UNCERTAINTY OF THE DEPENDENT VARIABLE IN GENOME WIDE ASSOCIATION STUDIES

by

SHANNON NICOLE SMITH

(Under the Direction of Romdhane Rekaya)

ABSTRACT

The use of high throughput technology has the potential to provide insight into the underlying biological mechanisms of several important complex traits. Unfortunately this information comes at a cost, sometimes in terms of low accuracy and poor quality of the data. One application of genomics in humans is the diagnoses of diseases such as bipolar disorder, Alzheimer’s disease, and cancer. However in order to correctly predict disease statuses, methods should be trained on datasets containing no errors. In the case for most conditions this is seldom the case, as mis-diagnostic errors are prevalent due to overlapping symptoms and lack of precise diagnostic tools. Therefore, a new approach for dealing with misclassification was developed and applied to simulated data sets where case and control observations were randomly switched and with varying odds ratios of influential SNPs, to examine the effects of potential misclassification on diagnostic accuracy. The cases when misclassification was ignored resulted in limited predictive power of the model. When the misclassification algorithm was applied, the predictive power increased across all scenarios demonstrating the effectiveness of the misclassification algorithm.
Additionally in livestock applications, genomic technology is used to detect genetic variants associated with economically important traits as well as to estimate genomic enhanced breeding values to be used in animal selection. For genome wide association studies (GWAS) in animal applications the dependent variable is often a ‘pseudo’ phenotype (estimated breeding values, de-regressed breeding values, etc.). Being estimates, these pseudo-phenotypes carry a certain level of inaccuracy or uncertainty. In some situations, such uncertainty is large and it is not constant across observations. Consequently, using these estimates directly as dependent variables in the GWAS can be problematic because the residual variance of the model is composed of two components (sampling variance and the error variance) that current methods are unable to accommodate. Thus, we developed and implemented a new procedure that correctly accounts for both components of the residual variance leading to an increase in accuracies of the estimated genomic breeding values. The proposed method was evaluated with real and simulated data.

INDEX WORDS: Genomics, misclassification algorithm, discrete responses, estimated breeding values, accuracy
UNCERTAINTY OF THE DEPENDENT VARIABLE IN GENOME-WIDE ASSOCIATION STUDIES

by

SHANNON NICOLE SMITH
B.S., The University of Tennessee, 2006
M.S., The Ohio State University, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013
UNCERTAINTY OF THE DEPENDENT VARIABLE IN GENOME-WIDE ASSOCIATION STUDIES

by

SHANNON NICOLE SMITH

Major Professor: Romdhane Rekaya
Committee: Joseph Keith Bertrand
            Ignacy Misztal
            Samuel Aggrey

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2013
DEDICATION

I would like to dedicate this work to my mother and grandmother, Beverly and Karen Eushery, whose endless love and encouragement served as the support needed to carry out this journey.
ACKNOWLEDGEMENTS

I express my appreciation to Dr. Romdhane Rekaya for his guidance, and financial and academic support throughout my course of study and especially for the assistance provided in the preparation of this dissertation.

I wish to thank my committee members, Drs. Keith Bertrand, Sammy Aggrey, and Ignacy Misztal for their intellectual support, encouragement, and enthusiasm, which made this dissertation possible, and for their patience in correcting both my stylistic and scientific errors.

I would like to give a special thank you to El Hamidi Hay for helping me with the calculations and data analysis, and thank you for being a listening ear and a comedic outlet that helped in keeping me sane these past few years.

To my mother Beverly, my grandmother Karen, and my Godbrother Tarvin, words cannot express how much I appreciate them for supporting and keeping me focused when times became overwhelming. I thank them for their unconditional love and unwavering faith in me as I continue to strive for excellence and make them proud. Lastly, to my Aunt Carolyn, though she is no longer with us, her encouraging words are forever engraved in my heart and memories of her gave me strength at times they were needed.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2 REVIEW OF LITERATURE</td>
<td>3</td>
</tr>
<tr>
<td>3 GENOME WIDE ASSOCIATION IN THE PRESENCE OF MISCLASSIFIED BINARY</td>
<td>24</td>
</tr>
<tr>
<td>RESPONSES</td>
<td></td>
</tr>
<tr>
<td>4 ANALYSIS OF BINARY RESPONSES WITH OUTCOME SPECIFIC MISCLASSIFICATION</td>
<td>50</td>
</tr>
<tr>
<td>PROBABILITY IN GENOME WIDE ASSOCIATION STUDIES</td>
<td></td>
</tr>
<tr>
<td>5 UNCERTAINTY IN CONTINUOUS DEPENDENT VARIABLES AND ITS EFFECT ON</td>
<td>76</td>
</tr>
<tr>
<td>GENOME WIDE ASSOCIATION STUDIES</td>
<td></td>
</tr>
<tr>
<td>6 CONCLUSIONS</td>
<td>91</td>
</tr>
<tr>
<td>APPENDIX</td>
<td></td>
</tr>
<tr>
<td>A RESULTS FOR THE SIMULATION SCENARIO WHERE MISCLASSIFICATION Rates</td>
<td>93</td>
</tr>
<tr>
<td>WERE SET TO 5% AND 5%</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

Over the past decade there has been rapid advancement in technologies that enable the collection of genomic data for both livestock and human populations. These technologies allow for capturing thousands of single nucleotide polymorphisms (SNPs) for better understanding of the genetic basis underlying common phenotypes. This led to the development of genome wide association studies (GWAS) which aim at testing thousands of common variants across the genome among groups of individuals in order to identify their association with the trait of interest. The success of these techniques is dependent on how well the amount of error is controlled in the data.

Issues in data quality are a common problem in both human medicine and animal production. These problems can arise in the genotypic data, and more specifically, misclassification and high uncertainty can occur in the dependent variables (phenotypic measurements) as well. In human medicine practices these misclassification errors are typically prevalent in neurological disorders due to their overlapping symptoms and subjective diagnoses. In animal agriculture application the dependent variables are usually pseudo-observations such as estimated breeding values (EBVs). These observations therefore have an additional level of uncertainty attached to them adding another component to the residual term of the model which can decrease the accuracy of the estimate if not adjusted for.

If these concerns are disregarded, it will result in damaging effects of the overall goal of GWAS. Many variants have small effects on the phenotype accounting for a small fraction of the
genetic variance. Consequently, any errors that exist in the data can easily lead to a decrease in prediction power of the model, as well as failure to identify or select the true causal variants associated with the trait of interest due to added noise to the system.

It is well known that accuracy of the phenotypes is essential for the validity of GWA studies in addition to allocating the correct protocol for treatment. Therefore, proper data quality control measures should be applied to reap the full benefits of GWAS. Unfortunately, accurately assigning individuals to disease classes is difficult and sometimes costly. Researchers have suggested SNPs can be useful in evaluating the association between discrete responses and genetic variations but published results have shown inconsistent results. The top potential reasons for these results include heterogeneity and potential misclassification of the discrete observations. Thus, using these phenotypes to determine genes which distinguish between disease classes could lead to erroneous results.

These issues are ones that need to be addressed in order to capitalize and take full advantage of the benefits of GWAS and prevent erroneous results. Computationally efficient methodologies should be developed to account for potential misclassification of traits which are difficult to measure. Statistical procedures should also be implemented in order to increase the accuracy of estimated dependent variables. Therefore the goals of the current studies are to 1) evaluate the effects of misclassification of binary responses on the results of association studies 2) develop and implement a new procedure for dealing with misclassification of binary responses in GWAS 3) deal with the uncertainty in continuous dependent variables and its effect on GWAS results.
CHAPTER 2
REVIEW OF LITERATURE

Genomic Information

Most traits of importance in human medicine and agricultural industries are complex (quantitative) in nature. Unlike simple traits (influenced by a single or a few genes), complex traits are under the collective control of environmental influences and the additive effect of many genes and sometimes the interaction between these two factors. Thus, they do not possess a one to one, genotype: phenotype relationship. Examples of these traits are mental illnesses such as bipolar disorder and Alzheimer’s disease in humans; milk production and reproductive traits in livestock; and corn ear length and seed weight in plant species. Due to these traits complexity, identifying genes associated with them would be beneficial in forming an understanding of the genetic architecture underlying them.

Genetic information can also be incorporated into selection methods. Traditionally, selection is carried out through the use of estimated breeding values (EBVs), which are measurements describing the genetic merit of an individual calculated from its phenotypic records and pedigree relationships. The use of these values has found success but comes with its disadvantages during data collection. For instance, measuring milk yield is only possible through females in dairy cattle creating a large generation interval, which in turn decreases the rate of genetic change. Therefore, identifying genes related to the trait would be beneficial as selection of animals for the desirable alleles will be possible at an early age.
Once the sequencing of the human genome was complete in 2003, genotyping techniques became fairly simple and inexpensive to implement creating the availability of large SNP panels across several species. All species have considerable variation with single nucleotide polymorphisms being the most frequent occurring across the genome, as more than 9 million have been reported in public databases (Rocha et al., 2006; Altshuler et al., 2005). SNPs occur in both coding regions, which alter the encoded proteins, and non-coding regions which can be used as physical markers in evolutionary and comparative genomic studies (Nickerson et al., 1998; Li and Sadler, 1991). In human applications researchers aim to identify SNPs which are associated with a certain disease status using case-control studies comparing differences in allele frequencies between the healthy and diseased groups (Tost and Gut, 2005; Emhazion et al., 2001; Schork et al, 2000). SNP identification is also implemented in pharmacogenomics studies with the purpose of clarifying differences in drug responses of individuals with different genotypes (Pirmohamed and Park, 2001; McCarthy and Hilfiker, 2000).

In plant and animal breeding, SNP information is used for genetic improvement through genomic selection leading to increases in the rate of genetic progress (Heffner et al., 2009; Meuwissen et al., 2001). Genomic selection is the selection of animals, or plants, based on their genomic enhanced breeding values (GEBVs) which have been calculated from the joint effects of genetic markers across the whole genome. Fernando and Grossman (1989) first demonstrated the inclusion of marker information in breeding values and Meuwissen and Goddard (1996) later predicted that this addition would increase genetic gain by 8-38%.
Use of the genetic information

Candidate Gene Approach

Before the development of GWA studies, researchers commonly carried out two different methods for gene discovery and association studies; the candidate gene approach and QTL mapping. Candidate gene approach studies are based on hypotheses about a relationship between known loci and phenotypes. It has been applied to gene-disease research, genetic association studies, and drug target selection in both animal and human populations (Tabor et al., 2002). Some prior knowledge of the gene’s biological functional impact on the trait of interest is needed in order for this process to be implemented. For instance, researchers found when studying alcoholism it is known that the enzyme alcohol dehydrogenase is involved in the pathways of alcohol metabolism (Kwon and Goate, 2000). Therefore it is reasonable to use genes which code for these enzymes when studying alcoholism. Another example, Mazeyrat et al. (1998) identified genes that are induced during plant defense and these genes were then considered as disease resistance loci.

A limitation surrounding the candidate gene approach is that it relies on the assumptions of the underlying molecular mechanisms that influence the development of a trait. Initial ideas about these mechanisms are often proven wrong because usually not much is known about causes of the disease which prompts the reason for them to be studied. This lack of knowledge leads to an information bottleneck as well as subjective decisions of which candidate genes to include in the analysis (Zhu and Zhao, 2007). In addition, the candidate gene approach is very time consuming and has rarely been successful.
QTL Mapping

Researchers then moved to an alternate approach where genes affecting a trait are mapped to a chromosomal region using genetic markers in order to locate the quantitative trait loci (QTL; Dekkers and Hospital, 2002; Van Laere et al., 2003; Andersson and Georges, 2004). This process is carried out through the use of linkage analysis and capturing the potential linkage disequilibrium (LD) between the markers and QTLs. Early efforts at mapping QTLs began with using blood groups as genetic markers (Neimann and Robertson, 1961; Rendel, 1961). However, identification of variable microsatellite markers led to a more powerful approach.

When performing linkage analysis using microsatellite markers large half-sibling families are used (Georges et al., 1995). This approach was more feasible and powerful in livestock applications than for use in human populations due to the ability of producing hundreds of offspring per sire. There have been a few quantitative traits completely analyzed from QTL mapping to gene identification. For example, in tomatoes, the ORFX gene was found to have an effect on fruit weight (Fray et al., 2000) and the MOM1 gene is known to affect the multiplicity and size of tumors in mice (Cormier et al., 1997).

The problem with linkage analyses is the difficulty in detecting and localizing the underlying mutation. For instance, if a particular region of the genome contains many genes that are highly correlated with one another and are associated with a trait, it becomes difficult to separate the genetic effect from its surroundings (Darvasi and Pisante-Shalom, 2002). As a result, QTLs become mapped to large confidence intervals of 20 centimorgans (cM) or more (Darvasi et al., 1993; Georges et al., 1995; Van Laere et al., 2003) making it infeasible to identify the location of the polymorphism.
In order to increase mapping precision, researchers began to exploit the LD between markers and QTLs. LD is caused by the inheritance of identical chromosome segments by multiple descendants from a common ancestor. Therefore, LD mapping is family dependent. The observed LD pattern of a population is an indication of its effective population size ($N_e$). The smaller the $N_e$, the less generations it takes for the current population to have an ancestor in common. So, fewer generations of recombination before LD extends to long distances creating large identical by descent (IBD) chromosmomse segments.

