Mathematical Statistics
Stockholm University

Genetic association studies with complex ascertainment

Maria Grünewald

Research Report 2004:5
Licentiate thesis

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2004:5**,
http://www.math.su.se/matstat

Genetic association studies with complex ascertainment

Maria Grünewald*

May 2004

Abstract

In genetic association studies outcome dependent sampling is often used in order to increase power. When analyzing the data, correction for the ascertainment scheme generally has to be made to avoid bias. Such correction is however not available in standard statistical methods when the data structure and/or the ascertainment scheme is complex. In this report three simulation based approaches that can be used for correction of known ascertainment schemes are described. These methods provide parameter estimates and are flexible in terms of what statistical models and ascertainment schemes can be handled. Some simulations are conducted to evaluate the methods.

KEY WORDS: genetic association study, diabetes, metabolic syndrome, ascertainment, outcome dependent sampling, importance sampling, stochastic EM

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: mariag@math.su.se.

Acknowledgements

Financial support from STINT, SSF, Nancy Pedersen and Peter Arner is gratefully acknowledged.

First of all I want to thank Keith Humphreys for supervising me. Keith has put a lot of effort into guiding me through the licentiate and he is a very nice person to work with. I would also like to thank my formal supervisor Juni Palmgren for introducing me to the area of statistical genetics, and Nancy Pedersen for co-supervising me. I got some very valuable feedback on the report from Ola Hössjer, and Iza Roos helped me to get the genetics right. Thank you!

I would like to thank all my colleges, both at Mathematical Statistics at Stockholm University and in the Biostatistics group at MEB, KI. Special thanks goes to Esbjörn Ohlsson for being a very good boss, to Christina Nordgren for making things work, to Roland Orre for always being fun and for using his super-powers to fix my computer, to my friends in the ”statbiol” group for proving how much fun work can be and to Gudrun Jonasdottir, Anna Svensson and Alexandra Ekman for cheering me up and for being such nice friends. ...and if anybody find that I have thanked them more than once, that is because you deserve it.

A warm thank you also goes to everybody at the Department of Biostatistics at Harvard School of Public Health for making me feel welcome when I was an exchange student in Boston.

...but most of all I would like to thank my family for all their love and support: my parents Arne Rydberg and Rakel Grünewald, my beloved fiance Erik Nilsson, and his parents Lars-Göran and Gun. I would also like to thank all my friends, especially Viveka Seitz and Linda Wickman for immense support an patience!

Contents

Notation

A	Ascertainment indicator
AA, Aa, aa	Genotype outcomes
df	Degrees of freedom
G	Genotype score (exposure)
H_0	Null hypothesis
H_A	Alternative hypothesis
L	Likelihood
l	Log likelihood
M	Number of simulated observations
n	Sample size
P	Probability
Ph	Phenotype (outcome)
w	Weight
Y	Data
α	Significance level
β_{0X}	Intercept for X
$\beta_{X_1X_2}$	Effect of X_1 on X_2
θ	True parameter value
θ'	Assumed parameter value or starting value
$\hat{\theta}$	Estimated parameter value
π	Probability of success
σ	Standard deviation

Abbreviations

BMI Body mass index

DNA Deoxyribonucleic acid

EM Estimation maximization

HWE Hardy Weinberg equilibrium

OR Odds ratio

RR Relative risk

SDPP Stockholm Diabetes Prevention Program

SEM Stochastic EM

SNP Single nucleotide polymorphism

SRS Simple random sample

TDT Transmission disequilibrium test

WHO World Health Organization

1 Introduction

Statistical efficiency is a major concern in both study design and analysis of data. This report is concerned with design and analysis of genetic association studies. Typically the costs of such studies are high and statistical power low, so making efficient use of available resources is of great importance. One approach to increase efficiency is to assign differential ascertainment probabilities according to specific characteristics of the sample units. An example of such a scheme is case-control studies, where ascertainment probabilities depends on disease status. The sampling scheme must however be taken into account in the analysis. If it is not corrected for bias will generally result.

The importance of correcting for the sampling scheme when estimating parameters in human genetics studies has been known for a long time. Fisher (1934) pointed out the need of ascertainment correction as early as 1934 and also commented on the lack of such correction in genetic studies. When the data structure is simple or only testing is of interest, ascertainment correction is sometimes straightforward or even not necessary, but for more complex problems there is still no consensus of how to proceed. Unfortunately the lack of easily accessible statistical tools to correct for ascertainment sometimes causes investigators to ignore the ascertainment scheme and perform analysis that are statistically biased. It may also prevent the investigator from using an ascertainment scheme that would be beneficial to the power of the analysis.

An aim of this report is to describe the benefits of using differential ascertainment probabilities in genetic association studies, and to stress the need of flexible methods for ascertainment correction. Some background on ascertainment is provided in Section 2. Different likelihoods for ascertained data are presented in Section 2.2 and some methods to correct for ascertainment are outlined in Section 2.3. The disadvantage of categorizing continuous variables to facilitate analysis is discussed in Section 2.4. Section 3 contains some background of genetic association studies. In Section 3.1 terminology and basic concepts of genetics are described and in Section 3.2 genetic association studies are introduced. Study design and ascertainment schemes in genetic association studies are discussed in Section 3.3 and some commonly used methods to analyze association studies are presented in Section 3.4. A data-set concerning the metabolic syndrome is described in Section 3.5; this data-set was collected using a complex ascertainment scheme and is presented in order to emphasize the need for statistical methods which can handle complex data structures.

A further aim of this report is to present and evaluate some estimation methods which correct for complex ascertainment. Three simulation based methods are described in Section 4. These methods can handle more general data structures and ascertainment schemes than the methods described in Section 2.3. In Section 4.4 some simulations which were performed to evaluate these methods are presented. The results of these simulations are discussed in Section 4.5.

The work in this report was inspired by sampling schemes which have been used in genetic association studies. The methods presented in Section 4 are however not restricted to the analysis of genetic association studies, but could be useful whenever sampling is performed in a complex manner. For readers interested in other areas than genetic association studies it is possible to skip Section 3 without any major loss of understanding of the statistical reasoning.

2 Background of ascertainment

The word *ascertain* means "to discover with certainty, as through examination or experimentation" (Houghton Mifflin Company 1993). In statistics ascertainment refers to measuring variables on study units. The *ascertainment probability* is the probability for a study unit to be included in a particular sample. This probability may be the same for all units in the study population, as in a simple random sample (SRS), or may depend on some characteristics of the study units. In this report ascertainment probabilities will depend on the value of the outcome variable, but not on the exposure. Ideally the ascertainment probabilities are known, and even controlled, by the investigator. In this report the word ascertainment refers to selection with un-equal selection probabilities where these probabilities will be assumed to be known to the investigator, and is considered to be distinct from situations where subjects that are selected are not observed. Such non-response may be the case for example when working with human volunteers, since they are free to refuse to participate. The failure to observe data on selected units can sometimes be handled by missing data methodology such as multiple imputation, see for example Rubin & Schenker (1991) for an overview. Heckman (1979) handles similar problems in the area of econometrics.

The concept of ascertainment was originally introduced into genetics in family studies where selection probability typically depended on the number of affected children in families. This sampling scheme increases the efficiency whilst retaining a valid test, since under the null the test statistic distribution is invariant to the sampling scheme. But, as Fisher (1934) pointed out for segregation analysis, the ascertainment scheme will bias estimates if not accounted for. This is true for association studies as well, if the ascertainment scheme is ignored in the analysis, it will generally lead to both biased prevalence estimates, biased effect estimates and biased variance estimates. Distributional assumptions for the residuals are also likely to be unrealistic. In some testing situations the consequence of non-random ascertainment can be an increased false positive rate. For an example in the genetics context see Smoller, Lunetta & Robins (2000).

How to select subjects to increase power in genetic studies has been discussed by several authors. Morton & Collins (1998) discuss the benefits of different approaches of how to select cases and controls in genetic association studies. For example can cases and controls be chosen from the extremes of the distribution, or selection probability can be decided by a combination of disease status and some other variable, such as age at onset of the disease, family history of the disease or some environmental exposure. What selection scheme is optimal depends on what the relationship is between the gene and the studied trait. Selection on extreme values of the outcome in genetic studies has also been discussed by for example Purcell, Cherny, Hewitt & Sham (2001) who investigate optimal selection of sib-pairs for linkage analysis, and Schork, Nath, Fallin & Chakravarti (2000) who concludes that sampling from extremes may give substantial increase in power in studies with unrelated individuals. Allison, Heo, Schork, Wong & Elston (1998) do however argue that selection on extreme values is not always optimal in genetic studies. In simulations they observe that under some genetic models sampling on extremes reduces power. There may also be biological reason for not sampling on extreme values of the outcome, sometimes there may be reason to believe that really extreme values may be the result of rare environmental factors, such as accidents.

In this section some background for the statistical issues of analyzing non-randomly ascertained samples will be presented. First, in Section 2.1, some background on sampling strategies in epidemiology will be provided. Different likelihoods for the ascertained data will then be presented in Section 2.2 and then some methods for ascertainment correction are presented in Section 2.3. These methods do however only work in special cases or are not very efficient. Continuous outcomes often complicate analysis and it is often tempting to categorize variables to facilitate the analysis. Categorizing continuous outcomes does however have disadvantages, such as loss of power. In Section 2.4 the consequences of categorizing continuous outcomes are described in more detail. Another complication in the analysis is comorbidity, the association of diseases. The implications of comorbidity will be described briefly in Section 2.5.

The diabetes data-set that will be presented in Section 3.5 is an example of a situation where both continuous outcomes and comorbidity complicate the analysis. The methods of analysis for association studies that typically appear in the literature can not handle this level of complexity.

The statistical issues presented in this section are not specific to genetic association studies and the section can be read without knowledge about such. For notational coherency some notation referring to genetic terminology will however be used, G will be used to denote exposure (genotype score) and Ph will be used to denote outcome (phenotype). The words genotype and phenotype will be defined in Section 3.1.

2.1 Sampling strategies in epidemiology

Epidemiology is "the branch of medicine that deals with the causes, distribution, and control of disease in populations" (Houghton Mifflin Company 1993). Genetic association can be considered as a particular subclass of epidemiological studies. In epidemiology the outcome of interest is often a rare disease, such as a specific kind of cancer. A simple random sample (SRS) would thus have to be very large to attain reasonable statistical efficiency and power. To reduce costs but maintain power a study can be designed so that the inclusion probabilities of individuals depend in some way on their particular characteristics. The most common non-random ascertainment scheme that is employed in epidemiology is the case-control study, where affected individuals are assigned probabilities of being ascertained that are higher than those assigned to unaffected individuals. One advantage of the case-control design is that under a specific statistical model the non-random ascertainment scheme can be ignored at the data analysis stage without getting biased effect estimates; this is further discussed in Section 2.3. Disease status is often straightforward to assess, as is typically the case when registers are used for sampling from. If the disease status of subjects is unknown in the sampling stage, another variable, that is associated with the disease, is sometimes used to determine sampling probabilities. Studies where ascertainment probabilities are determined by more than one variable are not uncommon.

Sometimes non-random ascertainment is not a result of design but of inability to sample according to a desired sampling scheme. An example of this is when cases and controls are sampled from different populations. In an ideal situation the study population can be chosen directly from a well defined study base. In some countries, such as Sweden, there are population registries that facilitate sample selection, but often such registries are not available, or the cost of sampling from them would be too large. Cases are sometimes sampled from hospital admissions for a specific disease, and it is then very difficult to define the study base in order to select controls. A variant of the case-control study that is often used to avoid selection bias is the matched case-control study, where cases and controls are matched in pairs or bigger groups on the basis of some characteristic of the subjects. Common matching variables in epidemiology are gender and age. Ethnicity is a relevant matching variable in genetic studies if available. One way to match for ethnicity is to choose controls within the same family as the case. The matching does always have to be taken into account in the analysis and it is not possible to estimate the effect of a matching variable. A problem in case-control studies is that controls are often chosen from patients having other diseases, meaning a discovered disease marker could as well be a marker for not having the other disease. There is also the possibility that there is differential non-response in for example different ethnical groups. Since non-response often is larger in controls than in cases this can also introduce bias. Non-random sampling due to anything but a planned and well-documented sampling scheme can not be corrected for by statistical analysis. The importance of documentation of the sampling procedure should be stressed since there is often neither practical nor economical reason to prevent doing this, yet it is often not done.

2.2 Likelihoods of data under non-random ascertainment

In likelihood-based analysis of data, ascertainment can be handled in a number of different ways: four different likelihoods for data with outcome-dependent ascertainment are written below as (2.1)-(2.4). The genetic exposure is denoted G, the outcome is denoted Ph, and A is an indicator variable signifying whether ascertainment has/has not occurred. The conditional likelihood, (2.4), is appropriate only for discrete outcomes whilst likelihoods (2.1), (2.2) and (2.3) are appropriate for continuous as well as discrete outcomes.

Prospective likelihood

$$L(\theta; Ph, G) = P(Ph|G, A = 1, \theta) \tag{2.1}$$

Retrospective likelihood

$$L(\theta; Ph, G) = P(G|Ph, A = 1, \theta) = P(G|Ph, \theta) \tag{2.2}$$

Joint likelihood

$$L(\theta; Ph, G) = P(Ph, G|A = 1, \theta) \tag{2.3}$$

Conditional likelihood for matched data

$$L(\theta; Ph, G) = P(Ph|G, \sum Ph, \theta) \tag{2.4}$$

Kraft & Thomas (2000) compare the efficiency of the different likelihoods for family based case-control studies and conclude that the conditional likelihood is the least efficient of the four, and that the joint likelihood is the most effective. The relative efficiency of the prospective and the retrospective likelihoods varied with data structure and genetic model. Kraft & Thomas (2000) also describe properties of the likelihoods: the prospective likelihood and the joint likelihood demand explicit modelling of the ascertainment rule while the conditional likelihood and the retrospective likelihood do not. For the retrospective likelihood this is however only true when the ascertainment probability does not depend on the gene or on modelled covariates. Another disadvantage of the retrospective likelihood is that it is not always possible to obtain parameter estimates of the effect of exposure on outcome. It is possible to write the retrospective likelihood in terms of $P(Ph|G, \theta)$ as in (2.5) but the model is only identifiable under specific parameterizations (Chen 2003).

$$P(G|Ph, \theta) = \frac{P(Ph|G, \theta)P(G|\theta)}{P(Ph|\theta)} = \frac{P(Ph|G, \theta)P(G|\theta)}{\sum_G P(Ph|G, \theta)P(G|\theta)} \tag{2.5}$$

The conclusion is that, in general, the two likelihoods, (2.1) and (2.3), that require information about the ascertainment scheme, are unfortunately the only ones that are flexible enough to handle complex data structures and complex ascertainment schemes.

