administrative and budget practicalities. The two-years interval

used at the Honolulu Heart Study showed variation in repeat correlation coefficients that makes sense in the light of what we know about many of these variables and was also of interest in showing that variables with high repeatability showed less subsequent variability than did those with low two-years repeatability.

4.7 Repeatability of Specific Variables

4.7.1 Blood Pressure

It is evident from Table 4.1 that systolic blood pressure (SBP) is better measured than diastolic blood pressure (DBP). This is widely accepted as the case by clinicians because SBP is determined by the cuff pressure at which blood flow sounds are first heard as the pressure is released, whereas diastolic is determined by a "muffling" of the sound which is not as clearly defined. This difference is discussed further in section 4.8 below.

4.7.2 Serum Cholesterol

Serum cholesterol has been known to be associated with CHD since the early reports from the Framingham Study. It can be measured satisfactorily without fasting as is demonstrated by the reasonable reproducibility between values at the first and the second examinations (r=0.73). The correlation across 25 years was still 0.4.

4.7.3 Serum Triglyceride

Serum triglyceride is easily affected by ingestion of a fatty meal in the 8 hours prior to blood draw. Nevertheless, it was measured without fasting at Exam 1. Although the repeatability was lower than serum cholesterol at 2 years, it turned out to be nearly as repeatable after 25 years when the measurement was done after an overnight fast. Triglyceride has not been as regularly found to be a predictor of CHD in prospective studies as has serum cholesterol.

4.7.4 Blood Sugar

Blood sugar is best measured after an overnight fast either as a straight fasting determination or after an oral glucose load. HHP subjects were not required to fast at Exam 1 and blood sugar was measured one hour after a 50g oral glucose drink. At Exam 2 blood glucose was measured in a modest subsample without the either fasting or a glucose load, while at Exam 4 the men were required to fast. Because of the non-comparability of these determinations, blood glucose has been omitted from the data presented here.

4.7.5 Serum Uric Acid

Uric acid is a metabolite of purines which are building blocks for DNA and are found in protein rich foods. Levels in the blood are influenced by diet and tend to rise in persons with kidney disease and certain other medical conditions. Uric acid was studied at Framingham and has been grandfathered into many subsequent cardiovascular studies. In general, however, it has not been found to be a strong risk factor for CHD. The reproducibility of the measurement was comparable to cholesterol. Elevated serum uric acid is a primary cause of gout.

4.7.6 Smoking

We have presented two kinds of smoking data: Number of cigarettes smoked per day and years smoked. Number of cigarettes per day was recorded in a comparable manner at exams 1,2, and 4. Non-smoking was counted as 0. Across the 25 years the smoking variables were the most consistent of the well-established, modifiable coronary risk factors.

4.7.7 Physical Activity Index (PAI)

Even though physical inactivity is a well-established risk factor for CHD [26], physical activity levels are difficult to reliably measure in individual patients[27]. In Honolulu subjects were asked how many hours in a 24-hour day they usually spent sleeping, sitting, and in light, moderate, or heavy activity. The hours in each category were multiplied by a coefficient meant to reflect the number of METS per hour the activity required and these were summed to create a physical activity index (PAI). The PAI was measured at Exam 1 and Exam 4 with the correlation across the 25 years (r=0.10) being the lowest shown in Table 4.1. Given the tracking of repeatability across the various exams and this very low correlation between exams 1 and 4, it seems likely that two-year repeatability for PAI was also low relative to the other risk factors.

4.7.8 Height, Weight and Body Mass Index (BMI)

Height is the classic example of a well measured clinical variable that is highly reproducible. Weight was also surprisingly consistent across the 25- years period, and so, of course, was BMI.

4.7.9 Skinfolds

Skinfolds were measured over the triceps muscle in the upper arm and below the tip of the shoulder blade (subscapular) by grasping and gently raising the loose skin and underlying adipose tissue and measuring its thickness with special calipers. Skinfolds are believed to be a more direct measure of adipose tissue than is BMI, which is influenced by body build, muscularity and bone. Although subscapular skinfolds are measured about as well as blood pressure and serum cholesterol, they are not as reproducible as BMI.