One issue with implementing LD is that it does not apply across breeds in livestock animals due to different breeds not having a recent common ancestor. This issue also arises in human populations as there is minimum LD at long distances. Due to this observed pattern in domesticated animals, it is possible for a marker to be in LD with a QTL some distance away but still capable of showing association to the trait preventing precise mapping. Meuwissen and Goddard (2004) published results which incorporated both linkage analysis to map a QTL to a larger region and LD to map it more precisely (LDLA). This technique has successfully been applied in dairy cattle (Druet et al., 2008; Meuwissen et al., 2001) however estimating a confidence interval surrounding the QTL was problematic. This became more practicable with the advancement of genome-wide SNP panels.

*Genome-wide association studies*

It was proposed in order to identify low penetrance genes, population-based association studies of unrelated individuals would be more powerful than family-based linkage methods (Risch and Merikangas, 1996). In order to accomplish this goal, individuals should be genotyped for a large number of variants across the genome which will be tested for association with the
phenotype. Over the past decade, through the use of SNPs, it is now possible to saturate the genome with markers. These SNPs are mutations which occur at a single base and accounts for approximately 90% of genetic variation. They occur every 100 to 300 base pairs allowing for high resolution mapping and implementation of GWA studies.

There are several essential components necessary to carry out a GWA study. The first element is an adequate number of samples that will be representative of the genetic information of the targeted population. This is necessary as sample size is one of the influencing factors of statistical power and a sizable number is needed for identification of causal variants with modest effects (Ioannidis, 2010; Spencer et al., 2009). In human genetic studies, most associations have effect sizes with odds ratios ranging between 1.1 and 1.3 which means the risk of getting the disease is 10% and 30% higher respectively, compared to individuals without the disease associated allele (Goldstein, 2009). Researchers have since determined that GWA studies are underpowered to detect modest effects which raise the risk of developing a disease by smaller values (less than 5%; Risch and Merikangas, 1996; Spencer et al. 2009; Stahl et al. 2010). However, these small associations may be of importance in understanding the biological basis underlying complex phenotypes.

In order to establish genetic associations, enough polymorphic alleles that cover the whole genome and that can be simply and reasonably genotyped are needed. Due to the advancements in high-throughput technologies and contributions to HapMap projects, there have been substantial numbers of SNPs identified throughout the entire genomes across several species. For example, product development by companies have allowed commercially available SNP chips for cattle (50,000 SNPs), dogs (22,362; Illumina CanineSNP20 BeadChip), sheep (56,000 SNPs), chickens (60,000 SNPs), pigs (60,000 SNPs; Illumina PorcineSNP60 Beadchip).
and increases to an exceedingly large number (1,000,000 SNPs) in humans. These tools have evolved making genotyping more affordable in addition to providing dense coverage across the genome.

As there will be thousands of samples collected and genotyped for the use of examining thousands of genetic variants, analytical methods which are statistically powerful and unbiased are needed in determining genetic associations (Risch, 2000; Bush and Moore, 2012). There are two different approaches; frequentist and Bayesian. Frequentist inferences of population associations between genetic variants and phenotypes are based on the computation of p-values for the null hypothesis that there is no association. Bayesian methods use an alternative approach from p-values by implementing additional model assumptions about effect sizes of truly associated SNPs (Beaumont, 2004; Marjoram and Tavare; 2006; O’Hara et al., 2008).

An example of frequentist approach is the single locus analysis. This involves a series of single-locus statistic tests examining each variant independently for association to a specific trait. The statistical test carried out is dependent on the nature of the phenotype. For instance, quantitative traits are analyzed using generalized linear model approaches most commonly the Analysis of Variance (ANOVA). The null hypothesis under ANOVA using a single SNP is that there is no difference between the trait means of any of the genotype groups (Balding, 2006). Binary traits in case/control studies are generally analyzed through the use of logistic regression which extends from linear regression. The outcome is transformed using a logistic function which predicts the probability of having a case status given a genotype class. This procedure adjusts for covariates and provides adjusted odds ratios as a measure of effect size (Prentice and Pyke, 1979).
In the case of Bayesian analyses, models assume prior distributions for the parameters. Priors are dependent on hyperparameters which are also drawn from some distribution. The goal is to determine the posterior probability of association for each SNP and use them to rank the SNPs for further analysis. Calculating the marginal likelihood of the genotypes and covariables to produce the most probable hyperparameters, then by conditioning on these values and the observed data the probability of association of each SNP is calculated (Stephens and Balding, 2009).

Genomic Selection

In animal and plant breeding, the main objective is to better select and predict genetic merit of potential breeding candidates. Once data collection and statistical analyses are run, and genetic associations are found, plant and animal breeding programs implement selection strategies. Traditionally, superior animals (plants) are selected based on their estimated breeding values calculated using a “black box” approach through the use of best linear unbiased prediction (BLUP; Henderson, 1984) procedures. BLUP assumes an infinitesimal model (Fisher, 1918; Bulmer, 1980); implying traits are determined by an infinitely number of additive genes of infinitesimal effect. Therefore, rather than targeting specific genes, it takes all genes into consideration when estimating genetic merit.

In 2001, Meuwissen et al. first suggested the use of genomic selection. Under this process genetic information is included in the calculation of breeding values by directly making use of the SNPs in the genome-wide panel that are in LD with the QTLs. This is an advantage over classical BLUP as the estimation of genetic merit of potential candidates includes pedigree, marker, and phenotypic information, and is currently successfully being used in livestock and
plant applications (Dekkers, 2004; Schaeffer, 2006; Jannink et al., 2010). Another advantage is being able to estimate breeding values at an earlier age, reducing the generation interval. And in comparison to linkage mapping, the estimated marker allele effect can be applied population wide rather than restricted to within families.

Two approaches can be implemented to carry out genomic selection, multiple-step and single-step. Genomic evaluations calculated through the multiple-step procedure are broken into several stages (VanRaden, 2008; Hayes et al., 2009). First a large sample of animals (plants) is needed that have been recorded for the trait of interest as well as genotyped for the panel of markers. This group of individuals is the reference population used to develop an equation which predicts the breeding values from the marker genotypes (GEBVs). The genotype variables for each individual can take on values 0, 1, or 2 corresponding to the number of minor alleles present at each SNP. Statistical analysis of the reference population estimates the effects of each marker to combine with the marker genotypes to predict the GEBV of each observation. The estimation of the SNP effects can occur in several different ways based on the assumptions of their prior distributions. For example, one can assume all SNPs in the panel are influential but have different variances (Bayes A) or it can be assumed with probability $\pi$ that some SNPs are relevant while others are not (Bayes B). Afterwards, the prediction equation is applied to ungenotyped individuals to select the best candidates for breeding.

The single-step procedure performs a joint evaluation using all phenotypic, pedigree, and genomic information (Legarra et al., 2009; Misztal et al., 2009) bypassing the estimation of SNP effects and calculating GEBVs directly. The numerator relationship matrix, $A$, from Henderson’s mixed model equations that includes pedigree relationships, is modified in order to incorporate genomic information. This allows for the inclusion of both genotyped and ungenotyped
individuals, where the genetic values of ungenotyped animals are a result from the conditioning on genotyped individuals using the relationship matrix. This gives rise to a pedigree-genomic matrix $H$, which is a joint distribution of genotyped and ungenotyped genetic values.

**GWAS Results**

**Achievements**

There has been much success following the introduction of GWA studies and genomic selection. In human applications, SNP panels have been used to detect disease related genes for a variety of conditions. Hugot et al. (2001) used microsatellite markers and SNPs to examine a region on ch16 believed to be associated with Chrohn’s disease. They detected 3 SNP mutations associated with Chrohn’s disease located in the coding region NOD2, a gene known to be involved in the activity of microbial pathogen receptors. It is well known that the apolipoprotein E (APOE) allele is a risk factor for late-onset Alzheimer’s disease (AD). Coon et al. (2007) used ultra-high density SNP genotyping (502,627 SNPs) and found SNP rs4420638 on chromosome 19 in linkage disequilibrium with to the APOE variant supporting the locus is the major susceptibility gene for late onset AD. Progress has even been made in understanding the progression and susceptibility of infectious diseases such as HIV/AIDS. A 32 base-pair deletion was found in the CCR5 gene that led to restriction of HIV in individuals homozygous for this mutation. It introduces a premature stop codon blocking cell entry of HIV due to truncating the protein (Agrawal et al., 2007; Rusconi et al., 2007).

Although the GWAS movement has been established by human geneticists, it and genomic selection are gradually being used in plant breeding programs. It is an attractive method for plant breeders as most plant species are inbred lines and once genotyped, they can be
phenotyped multiple times allowing for studying of different traits in different environments.

Atwell et al. (2010) applied GWAS to *Arabidopsis thaliana*, a self-fertilizing plant that is known to contain a substantial amount of genetic variation. They took a sample size of 192 *Arabidopsis* lines and detected many QTL and effect sizes were large for 44 out of 50 phenotypes. GWAS analysis was also ran 14 agronomic traits in rice landraces (Huang et al., 2010). They constructed a high-density haplotype map from the identification of approximately 3.6 million SNPs and were able to detect six loci which explained 36% of the phenotypic variation.

In animal breeding, SNP marker maps have been used to identify genomic regions associated with economically important traits including fertility, growth, milk quality and yield, meat quality and carcass traits. In dairy cattle, several reports have been published identifying 734 SNPs significantly effecting milk yield (Hayes et al., 2009; Mai et al., 2010; Bolormaa et al., 2010). In pigs, Duijvesteijn et al. (2010) conducted a GWA study using Illumina Porcin 60K SNP Beadchip and genotyped 987 pigs divergent for androstenone concentration, and found 37 SNPs, mainly on porcine chromosomes 1 and 6, significantly affected androstenone levels in fat tissue. In layer chickens, egg production and quality are of substantial importance. Lieu et al. (2011) carried out a study using White Leghorn and Brown-Egg Dwarf layers and resulted in 8 SNPs significantly associated with egg production located in genes (GRB14 and GALNT1) known to affect the function and development of the ovary. Stone et al. (2005) examined bovine chromosome 5 for association with carcass traits in beef cattle. After correcting for multiple testing, 2 haplotypes at the phosphodiesterase 1B locus were found to be significantly associated with fat thickness and rib fat.

In addition to identifying causative variants, evaluations have been able to estimate genomic breeding values which may allow the selection of younger animals, consequently
reducing generation intervals and increasing response to selection. Meuwissen et al. (2001) simulated a population with markers placed every 1 CM with each flanked region containing a QTL. They applied a Bayesian methodology and found the accuracy (correlation between genomic and true breeding values) of genomic breeding values to be 0.85. VanRaden et al. (2009) applied genomic selection to a real data set of Holstein dairy cattle using a reference population of 3,576 bulls genotyped for 38,416 SNPs and found the accuracy to be 0.71. Several other researchers have reported similar findings (Hayes et al., 2008; Gonzalez-Recio et al., 2009).

Challenges

Though there has been rapid advancement in genotyping technologies and much success has been made through association studies, as the number of SNPs available in public databases for the human genome reaches over 1,000,000 (Sachidanandam et al., 2001), there are still several challenges and obstacles that need to be overcome in order to fully understand susceptibility and biological basis underlying complex phenotypes. Hirschhorn et al. (2002) conducted a comprehensive review of genetic association studies for the most common human diseases, and found that of the 166 published putative associations only six have been replicated consistently. They listed several leading causes behind these non-replicative and biased results including population stratification, lack of power, varying linkage disequilibrium between the SNP and the true causal variant, and misclassification and uncertainty surrounding the dependent variable. Some of these complications also arise when being applied to animal and plant breeding programs.
Association studies for human applications are generally conducted through a case/control design which can unintentionally lead to ethnic admixture in the sample population. This occurs when individuals are being drawn from different ethnic groups or subgroups. Admixture also occurs in animal breeding studies as study samples can be from a mixture of breeds or consisted of crossbred animals. The issue with running GWAS when admixture exists in the sample is that the pattern and level of LD varies across populations, as well as allele and genotypic frequencies (Reich et al., 2001; Smith et al., 2001). Also, some mutations may not exist in certain populations, such as the NOD2 mutation known to affect more than 30% of Crohn’s disease patients of European descent but absent in Asian populations (Mathew and Lewis, 2004). In admixed human populations, if one of the groups has higher disease prevalence than the others, it will be overrepresented in cases and underrepresented in the control group. Therefore, if a variant is of high risk in that group it will exhibit association to the disease leading to false positive results (Teo et al., 2010; Marchini et al., 2004; Smith et al., 2004).

Linkage disequilibrium is important when dissecting the biological basis of complex traits as high-density SNP panels are designed to capture the potential LD between the SNP and the QTL. Therefore, the polymorphisms being tested are not necessarily the causal variants. As mentioned before, the genetic architecture changes across different populations, so there’s a possibility for an SNP to be in LD with a nearby disease allele in one population (breed) and not in the other.

