THE UNIVERSITY *of* EDINBURGH

Development of a Statistical Method for the Identification of Gene-Environment Interactions

Pauline L. Golding

Centre for Population Health Sciences

Thesis Presented for the Degree of Doctor of Philosophy

University of Edinburgh

2011

Declaration

I hereby declare,

(i) That this thesis is composed by myself

(ii) That the work presented within this thesis is my own unless otherwise stated

(iii) That this work has not been submitted for any other degree or professional qualification

Pauline L. Golding

May 2011

Acknowledgements

My first thanks go to Niall Anderson and Sarah Wild, for their supervision, encouragement, help and support over the course of my studentship. I have been honoured to be based in the Centre of Population Health Sciences, surrounded consistently by so many interesting people and intellectually stimulating projects. In particular I would like to thank Rosa Bissett for all her organisational help and support.

For computer support and tuition, I would like to thank David Reitter, formerly of the department of bioinformatics and the University of Edinburgh, currently of Carnegie Mellon University and my extremely patient husband, Ben Golding, software engineer and former teaching assistant in computer science, University of Edinburgh. I would also like to thank the makers of R and in particular the creators of the packages Rpart, randomForest, relimp, Hmisc, ORMDR, glm, stats, Rcmdr and Tinn-R.

I would like to thank Evropi Theodoratou, Harry Campbell, Albert Tenesa and Malcolm Dunlop for both allowing me access to real genetic data and for their help in collating and understanding my dataset prior to analysis.

Finally I would like to thank the Chief Scientist Office, for funding my work.

Abstract

In order to understand common, complex disease it is necessary to consider not just genetic risks and environmental risks, but also the interplay between them. This thesis aims to develop methodology for the detection of gene-environment interactions specifically; both by looking at the strengths and weaknesses of traditional approaches and through the development and testing of a novel statistical method. Developments in genotyping technology enable researchers to collect large volumes of polymorphisms in human genes, yet very few statistical methods are able to handle the volume, variation and complexity of this data, especially in combination with environmental risk factors. Interactions between genes and the environment are often subject to the curse of dimensionality, with each new variable increasing the potential number of interactions exponentially, leading to low power and a high false positive rate.

The Mixed Tree Method (MTM) exploits the differences between environmental and genetic variables, by selecting the most appropriate features from conventional methods (including recursive partitioning, random forests and logistic regression) and combining them with new comparison algorithms which rank the genetic variables by the likelihood that they interact with the environmental variable under study.

Results show the MTM to be as effective as the most successful current method for identification of interactions, but maintaining a much lower false positive rate and computational burden. As the number of SNPs in the dataset increases, the success of MTM compared to other methods becomes greater while the comparator approaches exhibit computational problems and rapidly increasing processing times. The MTM is also applied to a colorectal cancer dataset to show its use in a practical setting. The results together suggest that MTM could be a useful strategy for identifying gene-environment interactions in future studies into complex disease.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
CART	Classification and Regression Trees
CI	Confidence Interval
COPD	Chronic Obstructive Pulmonary Disease
CRC	Colorectal Cancer
DNA	Deoxyribonucleic Acid
FAP	Familial Adenomatous Polyposis
FDR	False Discovery Rate
FFN	Feed Forward Network
GAM	Generalised Additive Models
GLM	Generalised Linear Models
GPNN	Genetic Programming Neural Networks
GWAS	Genome Wide Association Study
GWS	Genome Wide Scan
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
LD	Linkage Disequilibrium
MDR	Multifactor Dimensionality Reduction
MSA	Multiallelic Set Association
MTM	Mixed Tree Method
NSAID	Non Steroidal Anti Inflammatory Drug
OR	Odds Ratio
ORMDR	Odds Ratio Multifactor Dimensionality Reduction
PDM	Parameter Decreasing Methods
PKU	Phenylketonuria
RP	Recursive Partitioning
RR	Relative Risk
SNP	Single Nucleotide Polymorphism
SOCCS	Study of Colorectal Cancer in Scotland

Chapter 1

Introduction

1.1 Background

Epidemiology is the study of patterns and trends of disease in populations and the factors that might influence these trends. The occurrence of a disease is compared by time, place and person to try and quantify the difference in disease frequency over time, between different geographical populations and between members of the same population. The results from epidemiological work can be combined with other disciplines to elucidate the aetiology of a disease or to try and develop public health interventions to reduce the disease frequency.

The fundamental concepts of epidemiology originated from studying the characteristics and trends of infectious diseases. However, many of the principles are equally applicable to common, non infectious, complex diseases and conditions. The field of epidemiology has now evolved to the point, in Western societies, where lifestyle choices are an essential area of research – an arena far removed from the infectious disease model.