4.7.10 Total Vital Capacity and 1 sec. Forced Expiratory Volume (FEV1)

These lung measurements are collected by having subjects blow into a spirometer which measures the total volume of air the subject can blow out after taking a deep breath as well as the amount that is exhaled in one second. The best of three tries was recorded. The measures are surprisingly repeatable, partly because total vital capacity is related to height.

4.8 Repeatability as a Predictor of Risk Factor Status

In Chapter 3 we showed that random misclassification of exposure variables reduces the power of a study and increases the probability that an apparently significant result will be a false positive. It follows from this that poorly measured variables that confer modest relative risks will be difficult to validate in studies of ordinary size and design. To test this inference, we identified several attributes measured at HHP that are known to be related to coronary heart disease and divided them into those that were found in Honolulu and in many of the older cohort studies, and those for which there is substantial other evidence of a causal or protective role, but that have not been easy to demonstrate in cohort studies with under 10,000 subjects. In Table 4.2 we list seven variables that fall into the first category in the left hand column - measures of blood pressure, cholesterol, obesity, smoking and alcohol. In the right hand column, we list dietary and physical activity variables that are believed by most experienced observers to have contributed substantially to the CHD epidemic, but have required very large, or specialized cohort studies to demonstrate their significance at the individual level[28,29,30]. Repeatability coefficients for the first group are mostly available from the present data while data from the literature have been used to estimate repeatability of the dietary variables. We infer low repeatability for the physical activity index from its very low value at 25 years of follow-up.

Table 4.2 CHD Risk Factors Grouped by Ease with Which Their Association with CHD Has Been Demonstrated in Cohort Studies with N<10,000

Risk Factors Shown in Cohort Studies with N<10,000	Pearson Corr. 2-years apart		Risk Factors Identified from Other Evidence but Not Easy to Show in Cohort Studies with N<10,000	Pearson Corr. 2-years Apart
Systolic blood pressure	0.75		Total fat consumption	~0.4
Diastolic blood pressure	0.70		Saturated fat consumption	~0.4
Serum cholesterol	0.73		Polyunsaturated fat consumption	~0.1
HDL Cholesterol (protective)			Physical activity (protective)	low
Body Mass Index	0.93			
Subscapular Skinfold	0.77			
Cigarettes/day	0.85			
Alcohol/day (protective)	0.75			

There is a clear difference between the repeatability of the two groups of variables, which is consistent with the hypothesis that poor measurement is a likely explanation for the failure of many population-based cohort studies to demonstrate the importance of dietary fat, even

though it has repeatedly been shown experimentally to have a substantial effect on LDL cholesterol and, therefore, on coronary risk. Likewise, the evidence that physical activity has an important role in reducing risk of coronary disease is strong and consistent, and yet it has been difficult to demonstrate persuasively in population based cohort studies. The implication is that average size studies of average strength exposures that are not repeatable with correlations of at least 0.50 are unlikely to lead to clear results that hold up across multiple studies.

4.9 Conclusion

The Honolulu Heart Program data illustrate and support the practicality of applying the concepts regarding exposure misclassification that are developed in the earlier chapters. They suggest that for chronic diseases with long incubation periods that repeatability coefficients are easily calculated, that a two-years interval works well to separate repeatable risk factors as compared to less repeatable ones, and that even lengthening this to six years does not deflate the correlations excessively. Presumably a somewhat shorter time interval would also work well, although that should be explored in other data sets.

A somewhat surprising finding was the very close agreement that was seen between Pearson and Spearman correlations across the many variables that were studied. Since usual transformations of skewed variables does not change their rank order it is unlikely that transforming the data to achieve a roughly normal distribution would change the extent of this agreement substantially. The correlations across time appear to be quite robust for the personal attributes studied in Honolulu. Whether this would be the case for environmental or nutritional variables deserves some investigation.

Finally, the Honolulu data support the inference that risk factors that are reproducible and well established are repeatable. If they were not, the studies would have compromised power such

significance tests that exceeded critical values would have an enhanced probability of being false positives.

Chapter 5

Summary and Conclusions

In many epidemiological studies the risk factor or exposure of interest is measured with significant error. While differential misclassification can usually be remedied by excellent study methods, nearly all observational studies will nevertheless include some non-differential measurement error. This reflects the imperfect measurement methods that are available for many clinical and epidemiological variables.