Another fundamental difficulty to overcome of GWA studies is their inability to detect the true association of small effects of rare alleles due to lack of sufficient power. For instance, an SNP with 10% frequency with a relative risk of 1.3, in order to achieve 80% power 1146 cases and controls are needed (Purcell et al., 2003). Underpowered studies will fail to reliably
detect weak genetic effects. If a weak effect is found in a study ran with a small sample size, it most likely due to overestimating the strength of the effect due to sampling variation. Therefore, to increase power of detection meta-analyses or large studies will be required to determine genetic associations.

Also, contributing to low statistical power is the multiple-testing problem. It is common to test hundreds of thousands of genetic markers simultaneously. Each statistical test has a type I error probability attached to it, thus when testing a large number of SNP associations true associations will be attached with a large number of spurious associations and white noise increasing the false discovery rate (Johnson et al., 2010; Hoggart et al., 2008). Many techniques have been used to control the error rate, such as Bonferroni correction and Holm’s step-down procedure (Holm, 1979), but the stringency of these methods increases the frequency of type II errors (false negatives). Therefore, practical statistical procedures for controlling false discovery and family-wise errors need to be explored.

During data collection, proper quality control measures should be taken in order to help maintain accuracy of the phenotypic and genotypic data. Even minor errors in the measurements can compromise statistical associations leading to false negative or false positive results. Several allele calling algorithms and genotyping technology are available for continued improvement of genetic data quality strategies (Laurie et al., 2010; Weale, 2010; Chanock et al., 2007). Misclassification and uncertainty of the dependent variables (phenotypes) are also cause for errors. In clinical data, identifying errors is not simple and even if misdiagnosis is suspected retesting can be costly and time exhaustive. Also, if statistical analysis is run on estimated variables, as is the case for breeding values used in genomic selection, it is necessary to account for multiple sources of variation. Hence, implementing algorithms that can account for probable
misclassification and statistical procedures that will increase accuracy of continuous traits will be advantageous.

**Literature Cited**


CHAPTER 3

GENOME WIDE ASSOCIATION IN THE PRESENCE OF MISCLASSIFIED BINARY RESPONSES

1

Abstract

Misclassification has been shown to have a high prevalence in binary responses in both livestock and human populations. Leaving these errors uncorrected before analyses will have a negative impact on the overall goal of genome-wide association studies (GWAS) including reducing predictive power. A liability threshold model that contemplates misclassification was developed to assess the effects of mis-diagnostic errors on GWA studies. Four simulated case-control datasets consisting of 2000 individuals each were analyzed with varying odds ratios of the influential SNPs and misclassification rates of 5% and 10%. Analyses of binary responses subject to misclassification resulted in underestimation of influential SNPs and failed to estimate the true magnitude and direction of the effects. Once the misclassification algorithm was applied there was a 12% to 29% increase in accuracy, and a substantial reduction in bias. The proposed method was able to capture the majority of the most significant SNPs that were not identified in the analysis of the misclassified data. In fact, in one of the simulation scenarios, 33% of the influential SNPs were not identified using the misclassified data. However, using the proposed method, only 13% were not identified. Furthermore, the proposed method was able to identify with high probability a large portion of the truly misclassified observations. The proposed model provides a statistical tool to correct or at least attenuate the negative effects of misclassified binary responses in GWAS. Across difference levels of misclassification probability as well as odds ratios of significant SNPs, the model proved to be robust. In fact, SNP effects, and misclassification probability were accurately estimated and the truly misclassified observations were identified with high probabilities compared to non-misclassified responses. This study was limited to situations where the misclassification probability was assumed to be the same in cases and controls which is not always the case based on real human disease data. Thus, it is of interest
to evaluate the performance of the proposed model in that situation, which is the current focus of our research.

**Keywords**

Misclassification, Genome wide association, Discrete responses

**Introduction**

Misclassification of dependent variables is a major issue in many areas of science that can arise when indirect markers are used to classify subjects or continuous traits are treated as categorical. Binary responses are typically subjective measurements which can lead to error in assigning individuals to relevant groups in case-control studies. Many quantitative traits have precise guidelines for measurements but in qualitative diagnosis different individuals will understand conditions in their own way [1]. Some disorders require structured evaluations but these can be time consuming and very costly and not readily available for all patients [2]. This sometimes requires clinicians to use heuristics rather than following strict diagnostic criteria [3], leading to diagnoses based on personal opinions and experience. It was found that physicians will disagree with one another one third of the time as well with himself (on later review) one fifth of the time. This lack of consistency leads to large variation and error [4, 5]. Researchers indicated that there is a common assumption under most approaches that disorders can be distinguished without error which is seldom the case [6]. For instance, a longitudinal study was carried out over 10 years where 15% of subjects initially diagnosed with bipolar disorder were re-diagnosed with schizophrenia, whereas 4% were reclassified in the opposite direction [7]. Reports have shown an error rate of more than 5-10% for some discrete responses
In some instances, these rates have proven to be significantly higher. The frequency of medical misdiagnosis and clinical errors has reached error rates as high as 47% as documented in several autopsy studies [10]. Error rates in clinical practices have shown to be higher than perceptual specialties [11], but still these areas have demonstrated high rates as well. In radiology areas, failure to detect abnormalities when they were present (false negative) ranged between 25-30%, and when the cases were normal but incorrectly diagnosed as diseased (false positive) ranged between 1.5-2% [12]. Some stated that these errors are not due to failure of not showing on film but due to perceptual errors [13]. These findings are similar to recent published studies [2, 5, 14, 15].

Unfortunately, finding these errors in clinical data is not trivial. Even in the best case scenario when well-founded suspicion exists about a sample, re-testing is often not possible and the best that could be done is to remove the sample leading to power reduction. Recently, several research groups [16-18] have proposed using single nucleotide polymorphisms (SNPs) to evaluate the association between discrete responses and genomic variations. Genome-wide association studies (GWAS) provide researchers with the opportunity of discovering genomic variations affecting important traits such as diseases in humans, and production and fitness responses in livestock and plant species. Several authors have indicated that the precision and validity of GWAS relies heavily on the accuracy of the SNP genotype data as well as the certainty of the response variable [19-24]. Thus, analyzing misclassified discrete data without correcting or accounting for these errors may cause algorithms to select polymorphisms with little or no predictive ability. This could lead to varying and even contradictory conclusions. In fact, it was reported that only 6 out of 600 gene-disease associations reported in the literature were significant in more than 75% of the studies published [25]. In majority of cases,
heterogeneity, population stratification, and potential misclassification in the discrete dependent variables were at the top of the list of potential reasons for these inconsistent results [21,26-29]. In supervised learning, if individuals are wrongly assigned to subclasses, false positive and erroneous effects will result if these phenotypes are used when trying to identify which markers or genes can distinguish between disease subclasses. Researchers carried out a study of misclassification using gene expression data with application to human breast cancer [30]. They looked at the influence of misclassification on gene selection. It was found that even when only one sample is misclassified, 20% of the most significant genes were not identified. Further results showed that with misclassification rates between 3-13%, there could be unfavorable effect on detecting the most significant genes for disease classification. Furthermore, if some genes are identified as significant while misclassification is present, this will lead to the inability to replicate the results due to the fact it is only relevant to the specific data.

To overcome these issues it would be advantageous to develop a statistical model that is able to account for misclassification in discrete responses. There have been several approaches proposed on how to handle misclassification. Researchers have suggested Bayesian methods [31-33], some described a latent Markov model for longitudinal binary data [34], others proposed marginal analysis methods [35], and some considered two-state Markov models with misclassified responses [36,37].

In 2001, a Bayesian approach was proposed for dealing with misclassified binary data [33]. This procedure, with the use of Gibbs sampling, “made the analysis of binary data subject to misclassification tractable”. It was concluded that without accounting for errors results in adverse effects related to the parameters of interest including genetic variance. The analysis was applied to simulated cow fertility data and was later implemented with the use of real data which
resulted in similar findings [9,30]. One study found considering a potential for misdiagnosis in the data could increase prediction power by 25% [9]. To extend their ideas we simulated a typical case-control study to measure and understand the effects of misclassification on GWAS using a threshold model and misclassification algorithm. Three analyses were conducted: (M1) the true data was analyzed with a standard threshold model; (M2) the noisy (5% and 10% miscoding) data analyzed with standard threshold model ignoring miscoding; (M3) the noisy data analyzed with threshold model with probability of being miscoded (π) included.

**Materials and Methods**

**Detecting discrete phenotype errors**

Let, \( y = (y_1, y_2, \ldots, y_n)' \) be a vector of binary responses observed for \( n \) individuals and genotypes for a set of SNPs are available for each. The problem is being able to link these responses to the measured genotypes when miscoding or misclassification of the binary status is present in the samples. Specifically, the observed binary data is a “contaminated” sample of a real unobserved data \( r = (r_1, r_2, \ldots, r_n)' \), where each \( r_i \) is the outcome of an independent Bernoulli trial with a success probability, \( p_i \) specific to each response. Misclassification then occurs when some of the \( r_i \) become switched. Assuming this error happens with probability \( \pi \), the joint probability of observing the actual data given the unknown parameters is:

\[
y \mid p_i, \pi = \prod_{i=1}^{n} [p_i(1-\pi) + (1-p_i)\pi]^y_i [p_i\pi + (1-p_i)(1-\pi)]^{(1-y_i)}
\]

\[
= \prod_{i=1}^{n} (q_i)^{y_i} (1-q_i)^{(1-y_i)}
\]

With \( q_i = p_i(1-\pi) + (1-p_i)\pi \)

The success probability for each observation (\( p_i \)) is then modeled as a function of the unknown vector of parameters \( \beta \), which in this case is the vector of SNP effects. Assuming conditional independence, the conditional distribution of the true data, \( r \), given \( \beta \) becomes:
\[
Pr(\mathbf{r} | \mathbf{\beta}) = \prod_{i=1}^{n} p_i(\mathbf{\beta})^{r_i} [1 - p_i(\mathbf{\beta})]^{1-r_i}
\]

where \( p_i(\mathbf{\beta}) \) indicates that \( p_i \) is a function of the vector of parameters \( \mathbf{\beta} \).

Let \( \mathbf{a} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^\top \), where \( \alpha_i \) is an indicator variable for observation \( i \) that takes the value of one ( \( \alpha_i = 1 \) ) if \( r_i \) is switched and 0 otherwise. Supposing each \( \alpha_i \) is a Bernoulli trial with success probability \( \pi \), then \( \alpha_i | \pi = \pi^{\alpha_i} (1-\pi)^{(1-\alpha_i)} \), the joint distribution of \( \mathbf{a} \) and \( \mathbf{r} \) given \( \mathbf{\beta} \) and \( \pi \) can be written as:

\[
Pr(\mathbf{a}, \mathbf{r} | \pi, \mathbf{\beta}) = \prod_{i=1}^{n} \pi^{\alpha_i} (1-\pi)^{(1-\alpha_i)} [p_i(\mathbf{\beta})]^{r_i} [1 - p_i(\mathbf{\beta})]^{1-r_i}
\]

(1)

Furthermore, the true unobserved binary data could be written as a function of the observed contaminated binary responses and the vector \( \mathbf{a} \) as:

\[
r_i = (1-\alpha_i) y_i + \alpha_i (1-y_i)
\]

(2)

Notice that when \( \alpha_i = 0 \) (no switching), the formula in (2) reduces to \( r_i = y_i \)

Using the relationship in (2), the joint probability distribution of \( \mathbf{a} \) and \( \mathbf{y} \) given \( \mathbf{\beta} \) and \( \pi \) becomes:

\[
Pr(\mathbf{a}, \mathbf{y} | \pi, \mathbf{\beta}) = \prod_{i=1}^{n} \pi^{\alpha_i} (1-\pi)^{(1-\alpha_i)} [p_i(\mathbf{\beta})]^{1-\alpha_i + \alpha_i y_i} [1 - p_i(\mathbf{\beta})]^{1-(1-\alpha_i) + \alpha_i(1-y_i)}
\]

To finalize the Bayesian formulation, the following priors were assumed to the unknown parameters in the model

\[
\mathbf{\beta} \sim U[\mathbf{\beta}_{\text{min}}, \mathbf{\beta}_{\text{max}}] \quad \text{and} \quad \pi | a, b \sim \text{Beta}(a, b)
\]

(3)

where \( \mathbf{\beta}_{\text{min}}, \mathbf{\beta}_{\text{max}}, a \) and \( b \) are known hyper-parameters.

The resulting joint posterior density of \( \pi, \mathbf{\beta}, \mathbf{a} \) is:
Implementation of the model in (4) could be facilitated greatly by using a data augmentation algorithm as described by fellow researchers [33]. It consists in assuming the existence of an unknown continuous random variable, \( l_i \), that relates to the binary responses through the following relationship:

\[
y_i = \begin{cases} 
1 & \text{if } l_i > T \\
0 & \text{otherwise}
\end{cases}
\]

where \( T \) is an arbitrary threshold value.

In matrix notation, the model at the liability scale could be written as:

\[
l_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + e_i
\]

where \( \mu \) is the overall mean, \( x_{ij} \) is the genotype for SNP \( j \) for individual \( i \), \( \beta_j \) is the effect of SNP \( j \) and \( e_i \) is the residual term. To make the model in (4) identifiable, two restrictions are needed. It was assumed that the \( T = 0 \) and \( \text{var}(e_i) = 1 \).