2.3 Some special cases where correction for non-random ascertainment is straightforward

For binary outcome data odds ratios (OR) are commonly used to analyze data under non-random ascertainment. An odds ratio for exposure i compared with a reference exposure j is defined as

$$OR_i = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}} \tag{2.6}$$

where π_i is the probability of success for exposure i. If there is differential ascertainment probabilities depending on outcome odds ratios will still give unbiased estimates. If we denote probability of success for exposure i in the ascertained sample $\pi_i^\star = \pi_i P(A = 1|\text{success})$ then the odds ratio calculated from the sample will be

$$\frac{\frac{\pi_i^\star}{1-\pi_i^\star}}{\frac{\pi_j^\star}{1-\pi_j^\star}} = \frac{\frac{\pi_i P(A=1|\text{success})}{(1-\pi_i)P(A=1|\text{failure})}}{\frac{\pi_j P(A=1|\text{success})}{(1-\pi_j)P(A=1|\text{failure})}} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}} \tag{2.7}$$

which is the same as the population odds ratio.

Binary outcomes can also be modelled with logistic regression. The logistic link models the odds and shares the odds ratios convenient property of giving unbiased effect estimates in case-control sampling (Prentice & Pyke 1979). The intercept is biased under non-random ascertainment even when a logistic link is used and can thus not be used to estimate prevalence of disease. The logistic link is the only link function that will give unbiased effect estimates without taking the ascertainment into account. Kagan (2001) proves this by comparing the likelihood under simple random sampling (2.8) with the likelihood under an ascertainment scheme (2.9), for both of these sampling schemes a prospective likelihood is applied.

$$P(Ph|G = g) = \prod_{i=1}^{n} \{h(\beta_{0Ph} + \beta_{GPh} \times g)\}^{Ph_i} \{(1 - h(\beta_{0Ph} + \beta_{GPh} \times g)\}^{1-Ph_i} \tag{2.8}$$

$$\begin{aligned} P(Ph|G = g, A = 1) &= \frac{P(A = 1|Ph = ph)P(Ph|G = g)}{P(A = 1)} \\ &= \prod_{i=1}^{n} \frac{\{h(\beta_{0Ph} + \beta_{GPh} \times g)\}^{Ph_i} \{r(1 - h(\beta_{0Ph} + \beta_{GPh} \times g)\}^{1-Ph_i}}{h(\beta_{0Ph} + \beta_{GPh} \times g) + r(1 - h(\beta_{0Ph} + \beta_{GPh} \times g))} \end{aligned} \tag{2.9}$$

where $h(\beta_{0Ph}+\beta_{GPh}\times g)$ is the inverse of the link function and $r = \frac{P(A=1|Ph=0,g)}{P(A=1|Ph=1,g)}$. Kagan (2001) concludes that these two likelihoods are equal, except for the intercept, if and only if the link function is of the form $h(u) = \frac{\exp(\lambda+\mu u)}{1+\exp(\lambda+\mu u)}$ for some λ and μ. If $\lambda = 0$ and $\mu = 1$ this gives $h(\beta_{0Ph} + \beta_{GPh} \times g) = \frac{\exp(\beta_{0Ph}+\beta_{GPh}\times g)}{1+\exp(\beta_{0Ph}+\beta_{GPh}\times g)}$. The bias of the intercept will be $-\frac{1}{\mu}\log(r)$.

Neuhaus (2000) describes how link functions can be adjusted to correct for ascertainment in binary regression models. The correction is based on modelling a prospective likelihood as in (2.1). The link function is corrected by replacing the mean by a function of the mean and the sampling probabilities. Neuhaus (2002) also specify what the bias will be when ascertainment is ignored for some common non-logistic link functions, and conclude that this bias can be substantial.

A simple, and more general, way to get unbiased estimates from data under non-random ascertainment would be to weigh each observation with $w_i = 1/$ (inclusion probability of subject i), see for example Armitrage & Colton (1999). In likelihood terms the weighted log likelihood contribution of individual i then is

$$w_i \log(L(\theta; y_i))$$

This method works for continuous outcomes as well as categorical. Weighting will however not give fully efficient results since the observations contribute with an unequal amount of information to the estimates.

Example of weighting: Linear regression

Since using the usual estimation equation for linear regression

$$\sum_i (Ph_i - (\beta_{0Ph} + \beta_{GPh} g_i)) \tag{2.10}$$

will give biased results in data with non-random ascertainment, a weighted regression solving a weighted estimating equation is used:

$$\sum_i ((Ph_i - (\beta_{0Ph} + \beta_{GPh} g_i)) \times w_i) \tag{2.11}$$

where w_i is proportional to 1/(inclusion probability of subject i). This will give unbiased results but is not fully efficient since highest efficiency of a weighted regression is obtained when w_i is proportional to 1/(variance of subject i), see for example Armitrage & Colton (1999). If the variance is equal for all individuals, then the efficiency is highest if all weights are also equal. ■

2.4 Disadvantages of categorizing continuous variables

Continuous variables are often categorized for computational simplicity, reduced cost, or due to lacking understanding of nature of the variable of interest. Categorization is very common in case-control studies since treating the outcome as binary allows a logistic regression model to be used without further ascertainment correction. However, there is an information loss in the categorization of continuous variables and this often leads to an unacceptably high power loss. Cohen (1983) compares the product-moment correlation between two normally distributed variables with the correlation when one of the variables is categorized and concludes that the reduction in correlation is about 20 percent when the data is split at the median, and even larger when the categories are of unequal size. If more than one variable are dichotomized the reduction in correlation follows a more complicated pattern and Cohen's formula should not be used (Vargha, Rudas, Delaney & Maxwell 1996). The reduction in efficiency has also been investigated in applications, for example by Neale, Eaves & Kendler (1994) who compare the power of continuous and categorized traits in genetic twin studies.

If the dichotomized variable is a *confounder*, a variable that affects the outcome of interest and that is also associated with the exposure, the information loss due to the dichotomizing can lead to insufficient confounder correction. Insufficient correction due to dichotomizing is discussed by Vargha et al. (1996).

Since the focus of this report is on ascertainment, it would be of interest to see how efficiency is affected by categorization when data is non-randomly ascertained. If individuals are chosen from the extremes of the outcome distribution, as discussed by for example Morton & Collins (1998), dichotomization is not likely to make a big difference, since there is little variation within the groups. If individuals are instead chosen from the whole range of the outcome variable the result of categorizing is less obvious. To illustrate how dichotomizing outcomes can affect power in an association study with non-random ascertainment the following example is considered.

Example: Dichotomizing an underlying continuous phenotype

This example is designed so that the statistical model has similar structure to that commonly estimated in genetic association studies. Some genetic vocabulary will be used to complement the statistical explanation. The genetic concepts are defined in Section 3.1.

The model is represented by Figure 2.1 where arrows indicate a directed causal relationship.

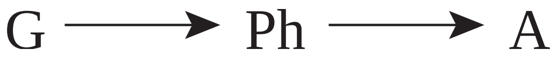

Figure 2.1: Data with ascertainment on phenotype

An explanatory variable, G, is assumed to be categorical and an outcome, Ph, is assumed to be conditionally normally distributed. Ascertainment, A, is dependent on the value of Ph according to a scheme described below. The explanatory variable has three categories based on values from a binomial distribution Bin(2,0.2). In genetic vocabulary this could be a biallelic SNP with an allele frequency of 0.2. The categories of the explanatory variable are then genotypes AA, Aa, aa. The relative effect of three genotypes on the outcome variable determine the genetic model. Here the scores that are used are $G = (0, 0, 1)$ for a *recessive* genetic effect and $(0, 0.5, 1)$ for an *additive* genetic effect. For a *dominant* effect $(0, 1, 1)$ are used.

The outcome Ph is assumed to be conditionally normally distributed where the mean depends on G, $Ph \sim N(\beta_{GPh} \times g, 1)$ where $\beta_{GPh} = 0.5$. In genetic terms we would call Ph a *phenotype*.

Individuals that have a phenotypic value over some cut-off are considered cases and all other individuals are controls. We suppose that the ascertainment scheme is such that there is an equal expected number of cases as controls regardless of cut-off, that is

$$P(A = 1|Ph < \text{cut-off}) = \frac{(1 - P(Ph < \text{cut-off}))}{P(Ph < \text{cut-off})} \tag{2.12}$$

and

$$P(A = 1|Ph \geq \text{cut-off}) = 1. \tag{2.13}$$

Simple random samples are also considered. The cut-off values are described in terms of quantiles, that is the proportion of the data that are controls for a SRS, $P(Ph < \text{cut-off})$.

The normal distribution of the phenotype is used for computational simplicity although in real data phenotypes do not always have this characteristic. It is also worth noting that if the genetic data is from a marker that is associated with the gene rather than a coding part of the gene, the phenotypes given the genetic data will be a mixture of distributions, as a result of the misclassification in the genotype. It is likely that the reduction in power due to categorization will be smaller when the data is not normally distributed and the results below should be interpreted with that in mind.

Power is calculated for detecting a genetic effect on the phenotype for both continuous phenotypes and dichotomized phenotypes. The power is calculated for a log likelihood ratio test with one degree of freedom, $T^2 = 2[l_A - l_0]$ where l_0 is the log likelihood under the null hypothesis ($\beta_{GPh} = 0$) and l_A the log likelihood under the alternative ($\beta_{GPh} \neq 0$). Since the data has a simple structure, modelling of the joint likelihood, defined in (2.3), is possible. For the details of how the power calculations were performed, see Appendix A.

In Figure 2.2 it can be seen that the power using the dichotomized phenotype is lower than using the continuous. This reduction in power is seen for each of the applied ascertainment schemes but the difference is most pronounced in the simple random sample, especially when a high cut-off is used. For both continuous and dichotomized phenotypes there is a gain in power in using the differential ascertainment probabilities compared with the SRS. This gain is larger the more extreme cut-off point is chosen. For high cut-off points an ascertainment sample with a dichotomized phenotype will give a higher power than a continuous variable in a SRS, but for a lower cut-off this will not be the case. ■

2.5 Comorbidity

A complication when correcting for ascertainment is comorbidity, the association of two or more diseases. Robins, Smoller & Lunetta (2001) argue that if comorbidity is present in data with non-random ascertainment it can result in non-valid tests. Robins et al. (2001) use causal directed acrylic graphs to show when tests are valid/not valid, and to illustrate how conditioning will affect the validity of tests. The tests discussed are *TDT* tests (Terwilliger &

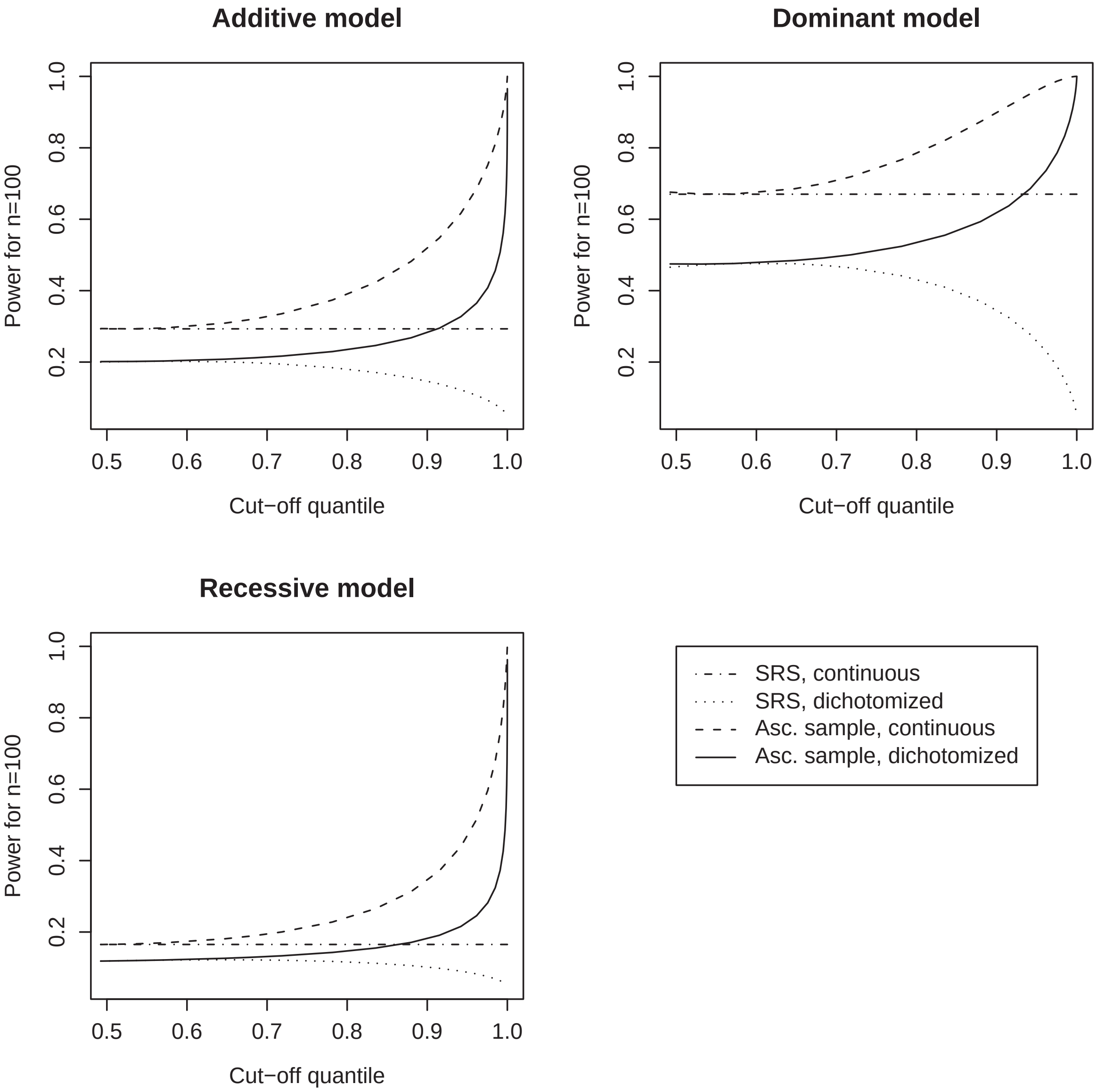

Figure 2.2: Power for different cut-off values for Example: Dichotomizing an underlying continuous phenotype

Ott 1992), which are tests of whether parents genetic material is inherited in different proportions in cases and controls.

An example of comorbidity is the metabolic syndrome which will be described in Section 3.5. In the metabolic syndrome there are many outcomes that may affect each other in different extents and directions. Here we will however only consider an extremely simplified model of the metabolic syndrome, as in Figure 2.3, incorporating only BMI (body mass index) and plasma glucose level. We will also assume that there is no causal effect of plasma glucose level on BMI, but only of BMI on plasma glucose level. Let the genotype score G affect both outcomes and let ascertainment probability A depend on both outcomes.