This non-differential misclassification makes it more difficult to detect associations and it biases estimates of effect such the risk ratio or the risk difference, toward the null. As a consequence, when a statistically significant result is reported for a poorly measured variable, the claim is sometimes made that the true effect is likely to be larger than the observed effect. In one sense these claims are logical but from another perspective they are paradoxical since they imply that when a statistically significant result is found (and such results are found in most published studies), it is potentially more important if the measurement is poor!

Intuitively, one recognizes that a statistically significant result found for a very poorly measured variable must be less meaningful than a result found for a well measured variable. We point out that the reason for this is that substantial misclassification increases the probability that a significant result, when found, will be a false positive. This is obvious if one considers the extreme example where the observed measure is so bad as to be essentially random, in which case any significant result would have to be an alpha error.

Thus, the quality of exposure measurement has important implications for the probability of a statistically significant result being a false positive. This probability ranges from the nominal

significance level in the absence of measurement error, usually 0.05, to 1.0 for a variable that is known to be random. This implies that with substantial misclassification findings of modest statistical significance (e.g. $p<0.05$) are quite likely to be false positive and should be viewed with a skepticism that is not always observed in the literature. Moreover, the classic assumptions that measurement error can be modeled as an unbiased additive error term is unlikely to hold when measurement error is extensive, because, with a fixed observed variance, the large error variance can imply unrealistically constrained variation in the true value.

To address this concern, we use a bivariate normal distribution to model the relationship between the observed variable in the study sample and the true variable in the study sample. With this model the true variance is not constrained and, under the null hypothesis, the expected true z value in the sample ranges from being identical to the observed z value if the correlation between them is 1.0 to an expectation of 0 if there is no association between observed and expected (observed being essentially random). This fits the concept that the more misclassification there is, the less information about the true variable is provided by the measurement.

To implement the bivariate normal model in practice, it is necessary to have an estimate of the correlation between the observed variable and the truth. Rarely there may be a gold standard with previous duplicate assessments that can provide a correlation of the study measurements with the standard. More commonly this is not the case. In that case a minimal estimate of misclassification can be obtained from the correlation between repeated measurements. If repeated measurements are independent, then the correlation between them should result only from both being correlated with the true value. It is easily shown that under these

assumptions the correlation between the observed and true values is the square root of the correlation between independent repeat variables.

Recognizing that repeat measurements may be correlated for reasons besides their association with the truth (i.e. not completely independent) the correlations between repeated measurements may exaggerate their correlation with the underlying truth, so the square root may be too high. Then adjusting for it would be minimal adjustment for the extent of misclassification that is likely to be present.

We develop estimates of stricter requirements for observed significance levels that are needed to reduce the chance of a false positive result to acceptable levels and describe related power and sample size implications. When the correlation between repeat observations is low it implies a low correlation of the measurement with the true value and the need for large sample size increases that may make the use of such variables impractical. The model implies the need for larger sample sizes to assure that an effect associated with a misclassified variable is sufficiently unlikely to have occurred by chance that it implies the underlying true variable also shows the effect.

If it is true that significant results found for misclassified variables are likely to be false positives, then one would expect such variables, even if they are causally important, to fail to hold up repeatedly in studies of ordinary size. We use data from the Honolulu Heart Program, a large prospective study of cardiovascular disease to show that risk factors for heart attacks that have stood the test of time mostly are repeatable across a two-year period with correlations exceeding 0.7. There is extensive evidence from metabolic experiments and targeted epidemiological studies that saturated fat intake and physical activity are important underlying causes of CHD but it has been difficult to demonstrate this in general population cohort studies.

In Honolulu the repeatability coefficients for these two variables were considerably lower than 0.7, providing an illustration of the likely effects of misclassification. Thus, the Honolulu data support the inference that risk factors that are reproducible and well established tend to be repeatable. If they were not, the studies would have compromised power and significance tests that exceeded critical values would have an enhanced probability of being false positives.

In conclusion, we believe that non-differential exposure misclassification is a common problem that is all too frequently ignored in the interpretation of epidemiological studies. It would be possible in most large, well-funded field studies to recycle some participants so that measures of repeatability of all study variables would be obtained. Having this information would make investigators more demanding of quality data and would enable readers to better assess the value of the statistical inferences made on misclassified variables.