At the liability scale and using the prior distributions specified in (3), the full conditional distributions needed for a Bayesian implementation of the model via Gibbs sampler are in closed form being normal for the position parameters and a binomial distribution for \( \alpha_i \):

\[
p(\alpha_i \mid \beta, \pi, a_{-i}, y) \propto \left[ p_i(\beta)^{(1-\alpha_i)\gamma_i, a_i (1-\gamma_i)} [1 - p_i(\beta)]^{(1-\alpha_i)\gamma_i, a_i (1-\gamma_i)} \right] \prod_{i=1}^{n} \pi^{\alpha_i} (1 - \pi)^{(1-\alpha_i)}
\]

where \( a_{-i} \) is vector \( a \) without \( a_i \).

For the misclassification probability, it conditional distribution is proportional to

\[
p(\pi \mid \beta, a, y) \propto \prod_{i=1}^{n} \pi^{\alpha_i} (1 - \pi)^{(1-\alpha_i)} p(\pi \mid a, b)
\]
Hence, \( \pi \) is distributed as \( \text{Beta}(a + \sum \alpha_i, b + n - \sum \alpha_i) \) with \( \sum \alpha_i \) is the total number of misclassified (switched) observations.

**Simulation**

PLINK software [38] was used to simulate a case-control type data sets. Four simulation scenarios were generated to determine the effects of misclassification of binary status on GWA studies. In each scenario, a dataset of 2000 individuals consisting of 1000 cases and 1000 controls was simulated. All individuals were genotyped for 1000 SNPs with minor allele frequencies generated from a uniform distribution between 0.05 and 0.49. Of the 1000 SNPs, 850 SNPs were assumed non-influential and the remaining 150 SNPs were assumed to be associated with the disease status. To mimic realistic scenarios, a series of bins were specified for the 150 influential SNPs to build a spectrum of odds ratios (OR) for disease susceptibility. Two different series of odds ratios were considered. The first group was generated with “moderate” ratios where 25 of the 150 disease associated SNPs were assumed to have an odds ratio of 1:4, 35 with OR of 1:2, and 90 with OR of 1:1.8. The second group was generated using the same distribution except the ratios increased to a more extreme range; 25 with OR 1:10, 35 with OR of 1:4, and 90 with OR of 1:2. Once these parameters were established, PLINK generated a quantitative phenotype based on the disease variants and a random component. Then a median split of that trait was performed thereafter each individual was assigned a binary status. When the “true” binary data were generated as described above, randomly 5 or 10% of the true binary records were miscoded, meaning binary records from cases were switched to controls and vice versa. Based on the OR distribution (moderate and extreme) and the level of misclassification (5 or 10%), four data sets were generated: 5% misclassification rate and moderate OR (D1); 5%
misclassification and extreme OR (D2); 10% misclassification rate and moderate OR (D3); and 10% misclassification rate and extreme OR (D4). For each dataset, 5 replicates were generated.

Results and Discussion

For moderate OR of influential SNPs, and under both simulation scenarios, the true misclassification probability was slightly underestimated. In fact, the posterior mean (averaged over 5 replicates) was 3 and 6% for D1 and D3, respectively. However in both cases, the true misclassification probability values still lie within their respective HPD95% interval indicating the absence of systematic bias (Table 3.1). As the average odd ratios of influential SNPs increased, the estimated misclassification probability increased to 4 and 7% for D2 and D4, respectively. In both cases the estimated misclassification probability was close to the true value used in the simulation. To further test the ability of our procedure to correctly estimate potential misclassification, a null analysis was performed. A true data set (without any misclassification) was analyzed with our proposed model that contemplates misclassification. As expected, the estimated misclassification probability was very close to zero (0.001) indicating, thus, absence of erroneous observations. Across all simulation scenarios, these results indicate the ability of the algorithm to efficiently distinguish between miscoded and correctly coded samples. Similar results were observed when dairy cattle fertility data subject to misclassification was analyzed [33] as well as when applied using cancer gene expression data [30].

Table 3.2 presents the correlation between the true and estimated SNP effects, where the true SNP effects were calculated based on the analysis of the true data (M1). As expected, across all simulated scenarios, the use of the proposed methods (M3) to analyze misclassified data has increased the correlation and consequently reduced any potential bias in estimating SNP effects. For instance, when D1 was used, the correlation between true and estimated SNPs effects
increased from 0.83 when M2 was used to 0.93 using M3 or an increase of around 12%. As the OR of influential SNPs increased, the difference in predicting the true SNP effects between M2 and M3 increased substantially. In fact, using D2 the accuracy increased by 29% from 0.67\(\text{M2}\) to 0.86\(\text{M3}\). The same trend was observed when the probability of misclassification increased from 5 to 10% with an increase in correlation of 0.15 and 0.27 for D3 and D4, respectively. These results indicate not only the superiority of our proposed method compared to a model that ignores potential misclassification (M2) but more importantly is that our methods seems to be robust to the level of misclassification rate or the OR of significant SNPs. Specifically, when the misclassification rate was increased from 5 to 10%, the accuracy of M2 decreased in average by 15\% whereas it decreased only by 4\% using our method. Furthermore, it is worth highlighting that even on the extreme case scenario (D4), our method still produces consistent results as the correlation between true and estimated SNPs effects was 0.84 (Table 3.2).

It is clear that across all four simulation scenarios our proposed method (M3) showed superior performance. Accounting for misclassification in the model increases the predictive power by eliminating or at least by attenuating the negative effects caused by these errors, allowing for better estimates of the true SNP effects. This is essential in GWA studies for correctly estimating the proportion of variation in cause of disease associated with SNPs. Complex diseases which are under the control of several genes and genetic mechanisms are of moderately to highly heritable [38-40].

To further investigate the consequences of misclassification errors on estimating SNP effects we observed the changes in magnitude and the ranking of influential SNPs. As mentioned before the value of GWAS lies in its dependency on being able to correctly detect
polymorphisms associated with a disease. This is driven by how well the model can estimate SNP effects so that the polymorphisms with significant associations will have the largest effects. Figure 3.1 presents SNP effects ordered in a decreasing order based on their estimates using M1 (no misclassification) for scenarios D1 (Fig 3.1A) and D2 (Fig 3.1B). It is clear that in both cases, the M2 method under-performed M3 in estimating the true magnitude and direction of the SNP effects. Even more pronounced results were observed when the misclassification rate was 10% as indicated in Figure 3.2. In fact, this underestimation effect is has been reported as one of the downfalls of GWA studies. When approximating SNP effects, there is an estimation error attached to them adding noise and weakening the strength of the effect [41]. In the presence of misclassification this “noise” is inflated which can lead to underestimating the effects of truly significant SNPs. It has been reported this is most severe when the diseases are influenced by numerous risk variants [42].

In addition to an inaccurate estimation of significant SNPs, M2 tends to report non-zero estimates for truly non-influential SNPs, especially under scenario D2, contrarily to M1 and M3. For example, under scenario D1, 3 out of the 15 most influential SNPs (top 10%) were not identified by M2. However, only one SNP was not identified using M3. This 20% loss of the most significant polymorphisms exhibited by M2 reduces the power of association. Accounting for potential misclassification as observed with our method aids in reducing false discovery rates which is essential in association studies. Similar results were found under D2 as M2 failed to identify 33% of the top 10% SNPs whereas M3 failed to identify only 13%.

To further evaluate the effectiveness of our proposed methods, we looked at its ability of correctly identifying misclassified observations. For that purpose, we calculated the posterior probability of misclassification of each observation in all four scenarios. Figure 3.3 presents the
average posterior misclassification probability for the 113 miscoded observations (Figs 3.3A and 3.3C) and the 1887 correctly coded observations (Figs 3.3B and 3.3D) when the misclassification rate was set to 5%. For scenario D1, the miscoded group exhibited a higher misclassification probability with a mean of 0.42 compared to a mean of 0.004 for the correctly coded group (Figs 3.3A and 3.3B). The lowest misclassification probability observed for the miscoded group was 0.20 far greater than the largest probability calculated for the correct miscoded group which was 0.07 (Fig 3.3B). This is important as it shows that the algorithm was able to distinguish between the two groups and the miscoded records were detected with a high probability. In fact, when the odd ratios increased (D2) this difference became more sizable, as the averages increased to 0.75 and 0.002 for the miscoded and correctly coded individuals, respectively (Figs 3.3C and 3.3D). The same trend held as misclassification increased to 10% as indicated in Figure 3.4. When D3 (D4) was used the average probability of the miscoded group was 0.41 (0.65) and 0.009 (0.006) for the correctly coded observations.

In real data set application, the miscoded observations will be unknown and a reliable cutoff probability is desired. Table 3.3 presents the percent of misclassified individuals correctly identified based on two classification probabilities. We first applied a hard cut off probability set at 0.5. At this limit, our proposed method (M3) was able to account for 28 and 22% of the misclassified individuals based on D1 and D3, respectively (Table 3.3). This is mostly due to the fact that setting such a strict cutoff does not allow for much variation around the threshold. In this case individuals with probabilities very close to 0.5 were not accounted for. As the odds ratios increased, even with the strict cutoff applied, 96 and 91% of the misclassified groups were identified for D2 and D4, respectively (Table 3.3). In order to relax the restrictions of a hard cut off probability, a soft classification approach was used where observations are declared to be
misclassified if their exceeded a heuristically determined threshold. In this study, the threshold was set based on the overall mean of the probabilities of being misclassified over the entire dataset plus two standard deviations. Both moderate scenarios, D1 and D3, showed better results compared to the strict cutoff as M3 correctly identified 94 and 78% of the misclassified observations. As the odds ratios increase, the genetic differences between cases and controls become more distinguishable allowing for better detection. This can be seen when the extreme case scenarios are used as 100% of the misclassified individuals were identified for D2 and 95% for D4 (Table 3.3). Furthermore, across all four scenarios and both cutoff probabilities, no correctly classified observation has a misclassification probability exceeding the cut off threshold and therefore was not incorrectly switched (Table 3.3). This further shows a tendency for misclassified individuals having higher probabilities compared to the correctly coded groups.

**Conclusions**

Misclassification of discrete responses has been shown to occur often in datasets and has proven to be difficult and often expensive to resolve before analyses are run. Ignoring misclassified observations increases the uncertainty of significant associations that may be found leading to inaccurate estimates of the effects of relevant genetic variants. The method proposed in this study was capable of identifying miscoded observations, and in fact these individuals were distinguished from the correctly coded set and were detected at higher probabilities over all four simulation scenarios. This is essential as it shows the capability of our algorithm to maintain its superior performance across different levels of misclassification as well as different odds ratios of the influential SNPs.

More notably, our method was able to estimate SNP effects with higher accuracy compared to estimation using the “noisy” data. Running analyses on data that do not account for
potential misclassification of binary responses, such as M2 in this study, will lead to non-replicative results as well as causing an inaccurate estimation of the effect of polymorphisms which can be correlated to the disease of interest. This severely reduces the power of the study. For instance, it was determined that conducting a study on 5000 cases and 5000 controls with 20% of the samples being misdiagnosed has the power equivalent to only 64% of the actual sample size [6]. Implementing our proposed method provides the ability to produce more reliable estimates of SNP effects increasing predictive power and reducing any bias that may have been caused by misclassification. Our results suggested that the proposed method is effective for implementation of association studies for binary responses subject to misclassification.

Acknowledgements

The first author was supported financially by the graduate school and the department of Animal and Dairy Science at the University of Georgia.

References


SW: Diagnostic shifts during the decade following first admission for psychosis. *Am J

8. West, M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Marks
JR, Nevins JR: Predicting the clinical status of human breast cancer by using gene

patients using gene expression data: Dealing with potential misdiagnosis. Online J.

10. Anderson RE, Hill RB, Key CR: The sensitivity and specificity of clinical diagnostics during
five decades: toward an understanding of necessary fallibility. *JAMA* 1989, 261:1610-
1617.


classification and lessons in 182 cases presented at a problem case conference. *Radiology


15. Plebani, M: Errors in clinical laboratories or errors in laboratory medicine? *Clin Chem Lab


38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC: PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* 2007, 81.


Table 3.1. Summary of the posterior distribution of the misclassification probability ($\pi$) for the four simulation scenarios (averaged over 5 replicates)

<table>
<thead>
<tr>
<th></th>
<th>Moderate</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $\pi$</td>
<td>PM$^2$</td>
<td>HPD95%</td>
</tr>
<tr>
<td>5%</td>
<td>0.03</td>
<td>0.01-0.05</td>
</tr>
<tr>
<td>10%</td>
<td>0.06</td>
<td>0.04-0.09</td>
</tr>
</tbody>
</table>

$^1$ Moderate effects for influential SNPs; $^2$PM=Posterior mean; $^3$HPD95%=High probability density interval
Table 3.2. Correlation between true\(^1\) and estimated SNP effects under four simulation scenarios using noise data (M2) and the proposed approach (M3)

<table>
<thead>
<tr>
<th></th>
<th>5% Moderate</th>
<th>5% Extreme</th>
<th>10% Moderate</th>
<th>10% Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>0.826</td>
<td>0.671</td>
<td>0.699</td>
<td>0.572</td>
</tr>
<tr>
<td>M3</td>
<td>0.930</td>
<td>0.867</td>
<td>0.849</td>
<td>0.840</td>
</tr>
</tbody>
</table>

\(^1\)True effects were calculated based on analysis of the true data (M1); \(^2\) Moderate effects for influential SNPs

Table 3.3. Percent of misclassified individuals correctly identified based on two cutoff probabilities across the four simulation scenarios.