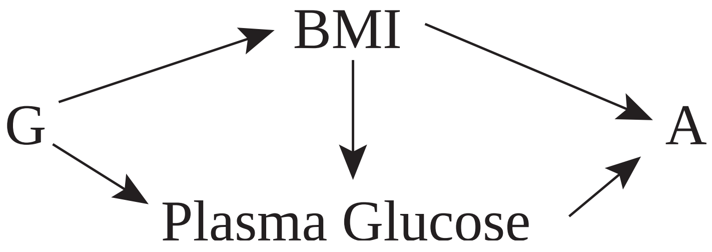

Figure 2.3: Data with comorbidity

Now consider a testing situation, investigating if there is an effect of the gene on plasma glucose. Figure 2.4 illustrates how the model would look without such an effect. Even when there is no causal effect of the gene on plasma glucose the gene and plasma glucose will be dependent through BMI, so we will have to condition on BMI to obtain a valid test for the direct relationship.

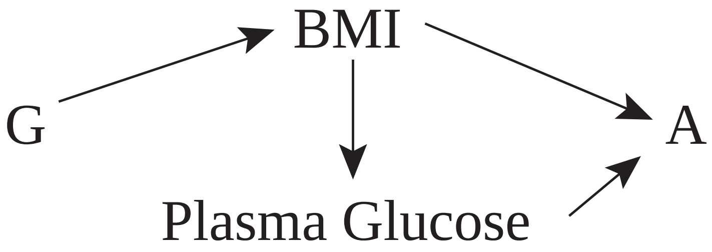

Figure 2.4: Data under H_0: No effect of G on plasma glucose

If we instead are testing if there is a causal effect of the gene on BMI the situation is more complicated. If there is no such effect, as in Figure 2.5, the gene and BMI will be dependent through plasma glucose and the ascertainment scheme. Conditioning on plasma glucose is however not a solution, such a conditioning would instead introduce dependence between the gene and BMI. In situations like this other methods could instead be considered, such as modelling the ascertainment corrected joint or prospective likelihood as suggested in Section 2.2.

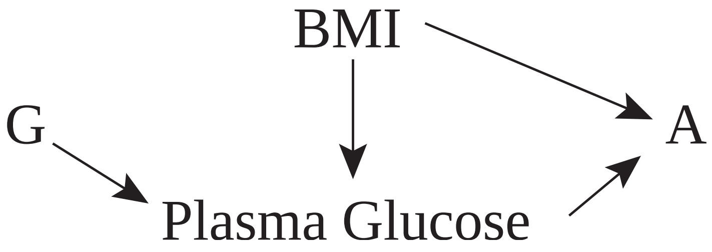

Figure 2.5: Data under H_0: No effect of G on BMI

3 Background of genetic association studies

3.1 Terminology and basic concepts of genetics

In this section some fundamental concepts concerning the structure of DNA, chromosomes and genes, which are relevant to the central theme of this report, are described.

The genetic code consists of strands of *nucleotides*, which consists of phosphate, sugar and *bases*. The bases in the nucleotides bind in pairs so that the nucleotides build a *double helix* that form a *chromosome*. There are four bases in the DNA, *adenine* (a), *guanine* (g), *cytosine* (c) and *thymine* (t). In the base-pairs *adenine* always binds to *thymine* on the opposite strand and *guanine* binds to *cytosine*. The human genome consists of more than 3,000,000,000 base-pairs, see for example Strachan & Reed (1999). Triplets of these base pairs code for the 20 amino acids that are used to build the about 90,000 different kinds of proteins that our bodies produces. A segment of bases that code for a protein is called a *gene*. Most gene products in the human genome are identical in all individuals. Sometimes when gene products vary between individual some variants are harmful and may cause disease. In this report statistical methods that can be used for identifying and characterizing genes with disease causing variants are considered.

Humans have 23 chromosome pairs. A location on a chromosome is called a *locus* (pl. loci). The locus can be either a single nucleotide or a string of nucleotides. Different variants that are present in the population at a specific locus are called *alleles*, if there are only two variants the locus is *biallelic*.

When an individual inherits DNA from his/her parents one copy of each chromosome is inherited from each parent. This means that for every individual there are two alleles at each locus, the number of a specific allele observed at a locus is thus 0, 1 or 2. The combined outcome of the two alleles at a locus is called a *genotype*. The two allele outcomes are often assumed to be independent and the number of a specific allele at a biallelic locus can be considered to be binomially distributed with two trials, p denoting the population allele frequency. In the case of more than two alleles genotypes will be multinomially distributed under independence. The independence property is referred to as Hardy-Weinberg equilibrium (HWE). It can be shown that genotype frequencies stabilize at Hardy-Weinberg proportions after a single mating, starting with any genotype frequencies that are equal in males and females if assumptions are met about of infinite population size, discrete generations and no selection, migration or mutation. Exceptions to HWE can result if a sample is derived from a mixture of different populations (population stratification), if there is non-random mating or if the probability of ascertainment or survival is not independent of genotype.

If a genotype consists of two copies of the same allele it is *homozygous* and otherwise *heterozygous*. Traits resulting from genotypes are called *phenotypes*. In this report the word phenotype will be used to indicate an observed variable in a broad sense. The phenotype can for example appear in a causal pathway for some final endpoint. Phenotypes that are considered final endpoints could be for example disease status and phenotypes in a causal pathway of the endpoint could for example be produced level of a specific protein.

When an individual inherits a chromosome from a parent it is not one of the two parental copies, each parental chromosome *recombine* on average about 1.5 times, see for example Strachan & Reed (1999), so that the inherited chromosome is a patchwork of the parental chromosomes. The location at which parental chromosomes recombine differ from generation to generation so that after several generations only small fragments of the original chromosomes remain and only bases that are located close together are inherited together. Recombination does not occur uniformly over chromosomes and distances between loci are often described in terms of recombination rather than physical distance, see for example Strachan & Reed (1999). Dependence between loci is called *linkage disequilibrium*.

It is typically not feasible to collect information on whole areas of the genome. Loci are instead selected which have previously been confirmed to be *polymorphic*, that is, where there exists variation between individuals. Areas that are segregating, but not necessarily coding for the gene of interest, are called *genetic markers*. When searching for a gene the hope is that markers are either in a coding part of the gene or are in linkage disequilibrium with the gene. Commonly used genetic markers are *single nucleotide polymorphisms* (SNPs) and *microsatellites*. A SNP is a variation between individuals in a single nucleotide base. The mapping of SNPs has been in rapid progress and currently approximately 2.7 million SNPs have been mapped (Carlson, Eberle, Rieder, Smith, Kruglyak & Nickerson 2003), most having been discovered in recent years. Microsatellites consist of tandem repeats of between one and five base-pairs. Microsatellites are more informative than SNPs since there are many possible allele but there are not as many microsatellites in the genome as there are SNPs, see for example Strachan & Reed (1999).

Combinations of alleles from different loci which reside on the same chromosome are called *haplotypes.* Typically genetic markers are measured one-at-a-time so that it is not always possible to infer *haplotype phase* with certainty, that is, which alleles belong together on the same chromosome. When information about parental phenotype is available phase can sometimes be inferred with certainty by observing which combinations are possible to inherit. Otherwise when phase is unknown the estimation of haplotypes can be viewed as a missing data problem and classical statistical methods can be used, see for example Excoffier & Slatkin (1995).

The way in which phenotype is related to genotype is referred to as *penetrance.* In the early days of genetics a few traits where found to be determined by single uncommon genes with full penetrance, that is the gene fully determined the phenotype. Genes like these are often called *Mendelian* after the monk Gregor Mendel who is often referred to as the father of genetics. Mendel studied traits in Moravian peas, such as color and texture (Mendel 1865), which are typical examples of traits that are determined by full penetrance genes. The full penetrance genes are relatively easy to identify since they create recognizable inheritance patterns in families and the gene locations were often found via *linkage analysis.* Linkage analysis is based on samples of related subjects and information about gene location is obtained by observing if the marker and the phenotype is inherited together in relatives. Most genes in humans do however not have full penetrance. A trait is called *complex* if multiple genes and/or environmental factors determine the trait. To be of clinical relevance a gene that affects a complex trait typically has to have a disease predisposing variant with a higher allele frequency of the than a full penetrance gene does. The relative efficiency of the study designs depend among other things on frequency of the disease causing allele and the penetrance. Rare alleles with a big effects are detected most efficiently in related individuals while for common, low penetrance genes, unrelated individuals are well suited.

As mentioned there are three possible genotypes, (AA, Aa, aa), at a single biallelic loci. If AA and Aa have the same effect on the phenotype the genetic model is referred to as *recessive* while if Aa and aa have the same effect on the phenotype the genetic model is referred to as *dominant.* There are also models where Aa has an intermediate effect of AA and aa. Such a genetic model is called *co-dominant.* In statistical analysis the co-dominant model will also be referred to as an additive or multiplicative model, depending on how effects are parameterized. To avoid making assumptions about the nature of the relationship between genotype and phenotype two variables can be used to describe the dependence instead of one. In this report the co-dominant model will be assumed most of the time but the methods described can be used for any genetic model.

3.2 Genetic association studies

This report covers some methods which are potentially of use for fine-mapping and characterization of genetic factors. We focus on what are known as genetic association studies. For a review of genetic association studies see Hirschhorn, Lohmueller, Byrne & Hirschhorn (2002). We concentrate on studies of unrelated individuals, but will discuss the use of related individuals in Section 3.3.1. We do not cover linkage analysis, which is more relevant for mapping genes to larger regions of the genome. The aim of genetic association studies is to find or characterize relationships between genes and phenotypes. The analysis is carried out by comparing phenotype distributions between persons with different genotypes. Typically the exact location of interest on the genome is not known and genetic markers, such as SNPs, are measured instead of the gene of interest.

To interpret the results of an association study it is important to understand which mechanisms can lead to association between marker and phenotype. The most desired reason is that the marker is causally related to the phenotype. It is also possible that the marker is in linkage disequilibrium with a causal gene. There may also be sources of bias in the data, population stratification, which was mentioned in Section 3.1, is an example of this. Comorbidity, which was discussed in Section 2.5 may also bias the analysis. Some of the sources of bias in epidemiological studies, like *recall bias*, differential recollection of events in cases and controls, are however not an issue, since genotype is constant over life and since measurement is not affected by subjective judgement.

3.3 Study design of genetic association studies

In recent years the fast development in the technology for analyzing genetic samples has created great optimism for finding and characterizing genetic causes to diseases. Statistical power and precision of genetic studies of complex diseases in humans are however typically low, since the gene of interest is only one of several possible reasons for getting the disease. Genetic factors are likely to be individually of small importance and as a consequence it is important that study designs are well thought through. In gene characterization estimates of gene effects are relevant and it may also be of importance to disentangle the effect of related phenotypes and environmental factors to understand the biological effect of the gene. If gene characterization is of interest, the study should be designed with this in mind so that power is calculated for realistic scenarios, all variables of interest are measured and so that the structure of the data allows estimation of effects. How sampling strategies can be used to increase power was described in Section 2.1 and in Section 3.3.1 designs with unrelated individuals will be compared with designs with related individuals.

3.3.1 Related or unrelated individuals?

The difficulty of gene characterization and identifying genetic association are clearly apparent, as is pointed out by for example Terwilliger & K.M. (2003). The structure of the data material will determine how well these difficulties are dealt with, in particular it is important if the selected individuals are known to be related to each other or if they are 'unrelated' subjects from the population.

If the studied population consists of more than one ethnical subgroup population stratification may be a problem in unrelated individuals. Bias can be introduced by population stratification if the subgroups differ in prevalence of the studied disease and in allele frequency of the studied marker even though the marker is not in linkage disequilibrium with the gene. In family data population stratification will not pose a problem since the information about association between gene and trait comes from within families.

Practical considerations will also often affect the choice of design. In family based studies it is desirable to have multiple affected individuals in each family. If the phenotype is complex with multiple measurements this is practically difficult to obtain. For late onset diseases it is also often hard to find family data since the parents and siblings of the affected individual are unlikely to be alive. On the other hand, relatives of persons affected by a disease are often more willing to participate in a study than unrelated individuals.

In reality persons labelled as unrelated do have common ancestors, generally so many generations back that the nature of the kinship is not known. In family data the shared regions in the genome will be large since few recombination events will have happened in closely related individuals while in unrelated individuals the shared regions will be small since many recombination events will have occurred in the chromosomes since the common ancestors. For that reason related individuals studies are used to search big areas for a gene location while unrelated individuals are used for fine-mapping when candidate areas are already identified.

In family studies effect estimates are generally not meaningful since subjects are on purpose sampled in such a manner that individuals that are likely to have a large gene effect are included in the study (Burton 2003). In unrelated individuals effect estimates can be computed but for multiple-testing reasons there is reason to be cautious about the interpretation of the results if combined with testing. If several studies investigate the same effect, the studies that by random had the highest effect estimates are most likely to get significant results. Given that almost only significant results get published it is likely that size of the effects in published studies are somewhat inflated.

3.4 Statistical analysis of association studies

In this section some fundamentals of statistical methods used in the analysis of association studies are briefly described. Let the probability of disease in a case-control setting be π_{AA}, π_{Aa} and π_{aa}, given genotype (AA, Aa, aa). The effect of the genotype on the risk of disease can for example be described either with a *relative risk* or an odds ratio. For both measures a reference group has to be chosen. If we let AA be the reference group the relative risk for genotype kl compared with the reference is

$$RR_{kl} = \frac{\pi_{kl}}{\pi_{AA}} \tag{3.1}$$

for π_{kl} = probability of disease given genotype kl. A multiplicative model in terms of relative risks would mean $RR_{aa} = RR_{Aa}^2$ The odds ratio, which was mentioned in Section 2.3, is in similar notation as the relative risk

$$OR_{kl} = \frac{\frac{\pi_{kl}}{1-\pi_{kl}}}{\frac{\pi_{AA}}{1-\pi_{AA}}}. \tag{3.2}$$

Asymptotic confidence intervals can be calculated for the odds ratios if a normal distribution is assumed for the logarithm of the odds ratio (Balding, Bishop & Cannings 2001). While relative risk is a somewhat more intuitive measure, odds ratios are convenient since the estimate of the odds ratio will be unbiased in a case-control setting, as described in Section 2.3. If the disease is rare for all genotypes odds ratios in the sample will be give very similar values to relative risks in the population (Balding et al. 2001).

Discrete phenotypes are often modelled with logistic regression. Most commonly a logistic link function is used but other link functions can be considered, for example if the outcome is a dichotomized normal variable a probit link would be appropriate. The advantage of the logistic link in case-control data was discussed in Section 2.3. When phenotypes are continuous standard statistical methods such as linear regression are often used, but since these methods require ascertainment correction continuous phenotypes are sometimes categorized to avoid this problem. This approach is however not efficient, as was discussed in Section 2.4.

Another way of testing for association is to use a log likelihood ratio test to test if a null hypothesis, such as $H_0 : \beta_{GPh} = 0$ where β_{GPh} is the effect of the gene of the phenotype, can be rejected in favor of some alternative hypothesis, such as $H_A : \beta_{GPh} \neq 0$. The test statistic is of the usual form

$$2[l_A - l_0] \tag{3.3}$$

where the log likelihoods l_A and l_0 are computed using the maximum likelihood estimates of the parameters. In the alternative hypothesis all parameters are estimated while under the null the parameters hypothesized about are fixed. The test statistic will be asymptotically chi-square distributed with degrees of freedom equal to the number of parameters used to describe the genetic effect (Balding et al. 2001).