Traditional epidemiology of transmissible disease is concerned with three main factors: agent, host and environment and the interplay between them [1]. The relationships between these factors are also present between non-infectious aetiological agents, hosts and environment. Examples of agents in complex disease could be radiations, carcinogens or cholesterol. Host factors, also called intrinsic factors, would be the same as for a complex disease as for an infectious disease. Examples of host factors include genetics, age, ethnicity and physiological state. Environmental factors, also called extrinsic factors, can affect the existence, exposure and susceptibility of the agent or host. Environmental factors include human population density, climate, socioeconomic factors and rare events such as floods or earthquakes.

There are specific fields of scientific research concerned with each factor, or a subgroup of a factor, individually. The task for epidemiology is to combine data from a variety of disciplines in a way that attempts to explain the trends or causes of a

disease. It is important to analyse how the environmental factors interact with the agents and how both interact with the host.

Gene-environment interactions are relevant to such a model and to the analytical methods associated with such a premise. In studying gene environment interactions, this triangular model of interaction is always relevant and suitable analysis will always take this into account. For example, in studying coronary heart disease, cholesterol (an agent) is affected by both genetic (host) and dietary (environmental) factors. It is important also to consider the interplay between the host and environmental factors, for example the effect of diet on physiological characteristics. There are different forms of interaction, involving a combination of main effects enhanced or decreased by the presence of a different factor and interacting effects only present as such.

In the 1970s there was a debate within the epidemiological community around the definition of interaction [2] and whether or not the terms interaction and synergy are synonymous [3]. However, in 1980, a paper by Kenneth J Rothman [4] found a compromise by dividing the term interaction into four categories: statistical, biological, public health and individual decision making. Therefore, it is important to define the type of interaction being considered before commencing a study.

There has been a large volume of epidemiological research looking at the agent or environmental risk factors, there has been genetic research looking directly at candidate disease genes but there is potential to study how the two fields interact. Understanding the interactions between genetics and the environment could both open up new avenues of research through identifying novel genes only involved in interactions or could help resolve some of the confusions where the same environmental factor confers different risk to different populations.

1.2 Public Health Implications of Gene-Environment Interactions

> "Some vegetarians with acceptable levels of cholesterol suffer myocardial infarction in their 30s. Other individuals seem to live forever despite personal stress, smoking and poor adherence to a Heart Association approved diet"
>
> R.A. Hegel (1992)

1.2.1 Historical Perspectives

The history of genetic epidemiology predates the recognition of genetics or epidemiology as fields of scientific interest, as throughout history people have observed similarities and differences among individuals, families, tribes and communities. Genetics and evolution help us understand these differences; public health techniques will help us extrapolate from these differences ways to improve population health.

Documented examples of historic writing that unknowingly connects genetics with public health include Plato, who in 360BC, in "The Republic" muses on mate choice and the fact that people selecting mates similar to their own may cause them to lose the "balancing characteristics" [5] reflects a similar sentiment to more recent work on mate choice and the Major Histocompatibility Complex (MHC) [6]. Many Victorian authors connect people with fair complexion as being fragile and prone to fainting. Phrases like "child bearing hips" have found their way into our everyday language without many people genuinely considering there might be a relationship between body shape and complications during childbirth.

The swing towards the scientific style of genetics we would recognise today began in the 1860s with Francis Galton, who was Charles Darwin's half cousin, creating the statistical concepts of regression and correlation; he was the first to apply statistical methods to the study of human differences. At the same time in Austria, Gregor Mendel was developing his "Laws of Inheritance". In the 1900s, Karl Landsteiner

classified the blood groups ABO and in 1953, Watson and Crick proposed the double stranded DNA helix as the unit of inheritance.

The progress of public health research, meanwhile, has been less consistent, constantly adapting to new and varied health risks. The concepts of quarantine, sanitation and medical interventions (albeit shaman and witch style interactions) have been present in many past civilisations. Major advances in public health include Girolamo Fracastoro proposing transferable infection in 1546, the development of the smallpox vaccine in 1796 by Edward Jenner (heralding an age of preventable medicine), and Louis Pasteur confirming germ theory in 1862.

It was not until the 1940s and 1950s, however, that programmes of epidemiological research for genetic diseases were pioneered and that neonatal screening programmes for inborn errors of metabolism, such as sickle cell anaemia or Tay-Sachs carrier status, were established. This was the first example of a public health intervention based on genetic epidemiology.

1.2.2 Opportunities in Public Health Genetics

Current focus in public health research is on environmental variables, however, there are five main areas of research where genetics and public health research methods could be used together for population benefit, these are: Ecogenetics, pharmacogenetics, nutritional genetics, behavioural genetics and infectious disease genetics. The more personalised medical treatments and screening, although primarily aimed at improving individual health, could also have a knock on benefit on population health. Although these areas are considered primarily genetic, the theoretical basis is quite often a gene-environment interaction, with an assigned environmental factor.