<table>
<thead>
<tr>
<th></th>
<th>D1 Misclass</th>
<th>D1 Correct</th>
<th>D2 Misclass</th>
<th>D2 Correct</th>
<th>D3 Misclass</th>
<th>D3 Correct</th>
<th>D4 Misclass</th>
<th>D4 Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard</td>
<td>0.28</td>
<td>0</td>
<td>0.96</td>
<td>0</td>
<td>0.22</td>
<td>0</td>
<td>0.91</td>
<td>0</td>
</tr>
<tr>
<td>Soft</td>
<td>0.94</td>
<td>0</td>
<td>1.00</td>
<td>0</td>
<td>0.78</td>
<td>0</td>
<td>0.95</td>
<td>0</td>
</tr>
</tbody>
</table>

¹Hard: cut off probability was set at 0.5. Soft: cut off probability was equal to the overall mean of the probabilities of being misclassified over the entire dataset plus two standard deviations; ²Misclass: individuals which were misclassified. Correct: Correctly coded individuals.
**Figure 3.1.** Distribution of SNP effects for 5% misclassification rate. The effects are sorted in decreasing order based on estimates using M1 when odds ratios of influential SNPs are moderate (A) and extreme (B). M1: misclassification was not present in the data. M2: misclassification was present in the data set but was not addressed. M3: misclassification was addressed using the proposed method.
Figure 3.2. Distribution of SNP effects for 10% misclassification rate. The effects are sorted in decreasing order based on estimates using M1 when odds ratios of influential SNPs are moderate (A) and extreme (B). M1: misclassification was not present in the data. M2: misclassification was present in the data set but was not addressed. M3: misclassification was addressed using the proposed method.
Figure 3.3. Average posterior misclassification probability for the 113 miscoded observations (a: moderate and c: extreme) and the 1887 correctly coded observations (b: moderate and d: extreme) when the misclassification rate was set to 5%.
Figure 3.4. Average posterior misclassification probability for the 205 miscoded observations (a: moderate and c: extreme) and the 1795 correctly coded observations (b: moderate and d: extreme) when the misclassification rate was set to 10%.
CHAPTER 4

ANALYSIS OF BINARY RESPONSES WITH OUTCOME SPECIFIC

MISCLASSIFICATION PROBABILITY IN GENOME WIDE ASSOCIATION STUDIES

---

Abstract

Misclassification has been shown to have a high prevalence in binary responses in human populations. These error rates tend to differ between cases and controls as diagnostic and screening tests have different levels of sensitivities and specificities. This increases the inaccuracies of classifying individuals into correct groups, giving rise to both false positive and false negative rates. Leaving these errors uncorrected before analyses will have a negative impact on the overall goal of genome-wide association studies (GWAS) including reducing predictive power. A liability threshold model that contemplates different misclassification rates for cases and controls was developed to assess the effects of misdiagnostic errors on GWA studies. Four simulated balanced case-control datasets consisting of 2000 individuals each were analyzed with varying odds ratios of the influential SNPs and misclassification rates of 5 and 0% and 7 and 3%. Ignoring misclassification resulted in inflated estimates of true non-influential SNPs and failed to estimate the true magnitude and direction of the effects. Once the misclassification algorithm was applied there was a 12% to 32% increase in accuracy, and a substantial reduction in bias. The proposed method was able to capture the majority of the most significant SNPs that were not identified in the analysis of the misclassified data. Furthermore, the proposed method was able to identify with high probability a large portion of the truly misclassified case observations without incorrect switching of non-misclassified observations. The proposed model provides a statistical tool to weaken the negative effects of misclassified binary responses in GWAS. Across difference levels of misclassification probabilities as well as odds ratios of significant SNPs, the model proved to be robust.

Keywords

Misclassification; Genome wide association; Discrete responses; False negative rate; False positive rate
Introduction

Misclassification of the dependent variables is known as one of the leading causes for non-replicative and biased results of genome-wide association studies (GWAS). Classifying individuals into different disease classes has proven to be erroneous as binary responses are subjective measurements with no precise or quantifiable guidelines. Consequently, the outcomes from implementing GWAS using case-control studies can be misleading if the observations are inaccurate. Screening and diagnostic tests are used to identify unrecognized disease or defects and have shown to exhibit potential for bias (Abram and Valesky, 2013). These testing activities are used to characterize and sort individuals into two groups, for example a high risk group for having or developing a disease and a low risk group or classifying them into different subclasses of the same disease or disorder. This screening process typically relies heavily on human perception therefore, the non-diseased group (controls) will always contain some people with the condition, and the diseased group (cases) will contain some of those without it.

The relationship between the test results and the disease status can be described using probabilistic measures such as sensitivity (proportion of those individuals in the screened population with the disease who are classified accurately as diseased) and specificity (proportion of those individuals without the disease who are classified accurately as healthy; Bland and Altman, 1994). Thus, a test with low sensitivity will result in a high false negative rate and a low specificity will lead to a high false positive result. Several reviews have been published in order to assess the variation among studies and evaluate test performances (Irwig et al., 1994; Vamvakas, 1998; Smith-Bindman, 1998). Deeks (2001) performed a systematic review on several diagnostic and screening tests. They pooled together estimates for sensitivities and specificities and found the average sensitivity to be 0.96. The average specificity was
exhibiting considerable variation around the mean ranging between 0.21 and 0.88. Such inaccuracy of screening tests will lead to high mis-diagnostic rates in disease classification across both clinical practices and perceptual specialties.

It was found in radiology areas, failure to detect abnormalities when they were present ranged between 25-30%, and when the cases were normal but incorrectly diagnosed as diseased ranged between 1.5-2% (Renfrew et al., 1992). False negative rates in cancer detection have been documented as one the most difficult limitations (Destounis et al., 2004). Published false negative rates have ranged between 10% and 25% for breast cancer detection (Warren et al., 2000; Birdwell et al., 2001). Goyal et al. (2006) examined the detection and false negative rate of the sentinel lymph node biopsy for early stage breast cancer using 282 samples and found 19 false negatives resulting in a sensitivity of 93.3%. Stock et al. (2012) evaluated cervical cancer screening tests and found false positive estimates ranging between 0.056 and 0.269. Crosswell et al. (2009) analyzed the cumulative risk of a false positive screening for participants in a multimodal cancer screening program. It was found after 14 tests, for men the risk of at least one false positive result was 60.4% and 48.8% for women.

False positive and negative rates are also prevalent in psychological disorders as it is often difficult for clinicians to distinguish between disorders due to overlapping or late development of symptoms. In the case of Alzheimer’s disease (AD), symptoms are more pronounced during later stages therefore diagnosis of incipient AD patients is more difficult. Two cognitive tests are generally administered for diagnosis, neurofibrillary tangles (NFT) and the Mini-Mental State Exam (MMSE). Reviews of NFT have shown it may be more associated with normal aging and may not be sufficient for accurate diagnosis of AD (Haroutunian et al., 1999; Price and Sisodia, 1998; Schmitt et al., 2000).
Unfortunately, finding these errors is not simple. Even in the best case scenario, if misclassification is suspected before analysis re-testing is often not possible or is costly and the sample must be removed reducing power of the study. Extensive research has been carried out in order to assess the effects of misclassification on both the well-being of the patient (Hirschfield et al., 2003; Bhattacharya et al., 2008) as well as its effects on studies including GWAS. GWAS aims to statistically associate genetic variants with disease status therefore it relies on the accuracy of both the genotypic and phenotypic data. Implementing association studies without proper data quality control measures can lead algorithms to select polymorphisms with no true significant relationship to the disease. This could lead to varying and even contradictory conclusions. In fact, it was reported that only 6 out of 600 gene-disease associations reported in the literature were significant in more than 75% of the studies published (Hirschhorn et al., 2002). In majority of cases, heterogeneity, population stratification, and potential misclassification in the discrete dependent variables were at the top of the list of potential reasons for these inconsistent results (Avery et al., 2009; Skafidas et al., 2012; Li and Meyre, 2012; Wu et al., 2011).

Studies examining the effects of uncertainty found it can lead to biased parameter estimates (Sapp et al., 2005; Spangler et al., 2006). To overcome these issues it would be advantageous to develop a statistical model that is able to account for misclassification in discrete responses. In 2001, a Bayesian approach was proposed for dealing with misclassified binary data (Rekaya et al., 2001). This procedure, with the use of Gibbs sampling, “made the analysis of binary data subject to misclassification tractable”. They found, using simulated binary data with a 5.6% misclassification rate, that failure to account for classification errors resulted in
biased parameter estimates, with the true values falling outside the 95% high density posterior interval.

Robbins et al. (2006) found considering a potential for misdiagnosis in the data could increase prediction power by 25%. To extend their ideas we simulated a typical case-control study and incorporated different values of misclassification to account for both false positive and false negative rates. We carried out the analysis to measure and understand the effects of misclassification on GWAS using a threshold model and misclassification algorithm. Three analyses were conducted: (M1) the true data was analyzed with a standard threshold model; (M2) the noisy (5 and 0% and 7 and 3% miscoding) data analyzed with standard threshold model ignoring miscoding; (M3) the noisy data analyzed with threshold model with two probabilities of being miscoded ($\pi_1$ and $\pi_2$) included.

Materials and Methods

The methodology first presented by Rekaya et al. (2001) and later extended and applied by Smith et al. (2013) was adopted in this study to analyze binary data subject to misclassification where the probability of miscoding is different between cases and controls. In presence of misclassification, the vector of observed binary responses $\mathbf{y} = (y_1, y_2, ..., y_n)'$ observed for $n$ individuals and genotypes for a set of SNPs, isa “contaminated” sample of a real unobserved data $\mathbf{r} = (r_1, r_2, ..., r_n)'$, where each $r_i$ is the outcome of an independent Bernoulli trial with a success probability, $p_i$ specific to each response. The problem is being able to link these responses to the measured genotypes when miscoding or misclassification of the binary status is present in the samples. Misclassification then occurs when some of the $r_i$ become switched. In this study a more realistic scenario is adopted regarding the misclassification rates of cases and controls. Contrarily to a common misclassification rate for both cases and controls assumed by
Rekaya et al. (2001) and Smith et al. (2013), specific misclassification rates for each was adopted in this study and to best of our knowledge this is the first time such distinction was assumed. Assuming misclassification happens with probability $\pi_1$ and $\pi_2$ for cases and controls, respectively, the joint probability of observing the actual data given the unknown parameters is:

$$y \mid p_i, \pi_1, \pi_2 = \prod_{i=1}^{n_1} [p_i(1 - \pi_1) + (1 - p_i)\pi_2]^{y_i} [p_i\pi_1 + (1 - p_i)(1 - \pi_2)]^{(1-y_i)}$$

$$= \prod_{i=1}^{n_1} (q_i)^{y_i} (1 - q_i)^{(1-y_i)}$$

With $q_i = p_i(1 - \pi_1) + (1 - p_i)\pi_2$

The success probability for each observation ($p_i$) is then modeled as a function of the unknown vector of parameters $\beta$, which in this case is the vector of SNP effects. Assuming conditional independence, the conditional distribution of the true data, $r$, given $\beta$ becomes:

$$\Pr(r \mid \beta) = \prod_{i=1}^{n_2} p_i(\beta)^{y_i} [1 - p_i(\beta)]^{(1-y_i)}$$

where $p_i(\beta)$ indicates that $p_i$ is a function of the vector of parameters $\beta$.