3.5 Example of complex ascertainment in genetic association studies of unrelated individuals

There are numerous examples of genetic association studies where a complex ascertainment scheme has been used. Here one such study, concerning the *metabolic syndrome*, will be described briefly. The work presented in this report is inspired by this data-set and the goal of the evaluation of the methods in Section 4 is to find ways in which this kind of data can be analyzed in practice. The intention has been that the methods should be flexible enough to handle both discrete and continuous variables, correction for confounding and complex ascertainment.

The metabolic syndrome consists of a number of co-dependent phenotypes like insulin production and sensitivity, glucose levels, cholesterol levels, BMI, body fat distribution and hypertension. Diseases closely connected with these phenotypes are for example diabetes and coronary heart disease. The dependence between the phenotypes is complex and yet not fully disentangled, they may both affect each other and they may have common causes. Common causes could be for example fetal malnutrition or genetic effects (Stern 1995). While some relationships between phenotypes are likely to be causal, like the effect of BMI on diabetes, other are more controversial. Jarrett (1984) argue that while there is a statistical association between diabetes and risk for coronary heart disease it is more likely that the two diseases have a common cause than that diabetes affect the risk for coronary heart disease. Lifestyle factors such as diet and exercise also have a big impact on the metabolic syndrome. Since diabetes and coronary heart disease are common diseases research about the metabolic syndrome is of high relevance to public health.

One study concerning the metabolic syndrome is the Stockholm Diabetes Prevention Program (SDPP). For a more detailed description of the SDPP see for example Agardh, Ahlbom, Andersson, Efendic, Grill, Hallqvist, Norman & Ostenson (2003) or Gu, Abulaiti, Ostenson, Humphreys, Wahlestedt, Brookes & Efendic (2004). A part of the SDPP is to study genes which are believed to affect the metabolic syndrome and also to describe the effect of the genes on the different phenotypes to increase understanding of the biological mechanisms. As indicated in Figure 3.1, ascertainment probability depended on BMI, fasting glucose and 2 hour fasting glucose. The selection on the two last variables was done in two stages, first persons with known diabetes or impaired glucose tolerance were excluded since they were likely to be on medication affecting these values, then a selection was made oversampling persons that qualify for a diabetes diagnosis or an impaired glucose tolerance diagnosis based on measurements made in the study. Fasting glucose and 2 hour fasting glucose are used to diagnose diabetes and impaired glucose tolerance according to the WHO diagnostic criteria for diabetes, a plasma glucose level of at least 7.8 or a 2 hour fasting glucose level of at least 11.1 gives a diabetes diagnosis and while a fasting glucose level of less than 7.8 combined with a 2 hour fasting glucose level between 7.8 and 11.1

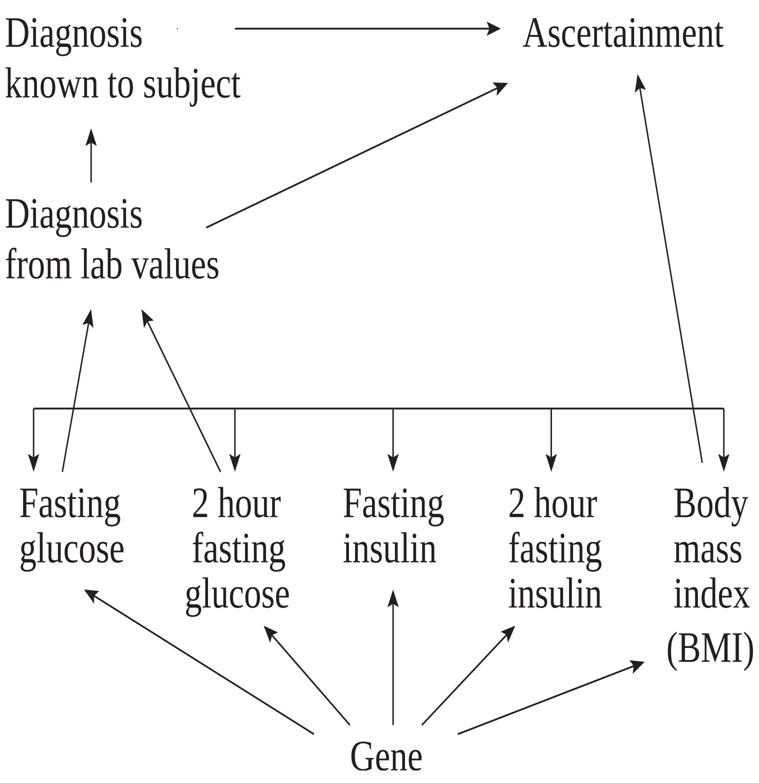

Figure 3.1: Study on the metabolic syndrome

gives an impaired glucose tolerance diagnosis. For controls there was an over-sampling on persons with a low BMI. Using this ascertainment scheme 500 controls, 339 persons with impaired glucose tolerance and 106 persons with diabetes were selected. Fasting plasma glucose level and 2 hour plasma glucose level are measured by separating plasma from blood taken from a fasting subject and then measuring the amount of glucose in the plasma, and BMI is calculated by dividing body weight in kilo by the squared height in meters. Other phenotypes, for example fasting insulin and 2 hour fasting insulin were also of interest in the analysis.

4 Estimation under complex ascertainment schemes

In this section some methods that can be used for estimation of parameters in statistical models, accounting for ascertainment, will be described. We will here assume that the probability that a unit is ascertained is known given phenotype, but typically this probability is estimated based on external information, for example registry data. As before, we use G to denote genotype score, Ph to denote phenotype and A to represent an indicator variable signifying that ascertainment has/has not occurred. Ph can be multivariate. We assume that the distribution of Ph conditional on G is parameterized by θ and that the probability of ascertainment is independent of θ given the observed data, that is $P(A = 1|G, Ph, \theta) = P(A = 1|G, Ph)$. Furthermore, we assume that ascertainment is independent of G conditional on Ph, $P(A = 1|G, Ph) = P(A = 1|Ph)$. The extension to let ascertainment depend on G is straightforward. We represent the dependence graphically as in Figure 4.1.

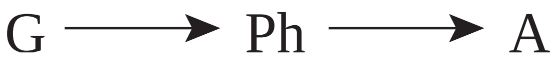

Figure 4.1: Data with ascertainment on phenotype

Where the likelihood of the data under non-random ascertainment is modelled we will deal with the joint likelihood, (2.3), rather than the prospective likelihood, (2.1). The methods described below are not restricted to the joint likelihood but it is convenient to model the data jointly for treatment of missing data on G, via the EM algorithm (Dempster, Laird, & Rubin 1977). An example of when this is of importance is in the estimation of haplotypes (Excoffier & Slatkin 1995). The possibility to extend the methods to handle haplotype estimation will however not be investigated in this report.

The joint likelihood of the data (G, Ph) that is ascertained is

$$L(\theta) = f(G, Ph|\theta, A = 1) = \frac{P(A = 1|Ph)f(G, Ph|\theta)}{P(A = 1|\theta)} \tag{4.1}$$

which corresponds to the log likelihood

$$\log(L) = \log(P(A = 1|Ph)) + \log(f(G, Ph|\theta)) - \log(P(A = 1|\theta)) \qquad (4.2)$$

where $\log(P(A = 1|Ph))$ does not depend on the model parameters. The complicated form of this likelihood makes standard likelihood-based estimation difficult. The computational problem is essentially that $P(A = 1|\theta) = \int P(A = 1|Ph)f(G, Ph|\theta)dPh$ is typically intractable, which is for example the case when Ph, conditional on G, is normally distributed.

One way to solve this problem is to use simulation based methods and it is such methods that we concentrate on and describe in the sections following. The methods described below all use simulation to correct for ascertainment but they differ in the distribution, from which data is simulated. In what we refer to as the *stochastic EM-algorithm*, (Section 4.3), the missing data is simulated. In the *data augmentation method* due to Clayton (2003), (Section 4.2), the ascertained data is simulated and in the *importance sampler*, (Section 4.1), the data is simulated from the population distribution. A way in which the stochastic EM-algorithm differs from the importance sampling and Clayton's method is that the stochastic EM-algorithm is an iterative procedure while the other methods do not require to be iterated, even if it is possible to do so.

4.1 Importance sampling

As mentioned above the difficulty in calculating the likelihood of the ascertained data lies in the integration in

$$P(A = 1|\theta) = \int P(A = 1|Ph)f(Ph, G|\theta)dPh. \qquad (4.3)$$

Importance sampling (Hammersly & Handscomb 1964) is a Monte Carlo method used for numerical integration. The basic idea is to sample from one distribution to obtain the expectation of another. This is advantageous for sampling efficiently but also when drawing samples from the target distribution is difficult. In general terms, for a random variable X which has density $f_1(x)$, the expectation of some function of X, $g(x)$, can be written as

$$\begin{aligned} \mu = E_{f1}[g(x)] &= \int g(x) f_1 dx \\ &= \int \frac{f_1}{f_2} g(x) f_2 dx = E_{f_2}[\frac{f_1}{f_2} g(x)] \end{aligned} \tag{4.4}$$

for $f_2 > 0$ whenever $g(x) \times f_1 > 0$. This means that samples can be drawn from f_2 to obtain the expectation of $g(x)$. Two possible estimates of the expectation above are

$$\hat{\mu} = \frac{\sum_i^M w_i g(x_i)}{\sum_i^M w_i} \tag{4.5}$$

and

$$\tilde{\mu} = \frac{\sum_i^M w_i g(x_i)}{M} \tag{4.6}$$

where $w = \frac{f_1}{f_2}$ and M is the number of simulated observations. For the importance sampler to give a good approximation of f_1, M should be large. The estimate $\hat{\mu}$ is sometimes more effective than $\tilde{\mu}$ but while $\tilde{\mu}$ is unbiased $\hat{\mu}$ has a bias of order $1/n$, see for example Elston, Olson & Palmer (2002). For other estimates of $E_{f1}[g(x)]$ see Hesterberg (1995). The choice of f_2 does effect the efficiency of the estimates, for $\tilde{\mu}$ the theoretically best distribution of f_2 is $c|g(x)|f_1$ for some constant c, see for example Hesterberg (1995). If f_2 is badly chosen a large variance for the estimate may result. Ways of choosing f_2 efficiently have been proposed by for example Torrie & Valleu (1977), Green (1992) and Geyer (1993).

We can apply the importance sampling technique to approximate $P(A = 1|\theta)$ in (4.1). It may be advantageous if we choose a distribution to simulate from which is close to the target distribution. One way to implement importance sampling in this context is to draw observations from a distribution which has the same parametric form as the target distribution $f(y|\theta)$, where $Y =$

(G, Ph), but in the place of the unknown θ, use naive guesses of the values of θ, which we call θ'. In this case $P(A=1|\theta)$ is estimated by noting that

$$P(A=1|\theta) = \int P(A=1|y)f(y|\theta)dy = \int [P(A=1|y,\theta)\frac{f(y|\theta)}{f(y|\theta')}]f(y|\theta')dy, \tag{4.7}$$

so that if we draw M observations from $f(y|\theta')$ which we denote as $y'_1, \ldots, y'_m$, we can estimate $P(A=1|\theta)$ using the estimator in (4.6). We then get

$$\hat{P}(A=1|\theta) = \frac{\sum_{j=1}^{M} P(A=1|y'_j)\frac{f(y'_j|\theta)}{f(y'_j|\theta')}}{M}. \tag{4.8}$$

As a consequence we can approximate the log likelihood contribution of individual i,

$$\log(L) \propto \log(f(y_i|\theta)) - \log(P(A=1|\theta)), \tag{4.9}$$

up to a constant, by replacing $P(A=1|\theta)$ by (4.8), thereby obtaining

$$\log(f(y_i|\theta)) - \log(\sum_{j=1}^{M} P(A=1|y'_j)\frac{f(y'_j|\theta)}{f(y'_j|\theta')}). \tag{4.10}$$

Since the approximation of the likelihood is expressed in terms of θ an approximation of the information matrix can be computed as minus the second derivative of the log likelihood as usual.

It would also be possible to construct an importance sampler by drawing from $f(y|\theta, A=1)$ instead of $f(y|\theta)$, and basing estimation of $P(A=1|\theta)$ on noting that

$$\begin{aligned} P(A=1|\theta) &= \int P(A=1|y)f(y|\theta)dy \\ &= \int [P(A=1|y)\frac{f(y|\theta)}{f(y|A=1,\theta')}]f(y|A=1,\theta')dy. \end{aligned} \tag{4.11}$$

but since the denominator, $f(y|A = 1, \theta') = \frac{P(A=1|y)f(y|\theta')}{P(A=1|\theta')}$, requires calculation of $P(A = 1|\theta')$, it is not practical doing so. We mention this because this is advocated as a possible approach by Clayton (2003) to correct for ascertainment. Clayton introduces this approach in the framework described in Geyer & Thompson (1992) where what is essentially importance sampling is used in the approximation of exponential family likelihoods, although it is described as Monte Carlo likelihood approximation. Clayton uses the notation from this paper and derives a likelihood that also incorporates correction for ascertainment. In Clayton's likelihood $P(A = 1|\theta)$ is written as

$$c(\theta) = P(A = 1|\theta) = \int_{y \in A} f(y|\theta)dy = \int_{y \in A} [\frac{f(y|\theta)}{f(y|\theta')}]f(y|\theta')dy \tag{4.12}$$

which is the same as (4.11), when ascertainment probabilities are defined to be 0/1 only, since

$$(4.11) = \int [P(A = 1|y)\frac{f(y|\theta)}{f(y|A = 1, \theta')}]f(y|A = 1, \theta')dy$$

$$= \int [P(A = 1|y)\frac{f(y|\theta)}{\frac{P(A=1|y)f(y|\theta')}{P(A=1)}}]\frac{P(A = 1|y)f(y|\theta')}{P(A = 1)}dy$$

$$= \int [P(A = 1|y)\frac{f(y|\theta)}{f(y|\theta')}]f(y|\theta')dy = \int_{y \in A} [\frac{f(y|\theta)}{f(y|\theta')}]f(y|\theta')dy = (4.12).$$

Based on this Clayton then uses

$$\frac{\sum_{j=1}^{M} \frac{f(y_j|\theta)}{f(y_j|\theta')}}{M} \tag{4.13}$$

as an estimator of $P(A = 1|\theta)$. This estimator does however not concur with the estimator (4.6) since the sampling distribution f_2 was $f(y|A = 1, \theta')$ and not $f(y_j|\theta')$, as suggested by using (4.13) as an estimator of (4.12). Clayton's data augmentation method will be used in this report (see Section 4.2), but we will use a matched case-control likelihood described in Section 3 of Clayton's paper and not the likelihood resulting from the reasoning above.