Ecogenetics

The term ecogenetics was introduced in 1971 by George Brewer as a wide-ranging term for genes interacting with environmental factors [7]. This area has since diverged into more distinct fields for drugs and dietary variables with the term ecogenetics

taking the more specific role of studying variables present in the environment, including work environments. The National Research Council (http://www.nationalacademies.org/nrc/) in America recommends that: “individual variation be more carefully investigated and be used in standard setting and in risk management to protect subgroups in the population”. Despite the tendency towards smaller samples, being able to define those subgroups by their genetics would lead to more accurate risk assessments and smaller confidence intervals as the within group variance would be lower.

The regulatory agencies in America are interested in the prospect of being able to create more accurate and case specific risk estimates to replace the current estimates, that some find are so conservative that they limit other areas of industry [8]. They are also trying to change the current strategy where a single chemical or exposure is studied separately and each health outcome in isolation [9]. The tendency to study every chemical individually is valuable for regulatory purposes but does not help with analysis of subgroups that might be exposed to combinations of the same low risk substances [10].

In the field of research there is a group called the Environmental Genome Project (http://www.niehs.nih.gov/envgenom/home.htm), which is part of the National Institute of Environment Health Sciences. It covers a large number of areas of research including sequencing, mouse models, biostatistics and population based modelling. The population based projects funded vary from the effects of air pollutants and asbestos to susceptibility genes and cancer.

It is also recognised that risks may not only differ by genetic makeup but also by stage of life, and therefore which genes are being expressed. This is of particular concern regarding the protection of children or foetuses at crucial stages of development. Understanding which genes are involved in teratogenesis may help predict future health risks to newborns through environmental interaction.

A better understanding of how humans, as genetic beings, interact with their surroundings would also help our understanding of the biological mechanisms of

carcinogenic and neurotoxic agents with an aim to improving treatments and identifying other harmful substances.

Pharmacogenetics

Pharmacogenetics is a specific area of genetic epidemiology where the exposure is the drug dose and the aim is to measure how risk of adverse outcomes and levels of benefit from treatment vary across strata defined by genotype and exposure [11]. The idea is then that pharmacological therapies are optimised based on the genetic characteristics of the patient. It is also an important field for the development of biological targets for interventions, such as drugs or vaccines [12].

Using pharmacogenetics it would be possible to better judge the correct pharmaceutical treatment for subgroups of the population and to increase the therapeutic margin for drugs. Advice is given to clinicians about tailoring treatment to the patient but this currently involves using general doses adjusted by weight, age and renal clearance only. These general drug dosages are normally recommended based on their pharmacokinetics when being taken by a group of healthy people, with normal drug metabolism levels [13]. However, varying drug metabolism rates can lead to some people developing clinically significant adverse side effects from dosages that elicit no therapeutic response in others.

Currently, there are two sides of the problem of where to draw the line with licensing a drug regarding side effects. A dose too high for some people is dangerous for them, whereas recommending a drug at a low level, safe for this subgroup, may not be the most effective treatment for the rest of the population. Drugs have been withdrawn over side effects that could be predicted using genetic testing. For example, a test for the CYP2D6 polymorphism that leads to a reduced ability to metabolise debrisoquine, might have prevented the withdrawal of this drug from the market, thus helping those less likely to suffer side effects [13] .

As well as differences in drug metabolism, pharmacogenetics also includes the study of genes and mechanism of drug targets and the disease pathway. One such pharmacogenetic result that has been well publicised is the breast cancer treatment Herceptin, which is more effective for women who have the genetically determined

immunophenotype HER2 [14]. Adverse drug reactions are estimated to be responsible for 10,000 deaths per year in Britain [15]. A better understanding of the causes of these reactions would be beneficial, for example, even something as simple as a familial trend would help establish if there was a potential genetic basis for the reaction.

Nutritional Genetics

Nutritional research has a strong emphasis on biochemical and metabolic mechanisms, many of which are also areas of genetic research. Cholesterol fractions, lipids and other fatty acids are risk factors for coronary heart disease, however, individuals with cholesterol levels elevated to about the same level may require a number of different dietary and pharmacological interventions [16]. There is also evidence to suggest both dietary vitamin D, the genetics of the Vitamin D receptor and exposure to sunlight are associated with the incidence of Multiple Sclerosis [17].

One dietary variable that is required and processed differently in different people is iron. Iron is needed in our diets for the manufacture of haemoglobin and other tissue molecules, a deficiency of iron, anaemia, is the most widespread nutritional problem around the world [18]. It has been proposed in America that food should be fortified with iron to ensure that those most at risk, children and pregnant women, do not suffer the consequences of anaemia. However, such fortification would put a subgroup of people, who suffer from iron overload, haemochromatosis, at risk.

Identifying people at risk could have a twofold effect. For example, the sufferers of haemochromatosis could follow a special diet, receive regular health checks and blood removal to reduce their iron levels. Also, it may reassure the people worried about the side effects of food fortification with iron, if it were possible to identify the at risk subgroup. Therefore some such foods, well labelled, could be introduced to the market to make some progress towards reducing anaemia in the general population.