Appendices*

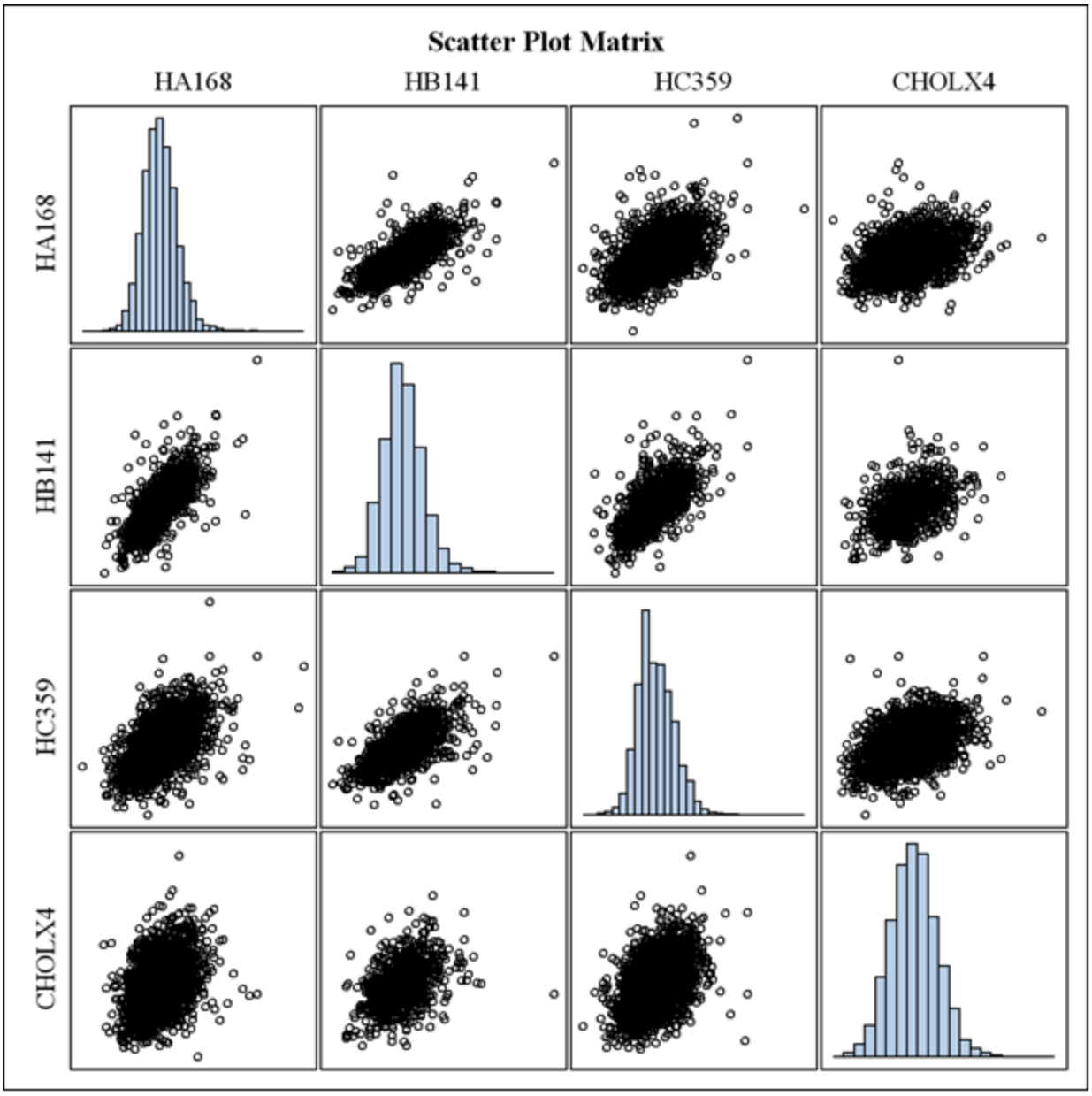
1. Examples of correlations that range from high to low