4.2 A data augmentation approach to the ascertainment problem

Clayton (2003) derives an ascertainment corrected likelihood by using an analogy to the conditional likelihood for matched case-control data. The idea behind this approach is to simulate a number of *pseudo-observations* for each real observation and use these in combination with the real data to build the likelihood. As in the importance sampling the true parameter values θ are unknown and are substituted by guesses, θ'. We will first review the conditional likelihood for matched case-control data and then describe how it applies to the ascertainment problem. For further description of the conditional likelihood for matched case-control data see Clayton & Hills (1996). The resulting likelihood resembles the importance sampling likelihood in Section 4.1, but a major difference is that data is simulated from the population distribution in the importance sampling but under the ascertainment scheme in the likelihood, (4.17), below.

The conditional likelihood is commonly used in the analysis of data from matched case-control studies. The data is matched based on characteristics that are believed to be potential confounders, such as age and gender. In each group there is a number of cases $y_1, \ldots, y_k = 1$ and a number of controls $y_{k+1}, \ldots, y_J = 0$. The notation $Z = \sum_1^J y_j = k$ will be used to indicate the number of cases in the set of J observations. For simplicity we let $k = 1$. The likelihood is based on the joint probability that $y_1 = 1$ and $y_2, \ldots, y_J = 0$, given that $Z = 1$

$$\begin{aligned} P(y_1 = 1, y_2, \ldots, y_J = 0 | Z = 1) &= \frac{P(y_1 = 1, y_2, \ldots, y_J = 0)}{P(Z = 1)} \\ &= \frac{P(y_1 = 1) P(y_2, \ldots, y_J = 0)}{\sum_{j=1}^J P(y_j = 1) P(y_{l \neq j}, \ldots, y_J = 0)} \\ &= \frac{\frac{P(y_1=1)}{P(y_1=0)} P(y_1 = 0) P(y_2, \ldots, y_J = 0)}{\sum_{j=1}^J \frac{P(y_j=1)}{P(y_j=0)} P(y_j = 0) P(y_{l \neq j}, \ldots, y_J = 0)} \\ &= \frac{\frac{P(y_1=1)}{P(y_1=0)}}{\sum_{j=1}^J \frac{P(y_j=1)}{P(y_j=0)}} \qquad (4.14) \end{aligned}$$

In Clayton's model it is not probability of being a case but probability for the observations to be real that is modelled. We can evaluate this probability using the same reasoning as that used in the case-control situation if we define an indicator variable R_j where $R_j = 1$ if y_j is a real observation, generated from $f(y|\theta, A)$, and $R_j = 0$ if it is a pseudo-observation, generated from $f(y|\theta', A)$. Let y_1 be the real observation. Then, for M pseudo-observations,

$$P(R_1 = 1, R_2, \ldots, R_{M+1} = 0|Z = 1, y) = \frac{\frac{P(R_1=1|y_1)}{P(R_1=0|y_1)}}{\sum_{j=1}^{M+1} \frac{P(R_j=1|y_j)}{P(R_j=0|y_j)}} = \frac{\frac{\frac{f(y_1|R_1=1)P(R_1=1)}{f(y_1)}}{\frac{f(y_1|R_1=0)P(R_1=0)}{f(y_1)}}}{\sum_{j=1}^{M+1} \frac{\frac{f(y_j|R_j=1)P(R_j=1)}{f(y_j)}}{\frac{f(y_j|R_j=0)P(R_j=0)}{f(y_j)}}}$$

$$= \frac{\frac{f(y_1|R_1=1)P(R_1=1)}{f(y_1|R_1=0)P(R_1=0)}}{\sum_{j=1}^{M+1} \frac{f(y_j|R_j=1)P(R_j=1)}{f(y_j|R_j=0)P(R_j=0)}} = * = \frac{\frac{f(y_1|R_1=1)}{f(y_1|R_1=0)}}{\sum_{j=1}^{M+1} \frac{f(y_j|R_j=1)}{f(y_j|R_j=0)}} = \frac{\frac{f(y_1|\theta)}{f(y_1|\theta')}}{\sum_{j=1}^{M+1} \frac{f(y_j|\theta)}{f(y_j|\theta')}}. \tag{4.15}$$

* The probability of $R = 1$ for an observation in the set of m observations is a constant $(1/m)$ if no information is provided about the observation.

Given the pseudo-observations The log likelihood contribution of individual i is

$$\log(\frac{f(y_i|\theta)}{f(y_i|\theta')}) - \log(\sum_{j=1}^{M+1} \frac{f(y_{ij}|\theta)}{f(y_{ij}|\theta')}) \tag{4.16}$$

which, up to a constant, can be written as

$$\log(f(y_i|\theta)) - \log(\sum_{j=1}^{M+1} \frac{f(y_{ij}|\theta)}{f(y_{ij}|\theta')}). \tag{4.17}$$

Since an expression for the likelihood is available parameter estimates can be obtained using maximum likelihood. Variances of these estimates are obtained as usual by calculating the information matrix from the likelihood For details see Appendix B.

The likelihood (4.17) is similar to the likelihood approximated with the importance sampler, (4.10), especially when ascertainment probabilities are 0/1. The essential differences are that

- Data is drawn under non-random ascertainment in (4.17), using Clayton's method, while it was drawn from the population distribution in (4.10), using the importance sampler.

- The sum in the second term is over the pseudo-observations only in (4.10) while the real observation are also included in (4.17).

- In (4.17) a separate estimate of $P(A = 1)$ is calculated for each real observation while in (4.10) $P(A = 1)$ is calculated only once.

The last of these differences means that while M pseudo-observations are produced in the importance sampler, for a sample size of n real observations, $M \times n$ pseudo-observations are produced in Clayton's method.

As Clayton suggests the data augmentation method can be iterated to refine the values of θ' in each step by using the parameter estimates from the previous iteration. There is however no guarantee that this will overcome problems of convergence that result from poor choices of θ', the point estimates may diverge if θ' is to far from θ in the first iteration.

As we will illustrate in Section 4.4 Clayton's method suffers, not surprisingly, from problems common to other simulation based methods, namely that a poor choice of sampling distribution may result in large variability in the estimates. Approaches that have been proposed in other contexts, such as importance sampling, may be useful.

4.3 Other sampling based algorithms, missing data

Although it does not completely fit into the classical framework of missing data problems (Little & Rubin 1987), non-random ascertainment can still be viewed in terms of a missing data problem. In the classical framework of a missing data problem there is a well-defined set of observations of which some the values are not observed and the data is partitioned into observed data, Y^{Obs}, and missing data ,Y^{Miss}. Usually there is partially complete information on each sample unit. In the ascertainment problem the unobserved observations are often of a different nature. If data is missing on an individual it is usually missing altogether, and it is not always even obvious how many individuals are unobserved. Nevertheless it is useful to consider algorithms used in missing data problems, such as the Estimation Maximization (EM) algorithm (Dempster et al. 1977) and it's extensions. We start by briefly describing the EM-algorithm as a background. An algorithm similar in spirit to the stochastic EM-algorithm is then described and is applied it to the ascertainment problem. These methods are frequentist. Bayesian approaches to the missing data problem will not be covered in this report.

The EM algorithm has been applied extensively to missing data problems. The EM algorithm can be used to obtain maximum likelihood estimates using a numerical technique. First starting values for the parameter estimates are decided upon and then the following two steps are iterated: In the E-step the expectation of the complete data is calculated using the parameter values, $\hat{\theta}$, from the previous M-step. In the M-step the maximum likelihood estimates, $\hat{\theta}$, from the complete data, created in the E-step, are calculated. It can be shown that if the likelihood has a unique maximum the EM-algorithm converges to that value (Wu 1983).

A problem that sometimes occurs when using the EM-algorithm is convergence to local maxima, the algorithm is therefore sensitive to what starting values are chosen. Another disadvantage of the EM-algorithm is that there is no direct way to calculate standard errors. One way to tackle the problem is to compute an asymptotic covariance matrix (Louis 1982). This approach uses the property that

$$-l''_{Obs}(\theta, y) = E_\theta[-l''_C(\theta, x)|y] - cov_\theta[l'_C(\theta, x)|y]. \tag{4.18}$$

The variance is obtained by taking the inverse of the observed information matrix $-l''_{Obs}(\theta, y)$.

If calculating the expected value of the missing data requires computationally demanding numerical integration one way to side-step the problem is to simulate the missing data and to use the value the observed mean instead of the calculated expectation. This is the Monte Carlo EM-algorithm (Wei & Tanner 1990). The algorithm is performed in two steps, in the S-step the missing data is simulated M times and in the M-step maximum likelihood estimates $\hat{\theta}$ are calculated using the combined data set containing observed and simulated data. Since the maximum likelihood estimates are calculated using the combined data set, the likelihood for the full data is used.

The stochastic EM-algorithm (SEM) (Celeux & Diebolt 1985) is a special case of the Monte Carlo EM-algorithm with only one simulation step per maximization step (McLachlan & Krishnan 1997). In iteration i the calculations are performed according to the following algorithm:

S-step: Simulate $M = 1$ set of the missing data Y^{Miss} using current parameter estimates $\hat{\theta}_{i-1}$.

$$\downarrow$$

Construct the complete data likelihood using the observed data and the simulated data:

$$\begin{aligned} L(\theta; Y^{Complete}) &= L(\theta; Y^{Obs}_{\in A=1}, Y^{Miss}_{\notin A=1}) \\ &\approx L(\theta; Y^{Obs}_{\in A=1}, Y^{Sim}_{\notin A=1}) \\ &= \prod f(Y^{Obs}_{\in A=1}, Y^{Sim}_{\notin A=1}|\theta) \\ &= \prod_{Obs} f(Y^{Obs}_{\in A=1}|\theta) \prod_{Sim} f(Y^{Sim}_{\notin A=1}|\theta) \end{aligned} \tag{4.19}$$

$$\downarrow$$

M-step: Get new parameter estimates $\hat{\theta}_i$ from (4.19) using maximum likelihood.

$$\downarrow$$

Repeat: Go to iteration $i + 1$ and repeat the steps above.

The stochastic EM will not converge to a single value but will have random variation, induced by the simulated data, around the estimate, and the result will be similar to that of a stationary Markov Chain Monte Carlo. See for example Gilks, Richardson & Speigelhalter (1996) for a description of Markov

Chain Monte Carlo. We will use the word convergence in a non-stringent manner to denote the process of the estimates moving towards a distribution around the correct value. In analogue with the terminology of Markov Chain Monte Carlo we will use the word *burn-in* to refer to the initial iterations of the chain that should be excluded from analysis in order to ensure that the estimates are produced by the right distribution.

Missing data problems such as censored data sets are applications where the stochastic EM-algorithm is useful (Ip 1994). Starting values are required for the stochastic EM-algorithm but it is more robust to misspecified starting values than the deterministic EM-algorithm (Gilks et al. 1996). One way of estimating parameters in the stochastic EM algorithm is to choose the set of parameter values in the iteration that gives the highest value of the likelihood for the observed data. The likelihood for the observed data might however be so complicated that this is unfeasible. A simpler way is to compute the mean, $\tilde{\theta}$, of the parameter values in the iterations after an appropriate burn-in period. An approximation of the variance of $\tilde{\theta}$ can according to Gilks et al. (1996) be computed using the method for the EM-algorithm by Louis that is described above. There are however some suspicions that this method may underestimate the variance (Gilks et al. 1996). For further reading about the stochastic EM-algorithm see for example Gilks et al. (1996) or McLachlan & Krishnan (1997).

We can implement an algorithm similar in spirit to the stochastic EM algorithm for the ascertainment problem. The non-ascertained data is considered missing and is imputed in the S-step using the parameter estimates from the previous M-step. Normally when the stochastic EM algorithm is used to fill in missing data there is a fixed sample size and data is filled in for those individuals where data is missing. Here we assume that the sample size of the full data is not known but only the ascertainment probabilities conditional on the data. This is however not a problem if the data is simulated as described below. In the S-step the missing data is filled in by rejection sampling (see for example Gilks et al. 1996), using a reverse ascertainment scheme:

Simulate: Simulate data from the population distribution $f(Y|\hat{\theta})$ and sort the observations into data that would have been ascertained, $Y^{Sim}_{\in A=1}$, and data that would not have been ascertained, $Y^{Sim}_{\notin A=1}$. Stop when n observations from $Y^{Sim}_{\in A=1}$ have been obtained.

↓

Reject: Throw out the observations in $Y^{Sim}_{\in A=1}$ and keep those in $Y^{Sim}_{\notin A=1}$.

The sample size, n, of the observed ascertained data, $Y^{Obs}_{\in A=1}$, is a known value but the size of the simulated data-set is random and will depend on $\hat{\theta}$. If population size is known, data is instead simulated until the combined data-set of $Y^{Obs}_{\in A=1}$ and $Y^{Sim}_{\notin A=1}$ has the appropriate sample size. In the M-step maximum likelihood estimates are obtained from the likelihood of the real ascertained data combined with the simulated non-ascertained data as described above.

If $Y^{Sim}_{\notin A=1}$ is large this algorithm will be slow. An alternative to sampling the whole set of missing data is to simulate only a portion of the data and weigh up the likelihood contribution of the simulated data. If a proportion of $1/k$ of the missing data is desired, data is simulated as above until n/k observations from $Y^{Sim}_{\in A=1}$ have been produced. If n/k is not an integer randomization can be used to determine if it should be rounded up or down. Alternatively n/k can be fixed to an integer value and the value of k calculated. Small values of n/k will cause large variability in the estimates so the choice of k is a balance between sample size and number of steps in the chain. Ripatti, Larsen & Palmgren (2002) suggest a rule for increasing the number of samples in a Monte Carlo EM-algorithm when approaching convergence. In this context simulating a proportion of $1/k$ of the missing data would correspond to simulating $1/k$ samples. The basic idea of altering the number of samples when approaching the estimate might however be used also in this context. If the size of the missing data is small it is of course also possible to choose $M > 1$, giving an algorithm similar to the Monte Carlo EM. In the simulations in Section 4.4 $M = 1$ will be used.

4.4 Simulations

To illustrate how the different methods perform in our context some simulations are performed. A simple model with only one phenotype is first investigated and then a few runs for a model with two co-dependent phenotypes are made.

The phenotypes are simulated to mimic two phenotypes in the metabolic syndrome, BMI and fasting plasma glucose level. According to WHO a BMI of at least 30 indicate obesity. About 10 percent of the Swedish population in the ages of 25-64 have such a BMI according to the WHO MONICA project (WHO 2000). A plasma glucose level of at least 7.8 or a 2 hour fasting glucose level of at least 11.1 gives a diabetes diagnosis according to the WHO diagnostic criteria for diabetes. We will use a BMI of 30 and a plasma glucose level of 7.8 as cut-offs in the ascertainment schemes.

For each method a number of simulations are performed where a new data-set is produced under the ascertainment scheme in each simulation, this is our 'real' data that is then analyzed. The sample-size in each data-set is 300 if another sample-size is not indicated. The parameter estimates presented in Table 4.2-4.9 are the mean of the estimates in 100 simulations under the same conditions and standard errors reported measure the variation between parameter estimates in these simulations.

Starting values are required for the stochastic EM-algorithm and parameter values for the simulated data have to be specified for the other methods. Simulations are run both for correctly specified values and for misspecified values to investigate how the methods perform both under ideal and not so ideal conditions. For simplicity we will denote both the parameter values for the sampling distributions in the importance sampler and in Clayton's method and the starting values in the SEM as θ'.