Careful studies of how people respond to different dietary and vitamin supplement interventions and their cost effectiveness in different subgroups could help guide public health policy.

Behavioural Genetics

This is an extremely controversial area of genetic study due to the difficulties in measuring human behavioural traits, such as personality and intelligence. The racial, socio-economic and educational biases inherent in such measures are also a problem for their analysis. Behavioural genetic studies both how human actions interact with environmental factors to influence risk and how genetic factors might affect behaviours. The genes that influence the likelihood of someone partaking in an unhealthy behaviour is sometimes called a "Gene Environment Correlation." These are hard to study as many families both share genes and are responsible for other family members' environmental exposures. A gene environment interaction is different from a correlation, as the increase in risk due to both genetic and environmental factors acting together is far greater than the predicted increase from either factor independently, but the factors are considered to be independent. Whereas for a correlation, the effects are not independent of one another, the environmental risk is related to the genetics and it is this relationship that provides the correlation.

Infectious Disease Genetics

Infectious diseases have influenced our genomes through natural selection for thousands, if not millions, of years. Before the development of antibiotics and vaccines, human populations were constantly battling diseases that led to the deaths of the most susceptible or weakest people. Evolutionary pressure from infections meant that the selective advantages of those carrying the haemoglobin variants S, C, D and E, the thalassemias, or glucose-6-phosphate dehydrogenase deficiency in the face of malaria [19] increased survival. Another possible example of a selection effect is CCR5 homozygosity, thought to have evolved from cholera exposure, providing a degree of resistance to HIV infection [20].

There are a number of ways that human genetics play a role in infectious disease susceptibility and progression. Taking Chronic Hepatitis B virus (HBV) infection as an example, human genetics plays a role in a number of categories, including: mediation of the viral entry (binding, membrane fusion and transportation) into cells; genes modulating the immune response to infection; the pathological alterations and symptoms in tissue; the disease development, including control of mother to child transmission and development of resistance to drug treatments [21]. It is also useful to

be able to study the genomes of infectious organisms, the complete sequence of the malaria parasite, *Plasmodium falciparum*, was finished in 2002 [22], for example, which can help to understand the organism and develop drug targets to fight it. Genetic variation at all these levels provides potential for clinical interventions or preventative strategies to benefit individuals and the population as a whole.

Screening

On a more personal health level, there are some highly penetrant genetic mutations which account for disease clusters within families. These tend to be rare and account for less than 5% of major cancers and coronary heart disease [23]. Examples of some well known, highly penetrant mutations include the LDL (low density lipoprotein) receptor, where mutations can lead to large increases in cholesterol or heart disease risk, or the BRCA1 and BRCA2 genes which confer an increased risk of breast cancer. For the individuals affected, the health gains from early and successful interventions are enormous. Although not targeted at the entire population, the knowledge gained about biological mechanisms and outcomes from interventions can benefit the population at large.

There are very successful screening programmes aimed at identifying diseases in adults by identifying early stage symptoms, for example pre-cancerous cells. However, these are a separate public health intervention from the genetic screening for susceptibility and such success, or lack, does not necessarily translate to evidence for or against genetic screening.

1.2.3 Are Genetic Studies Justified on a Public Health Basis?

There are two schools of thought that consider genetic epidemiology overrated and that it should not be such a focus for future medical research. The first think that all epidemiology is lacking in theory development and is not a progressive field of research [24, 25]. Then there are the epidemiologists who feel that environmental risk factors should be prioritised over genetic ones; that it is not useful to concentrate on risk factors that cannot be altered when there is still so much to do on the risk factors that can [26].

Relevance of Epidemiology

Over the last fifteen years, the usefulness of risk factor epidemiology has been called into question [27]. The concern is that too much research is focused on methodology and data manipulation and not enough on theory development, which might actually advance our field of knowledge. It is thought that new studies into the effects of smoking, drinking and obesity will add little to guidelines for public health policy. There are already interventions in place to try to reduce these risk factors and there is no requirement for additional evidence that they are unhealthy. Also, epidemiology is only relevant to the time the studies are carried out and can therefore not justify research for its own sake in the way of other scientific disciplines [25].

There are also problems inherent to risk factor epidemiology. The ecological fallacy means erroneously making assumptions about an individual based on the risk of the group or groups he belongs to. Measurement of environmental variables, usually done retrospectively, may have substantial amounts of error. If it is the desire of epidemiology to predict future outcomes, will the risks being measured be relevant and have similar effects in the future. There is a strong publication bias in favour of positive results, multiple potential confounders and the possibility that any results are the product of chance. All of which combined makes epidemiological studies hard work for small reward.