Let $\alpha = [\alpha_1, \alpha_2, ..., \alpha_{n_1}]$’ be a vector of indicator variables for the $n_1$ case observations, where $\alpha_i$ for observation $i$ that takes the value of one ($\alpha_i = 1$) if $r_i$ is switched from case to control (from one to zero) and 0 otherwise. Similarly, let $\lambda = [\lambda_1, \lambda_2, ..., \lambda_{n_2}]$’ be a vector of indicator variables for the $n_2$ control observations, where $\lambda_i$ for observation $i$ that takes the value of one ($\lambda_i = 1$) if $r_i$ is switched from control to case (from zero to one) and 0 otherwise. Further, each $\alpha_i$ and $\lambda_i$ was assumed to be a Bernoulli trial with success probability $\pi_1$ and $\pi_2$, respectively.
Thus, \\
\[ \alpha_i \mid \pi_1 = \pi_1^{\alpha_i} (1 - \pi_1)^{(1-\alpha_i)} \]

\[ \lambda_i \mid \pi_2 = \pi_2^{\lambda_i} (1 - \pi_2)^{(1-\lambda_i)} \]

the joint distribution of \( \alpha \) and \( \lambda \) given \( \beta \), \( \pi_1 \) and \( \pi_2 \) can be written as:

\[
Pr(\alpha, \lambda \mid \pi_1, \pi_2, \beta) = \prod_{i=1}^{n_1} \pi_1^{\alpha_i} (1 - \pi_1)^{(1-\alpha_i)} \prod_{i=1}^{n_2} \pi_2^{\lambda_i} (1 - \pi_2)^{(1-\lambda_i)}
\]

Furthermore, the true unobserved binary data could be written as a function of the observed contaminated binary responses and the vectors \( \alpha \) and \( \lambda \) as:

If \( y_i \) is a case then \( r_i = (1 - \alpha_i) y_i + \alpha_i (1 - y_i) \)

If \( y_i \) is a control then \( r_i = (1 - \lambda_i) y_i + \lambda_i (1 - y_i) \)

Notice that when \( \alpha_i = \lambda_i = 0 \) (no switching), the formulas in (2) reduce to \( r_i = y_i \)

Using the relationship in (2), the joint probability distribution of \( \alpha, \lambda \) and \( y \) given \( \beta, \pi_1 \) and \( \pi_2 \) becomes:

\[
Pr(\alpha, \lambda, y \mid \pi_1, \pi_2, \beta) = \prod_{i=1}^{n_1} \pi_1^{\alpha_i} (1 - \pi_1)^{(1-\alpha_i)} [p_i(\beta)]^{(1-\alpha_i)} y_i + \alpha_i (1 - y_i) [1 - p_i(\beta)]^{1-(1-\alpha_i)} y_i - \alpha_i (1 - y_i)
\]

\[
\times \prod_{i=1}^{n_2} \pi_2^{\lambda_i} (1 - \pi_2)^{(1-\lambda_i)} [p_i(\beta)]^{(1-\lambda_i)} y_i + \lambda_i (1 - y_i) [1 - p_i(\beta)]^{1-(1-\lambda_i)} y_i - \lambda_i (1 - y_i)
\]

To finalize the Bayesian formulation, the following priors were assumed for the unknown parameters in the model

\[ \beta \sim U[\beta_{\min}, \beta_{\max}], \pi_1 \mid a_1, b_1 \sim Beta(a_1, b_1), \text{ and } \pi_2 \mid a_2, b_2 \sim Beta(a_2, b_2) \]
where $\beta_{\min}$, $\beta_{\max}$, $a_1$, $a_2$, $b_1$, and $b_2$ are known hyper-parameters.

The resulting joint posterior density of $\beta$, $\alpha$, $\lambda$, $\pi_1$, and $\pi_2$ is:

$$p(\beta, \alpha, \lambda, \pi_1, \pi_2 | y) \propto \prod_{i=1}^{n_1} \pi_1^{a_1} (1 - \pi_1)^{1-a_1} \left[ p_1(\beta) \right]^{(1-a_1)y_i + a_1(1-y_i)} \left[ 1 - p_1(\beta) \right]^{1-(1-a_1)y_i - a_1(1-y_i)}$$

$$x \prod_{i=2}^{n_2} \pi_2^{\lambda_2} (1 - \pi_2)^{1-\lambda_2} \left[ p_2(\beta) \right]^{(1-\lambda_2)y_i + \lambda_2(1-y_i)} \left[ 1 - p_2(\beta) \right]^{1-(1-\lambda_2)y_i - \lambda_2(1-y_i)}$$

(4)

$$xp(\pi_1 | a_1, b_1) p(\pi_2 | a_2, b_2)$$

Implementation of the model in (4) could be facilitated greatly by using a data augmentation algorithm as described by fellow researchers (Rekaya et al., 2001). It consists in assuming the existence of an unknown continuous random variable, $l$, that relates to the binary responses through the following relationship:

$$y_i = \begin{cases} 1 & \text{if } l_i > T \\ 0 & \text{otherwise} \end{cases}$$

where $T$ is an arbitrary threshold value.

In matrix notation, the model at the liability scale could be written as:

$$l_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + e_i$$

(5)

where $\mu$ is the overall mean, $x_{ij}$ is the genotype for SNP $j$ for individual $i$, $\beta_j$ is the effect of SNP $j$ and $e_i$ is the residual term. To make the model in (4) identifiable, two restrictions are needed. It was assumed that the $T=0$ and $\text{var}(e_i) = 1$.

At the liability scale and using the prior distributions specified in (3), the full conditional distributions needed for a Bayesian implementation of the model via Gibbs sampler are in closed form being normal for the position parameters and a binomial distributions for $\alpha_i$ and $\lambda_i$. 
\begin{align*}
p(\alpha_i \mid \beta, \pi_1, \mathbf{a}_{-i}, \mathbf{y}) & \propto \left[ p_i(\beta) \right]^{(1-\alpha_i)\gamma_i + \alpha_i(1-\gamma_i)} \left[ 1 - p_i(\beta) \right]^{(1-(1-\alpha_i)\gamma_i - \alpha_i(1-\gamma_i))} \times \pi_1^{\alpha_i} (1 - \pi_1)^{(1-\alpha_i)} \\
p(\lambda_i \mid \beta, \pi_2, \mathbf{\lambda}_{-i}, \mathbf{y}) & \propto \left[ p_i(\beta) \right]^{(1-\lambda_i)\gamma_i + \lambda_i(1-\gamma_i)} \left[ 1 - p_i(\beta) \right]^{(1-(1-\lambda_i)\gamma_i - \lambda_i(1-\gamma_i))} \times \pi_2^{\lambda_i} (1 - \pi_2)^{(1-\lambda_i)}
\end{align*}

where \( \mathbf{a}_{i} (\mathbf{\lambda}_{-i}) \), is vector \( \mathbf{a}(\mathbf{\lambda}) \) without \( a_i (\lambda_i) \).

For the misclassification probabilities, there conditional distributions are proportional to

\begin{align*}
p(\pi_1 \mid \beta, \mathbf{a}, \mathbf{y}) & \propto \prod_{i=1}^{n_i} \pi_1^{a_i} (1 - \pi_1)^{(1-a_i)} \ p(\pi_1 \mid a_i, b_i) \\
p(\pi_2 \mid \beta, \mathbf{\lambda}, \mathbf{y}) & \propto \prod_{i=1}^{n_i} \pi_2^{\lambda_i} (1 - \pi_2)^{(1-\lambda_i)} \ p(\pi_2 \mid a_i, b_i)
\end{align*}

Hence, \( \pi_1 \) and \( \pi_2 \) are distributed as Beta\( (a_i + \sum a_i, b_i + n_i - \sum a_i) \) and

Beta\( (a_i + \sum \lambda_i, b_i + n_i - \sum \lambda_i) \) with \( \sum a_i \) and \( \sum \lambda_i \) are the total number of misclassified

switched cases and control observations, respectively.

**Simulation**

PLINK software (Purcell et al., 2007) was used to simulate case-control type data sets.

Four simulation scenarios were generated to determine the effects of misclassification of binary

status on GWA studies. In each scenario, a dataset of 2000 individuals consisting of 1000 cases

and 1000 controls was simulated. All individuals were genotyped for 1000 SNPs with minor

allele frequencies generated from a uniform distribution between 0.05 and 0.49. Of the 1000

SNPs, 850 SNPs were assumed non-influential and the remaining 150 SNPs were assumed to be

associated with the disease status. To mimic realistic scenarios, a series of bins were specified
for the 150 influential SNPs to build a spectrum of odds ratios (OR) for disease susceptibility. Two different series of odds ratios were considered. The first group was generated with “moderate” ratios where 25 of the 150 disease associated SNPs were assumed to have an odds ratio of 1:4, 35 with OR of 1:2, and 90 with OR of 1:1.8. The second group was generated using the same distribution except the ratios increased to a more extreme range; 25 with OR 1:10, 35 with OR of 1:4, and 90 with OR of 1:2. Once these parameters were established, PLINK generated a quantitative phenotype based on the disease variants and a random component. Then a median split of that trait was performed thereafter each individual was assigned a binary status. When the “true” binary data were generated as described above, randomly 5% or 7% of the cases or 0% and 3% of the controls were miscoded, meaning binary records from cases were switched to controls and vice versa.

Based on the OR distribution (moderate and extreme) and the levels of misclassification (5 and 0% or 7 and 3%), four data sets were generated: 5 and 0% misclassification rates and moderate OR (D1); 5 and 0% misclassification rates and extreme OR (D2); 7 and 3% misclassification rates and moderate OR (D3); and 7 and 3% misclassification rates and extreme OR (D4). For each dataset, 5 replicates were generated.

**Results and Discussion**

To test the capability of our method to distinguish between miscoded and correctly coded observations, the posterior means (averaged over 5 replicates) of the true misclassification probabilities for both cases and controls were calculated. Except for scenarios where misclassification was set at 0%, misclassification probabilities were slightly underestimated but still fell within their respective HPD95% interval (Table 1). For example when moderate OR of
the influential SNPs were used, posterior means were 0 and 4%, and 5 and 2% for D1 and D3, respectively. But as we increased the OR for the extreme cases these means increased to 0 and 5% (D2) and 6 and 2% (D4). Though our algorithm is designed to anticipate and account for potential misclassification, a null dataset was ran with no coding errors to ensure its ability to indicate no mis-diagnostic errors. As expected, this analysis resulted in misclassification probabilities close to zero, with estimates of 0.001 and 0.002.

Adequate sample size is one of the major contributing factors to obtaining sufficient power of a GWA study. Researchers estimate effect sizes are mostly 1.5 or smaller for genetic variants affecting human diseases (Burton et al., 2007) requiring an estimated 3,000 cases and 3,000 controls for reasonable power. Thus, it would be beneficial to identify and correct misclassified samples rather than removing them from the study. Therefore, to continue evaluating the effectiveness of our proposed method to detect miscoded individuals, we calculated the posterior probability of misclassification for each observation (averaged over 5 replicates) in all four scenarios. With moderate OR and misclassification rates set to 5% for cases and 0% for controls the 54 miscoded observations exhibited higher misclassification probability with a mean of 0.58 (Fig 4.1a) compared to a mean of 0.002 for the 1,946 observations of the correctly coded group (Fig 4.1b). As the odds are increased for the extreme scenario (D2), the distinction became more evident as the mean increased to 0.85 compared to 0.006. This is of importance as it shows our method is able to distinguish between miscoded and correctly coded samples at a high probability. In fact, the lowest misclassification probability observed for the miscoded group was 0.28 which was far greater than 0.06, the largest probability calculated for the miscoded group (D1). Similar estimates were obtained when misclassification rates increased
to 7% for cases and 3% for controls (Figure 4.2). For D3 (D4), the means were 0.43 (0.74) and 0.003 (0.002), for the miscoded and correctly coded groups, respectively.

Outside of a controlled study, there is no indication of which individuals are misdiagnosed. Table 4.2 presents the number of correctly identified true misclassified responses based on two cutoff probabilities (averaged over 5 replicates). Using our proposed method with a hard cutoff (p=0.5), 65 (D1) and 94% (D2) of the 54 misclassified binary responses were correctly identified. When the rate of misclassification increased to 7% for cases and 3% for controls, of the 98 miscoded observations 44 (D3; moderate OR) and 97% (D4; extreme OR) were correctly detected. Applying such a strict cutoff allows for little variation around the designated probability, but our method was still able to identify a considerable amount of the misclassification. Once we relaxed the restrictions of the cutoff, where the probability was equal to the overall mean of the probabilities of being misclassified plus two standard deviations, roughly 100% of the miscoded samples were identified across all scenarios except for D3 where 86% were detected. Across both cutoff probabilities for the two scenarios where the overall misclassification rate was 10%, there was a higher detection rate for cases compared to controls. This could be due to the fact that the misclassification rate for cases was set higher when compared to controls; 7% versus 3%.

In GWAS, the association between the frequency of thousands of genetic variants and a given phenotype are tested in hopes of elucidating the biology of a trait, and in human medicine, further lead to better therapeutics. In this instance there’s a need for unbiased and accurate estimates of the SNP effects. In order to assess the consequences of the presence of misclassified samples on estimating effects, we calculated the correlation between the true SNP (calculated based on analysis of the true data, M1) and the estimated SNP effects. Across all four scenarios
our method (M3) was capable of increasing the correlation compared to the ‘contaminated’ data (M2; Table 4.3). For example, for scenarios when OR of the influential SNPs were moderate accuracies increased by 8% for D1 and 12% for D3. As the OR increased for the extreme scenarios, the same trend was observed but correlations increased by a more substantial amount. When misclassification rates were 5 and 0%, correlation increased by 0.134 and 0.217 for D2 and D4, respectively. This is evidence of the superiority of our method being able to produce consistent results of decreasing any potential misclassification bias on estimation of SNP effects.

The effect sizes of SNPs with true association to the phenotype should be larger in magnitude compared to non-causal SNPs. We observed the ranking of the SNPs by monitoring the top 10% most influential, and in the presence of misclassified observations (M2) the non-influential SNPs tended to have non-zero estimates. Using scenario D4 (ignoring misclassification), 8 out of the 15 most influential SNPs were not accounted for but after correction, our method (M3) was able to capture 11 out of the 15 giving an increase of 20% in power of association. Even in the more modest case, when misclassification rates were set at 5% for cases and 0% for controls with moderate OR of the disease associated SNPs, M2 caused a loss of 20% in power but our method reduced it to 7%.