1.1 Results for *Measures of Association for HHP Cholesterol (Exam 1 to Exam 4)*

Simple Statistics							
Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
HA168	7961	218.34	38.26	216.00	51.00	537.00	SERUM CHOLESTEROL(MG%)(5:53-55)
HB141	1855	210.44	35.26	209.00	99.00	454.00	SERUM CHOLESTEROL (MG PCT) 45:22-24
HC359	6753	215.88	36.57	212.00	79.00	510.00	SERUM CHOL-MG%(932-934)
CHOLX4	3572	189.73	33.16	189.00	81.00	382.00	cholesterol at exam 4

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations				
	HA168	**HB141**	**HC359**	**CHOLX4**
HA168 **SERUM CHOLESTEROL(MG%)(5:53-55)**	**1.00** **7961**	**0.73** **<.0001** **1845**	**0.62** **<.0001** **6718**	**0.39** **<.0001** **3551**
HB141 **SERUM CHOLESTEROL (MG PCT) 45:22-24**	**0.73** **<.0001** **1845**	**1.00** **1855**	**0.66** **<.0001** **1644**	**0.44** **<.0001** **911**
HC359 **SERUM CHOL-MG**	**0.62** **<.0001** **6718**	**0.66** **<.0001** **1644**	**1.00** **6753**	**0.44** **<.0001** **3398**
CHOLX4 **cholesterol at exam 4**	**0.38976** **<.0001** **3551**	**0.44176** **<.0001** **911**	**0.43515** **<.0001** **3398**	**1.00000** **3572**

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations				
	HA168	**HB141**	**HC359**	**CHOLX4**
HA168 **SERUM CHOLESTEROL(MG%)(5:53-55)**	**1.00** **7961**	**0.74** **<.0001** **1845**	**0.63** **<.0001** **6718**	**0.40** **<.0001** **3551**
HB141 **SERUM CHOLESTEROL (MG PCT) 45:22-24**	**0.74** **<.0001** **1845**	**1.00** **1855**	**0.67** **<.0001** **1644**	**0.47** **<.0001** **911**
HC359 **SERUM CHOL-MG**	**0.63** **<.0001** **6718**	**0.67** **<.0001** **1644**	**1.00** **6753**	**0.44** **<.0001** **3398**
CHOLX4 **cholesterol at exam 4**	**0.40** **<.0001** **3551**	**0.47** **<.0001** **911**	**0.44** **<.0001** **3398**	**1.00** **3572**