Such arguments can be even more forcefully applied to the field of genetic epidemiology where interventions, if possible at all, will take years of further research following the identification of a gene.

Genetic Epidemiology

"If causes can be removed, then susceptibility ceases to matter"

- Geoffrey Rose

Within the epidemiological community, there are people that consider genetic research not to be worth the time, funding and prominence that it is given. Environmental risk factors are more easily modified and there is considerably more evidence of their importance.

Epidemiology and human genetics gained authority through efficient and successful identification of Mendelian diseases, infectious disease vectors and environmental risks. These discoveries were easily converted to public health recommendations in the form of prenatal testing, workplace regulations, vaccinations and other preventative measures. However, the lack of success of genetic studies to identify causes of complex common disease to date is seen by many as evidence that it is not a fruitful area of research. It has been proposed that statistical empiricism and epidemiology are not working and what is needed is a molecular based approach to understanding the biological mechanism of disease [26]. This does not include further genetic studies, as the nature of complex disease means they are susceptible to the same problems as basic epidemiological studies.

The Nature of Complex Disease

Any complex disease has multiple genetic and environmental components and pathways, the relevance of which is normally confined to the discussion section of research articles. The methods currently being employed were designed to detect risk factors with large effects, whereas complex diseases tend to have multiple risk factors, mostly with fairly weak effects.

One of the complications that arise from working with complex disease is phenotypic and genotypic heterogeneity. Phenotypic heterogeneity is when the definition of the disease is hard to classify and even a biologically specific definition can have different disease outcomes. Genotypic heterogeneity is when a variety of different alleles lead to the same disease outcome. This is relatively common as evolution acts on the phenotype of traits and not their genotypes, so a number of different components of a biological pathway could be under the same selective pressure. Even if some of the causal components are known, trying to relate the other potential risk factors to the web of causality and interactions is extremely difficult.

Population Risk

Even assuming genetic factors could be useful indicators of individual risk, it does not follow that they are informative regarding population risk [28]. It is extremely difficult to gather data from individuals, extrapolate from it conclusions about a population, and then from these conclusions apply interventions aimed at individuals.

Other than pharmacogenetics, it can been argued that targeted interventions based on genetics may have less impact than an intervention targeted at the whole population [29]. Interventions aimed at high risk groups have not been as effective at improving population health as general interventions, based in part on the fact that those in high risk groups find it more difficult to adapt their behaviours away from those of the general population.

Another reason it is difficult to make conclusions about population health is that genetic heterogeneity is greater within populations than between populations, the opposite of environmental risk factors. It is impossible to identify environmental exposures as risk factors if they are uniform across the study populations (i.e. if everyone smoked the same number of cigarettes every day, variation in lung cancer incidence would be mostly explained by genetics). This explains why people who emigrate tend to have similar disease rates to the country they moved to [30]. This problem is compounded when isolated populations are studied as the range of exposure variables, both genetic and environmental, is extremely limited, compared to an outbred population.

In terms of screening, there is some debate on whether genetic testing for complex diseases actually adds anything to current methodologies. It has been argued that if genetic profiling is to be considered as a public health intervention then the screening programmes should be evaluated by the same criteria as current programmes [31] The majority of genetic screening would fail by these standards as the excess risk identified is too low or the knowledge of the excess risk does not influence an already supported disease management strategy [28].

Even some of the more promising genetic studies are based on biologically plausible theories from epidemiological analysis and do not add any information with direct public health benefit. For example, the polymorphisms involved in oestrogen synthesis and metabolism could identify a prognostic index for breast cancer based on the accumulated risk of the different genotypes [32]. This however, is no more useful than measuring serum levels of oestrogen and may even be less useful as it does not account for known environmental risk factors, such as contraception. Genetic epidemiology does not need to prove that it could be useful in a world of research

where nothing is known. It needs to prove that it can add to or improve current methodologies.

Importance of Genetic Studies

When faced with trying to understand the epidemiology for a disease associated with both environmental and genetic risk, the environmental variables are more easily altered and more tangible. Why then investigate the genetic factors? It is because there is no alternative [33]. Many years have been spent making claims about macroscopic risk factors and refuting such claims. The inconsistency and instability of many claims may be due to the interplay of many more proximal factors: to study such factors would assist the research into environmental risks, not be in competition with it. Traditional risk factors, like smoking, obesity and diet, are composed of hundreds or thousands of very heterogeneous factors. The benefit associated with eating fruit results from a very complex mechanism involving taste, digestion, gut flora, antioxidants, oxidants, fibre and many more factors. The current measure, also fairly inaccurate, of portions per day is proving an insufficient measure. First fruit and vegetables were found to reduce breast cancer risk [34, 35], more recently large cohorts have found there to be no effect at all [36] claiming the results could be due to poor controls or confounding.