As previously mentioned, we noticed a change in rankings of the SNPs so we further investigated errors in estimation due to data misclassification by examining the magnitude of the SNP effects. SNPs were ordered in decreasing order based on their estimates calculated from the data with no misclassification present (M1). For scenarios D1 (Fig. 4.3a) and D2 (Fig. 4.3b), it is clear M2 failed to capture the true magnitude and direction of the SNP when compared to our proposed method (M3). This distinction became more evident when we increased the misclassification rates to 7% for cases and 3% for controls (Figure 4.4). In fact, imprecise
phenotyping leading to reduced estimates of effect sizes is reported as one of the limitations of GWAS (Pearson and Manolio, 2008). Accumulation of erroneous estimates from selection of non-significant SNPs leads to biased estimates of genetic parameters including the variance explained by SNPs, true genetic correlations between disorders, and lower estimates of heritabilities (Wray et al., 2012; Lee et al., 2011; Eichler et al., 2010). It is reported this is most severe for complex traits as these are influenced by numerous risk variants (Stringer et al., 2011).

Conclusions

High false positive and false negative rates of discrete responses occur often in datasets and correcting misclassified observations has proven to be difficult and often expensive to remedy. Ignoring them increases the uncertainty of significant associations that may be found leading to inaccurate estimates of the effects of relevant genetic variants. This in turn will lead to an increase of false positive results as non-influential SNPs will tend to have inflated estimates. Our method was capable of distinguishing and identifying miscoded observations for both cases and controls with high probabilities. Cases tended to have higher probabilities than controls in part due to having a higher prevalence of being misclassified.

Our method estimated SNP effects with higher accuracy compared to estimation using the “noisy” data which will aid in decreasing the rate of non-replicative results as well as reducing inaccurate estimation of the effect of polymorphisms which can be correlated to the disease of interest. This in turn will produce more reliable estimates of SNP effects increasing predictive power and reducing any bias that may have been caused by misclassification. Our procedure performed well even when one of the misclassification rates was set to zero which is important when diagnostic procedures have either a high sensitivity or specificity. Our results
suggested that the proposed method is effective for implementation of association studies for binary responses subject to differing misclassification rates of cases and controls.

References


Table 4.1. Summary of the posterior distribution of the misclassification probability (\(\pi\)) for the four simulation scenarios (averaged over 5 replicates)

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>Moderate(^1)</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\pi_1)</td>
<td>(\pi_2)</td>
<td>(\pi_1)</td>
</tr>
<tr>
<td>5%</td>
<td>0%</td>
<td>0.04 0.002</td>
<td>0.006 0.0003</td>
</tr>
<tr>
<td>7%</td>
<td>3%</td>
<td>0.05 0.02</td>
<td>0.008 0.004</td>
</tr>
</tbody>
</table>

\(^1\)Moderate effects for influential SNPs; \(^2\)PM=Posterior mean; \(^3\)PSD=posterior standard deviation
Table 4.2. Correlation between true\(^1\) and estimated SNP effects under four simulation scenarios using noise data (M2) and the proposed approach (M3)

<table>
<thead>
<tr>
<th></th>
<th>5 and 0%</th>
<th>7 and 3%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Moderate</td>
<td>Extreme</td>
</tr>
<tr>
<td>M2</td>
<td>0.894</td>
<td>0.777</td>
</tr>
<tr>
<td>M3</td>
<td>0.969</td>
<td>0.911</td>
</tr>
</tbody>
</table>

\(^1\)True effects were calculated based on analysis of the true data (M1); \(^2\) Moderate effects for influential SNPs
Table 4.3. Percent of misclassified individuals correctly identified based on two cutoff probabilities across the four simulation scenarios.

<table>
<thead>
<tr>
<th></th>
<th>D1 Misclass</th>
<th>Correct</th>
<th>D2 Misclass</th>
<th>Correct</th>
<th>D3 Misclass</th>
<th>Correct</th>
<th>D4 Misclass</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard</td>
<td>0.65</td>
<td>0</td>
<td>0.94</td>
<td>0</td>
<td>0.44</td>
<td>0</td>
<td>0.97</td>
<td>0</td>
</tr>
<tr>
<td>Soft</td>
<td>1.00</td>
<td>0</td>
<td>0.98</td>
<td>0</td>
<td>0.86</td>
<td>0</td>
<td>1.00</td>
<td>0</td>
</tr>
</tbody>
</table>

1Hard: cut off probability was set at 0.5. Soft: cut off probability was equal to the overall mean of the probabilities of being misclassified over the entire dataset plus two standard deviations;

2Misclass: individuals which were misclassified. Correct: Correctly coded individuals.
Figure 4.1. Distribution of SNP effects for 5 and 0% misclassification rates. The effects are sorted in decreasing order based on estimates using M1 when odds ratios of influential SNPs are moderate (A) and extreme (B). M1: misclassification was not present in the data. M2: misclassification was present in the data set but was not addressed. M3: misclassification was addressed using the proposed method.
Figure 4.2. Distribution of SNP effects for 7 and 3% misclassification rates. The effects are sorted in decreasing order based on estimates using M1 when odds ratios of influential SNPs are moderate (a) and extreme (b). M1: misclassification was not present in the data. M2: misclassification was present in the data set but was not addressed. M3: misclassification was addressed using the proposed method.
Figure 4.3. Average posterior misclassification probability for the 54 miscoded observations (a: moderate and c: extreme) and the 1946 correctly coded observations (b: moderate and d: extreme) when the misclassification rates were set to 5 and 0%.
Figure 4.4. Average posterior misclassification probability for the 98 miscoded observations (a: moderate and c: extreme) and the 1,902 correctly coded observations (b: moderate and d: extreme) when the misclassification rates were set to 7 and 3%. 

CHAPTER 5
UNCERTAINTY IN CONTINUOUS DEPENDENT VARIABLES AND ITS EFFECT ON GENOME WIDE ASSOCIATION STUDIES

Abstract

In genome-wide association studies (GWAS) using multiple-step procedures, the dependent variable (DV) is often a pseudo-observation such as estimated breeding values (EBVs) or de-regressed proofs. Thus, these “estimated” DVs include a certain level of uncertainty. If the sampling errors (SE) attached to these variables are constant across all records, regression analyses will accommodate this situation without major difficulty. However, when SE are heterogeneous across observations, the case with pseudo-records used in GWAS in livestock applications, the situation is more complex. The residual terms of the regression models include two components, the SE due to estimation of the DV, i.e. EBVs, and white noise that will exist even if it was possible to observe (measure) the DV. This type of data is often analyzed using ordinary least squares (OLS) assuming homogeneous residuals or weighted least squares (WLS) where both components of the residual are assumed to be heterogeneous when only the first component is. In this study we present a method for analyzing uncertain DVs when only one component (SE) of the residual term is heterogeneous (OWLS). Using real and simulated data, our proposed approach was superior based on the accuracies of the genomic enhanced breeding values (GEBVs). In fact the proposed methods increased accuracies by 2% and 8% using real data compared to OLS and WLS, respectively. Such increase was of the order of 2-5% using simulated data. The results also indicated that not only the magnitude but also the spread of the SE variance affect the performance of all three methods.

Keywords

Estimated breeding values; Prediction error variance; SNP, Association
Introduction

The availability of large SNP panels in domesticated species has given rise to the use of
genome-wide associated studies (GWAS) and genomic selection for understanding the
underlying genetic mechanisms of complex traits as well as enhancing genetic improvement
programs (Hayes et al., 2009; Dekkers, 2010). The goal of GWAS is to find statistically
significant relationships between single nucleotide polymorphisms (SNPs) and economically
important traits. Unlike other methods such as QTL mapping, GWAS allows for scanning the
entire genome rather than testing one or few regions at a time. The model generally used in
animal breeding programs involved in the estimation of SNP effects includes fixed effects, such
as contemporary group, as well as the breeding value for each animal (Goddard and Hayes,
2009). Breeding values cannot be directly measured therefore they are estimated. In GWAS
using multiple-step procedures the dependent variable is often a pseudo-observation, such as
these estimated breeding values (EBVs) or de-regressed proofs. As with any estimation a level of
uncertainty is attached to each value known as the prediction error variance.

The accuracy, a function of PEV, of the EBVs is an important factor in genetic
improvement and response to selection. It’s an indicator of the reliability of an EBV and the
extent of its possible change as more information is added, and reflects the quality and quantity
of the data (Hickey et al., 2008); the smaller the PEV, the more accurate the estimate. It also
influences the estimation of SNP effects, for instance if the EBVs are calculated with low
precision, it could result in statistical bias leading to the selection of SNPs which may have little
or no predictive ability of the trait of interest.

In GWAS, regression analysis is generally completed in two steps. During the first step
the phenotypic data is used to estimate breeding values which are then regressed against the
genotypes of the SNPs for each animal. When using estimated values in regression analysis the residual term of the model includes two components; the sampling errors due to the estimation of the dependent variable, in this case the PEV of the EBVs, as well as the white noise that will exist even if it was possible to directly observe or measure the dependent variable. This data is commonly analyzed by ordinary least squares (OLS) and weighted least squares (WLS), but each method inflicts its own issues. OLS assumes constant variance across all observations imposing further inefficiency and inconsistent standard error estimates (Lewis and Linzer, 2005). WLS assumes both components are heterogenous when in fact only PEV varies across individuals, and just as OLS may produce inconsistent parameter estimates.

Several attempts have been made at calculating and dealing with the uncertainty (PEV) surrounding EBVs. PEV can be computed from the inverse of the coefficient matrix of the mixed model equations. The drawback is that this becomes too computationally demanding due to the large number of equations. Meyer (1989) proposed a technique based on the absorption of information through the relationship matrix. This summarizes all information available on an individual into a single number and uses it to derive the accuracy. For example, if both parents were known a 4x4 matrix was constructed, inverted, and then approximation of the corresponding elements of $C^{-1}$ for each animal in the pedigree. The disadvantage is the process becomes time consuming.

Thompson et al. (1993) calculated PEV similar to Meyer using sparse matrix techniques. It was computationally more feasible as rather than saving the components of the entire coefficient matrix, only the non-zero elements are stored permitting the use of less memory.

Some researchers have chosen a different approach, rather than focusing on ways to invert the coefficient matrix they developed formulations to approximate the PEV directly.
Garcia-Cortes et al. (1995) proposed two approaches by resampling. One which assumed the covariance between the estimated and true breeding values is equal to the variance of the EBV, where PEV is equal to the difference between the additive genetic variance and the variance of the breeding value. The second, which stated the previous assumption is unnecessary, was calculated as the variance between the difference of the true and estimated values. Fouilloux and Laloe (2001) approximated PEV by subtracting the covariance between the true and estimated value from the genetic variance. They assumed the covariance was equal to the variance of the EBV and the variance of the true breeding value was equal to the genetic variance.

Traditionally, animal’s EBVs and PEVs are calculated based on Best Linear Unbiased Prediction (BLUP; Henderson, 1973). This method allows for maximizing the correlation between true and predicted breeding values therefore minimizing the prediction error variance, but this process comes with several disadvantages. First, a covariance exists between prediction errors and the breeding values. Secondly, BLUP is a shrinkage estimator which shrinks observations towards the parent average. To try and overcome these issues researchers have tried deregressing, or inflating, the EBVs as well as removing the parent average.

Garrick et al. (2009) implemented deregressed EBVs with parent average effect removed. The estimates also have heterogenous variances therefore the deregressed estimates should be weighted. They found that neither reliability nor the prediction errors were the appropriate weights and concluded a ratio that incorporates heritability, reliability, and the proportion of genetic variance not explained by the markers were more efficient. This influences the accuracy of individuals with more reliable information in comparison to individuals with less reliability. Ostersen et al. (2011) implemented this method in pure-bred pigs and produced similar results.
A study was conducted using the approximations of PEV above along with several others to test the accuracy and convergence rates via Monte Carlo sampling (Hickey et al., 2009). They found the convergence rates were dependent on the number of samples and the level of PEV. The formulations which approximated the PEV the best were those that made use of either the variance of the true breeding value, the variance of the EBV, or the covariance between the two.

As mentioned before the most recognized ways of dealing with uncertainty of estimates is using the OLS or the WLS approach. Since the variance is comprised of two components a method for incorporating both in a regression analysis, taking into consideration one is heterogenous and the other is not, would be ideal. The process would involve steps similar to WLS regression except the weightings attached to each estimate would be the inverse of both components of the residual variance. In this study four analyses were conducted to compare differences in accuracies of EBVs: (M1) analyzing the data using ordinary least squares method; (M2) weighted least squares; (M3) weighted least squares assuming only heterogeneous SE using the estimated prediction error variance.

Materials and Methods

GWAS in multiple-step procedures often uses pseudo-observations. In animal breeding applications usually estimated breeding values (EBVs), daughter yield deviation or deregressed proofs are used as dependent variables in the regression model. Contrarily to accurately measure continuous responses, these dependent variables are estimates and thus they carry a certain level of uncertainty. Furthermore, such uncertainty is heterogeneous between observations.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is a vector of true (non-observed) dependent variables and $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$ is an unbiased estimate of $\mathbf{y}$. Further, let $x_{ij}$ be the genotype for individual $i = 1, 2, \ldots, n$ at SNP $j = 1, 2, \ldots, p$. Without loss of generality, we assume that $p < n$. 
A general linear model for association between the true dependent variables and the SNP genotypes is:

\[ y_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + e_i \]  

(1)

Where \( \mu \) is the overall mean, \( x_{ij} \) is the genotype for animal \( i \) at SNP \( j \), \( \beta_j \) is the effect of SNP \( j \), and \( e_i \) is the error term.