The number M has to be decided for Clayton's method and the importance sampler, recall that in Clayton's method M is the number of pseudo-observations per 'real' observation while in the importance sampler it is the total number of pseudo-observations.

Clayton's model and the importance sampler are run under some different values of M for the simpler model to investigate the effect of M on the variability of the estimates. In the comparison between methods $M = 50$ pseudo-observations are used for Clayton's method and in the importance sampler $M = 30000$ has been used.

In the analysis the phenotype and the genotype frequency are modelled with a joint likelihood, as in (2.3). Genotype scores, here representing outcomes of SNP's, will be assumed to be binomially distributed $G \sim Bin(2, \frac{\exp(\beta_{0G})}{1+\exp(\beta_{0G})})$ and phenotypes will be normally distributed.

The simulations are carried out using the software R (The R Development

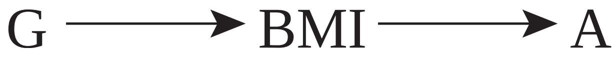

Figure 4.2: Model *i*

Core Team 2001). In all three methods the R optimizing algorithm **"optim()"** is used to calculate maximum likelihood estimates. The starting values used in the algorithm is θ' in the importance sampler and in Clayton's method. In the stochastic EM-algorithm the current value of $\hat{\theta}_{i-1}$ is used.

4.4.1 Model *i*

In this model we have a genotype represented by a biallelic SNP with an allele frequency of $\frac{\exp(\beta_{0G})}{1+\exp(\beta_{0G})} \approx 0.2$, ($\beta_{0G} = -1.4$), and the genotype score is $G \sim Bin(2, \frac{\exp(\beta_{0G})}{1+\exp(\beta_{0G})})$. The genotype is assumed to have an additive effect of $\beta_{GPh} = 4$ per copy of the rare allele, on a normally distributed phenotype (Ph), BMI, with standard deviation $\sigma_{Ph} = \sqrt{2}$. The intercept, the mean phenotype value for genotype AA, is $\beta_{0Ph} = 24$. That is; $Ph|G = g \sim N(\beta_{0Ph} + \beta_{GPh} \times g, \sigma_{Ph})$. The ascertainment is made on the phenotype. All individuals with a BMI larger or equal to 30 are selected while individuals with a lower BMI have an ascertainment probability of about 0.067. This will give an approximately equal number of cases as controls. The model is illustrated by the graph in Figure 4.2.

4.4.2 Model *ii*

The genotype in this model is also represented by a biallelic SNP, with the same allele frequency as in model *i*. Instead of one phenotype as in model *i* we here have two, Ph_1 =BMI and Ph_2 = plasma glucose level. BMI is here a *co-morbid* disease of plasma glucose level, that is, BMI is affected by the gene and will in turn affect the plasma glucose level. The genotype is assumed to have an additive effect on both phenotypes, and Ph_1 will have an additive effect on Ph_2. Given the genotype, Ph_1 has distribution $N(\beta_{0Ph_1} + \beta_{GPh_1} \times g, \sigma_{Ph_1})$ where $\beta_{0Ph_1} = 24$, $\beta_{GPh_1} = 4$ and $\sigma_{Ph_1} = \sqrt{2}$

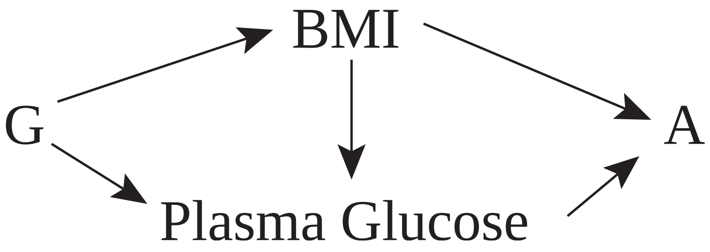

Figure 4.3: Model *ii*

while Ph_2 will, given genotype score, $G = g$, and BMI value, $Ph_1 = ph_1$, have distribution $N(\beta_{0Ph_2}+\beta_{GPh_2}\times g+\beta_{Ph_1Ph_2}\times ph_1, \sigma_{Ph_2})$, where $\beta_{0Ph_2} = 3$, $\beta_{GPh_2} = 1$, $\beta_{Ph_1Ph_2} = 1/15$ and $\sigma_{Ph_2} = 0.5$. The ascertainment probability is dependent upon both phenotypes, according to Table 4.1. Model *ii* is illustrated by the graph in Figure 4.3.

	$Ph_1 < 30$	$Ph_1 \geq 30$
$Ph_2 < 7.8$	0.1	0.3
$Ph_2 \geq 7.8$	0.3	1

Table 4.1: Ascertainment probabilities in model *ii*

4.5 Results of simulations

4.5.1 Results for model *i*

Clayton's model and the importance sampler for different values of M:

To investigate how the number of pseudo-observations affects the parameter estimates in Clayton's method and the importance sampler, these models were run for different values of M. The true parameter values θ were here used as θ'. In the tables standard errors of the estimates are presented in parentheses.

Clayton's method was run for M =2, 5, 10 and 50 as seen in Table 4.2. Since M pseudo-observations was produced for each real observation the total number of pseudo-observations was $M \times n$ = 600, 1500, 3000 and 15000. As Clayton (2003) points out, the information loss in the method seems to be of the order of $M/(M+1)$.

	True	M=2	M=5	M=10	M=50
$\hat{\beta}_{0G}$	-1.4	-1.388 (0.096)	-1.395 (0.092)	-1.408 (0.073)	-1.403 (0.062)
$\hat{\beta}_{0Ph}$	24	23.999 (0.174)	23.995 (0.141)	24.009 (0.146)	24.007 (0.110)
$\hat{\beta}_{GPh}$	4	3.996 (0.152)	3.996 (0.130)	3.988 (0.117)	3.991 (0.098)
$\hat{\sigma}_{Ph}$	$\sqrt(2) \approx 1.414$	1.420 (0.072)	1.415 (0.065)	1.415 (0.060)	1.416 (0.054)

Table 4.2: Clayton's method run for different number of pseudo-observations, $n = 300$

The importance sampler was also investigated with respect to M. For the Importance sampler M is the total number of pseudo-observations so $M =$ 600, 1500, 3000 and 15000 were chosen. Since the importance sampler seemed to need a larger total number of pseudo-observations than Clayton's method the importance sample was also run for $M = 30000$.

In the analysis below $M = 50$ will be used in Clayton's method and $M =$ 30000 will be used in the importance sampler.

Comparison of models:

Results of the three simulations based methods, calculated where true parameter values were used as θ', are presented in Table 4.4. Naive estimates, calculated by optimizing the likelihood of the data without ascertainment

	True	M=600	M=1500	M=3000	M=15000	M=30000
$\hat{\beta}_{0G}$	-1.4	-1.359	-1.384	-1.399	-1.404	-1.394
		(0.146)	(0.097)	(0.090)	(0.087)	(0.075)
$\hat{\beta}_{0Ph}$	24	24.031	24.025	24.022	24.022	23.996
		(0.200)	(0.138)	(0.131)	(0.139)	(0.131)
$\hat{\beta}_{GPh}$	4	3.979	3.987	3.991	3.991	4.008
		(0.214)	(0.159)	(0.131)	(0.115)	(0.119)
$\hat{\sigma}_{Ph}$	$\sqrt{(2)} \approx 1.414$	1.429	1.406	1.407	1.406	1.416
		(0.108)	(0.071)	(0.065)	(0.050)	(0.054)

Table 4.3: Importance sampler run for different values of M, $n = 300$

correction, are also presented. Standard errors of the estimates are presented in parenthesis.

	True values	Naive estimates	Importance sampling	Clayton's method	SEM
$\hat{\beta}_{0G}$	-1.4	-0.107	-1.394	-1.403	-1.400
		(0.083)	(0.075)	(0.062)	(0.068)
$\hat{\beta}_{0Ph}$	24	24.409	23.996	24.007	23.988
		(0.141)	(0.131)	(0.110)	(0.100)
$\hat{\beta}_{GPh}$	4	4.152	4.008	3.991	4.006
		(0.105)	(0.119)	(0.098)	(0.106)
$\hat{\sigma}_{Ph}$	$\sqrt{(2)} \approx 1.414$	1.589	1.416	1.416	1.408
		(0.059)	(0.054)	(0.054)	(0.055)

Table 4.4: Comparison of models when $\theta' = \theta$. $n = 300$

Any observed differences in variances of estimates between models in Table 4.4 should be interpreted with caution since the variance of an estimate based on Clayton's method and the importance sampler depends on M, and the variance of an estimate based on using the SEM depends on chain length.

Analysis with misspecified θ'

In Table 4.5 parameter estimates and standard errors of estimates for the importance sampler are presented for $\beta'_{GPh} = 4$, 2 and 0, the correct value of β_{GPh} is 4. For the other parameters $\theta' = \theta$ was used. In Table 4.6 corresponding results are presented for Clayton's model. Misspecified θ' affect the standard errors of the estimates in both importance sampling and in Clayton's model; the effect seems to be more pronounced in the importance sampler than in Clayton's model. The effect of the misspecification on the standard errors is not linear. It is worth noting that the standard error of $\hat{\beta}_{GPh}$ in the importance sampler actually seems to be larger for slightly misspecified β'_{GPh} than for more severely misspecified β'_{GPh}, as can be seen in Table 4.5, Table 4.7 or Table 4.8. The effect estimates seem however to be biased for misspecified parameter values so the standard errors may not be a satisfying tool for comparison of the methods.

Since the importance sampler estimator is claimed to be unbiased it may seem surprising that the parameter estimates are biased. As stated in Section 4.1, a condition for the importance sampler is that the sampling distribution f_2 should be positive whenever $g(x) \times f_1 > 0$. This condition is fulfilled in the simulations above, but when θ' is misspecified f_2 may be so small for some $g(x) \times f_1$, that no observations are actually sampled from these regions. Further investigation is however required determine if this is the cause of the bias in the parameter estimates.

	True	$\beta'_{GPh} = \beta_{GPh} = 4$	$\beta'_{GPh} = 2$	$\beta'_{GPh} = 0$
$\hat{\beta}_{0G}$	-1.4	-1.394 (0.075)	-1.189 (0.310)	3.213 (2.081)
$\hat{\beta}_{0Ph}$	24	23.996 (0.131)	23.850 (0.212)	26.342 (1.833)
$\hat{\beta}_{GPh}$	4	4.008 (0.119)	4.292 (0.352)	4.687 (0.304)
$\hat{\sigma}_{Ph}$	$\sqrt{(2)} \approx 1.414$	1.416 (0.054)	1.401 (0.156)	2.001 (0.370)

Table 4.5: Misspecified θ' in importance sampling. $n = 300$ M=30000

	True	$\beta'_{GPh} = \beta_{GPh} = 4$	$\beta'_{GPh} = 2$	$\beta'_{GPh} = 0$
$\hat{\beta}_{0G}$	-1.4	-1.403	-1.409	-1.316
		(0.062)	(0.098)	(0.344)
$\hat{\beta}_{0Ph}$	24	24.007	24.000	23.991
		(0.110)	(0.130)	(0.133)
$\hat{\beta}_{GPh}$	4	3.991	3.984	4.094
		(0.098)	(0.136)	(0.414)
$\hat{\sigma}_{Ph}$	$\sqrt{(2)} \approx 1.414$	1.416	1.392	1.434
		(0.054)	(0.070)	(0.093)

Table 4.6: Misspecified θ' in Clayton's model. $n = 300$ $M = 50$

For the SEM misspecified θ' does not have the effect of inflated standard errors of the estimates. If the SEM converges to a distribution around the parameter estimates the standard errors of the estimates after the convergence will be the same as for correctly specified starting values. The running time of the SEM will be longer when θ' is misspecified to allow for convergence and similar as to in Marcov Chain Monte Carlo simulations an appropriate burn in period has to be identified. There is a risk that the parameter estimates in the SEM will diverge when θ' is misspecified, this did however not happen in any of our simulations.

To illustrate the behavior of the SEM under misspecified starting values a chain was run under $\beta'_{0G} = 0$, $\beta'_{0Ph} = \beta_{0Ph}$, $\beta'_{GPh} = 0$ and $\sigma'_{Ph} = \sigma_{Ph}$ and the parameter estimates plotted in Figure 4.4. The values of θ' chosen here are meant to resemble a situation where there is some knowledge of the distribution of the phenotype in the population but no knowledge of the allele frequency or the gene's effect on the phenotype.

If Clayton's method is run for the same values of θ' as the SEM the estimates seem to be unbiased but with large variances while the Importance sampling did not give useful estimates at these values of θ'. A possible way to obtain better starting values is to use the naive parameter estimates as θ'. Running Clayton's method from naive estimates did however still give large variances.

For Clayton's model theoretical variance estimates were also calculated, for technical details see Appendix B. The theoretical variances seemed to concur with the observed variances and will therefore not be reported separately. If

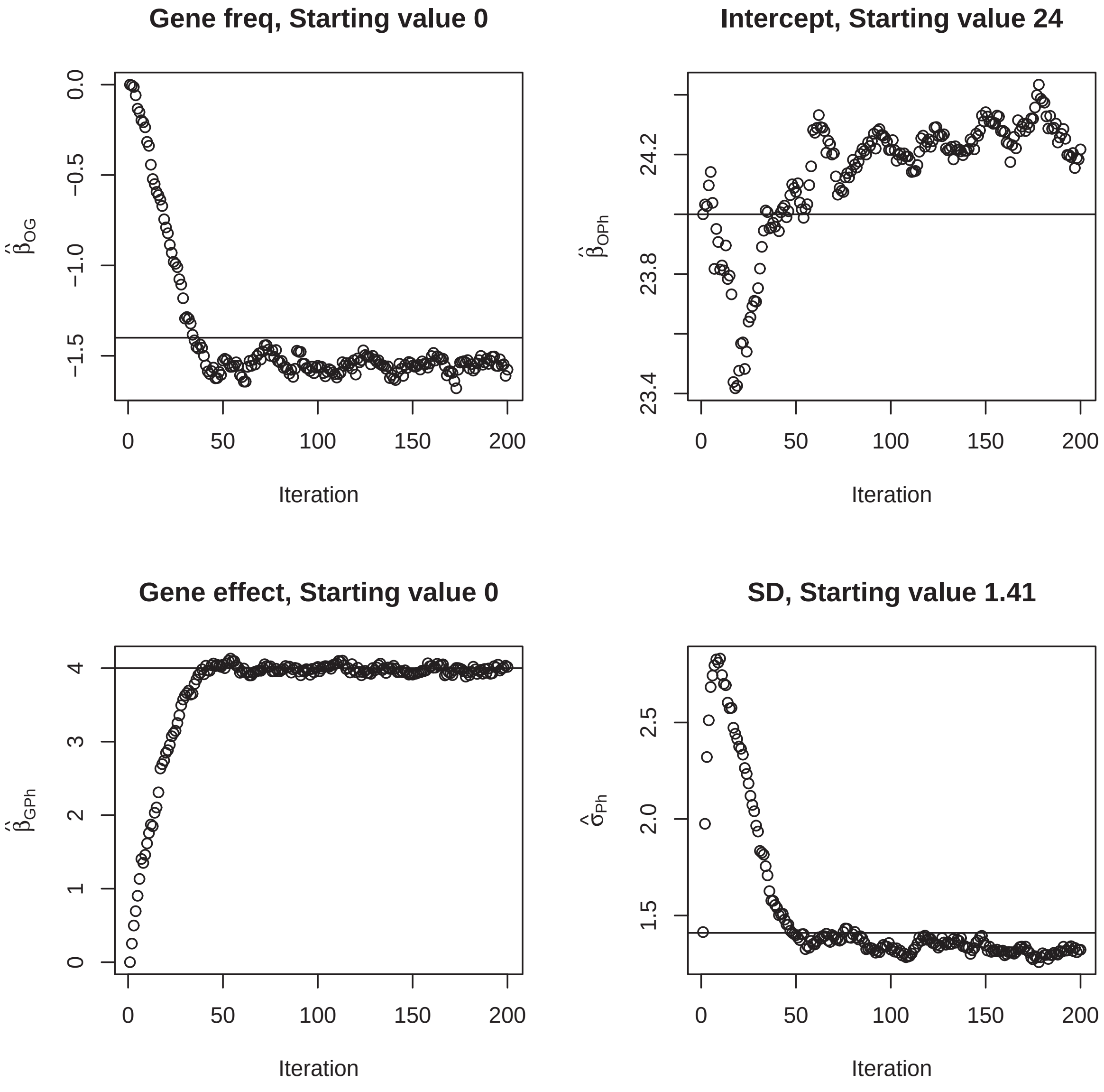

Figure 4.4: The first 200 iterations in SEM model i for misspecified θ', True parameter values as solid line, $n = 300$

θ' is severely misspecified so that the variances of the estimates are very large the information matrix needed for the theoretical variance estimates will sometimes be singular.