In a number of studies, it is implied that the population attributable risk (the percentage of the disease incidence that would be reduced if the risk factor was eliminated) is entirely dictated by the risk factor in question. This is quite often not the case; the attributable risks for a complex disease can add up to considerably more than 100%. This is a result of the interactions amongst the various risk factors. For example, it may be the case that both a gene and environmental factor are required for disease symptoms, both with 100% attributable risk, leading to 200% of the risk being accounted for. For phenylketonuria both the genetic mutation and a diet containing phenylalanine are necessary for the disease and could eliminate the disease if removed.

There are accusations that genetic epidemiology has not lived up to its expectations and that such slow progression is evidence that there is not much progress to be made. This is considered to be the weakness of epidemiology and not a sign of how

high expectations are set for any field of research with potential. Is the twelve years that has elapsed since the completed human genome mapping project a reasonable timeframe within which to expect unequivocal results [37]? Normally these expectations are set by people with only a token understanding of the area and should be treated with caution or specialists that are exaggerating the potential of their research to secure funding. Areas of research involving human subjects do not often live up to the preceding hype, humans are complex and complicated and often the most desirable research academically would be the most awful ethically. Providing a cure is not the sole target of epidemiology, showing cause can also alleviate blame. To imply that all mental health epidemiology has been useless because it cannot now cure everybody would be a ridiculous claim, yet more than 150 years have passed in which to find cures compared to the 6 years that passed between the completion of the human genome sequence and the first criticisms of genetic epidemiology. We are also in a very different research climate, where conclusions are now reached more slowly but more carefully. This may add to the impression that genetic research is not progressing as fast as other areas did. It is not slower, just more tentative.

There is also the argument that current environmental interventions are proving inadequate: for example, obesity levels are still rising despite campaigns for awareness of obesity related risks. Often a single contradiction to the published evidence is enough to negate the effect of a health promotion intervention. For example, smoking in teenage girls is often justified by them knowing someone or knowing of someone who smoked and is not sick [38]. Explaining that there may be a genetic reason that that person is not sick may remove this line of argument and add weight to the intervention, even if it is in some way conjecture.

Many of the arguments that are being used against genetic epidemiology do not prove that the area of research is useless but rather that we have to approach it slightly differently. Many models and statistical methods employed for genetic epidemiology are perfect for analysing Mendelian traits or single gene disorders, but are not appropriate for conditions where many variables have small effects, like complex diseases.

There are a number of factors that contribute to the inconsistency in gene identification studies, making genetic effects seem less reliable or even non-existent. Small study designs make it hard to identify effects within a sufficiently narrow confidence interval and occasionally inappropriate controls are used, for example, younger family members when studying a late onset condition, so their case status cannot be accurately determined. Publication bias towards positive results may encourage publication before verification. Most importantly though, most studies fail to collect data on environmental variables that might modify the relationship between genotype and disease status, meaning an association between gene and disease in one study would not necessarily show up in a study where the population have different environmental exposures. Instead of being contradictory, these differences may provide new study opportunities.

1.2.4 Conclusions

The public health implications of gene environment interactions may be important in the field of genetic epidemiology. Looking for interactions alongside gene identification studies could help elucidate some of the contradictions and residual variation that currently complicate genetic studies. Gaining knowledge about the interactions between risk factors would make studies and risk estimates more accurate and increase our understanding of the underlying biology, which in turn could benefit the health of the population. Studying one field, whether genetic, environmental or behavioural without acknowledging the others and possible interplay between the two is scientific bad practice. The best way to approach complex problems is to work together and not in competition.

1.3 Dimensionality and Power

The number of variables in gene-environmental interaction studies can lead to low power and a high error rate. Statistically, it is not possible to use traditional parametric methods, such as logistic regression, for data with as many exposure variables as genome wide association scans (GWAS). If there are insufficient outcome events relative to the number of independent variables being analysed, many multivariable methods, including logistic regression, produce problematic results [39]. In studying interactions, it is not simply the number of variables that are being studied but each possible pairing of these variables.

Therefore one of the most common problems in studying interactions between variables is the dimensionality of the problem. Every new variable added to the model adds a dimension, with the potential interaction terms increasing the hypothesis space exponentially. This adding of dimensions to a mathematical space can mean that the space is enormous and the data are sparse. This is often referred to by statisticians as the *curse of dimensionality*.

Taken together, the number of variables present in a GWAS and the further increase introduced by studying interactions, there is a need to be aware of the potential false positive rate or bias introduced through multiple testing. Initally the results from GWAS were intended to suggest candidates for further study [40]. However, the exceedingly large multiple testing used in these studies can also miss gene variants with small effects [41].

With such large volumes of data, an efficient and straightforward way to explore the data sets prior to statistical analysis would be beneficial both in saving time and preserving the value of the data.

1.4 Research Problem

1.4.1 Definitions and Key Assumptions

The definition of "novel" with respect to the method being developed can incorporate functions and theory from traditional methods, as long as the application and combination is new in some way. The computer code for the overall method will be unique to this project and thoroughly tested. The different definitions of interaction are covered and discussed throughout.