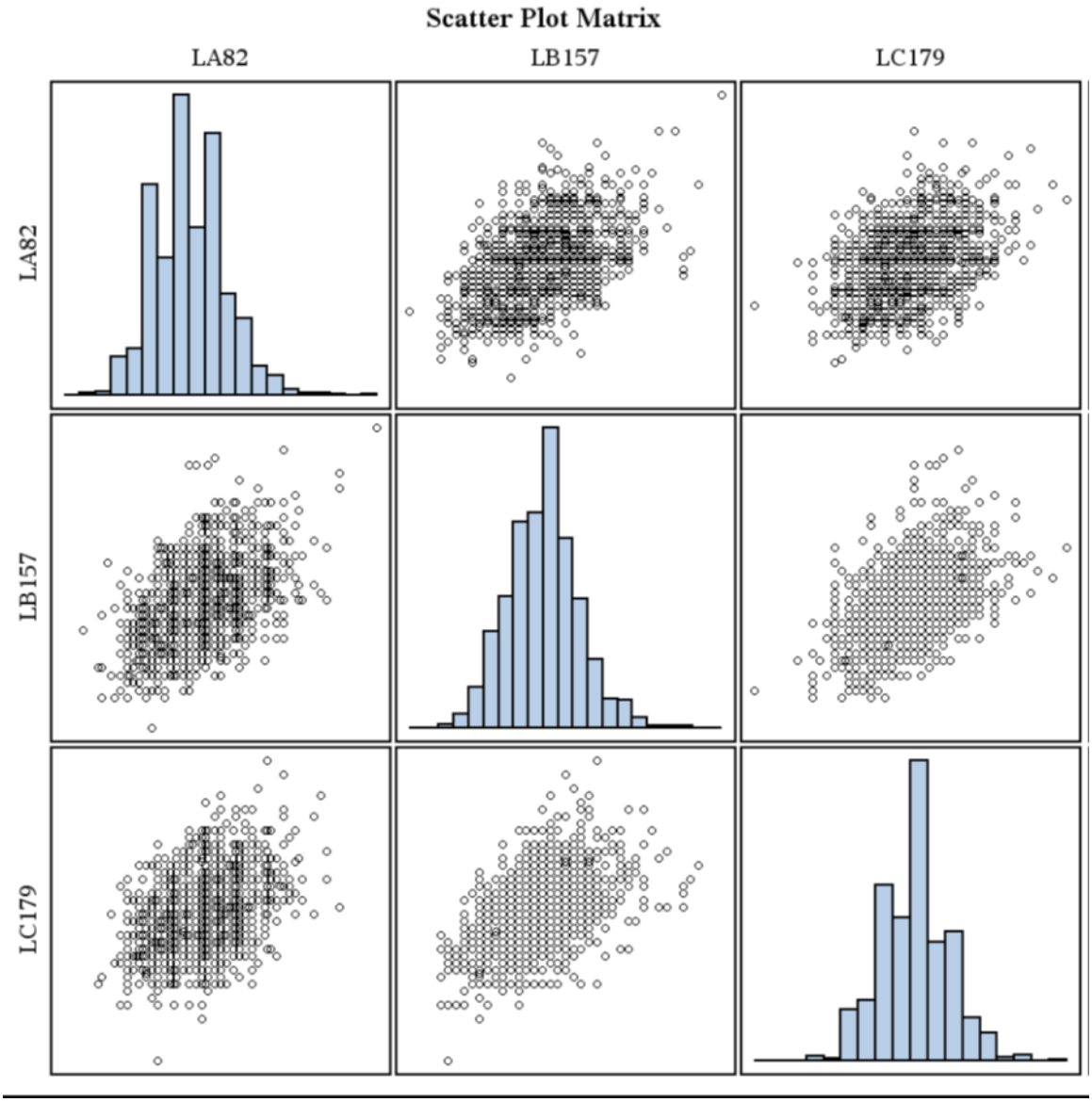

Scatter Plot Matrix
HA168
HB141
HC359
CHOLX4
HA168
HB141
HC359
CHOLX4

1.2 Results for *Measures of Association for HHP DBP (lipo 1-3)*

Simple Statistics							
Variable	N	Mean	Std	Media	Min	Max	Label
LA82	2779	82.72	11.40	82.00	46.0	140.	EXAM2 MD DIASTOL BLOOD PRES-L(74:30-32)
LB157	2385	83.16	9.64	84.00	50.0	130.	MD DIASTOLIC BLOOD PRESSURE 85:18-20
LC179	1963	80.70	9.88	80.00	38.0	124.	MD DIASTOLIC BP (MM HG) 98:26-28

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0			
	LA82	LB157	LC179
LA82 EXAM2 MD DIASTOL BLOOD PRES-L(74:30-32)	1.00	0.51 <.0001	0.42 <.0001
LB157 MD DIASTOLIC BLOOD PRESSURE 85:18-20	0.51 <.0001	1.00	0.53 <.0001
LC179 MD DIASTOLIC BP (MM HG) 98:26-28	0.42 <.0001	0.53 <.0001	1.00

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0			
	LA82	LB157	LC179
LA82 EXAM2 MD DIASTOL BLOOD PRES-L(74:30-32)	1.00	0.50 <.0001	0.40 <.0001
LB157 MD DIASTOLIC BLOOD PRESSURE 85:18-20	0.50 <.0001	1.00	0.53 <.0001
LC179 MD DIASTOLIC BP (MM HG) 98:26-28	0.40 <.0001	0.53 <.0001	1.00