As \( y_i \) was assumed non-observed and only an unbiased estimates of it, \( z_i \), is available, the following relationship could be established between \( y_i \) and \( z_i \):

\[ z_i = y_i + \varepsilon_i \]  

(2)

Where \( \varepsilon_i \) is the sampling error with mean equal to zero (\( E(\varepsilon_i) = 0 \)). Using the relationships in equations [1] and [2], the linear model for association between the estimated dependent variables and the SNP genotypes will be:

\[ z_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + e_i + \varepsilon_i \]  

(3)

Let \( w_i = e_i + \varepsilon_i \)

\[ z_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + w_i \]  

(4)

Assuming that the error terms follow a normal distribution with zero mean and a homogeneous variance

\[ e \sim N(0, \sigma^2_e) \]

The structure of the residual (co)variance in equation [4], depends on the assumption about the sampling errors, \( \varepsilon_i \). Assuming that the sampling errors are independent between observations, \( cov(\varepsilon_i, \varepsilon_j) = 0 \), the resulting residual (co)variance matrix for model in equation [4] will be:
where \( \text{var}(\epsilon_i) = \delta_i^2 \). In the vast majority of real world applications, the sampling error variance between observations is heterogeneous. Furthermore, only estimates of the sampling error variances are available.

In animal breeding and genetic applications, a similar situation arises when “pseudo” phenotypes are used in genome wide associations with SNP markers. In this study, estimated breeding values will be used as the “pseudo” phenotype. Thus, the sampling error variances, \( \delta_i^2 \), are the estimated prediction error variances (PEV).

**Analysis of estimated dependent variables**

As previously mentioned, estimated dependent variables are often analyzed using an ordinary least squares (OLS) or a weighted least squares approach (WLS). Using OLS, the estimated dependent variables were assumed known as if they were observed without uncertainty. Although in clear violation of the response distribution, this approach works reasonably well if the differences between the sampling error variances are small or/and when the error variance is much larger than sampling error variances \( \sigma_e^2 \gg \delta_i^2 \). For the WLS, the estimated dependent variable is weighted by the inverse of the standard deviation of the variance.

\[
\omega_i = \frac{1}{\sqrt{(\sigma_e^2 + \delta_i^2)}}
\]
When both variance components $\sigma_e^2$ and $\delta_i^2$ are known, the WLS approach yields the best results. Unfortunately, in the majority of the cases, only the sampling error variance $\delta_i^2$ is available and the weights are calculated as:

$$\omega_i = \frac{1}{\sqrt{\delta_i^2}}$$

In such cases, the performances of WLS are no more guaranteed to be optimum and depend crucially on the ratio between $\sigma_e^2$ and $\delta_i^2$. As $\sigma_e^2 / \delta_i^2 \to 0$, the closer the performance of the WLS tends to optimality.

The third approach (OWLS) is an extension of the method first presented by Hanushek (1974) and it consists in estimating the matrix $R$ assuming that the sampling error variances are known. The resulting estimates of $\sigma_e^2$ and $\delta_i^2$ will used to compute weight factors as:

$$\omega_i = \frac{1}{\sqrt{(\sigma_e^2 + \delta_i^2)}}$$  \hspace{1cm} (5)

**Estimating $\sigma_e^2$**

Using equation [5], the expectation of the sum of squared residual terms is equal to:

$$E(\tilde{W}'\tilde{W}) = E(W'W) - tr[(X'X)^{-1}X'RX]$$

$$= n\sigma_e^2 + \sum_i \delta_i^2 - tr[(X'X)^{-1}X'RX]$$
Let \( R = I\sigma_e^2 + \emptyset \) where,

\[
\emptyset = \begin{bmatrix}
\delta_1^2 & \cdot & 0 \\
\cdot & \delta_2^2 & \cdot \\
0 & \cdot & \delta_n^2
\end{bmatrix}
\]

Then,

\[
E(\hat{\mathbf{W}}'\hat{\mathbf{W}}) = n\sigma_e^2 + \sum_i \delta_i^2 - p\sigma_e^2 - tr[(X'X)^{-1}X'\emptyset X]
\]

\[
= (n - p)\sigma_e^2 + \sum_i \delta_i^2 - tr[(X'X)^{-1}X'\emptyset X]
\]

Finally,

\[
\hat{\delta}_e^2 = \frac{\sum_i \hat{\delta}_i^2 - \sum_i \delta_i^2 + tr[(X'X)^{-1}X'\emptyset X]}{n - p}
\]

The three methods (OLS, WLS and OWLS) were compared under different simulation scenarios.

**Simulation**

Simulated and real data were used. The simulated data consisted of 1200 individuals phenotyped and genotyped for 49,950 SNPs for a trait with a heritability of 0.25. QMSim software (Sargolzaei and Schenkel, 2009) was used to simulate a population of 10 generations of animals totaling 4,420. The historical generation consisted of 420 animals and the effective size
remained constant across all generations. The last three generations (8, 9, and 10) were chosen to be genotyped for roughly 50,000 markers. A genome was simulated with 30 chromosomes and 100 cM long. 1000 QTLs were chosen uniformly across the genome and the portion of the genetic variance explained by the QTLs ranged from 0.8 to 1.0. Once the genotypes were generated phenotypes were simulated following a model that included other than the genetic effect, two fixed effects with 2 levels for the first and 10 levels for the second. The residual terms were generated from a normal distribution with zero mean and variance set to result in a heritability equal to 0.25. Each simulation scenario was replicated five times.

The real data set consisted of 1989 animals genotyped for 50k SNPs. EBVs and associated prediction error variance (PEV) and reliability were computed using phenotypic and pedigree information. The most relevant SNPs for simulated (1000 SNPs) and real (209 SNPs) were pre-selected using single marker analyses and stepwise regression. Pre-selected SNPs were used for association with estimated breeding values (EBVs). Three methods were used to analyze both data sets: 1) OLS (M1); 2) WLS (M2); and 3) OWLS assuming only heterogeneous SE using the estimated PEV (M3).

**Results and Discussion**

Table 5.1 presents the accuracies, defined as the correlation between true BVs and estimated genomic breeding values, using simulated data and the three methods of analysis. Across all values of the error terms variances, it seems to be little to no difference between OLS and WLS accuracies. In fact as the $\sigma^2_e$ increases, the performance of the WLS approach decreases and even becomes lower than the accuracies obtained using the simple OLS. This behavior is somehow expected due to the fact that the WLS approach uses only the sampling error variance for the weights and thus as $\sigma^2_e$ increases, the relative weight of SE variance in the
total residual variance decreases. Similar results were observed by White (1980). On the other hand, when the OWLS was used (M3), the accuracies were higher across all values of $\sigma_e^2$.

Furthermore, the superiority of M3 increased with the increase of $\sigma_e^2$. This is also expected, as M3 accounts correctly for $\sigma_e^2$ contribution into the total residual variance contrarily to the other two methods. In order to further evaluate the performance of the three methods, we looked at the accuracies of GEBVs as a function of the spread of the sampling error variances between observations (Table 5.2). For animals with low PEV (low sampling variance), their accuracies were higher using all three methods compared to animals with high PEV. Furthermore, the superiority of the M3 was higher for animals with high PEV. This result is relevant for breeding applications when low reliabilities animals are used in the training population.

For the real data and using the 209 pre-selected SNPs, the superiority of OWLS ranged from 3% compared to M2 and 8% compared to M1. It is worth mentioning that reliability of EBVs in the real data set was in general low as animals with limited information were genotyped.

**Conclusions**

The increase of contribution of the sampling error in the total residual variance or/and its dispersion (variability between observations) decreases the performance of OLS based methods. Regular WLS seems not to be a good fix when using estimated dependent variables. The proposed methods seems to works in all the cases as it accounts correctly for the contribution of the error terms and the sampling errors to the total residual variance of the association model.

**References**


Table 5.1: Accuracies of GEBVs for different values for the error terms variance using three methods (Simulated data)

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma_e^2 = 0$</th>
<th>$\sigma_e^2 = 1$</th>
<th>$\sigma_e^2 = 2$</th>
<th>$\sigma_e^2 = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS (M1)</td>
<td>0.91</td>
<td>0.89</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>WLS (M2)</td>
<td>0.93</td>
<td>0.91</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>OWLS (M3)</td>
<td>0.93</td>
<td>0.93</td>
<td>0.89</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Table 5.2: Accuracies of GEBVs for all animals, low and high PEV animals using three methods (Simulated data)

\[ \sigma^2_e = 0.3 \]

<table>
<thead>
<tr>
<th>Method</th>
<th>Low PEV</th>
<th>All</th>
<th>High PEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS (M1)</td>
<td>0.84</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>WLS (M2)</td>
<td>0.80</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>OWLS (M3)</td>
<td>0.84</td>
<td>0.86</td>
<td>0.82</td>
</tr>
</tbody>
</table>
CHAPTER 6
CONCLUSIONS

This study showed the capability of our method to account for misclassification in a computationally efficient manner and increase the prediction accuracy and power to detect true influential genetic variants. For diseases where clinical diagnostic tools are ineffective and inaccurate, the misclassification algorithm can provide an approach of identifying potentially misclassified individuals. Given the results using the simulated datasets with varying odds ratios of the influential polymorphisms as well as different misclassification rates, it is evident the use of genetic information for identifying individuals potentially misdiagnosed is important in identifying genomic features. Results from the studies involving differing misclassification rates between cases and controls exhibited similar outcomes. In animal agriculture, it is believed that the advantage of using genomic information will be greater for qualitative and low heritability traits that are often discrete in nature. Unfortunately, these traits, as in human disease traits, are not free from errors. In fact, we expect a much higher rate of misclassification of discrete responses in animal applications for obvious reasons. Thus, our approach will provide tools to deal with or at least attenuate the negative effects of misclassification of discrete responses in genome wide association studies.

The problem of uncertainty of response variables is not limited to discrete data. In fact, in GWAS it is often the case that “pseudo” phenotypes, such estimated breeding values, are used as dependent variables. As non-observed quantities, these dependent variables have a certain level of uncertainty that needs to be accounted for in the association analyses. In this study, we
presented an approach that seems to work in all the cases as it accounts correctly for the contribution of the error terms and the sampling errors to the total residual variance. Furthermore, the proposed approach is simple to implement.
APPENDIX A

RESULTS FOR THE SIMULATION SCENARIO WHERE MISCLASSIFICATION RATES WERE SET TO 5% AND 5%

Table A.1. Summary of the posterior distribution of the misclassification probabilities ($\pi$) for the 5 and 5% simulation scenario (averaged over 5 replicates)

<table>
<thead>
<tr>
<th></th>
<th>Moderate(^1)</th>
<th>Extreme</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
<td>$\pi_1$</td>
</tr>
<tr>
<td>True</td>
<td>5%</td>
<td>5%</td>
<td>0.025</td>
<td>0.03</td>
<td>0.008</td>
<td>0.007</td>
<td>0.03</td>
<td>0.04</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\(^1\) Moderate effects for influential SNPs; \(^2\) PM=Posterior mean; \(^3\) PSD=Posterior standard deviation
Table A.2. Correlation between true\textsuperscript{1} and estimated SNP effects under 5 and 5% simulation scenarios using noise data (M2) and the proposed approach (M3)

<table>
<thead>
<tr>
<th></th>
<th>Moderate\textsuperscript{2}</th>
<th>Extreme</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>0.899</td>
<td>0.644</td>
</tr>
<tr>
<td>M3</td>
<td>0.966</td>
<td>0.820</td>
</tr>
</tbody>
</table>

\textsuperscript{1}True effects were calculated based on analysis of the true data (M1); \textsuperscript{2}Moderate effects for influential SNPs
Table A.3. Percent of misclassified individuals correctly identified based on two cutoff probabilities across 5 and 5% simulation scenarios for moderate (D5) and extreme (D6) OR.

<table>
<thead>
<tr>
<th></th>
<th>D5</th>
<th>D6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Misclass</td>
<td>Correct</td>
</tr>
<tr>
<td>Hard</td>
<td>0.27</td>
<td>0</td>
</tr>
<tr>
<td>Soft</td>
<td>0.79</td>
<td>0</td>
</tr>
</tbody>
</table>

1Hard: cut off probability was set at 0.5. Soft: cut off probability was equal to the overall mean of the probabilities of being misclassified over the entire dataset plus two standard deviations;

2Misclass: individuals which were misclassified. Correct: Correctly coded individuals.
Figure A.1 Distribution of SNP effects for 5 and 5% misclassification rates. The effects are sorted in decreasing order based on estimates using M1 when odds ratios of influential SNPs are moderate (a) and extreme (b). M1: misclassification was not present in the data. M2: misclassification was present in the data set but was not addressed. M3: misclassification was addressed using the proposed method.
Figure A.2. Average posterior misclassification probability for the 113 miscoded observations (a: moderate and c: extreme) and the 1,887 correctly coded observations (b: moderate and d: extreme) when the misclassification rates were set to 5 and 5%.