1.4.2 Aims

The overarching aim of this thesis is to develop a novel statistical method of exploring genetic data and identifying gene-environment interactions, given the high number of dimensions. In order to achieve this, there are three core research components addressing different aspects of the problem: the current methods, developing a new method and ensuring such a method can be used, or adjusted to be used, across a variety of data types gathered from different study designs. A final aim involves using the method on a real dataset to identify any gene-environment interactions present.

i) Assessment of methods currently used to identify gene-environment interactions

The first aim is to undertake a systematic literature review on the statistical techniques commonly used on data with many dimensions. This includes papers focused on gene-environment and gene-gene interactions, statistical theory of high dimensions, identification studies looking for interactions and papers discussing the public health implications of such research. This review can be found in Chapter 2, with background on colorectal cancer and potential public health benefits in Chapter 3.

ii) Development of Novel Method

Identifying interactions with statistically significant effects in large data sets, using a single traditional method, would require either an enormous sample or an unrealistic effect size. Exploring the data in order to generate hypotheses, then testing these hypothesis, would increase the chances of an interaction being found. Exploratory data analysis, also known as data mining, can take a number of forms including both visualisation and machine learning methods.

There are a group of data mining methods called tree based methods, elements of which have been combined within the novel method being proposed in this thesis. The primary tree method is Classification And Regression Trees (CART), which search the hypothesis space for the variable that, when dividing the data by that variable, creates the greatest difference between subsets. Another important tree based method is the random forests approach, which uses tree methodology, cross validation and boot strapping to grow large numbers of decision trees (hence the term forest) to identify the most significant variables.

These current tree methods enter all the data simultaneously and often the main effects of the environmental variables dominate the higher splits in the tree. However, if looking specifically for interactions between the genes and the environment, it is possible effectively to plant the environmental variable as the tree root and enter only the genetic data to build the tree from that root. Trees can be built from all the different environmental variable roots and the branches analysed. The position of a variable in a subtree and the comparison of variable positions between different subtrees can be quantitatively assessed. Once the computer programme has run trees for each of the environmental variables, it can assess which genetic variables show the least consistency in branch position and identify them as potentially interacting variables. A quantitative measure of association, such as logistic regression can then be used to determine the risk estimates.

Another aim is to compare the novel method with the traditional methods using identical artificial data sets with known effects sizes and a variety of underlying interaction models. Such comparison forms the basis of Chapter 7.

iii) Adapting the Method to Different Data Types

The artificial data will be used to assess the performance of the method for different data containing different characteristics and underlying, complex effects, before real data are analysed. The aim is for the model to be flexible, or able to be adjusted prior to analysis, for a variety of study and data types, explored in Chapter 6.

Case control data are the simplest to analyse with binary outcomes being well suited to the tree theory. Another advantage, for most statistical analysis, is the similar proportions of cases to controls. However, different case-control ratios will be tested and potential adaptations for data type explored.

Continuous outcomes variables can either be stratified by cut off points or a regression model can be fitted to the leaf of the branches. For a variable like BMI where there are well recognised categories, it may be suitable to group the outcomes into categories: underweight, healthy weight, overweight and obese. Such grouping in advance is a weakness of the tree methods and may lead to a loss of information. Therefore in this study both pre-determined and data driven cut-off points will be explored.

iv) Application to Real Data

Once the method has been developed and its performance maximised by identifying the parameters best suited to the real data set, an analysis of a genuine example is run. The results can be found in Chapter 8 are compared to the relevant literature and examined in a wider genetic epidemiological context.

The real data are from the epidemiological case control study of colorectal cancer: Study of Colorectal Cancer in Scotland (SOCCS). The aim was prospective recruitment of all incident cases, in people aged 16-79, of colon or rectal carcinoma in Scottish hospitals. In order to minimise survivor bias recruitment was as close as possible to the first admission. There were certain exclusions: patient death before recruitment or severe illness during participation; the cancer being recurrent; and the patient being unable to give informed consent – either due to illness or learning difficulties. The controls were matched for age, gender and geographical region and drawn randomly from a population based register.

The SOCCS study was not designed with this analysis in mind and therefore is perhaps not the way sampling would have been done if gene-environment interactions were the primary aim. For example, matching is not the ideal approach for non parametric methods and therefore the matched environmental variables cannot be analysed.

Chapter 2

Review of Statistical Methods

2.1 Introduction

Many complex, chronic, human diseases are the result of numerous genetic and environmental factors and interactions between the two. In the last ten years, there have been rapid developments in genomics, molecular biology, population genetics and genotyping technology. We now have large data sets available to study, yet the analysis of these data sets has been slow to yield results and rarely replicated. The genetic influences on many common, chronic diseases are extremely complex, they have a pattern of family history, yet do not follow the rules of simple Mendelian models of inheritance [42]. There is often no clear autosomal, X-linked, dominant or negative model. This could be due to the presence of the recognised, but yet unclassified, interplay between genetics and environmental factors.