The importance sampler for different sample sizes:

It would be of interest to see if the effect of misspecification of θ' is dependent upon the sample-size, n, of the data. To investigate this the importance sampler with misspecified β'_{GPh} was run for $n = 600$, and $n = 1200$. The results are presented in Table 4.7 and 4.8.

As can be seen by comparing Table 4.7 and 4.8 with Table 4.5, where $n = 300$, the standard errors seem to be of similar size regardless of sample-size when $\beta'_{GPh} = 0$. When β'_{GPh} is not misspecified the standard errors decrease with sample-size in the usual rate. An interpretation of the results is that the variability resulting from misspecification dominates over the variability from the data when $\beta'_{GPh} = 0$, so that the decrease of the standard deviation due to the increase in sample-size, n, is harder to detect.

	True	$\beta'_{GPh} = \beta_{GPh} = 4$	$\beta'_{GPh} = 2$	$\beta'_{GPh} = 0$
$\hat{\beta}_{0G}$	-1.4	-1.399	-1.197	3.039
		(0.050)	(0.301)	(1.676)
$\hat{\beta}_{0Ph}$	24	24.003	23.862	26.459
		(0.095)	(0.204)	(1.929)
$\hat{\beta}_{GPh}$	4	3.999	4.239	4.705
		(0.079)	(0.372)	(0.278)
$\hat{\sigma}_{Ph}$	$\sqrt(2) \approx 1.414$	1.414	1.389	2.083
		(0.039)	(0.156)	(0.386)

Table 4.7: Misspecified θ' in importance sampling. $n = 600$ $M = 30000$

4.5.2 Results for model *ii*

When using $\theta' = \theta$ in model *ii* both the importance sampler, Clayton's method and the SEM gave reasonable estimates . The SEM was somewhat time-consuming to run in for this model.

	True	$\beta'_{GPh} = \beta_{GPh} = 4$	$\beta'_{GPh} = 2$	$\beta'_{GPh} = 0$
$\hat{\beta}_{0G}$	-1.4	] -1.395	-1.186	2.968
		(0.045)	(0.286)	(1.657)
$\hat{\beta}_{0Ph}$	24	24.000	23.837	26.482
		(0.055)	(0.194)	(3.050)
$\hat{\beta}_{GPh}$	4	3.997	4.261	4.661
		(0.049)	(0.373)	(0.272)
$\hat{\sigma}_{Ph}$	$\sqrt(2) \approx 1.414$	1.415	1.390	2.068
		(0.026)	(0.124)	(0.445)

Table 4.8: Misspecified θ' in importance sampling. $n = 1200$ $M = 30000$

Model ii was also run for misspecified θ', the values of θ' were $\beta_{0G} = 0$ $\beta'_{0Ph_1} = \beta_{0Ph_1}$, $\beta'_{GPh_1} = 0$, $\sigma'_{Ph_1} = \sigma_{Ph_1}$, $\beta'_{0Ph_2} = \beta_{0Ph_2}$, $\beta'_{GPh_2} = 0$, $\beta'_{Ph_1 Ph_2} = 0$, and $\sigma'_{Ph_2} = \sigma_{Ph_2}$. As in model i these values were chosen as if some knowledge was available about the distribution of the phenotypes while no knowledge was available about the allele frequency or the effect of the gene. The effect of BMI on plasma glucose level was also assumed unknown. As can be seen in Table 4.9 these values of θ' do not give adequate parameter values in neither Clayton's method nor the importance sampler. The SEM does converge but does take longer to converge than in model i, a run of the SEM is shown in Figure 4.5.

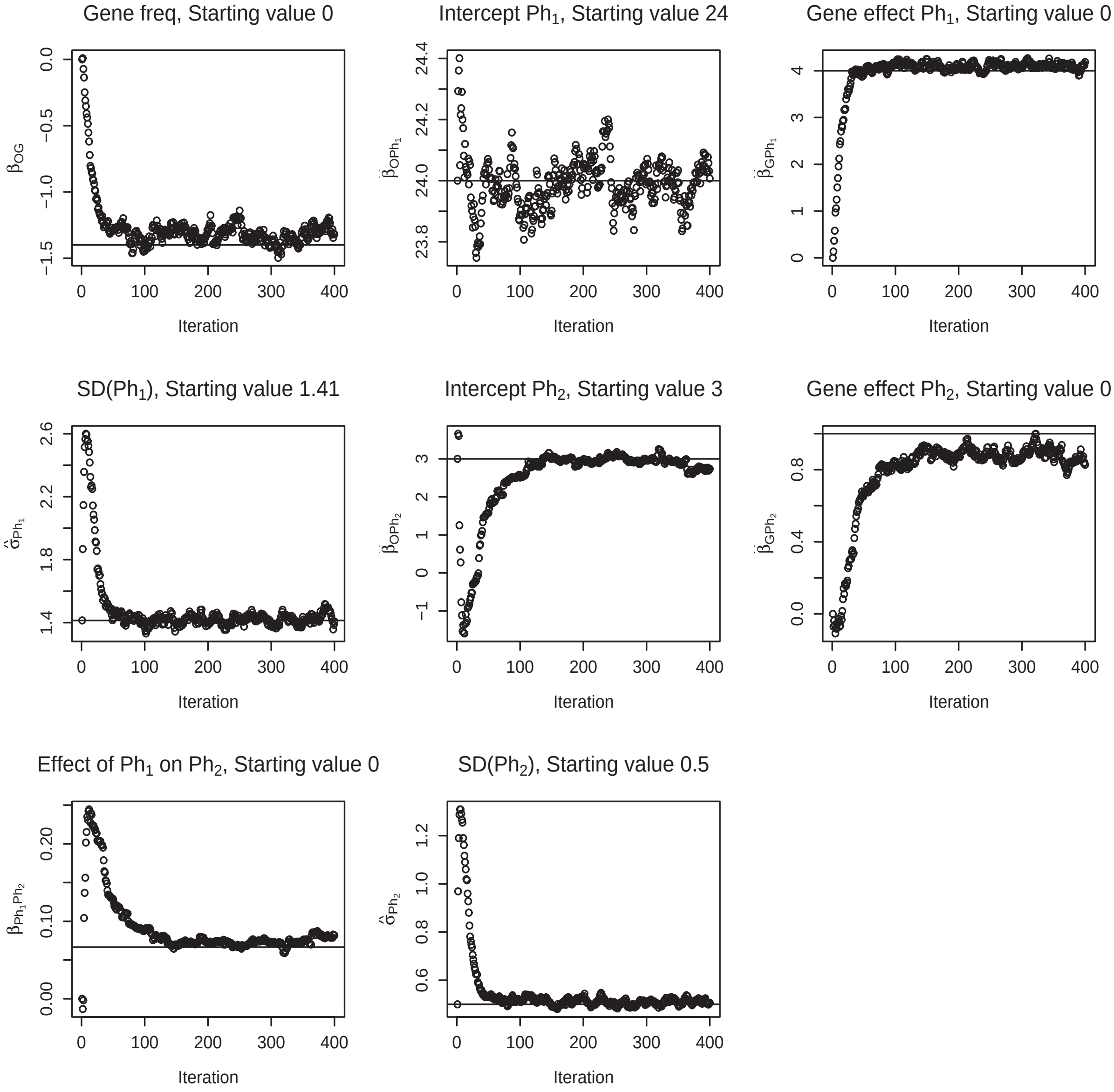

Figure 4.5: The first 400 iterations in SEM model *ii* for misspecified θ', True parameter values as solid line. $n = 300$

	True θ	θ'	Clayton	Importance sampling
$\hat{\beta}_{0G}$	-1.4	0	-0.376 (1.049)	-1.001 (0.935)
$\hat{\beta}_{0Ph_1}$	24	24	23.654 (1.094)	24.371 (0.898)
$\hat{\beta}_{GPh_1}$	4	0	0.304 (0.432)	0.584 (1.015)
$\hat{\sigma}_{Ph_1}$	$\sqrt(2) \approx 1.414$	$\sqrt(2)$	1.719 (0.200)	1.704 (0.382)
$\hat{\beta}_{0Ph_2}$	3	3	4.825 (0.325)	4.977 (0.515)
$\hat{\beta}_{GPh_2}$	1	0	0.557 (0.191)	0.645 (0.390)
$\hat{\beta}_{Ph_1Ph_2}$	$1/15 \approx 0.067$	0	0.061 (0.028)	0.015 (0.020)
$\hat{\sigma}_{Ph_2}$	0.5	0.5	0.064 (0.077)	0.002 (0.040)

Table 4.9: Clayton's method and importance sampling for model *ii* under misspecified θ'. $n = 300$

5 Discussion

The Stochastic EM algorithm converges in the tested examples. If the starting values are misspecified we have to run the algorithm for a longer time to allow it to converge but the starting values will not affect the estimates once convergence is achieved. If the model is complex the Stochastic EM algorithm therefore seems to be preferable to Clayton's method and the importance sampler since both of these methods seem prone to break down when θ' is misspecified. The importance sampler seems to be more sensitive to misspecified θ' than Clayton's method but it is not investigated here how much of that difference that is due to the choices of the size, M, of the simulated data. It would be interesting to investigate if a larger M would give adequate estimates in model *ii* with misspecified θ', where poor estimates were obtained in the simulations in Section 4.5.2. The importance sampler may perform better if another importance sampling estimate than (4.6) is used. Hesterberg (1995) describes such alternatives and argues that (4.6) is unreliable since the weights, w_i, do not sum to one. Hesterberg (1995) also points out that a mixture of sampling distributions can be used in importance sampling for better coverage of the sample space. This approach may be beneficial in the ascertainment problem since the choice of sampling distribution turned out to be a major difficulty. Using a similar approach in Clayton's method may also be considered.

Another possible strategy to avoid the effects of seriously misspecified θ' in Clayton's method and the importance sampler is to iterate the procedure using the estimates from the previous step as θ'. This however demands that the parameter estimates from the first step are somehow reasonable, otherwise the iterative procedure could diverge.

Both the importance sampler, Clayton's method and the Stochastic EM algorithm demands prior knowledge of sampling probabilities given the data. These probabilities are often not known and approximations may have to be made using for example registry data or prior knowledge about disease occurrence. If the sampling probabilities are known the complexity of ascertainment scheme does however hardly affect the complexity of the calculations. In Clayton's method and the SEM the ascertainment probabilities are used only when simulating data, and not in the likelihood, while in the importance sampler the sampling probabilities are used in the estimator of $P(A = 1|\theta)$ in a computationally simple manner. Another advantage of the three simulation based methods is that they are not restricted to any specific kind of model while some of the traditional methods handle only specific kinds of data.

The results of the simulation based methods described here will be sensitive to distributional assumptions. Phenotypes are here assumed to be normally distributed given genotype scores but in real data they are often not, so nonparametric extensions of the models would be of interest. It is not possible to check distributional assumptions using standard procedures such as normal QQ-plots since the ascertained data is not assumed to follow the distribution in the population. When missing data is filled in, as in the SEM, checks of distributional assumptions can be misleading since the combined data is a mixture of data from the population distribution and data simulated according to the distributional assumptions. If distributional assumptions are to be checked custom made checks have to be constructed.

Another useful extension would be to incorporate analysis of ambiguous haplotypes in the models.

A Power calculations for dichotomized data

Description of the power calculations presented in Section 2.4:

The log likelihood ratio is computed, but instead of real data the expected data under the ascertainment scheme is used, as described below. The procedure has been described for generalized linear models by Self, Mauritsen, & Ohara (1992) and applied to binary case-control data by Longmate (2001) who calls it *the exemplary data method*.

Under the null hypothesis the log likelihood ratio statistic T^2 is asymptotically χ^2 distributed and the null hypothesis is rejected when $T^2 > q \sim \chi^2(df, 1-\alpha)$ where *df* is the discrepancy in parameters between the alternative and the null hypothesis and α is the intended significance level.

The power is calculated by observing how large proportion of samples would be larger than q under the alternative hypothesis. The distribution of T^2 under the alternative hypothesis is asymptotically non-central χ^2. T^2 is also called the non-centrality parameter. The expectation of T^2 is obtained by calculating T^2 using the expected data. The power is calculated using the noncentral χ^2 probability function $F_{\chi^2}(q, df, \nu)$ where ν is the expected value of T^2 under the alternative. The non-central χ^2 distribution is available in for example the statistical software R (The R Development Core Team 2001).