Scatter Plot Matrix
LA82
LB157
LC179
LA82
LB157
LC179

References: *

1. Rhoads G.G., Reliability of diet measures as chronic disease risk factors. Am J Clin Nutr 1987; 45: 1073—1079
2. Mertens T., Estimating the effects of misclassification. Lancet 1993; 342: 418-421.
3. Dosemeci M., Wacholder S., Lubin J.H., Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epid 1990; 132:746-748.
4. Rothman K.J., Greenland S., Modern Epidemiology. 2nd Ed. Philadelphia: Lippincott-Raven; 1998
5. Jurek A.M., Greenland S., Maldonado G., Church T.R., Proper interpretation of non-differential misclassification effects: expectation vs observation Int J Epidemiol 2005; 34 (3): 680-687
6. Ellis, P.D., The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. London: Cambridge University Press; 2010
7. Tsang R.; Colley L.; Lynd L.D., "Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials". Journal of Clinical Epidemiology 2009; 62 (6): 609–616.
8. Thomas L., Retrospective power analysis. Conservation Biology 1997;11(1):276–280
9. Hoenig J.M., Heisey D.M., The Abuse of Power The American Statistician 2001: 55(1):19-24
10. http://www.encyclopediaofmath.org/index.php/Power_function_of_a_test
11. Aberson C.L., Applied Power Analysis for the Behavioral Science. New York: Taylor & Francis Group; 2010.
12. Cohen J., Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
13. Pocock S.J., Clinical trials: a practical approach. New York: John Wiley and Sons; 1995.
14. Gail M., Williams R., Byar D., Brown C., How many controls? J Chronic Dis 1976;29: 723–31.
15. Dosemeci M., Wacholder S., Lubin J.H., Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epid 1990; 132:746-748.
16. Rhoads G.G., Feinleib M., Serum triglyceride and risk of coronary heart disease, stroke, and total mortality in Japanese-American men. Arterioscler Thromb Vasc Biol. 1983; 3:316-322
17. Kagan, A., Rhoads G.G., Zeegen P.D., Nichaman M.Z., Coronary heart disease among men of Japanese ancestry in Hawaii. The Honolulu Heart Study. Israel. J. Med. Sci. 1971; 7: 1573-77

18. Gordon T., Kagan A., Garcia-Palmieri M., Kannel W.B., Zukel W.J., Tillotson J., Sorlie P., Hjortland M., Diet and its relation to coronary heart disease and death in three populations. Circulation.1981 Mar;63(3):500-15.

19. Rhoads G.G., Gulbrandsen C.L., Kagan A., Serum lipoproteins and coronary heart disease in a population study of Hawaii Japanese men. N. Engl. J. Med. 1976; 294: 293-298

20. Reed D., Yano K., Kagan A., Lipids and lipoproteins as predictors of coronary heart disease, stroke and cancer in the Honolulu Heart Program. Am. J. Med 1986; 80: 871-878

21. Rhoads G.G., Morton N.E., Gulbrandsen C.L., Kagan A., Sinking pre-beta lipoprotein in Japanese. Am. J. Epidemiol., 1978; 14: 207-12

22. Rhoads G.G., Dahlen G., Berg K., Morton N.E., Dannenberg A.L., Lp(a) lipoprotein as a risk factor for myocardial infarction. JAMA 1986; 256: 2540-2544

23. Patty W. Siri-Tarino, Qi Sun, Frank B. Hu, and Ronald M. Krauss, Saturated Fatty Acids and Risk of Coronary Heart Disease: Modulation by Replacement Nutrients Curr Atheroscler Rep. 2010 Nov; 12(6): 384–390.

24. MacMahon S., Peto R., Cutler J., Collins R., Sorlie P., Neaton J., Abbott R., Godwin J., Dyer A., Stamler J., Blood pressure, stroke, and coronary heart disease. Part I: Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet 1990; 335:765-774.

25. Grundy S.M.,Blackburn G., Higgins M., Lauer R., Perri M.G., Ryan D., Physical activity in the prevention and treatment of obesity and its comorbidities. Medicine and Science in Sports Exercise 1999 Nov; 31(11):1493-1508

26. Paffenbarger R.S.Jr., Hale W.E., Work activity and coronary heart mortality. N Engl J Med. 1975; 292:545-50

27. Stampfer M.J., Hu F.B., Manson J.E., Rimm E.B., Willett W.C., Primary prevention of coronary heart disease in women through diet and lifestyle. N Engl J Med. 2000; 343:16-22.

28. Sesso H.D., Paffenbarger R.S.Jr., Lee I.M., Physical activity and coronary heart disease in men: The Harvard Alumni Health Study. Circulation 2000; 102; 975-80

29. Paffenbarger R.S.Jr., Wing, A.L., Characteristics in youth predisposing to fatal stroke in later years. Lancet 1967; 1:753-4

30. Oguma Y., Sesso H.D., Paffenbarger R.S.Jr., Lee I.M., Weight change and risk of developing type 2 diabetes. Obes Res. 2005; 13: 945-51