Calculation of T^2:

The power calculation are performed by computing the expected value of the log likelihood ratio statistic testing if $\beta_{GPh} = 0$

$$T^2 = 2[l_A - l_0] \tag{A.1}$$

where l is the log likelihood. In the alternative hypothesis the parameters β_{0Ph} and β_{GPh} are estimated using maximum likelihood while under the null only β_{0Ph} is estimated. The test statistic will be asymptotically chi-square distributed with one degree of freedom. The expected value of the test statistic is

$$E(2[l_A - l_0]) = 2[E(l_A) - E(l_0)]. \tag{A.2}$$

For continuous data ascertainment correction has to be made and the likelihood for a single observation is thus

$$\begin{aligned} L_A = f(Ph, G|A, \hat{\theta}) &= \frac{P(A|G, Ph, \hat{\theta}) f(Ph|G, \hat{\theta}) P(G|\hat{\theta})}{P(A|\hat{\theta})} \\ L_0 = f(Ph, G|A, \theta_0) &= \frac{P(A|G, Ph, \theta_0) f(Ph|G, \theta_0) P(G|\theta_0)}{P(A|\theta_0)} \end{aligned} \tag{A.3}$$

where $\hat{\theta} = \theta$ for the exemplary data. The parameter values depend on the hypothesis in L but not in the other parts of (A.6). The expressions $P(A|G, Ph)$ and $P(G)$ do not depend on the parameter values at all and will be the same for both hypothesis.

The ascertainment probability A is calculated as

$$\begin{aligned} P(A) &= \int_{Ph} P(A|Ph) f(Ph) dy = \sum_G P(G) \int_{Ph} P(A|Ph) f(Ph|G) dy \\ &= \sum_G P(G)[P(A|Ph < c) \int_{-\infty}^{c} f(Ph|G) dy + P(A|Ph \geq c) \int_{c}^{\infty} f(Ph|G) dy]. \end{aligned} \tag{A.4}$$

Some numerical integration will have to be made to obtain $P(A)$.

For simplicity the dichotomized data is here assumed to fit a logistic regression model with a logistic link even though a probit link would be more appropriate since the outcome data is generated by a normal distribution. Using the logistic link function ascertainment correction does not have to be made, as argued in Section 2.3, and the likelihood will be

$$l \propto \frac{\exp(\theta)}{1 + \exp(\theta)}. \tag{A.5}$$

Note however that $\hat{\beta}_{0Ph}$ is affected by the ascertainment scheme and has to be calculated with this in mind.

The expectation of the log likelihoods are also calculated under the ascertainment scheme. Here the true value of θ should be used, even when calculation the expectation of l_0.

$$\begin{aligned}
E[l] &= \sum_G \int_{Ph} \log(L) \frac{P(A|G,Ph,\theta)f(Ph|G,\theta)P(G,\theta)}{P(A|\theta)} dy \\
&= \frac{1}{P(A|\theta)} \sum_G P(G) \int_{Ph} \log(L)P(A|Ph,\theta)P(Ph|G,\theta)dy \\
&= \frac{1}{P(A|\theta)} \sum_G P(G|\theta)P(A|Ph < c,\theta) \int_{-\infty}^{c} \log(L)P(Ph|G,\theta)dy \\
&+ \frac{1}{P(A|\theta)} \sum_G P(G|\theta)P(A|Ph \geq c,\theta) \int_{c}^{\infty} \log(L)P(Ph|G,\theta)dy \qquad \text{(A.6)}
\end{aligned}$$

Since the observations are iid the expected log likelihood of the whole data set obtained by multiplying the log likelihood with n.

B Information matrix in Clayton's method

The contribution of the ith individual to the information matrix is according to Clayton (2003)

$$\frac{w_{i.}^*}{w_i + w_{i.}^*}(I_i - \bar{I}_i^*) + \frac{1}{w_i + w_{i.}^*}(\frac{w_{i.}^*}{w_i + w_{i.}^*} w_i(u_i - \bar{u}_i^*)^2 + \sum_{j=1}^{m} w_{ij}^*(u_{ij}^* - \bar{u}_i^*)^2) \quad \text{(B.1)}$$

where $*$ indicates simulated data

$$
\begin{aligned}
w_i &= \frac{f(y_i;\theta)}{f(y_i;\theta')} \\
w_{ij}^* &= \frac{f(y_{ij}^*;\theta)}{f(y_{ij}^*;\theta')} \\
w_{i.}^* &= \sum_{j=1}^{m} w_{ij}^* \\
u_i &= \frac{\delta \log f(y_i;\theta)}{\delta\theta} \\
u_{ij}^* &= \frac{\delta \log f(y_{ij}^*;\theta)}{\delta\theta}
\end{aligned}
$$

and

$$\frac{w_{i.}^*}{w_i + w_{i.}^*}(I_i - \bar{I}_i^*) = 0.$$

If θ is a vector of parameters $\theta = (\beta_{0G}, \beta_{0Ph}, \beta_{GPh}, \sigma)$ then for row vectors u, $(u_i - \bar{u}_i^*)^2$ is replaced by $(u_i - \bar{u}_i^*)^T(u_i - \bar{u}_i^*)$ and $(u_{ij}^* - \bar{u}_i^*)^2$ by $(u_{ij}^* - \bar{u}_i^*)^T(u_{ij}^* - \bar{u}_i^*)$ in the expression above.

The derivatives of $\log(f(Ph|\theta)) = -\log(\sqrt{2\pi}) - \log(\sigma) - \frac{(y-(\beta_{0Ph}+\beta_{GPh}g))^2}{2\sigma^2} + g\log(\frac{\exp(\beta_{0G})}{1+\exp(\beta_{0G})}) + (2-g)\log(1 - \frac{\exp(\beta_{0G})}{1+\exp(\beta_{0G})})$ denoted

$u_\theta = (u_{\beta_{0G}}, u_{\beta_{0Ph}}, u_{\beta_{GPh}}, u_\sigma)$ are

$$
\begin{aligned}
u_{\beta_{0G}} &= \frac{\delta \log(f)}{\delta\beta_{0G}} = \frac{(g\exp(-\beta_{0G}) - 2 + g)}{(\exp(-\beta_{0G}) + 1)} \\
u_{\beta_{0Ph}} &= \frac{\delta \log(f)}{\delta\beta_{0Ph}} = \frac{(ph - (\beta_{0Ph} + \beta_{GPh}g))}{\sigma^2} \\
u_{\beta_{GPh}} &= \frac{\delta \log(f)}{\delta\beta_{GPh}} = \frac{g(ph - (\beta_{0Ph} + \beta_{GPh}g))}{\sigma^2} \\
\text{and } u_\sigma &= \frac{\delta \log(f)}{\delta\sigma} = -\frac{1}{\sigma} + \frac{(ph - (\beta_{0Ph} + \beta_{GPh}g))^2}{\sigma^3}.
\end{aligned}
$$

References

Agardh, E., Ahlbom, A., Andersson, T., Efendic, S., Grill, V., Hallqvist, J., Norman, A. & Ostenson, C. (2003), 'Work stress and low sense of coherence is associated with type 2 diabetes in middle-aged swedish women.', *Diabetes Care* **26**(3), 719–24.

Allison, D., Heo, M., Schork, N., Wong, S. & Elston, R. (1998), 'Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power.', *Human Hered.* **48**, 97–107.

Armitrage, P. & Colton, T. (1999), *Encyclopedia of biostatistics*, Vol. 4 & 6, John Wiley & sons, Great Britain, pp. 2791–2793, 4735–4738.

Balding, D. J., Bishop, M. & Cannings, C. (2001), *Handbook of statistical genetics*, Chichester : Wiley, Midsomer Norton, chapter 19 Population Association, David Clayton, pp. 519–540.

Burton, P. (2003), 'Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using gibbs sampling', *Genetic Epidemiology* **24**, 24–35.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Smith, J. D., Kruglyak, L. & Nickerson, D. A. (2003), 'Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans.', *Nat Genet.* **33**(4), 518–21.

Celeux, G. & Diebolt, J. (1985), 'The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem', *Computational Statistics [Formerly: Computational Statistics Quarterly]* **2**, 73–82.

Chen, H. Y. (2003), 'A note on the prospective analysis of outcome-dependent samples', *J.R. Statist. Soc. B* **65, Part 2**, 575–584.

Clayton, D. (2003), 'Conditional likelihood inference under complex ascertainment using data augmentation.', *Biometrika* **90**(4), 976–981.

Clayton, D. & Hills, M. (1996), *Statistical models in epidemiology*, Oxford University Press, New York, chapter 29, pp. 294–295.

Cohen, J. (1983), 'The cost of dichotomization', *Applied psychological measurement* **7**(3), 249–253.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–37.

Elston, R., Olson, J. & Palmer, L. (2002), *Biostatistical Genetics and genetic epidemiology*, Wiley reference series in biostatistics, Wiley, Chippenham Wiltshire, p. 399.

Excoffier, L. & Slatkin, M. (1995), 'Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population', *Mol Biol Evol* **12**(5), 921–927.

Fisher, R. (1934), 'The effects of methods of ascertainment upon the estimation of frequencies', *Annals of Eugenics* **6**, 13–25.

Geyer, C. J. (1993), 'Estimating normalizing constants and reweighting mixtures in markov chain monte carlo.', *Technical Report 568, School of Statistics, University of Minnesota* .

Geyer, C. J. & Thompson, E. A. (1992), 'Constrained monte carlo maximum likelihood for dependent data', *J.R. Statist. Soc. B* **54**(3), 657–99.

Gilks, W. R., Richardson, S. & Speigelhalter, D. J. (1996), *Markov Chain Monte Carlo in practice*, first edn, Chapman & Hall, London.

Green, P. (1992), 'Discussion on constrained monte carlo maximum likelihood for dependent data (by Geyer, C. J. and Thompson, E. A.)', *J.R. Statist. Soc. B* **54**, 683–684.

Gu, H., Abulaiti, A., Ostenson, C., Humphreys, K., Wahlestedt, C., Brookes, A. & Efendic, S. (2004), 'Single nucleotide polymorphisms in the proximal promoter region of the adiponectin (APM1) gene are associated with type 2 diabetes in swedish caucasians.', *Diabetes* **53**, Suppl 1:S31–5.

Hammersly, J. M. & Handscomb, D. C. (1964), *Monte Carlo methods*, Methuen, London.

Heckman, J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**, 153–161.

Hesterberg, T. (1995), 'Weighted average importance sampling and defensive mixture distributions', *Technometrics* **37**, 185–194.

Hirschhorn, J., Lohmueller, K., Byrne, E. & Hirschhorn, K. (2002), ‘A comprehensive review of genetic association studies’, *Genetics in Medicine* **4**, 45–61.

Houghton Mifflin Company (1993), *The American heritage college dictionary*, 3rd edn, Houghton Mifflin, Boston.

Ip, E. H. S. (1994), ‘A stochastic EM estimator in the presense of missing data- theory and applications.’, *Technical report, Department of Statistics, Stanford University* .

Jarrett, R. (1984), ‘Type 2 (non-insulin-dependent) diabetes mellitus and coronary heart disease - chicken, egg or neither?’, *Diabetologia* **26**, 99–102.

Kagan, A. (2001), ‘A note on the logistic link function’, *Biometrika* **88**(2), 599–601.

Kraft, P. & Thomas, D. C. (2000), ‘Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods’, *Am. J. Hum. Genet* **66**, 1119–1131.

Little, R. J. A. & Rubin, D. (1987), *Statistical analysis with missing data*, John Wiley & Sons, New York; Chichester.

Longmate, J. (2001), ‘Complexity and power in case-control association studies.’, *Am J Hum Genet.* **68**(5), 1229–1237.

Louis, T. (1982), ‘Finding the observed information matrix when using the EM algorithm’, *J.R. Statist. Soc. B* **44**, 226–233.

McLachlan, G. J. & Krishnan, T. (1997), *The EM algorithm and extensions*, John Wiley & sons Inc, chapter 6.

Mendel, J. (1865), ‘Verhandlungen des naturforschenden vereines in brünn’, *Abhandlungen* **4**, 3–47.

Morton, N. & Collins, A. (1998), ‘Tests and estimates of allelic association in complex inheritance’, *Proc. Natl. Acad. Sci. USA* **95**, 11389–11393.

Neale, M., Eaves, L. & Kendler, K. (1994), ‘The power of the classical twin study to resolve variation in threshold traits.’, *Behav Genet* **24**, 239–258.

Neuhaus, J. M. (2000), ‘Closure of the class of binary generalized linear models in some non-standard settings’, *J.R. Statist. Soc. B* **62**(Part 4), 839–846.

Neuhaus, J. M. (2002), 'Bias due to ignoring the sample design in case-control studies.', *Aust. N.Z.J. Stat* **44**(3), 285–293.

Prentice, R. L. & Pyke, R. (1979), 'Logistic disease incidence models and case-control studies', *Biometrika* **66**, 403–412.

Purcell, S., Cherny, S., Hewitt, J. & Sham, P. (2001), 'Optimal sibship selection for genotyping in quantitative trait locus linkage analysis', *Hum Hered.* **52**(1), 1–13.

Ripatti, S., Larsen, K. & Palmgren, J. (2002), 'Maximum likelihood inference inference for multivariate frailty models using an automated monte carlo EM algorithm', *Lifetime Data Analysis* **8**, 349–360.

Robins, J. M., Smoller, J. W. & Lunetta, K. L. (2001), 'On the validity of the TDT test in the presence of comorbidity and ascertainment bias', *Genetic Epidemiology* **21**(4), 326–336.

Rubin, D. B. & Schenker, N. (1991), 'Multiple imputation in health-care databases: An overview and some applications', *Statistics in Medicine* **10**, 585–598.

Schork, N. J., Nath, S. K., Fallin, D. & Chakravarti, A. (2000), 'Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects', *Am. J. Hum. Genet.* **67**, 1208–1218.

Self, S. G., Mauritsen, R. H., & Ohara, J. (1992), 'Power calculations for likelihood ratio tests in generalized linear models', *Biometrics* **48**, 31–39.

Smoller, J., Lunetta, K. & Robins, J. (2000), 'Implications of comorbidity and ascertainment bias for identifying disease genes.', *Am J Med Genet.* **96**(6), 817–22.

Stern, M. P. (1995), 'Perspectives in diabetes. diabetes and cardiovascular disease. the "common soil" hypothesis', *Diabetes* **44**.

Strachan, T. & Reed, A. P. (1999), *Human Molecular Genetics 2*, second edn, BIOS Scientific Publishers Ltd, Bath, pp. 40,271,273,303.

Terwilliger, J. & K.M., W. (2003), 'Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth", *Annals of Medicine* **35**, 1–13.

Terwilliger, J. & Ott, J. (1992), ‘A haplotype-based ’haplotype relative risk’ approach to detecting allelic associations.’, *Hum Hered.* **42**(6), 337–46.

The R Development Core Team (2001), ‘R’, Version 1.4.0.

Torrie, G. & Valleu, J. (1977), ‘Nonphysical sampling distributions in monte carlo free-energy estimation:umbrella sampling.’, *J. Comput. Phys.* **23**, 187–199.

Vargha, A., Rudas, T., Delaney, H. D. & Maxwell, S. E. (1996), ‘Dichotomization, partial correlation, and conditional independence’, *Journal of educational and behavioral statistics* **21**(3), 264–282.

Wei, G. C. G. & Tanner, M. A. (1990), ‘A monte carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms’, *Journal of the American Statistical Association* **85**, 699–704.

WHO (2000), ‘Monica project’, ”`http://www.ktl.fi/publications/monica/surveydb/title.htm`”.

Wu, C. (1983), ‘On the convergence properties of the EM algorithm’, *Annals of Statistics* **11**, 95–103.