2.1.1 Aims of Literature Review

There are a number of important aims to be considered in the review of the literature including: defining the question, reviewing previous work, the applications of such work and the role of new research in a wider context. In order to properly define the research question, the different possible interpretations of interaction- in terms of biology, statistics and gene-environment interactions- were explored. The main area of this review is an in-depth summary of current statistical methods for detecting gene-environment interactions, both to identify their strengths and weaknesses and to evaluate how appropriate they were for use as comparators for the methodology reported in this work. This section is the background to the method development stage of the project.

2.1.2 Definition of Interaction

To understand the definitions of interaction, also called synergy, it is first necessary to understand the definition of causality. A cause which inevitably produces an effect is termed a *sufficient* cause. A sufficient cause can be made up of a number of

component causes, where the lack of any component renders the remaining components insufficient. For example, the bacterium *Vibrio Cholerae* is the cause of cholera, whereas a sufficient cause might involve lack of immunity, poor sanitation, contaminated water or even blood group [43]. A *necessary* cause is one without which the disease would not happen, in the case of cholera, this would be exposure to the bacteria.

There are both biological and statistical definitions for gene-environment interactions, sometimes referred to as "plastic reaction norms". Biologically, an interaction is a situation where the qualitative nature of the action mechanism of a factor is affected by the presence or absence of another. The model underpinning the theory of biological definitions is additive, with interaction being suspected if the disease rates of the joint effect exceed the sum of the separate effects. If two risk factors are involved in the same sufficient cause, i.e. there is one pathway that confers disease risk in which both risk factors are required, they can be said to interact biologically.

Statistically, a gene environment interaction is defined in two ways, as: "a different effect of an environmental exposure on disease risk in persons with different genotypes," or "a different effect of genotype on disease risk of persons with different environmental exposures." For two risk factors to interact, they must not agree with the stipulations of conditional independence: that the relationship between two factors is the same across strata defined by a third factor. Assessing the effect of a genetic and environmental risk factor, G and E respectively, on disease risk would involve looking for Conditional Independence. This would mean that the effect of G on disease risk would be the same across a range of strata defined by E. If these conditions are not met and the effect of G on disease risk varies depending on E, an interaction can be said to exist.

A statistical interaction would imply that for the model to fit the data well it should include an interaction term. In simple linear regression, the non additive nature of adding an interaction term complies well with the biological definition. However, a logistic regression model is implicitly exponential and therefore multiplicative (only additive following logarithmic transformation). Therefore, including an interaction

term in a logistic regression model would imply that the interaction between the variables is non multiplicative [44].

Some epidemiologists use the term "Public Health Interaction" to denote the relevance of risk factors to the health of a population. In this case, if the disease rate is not dependent on two risk factors occurring together, they are considered independent from each other [4]. If the rate at which the two risk factors occur together is higher than expected, it is considered to be an interaction. This is very similar in definition to the biological definition and often has a basis in biological and statistical studies with results extrapolated to the entire population.

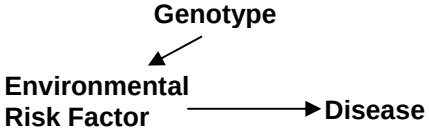
2.1.2.1 Biological Theories of Interaction

The magnitude and prevalence of interactions or joint actions of multiple factors in biology is still unknown. Canalisation, first identified in the 1930s, is the genetic capacity to buffer developmental pathways against mutational or environmental perturbations. More recent work has indicated that "cryptic" interactions or decanalisation may be the more common types of interaction [45, 46]. The possible mechanisms of interactions could include a number of possibilities. It is possible that the interaction is an inherent property of a network system and that the gene effect is context dependent, one example would be enzyme saturation in a multi step biological pathway. The second possibility is that there is a large network of interaction mechanisms, with positive and negative feedback regulation.

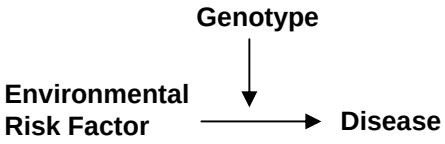
2.1.2.2 Biological Models of Interaction

There are five different biologically plausible models for how a risk genotype and an environmental variable might effect disease risk [42], they are:

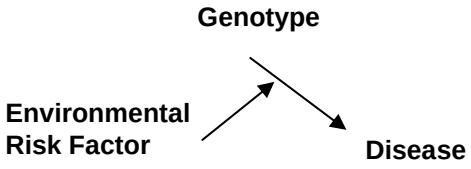
Model A: The genotype produces or increases the expression of an environmental risk factor, which can also be produced nongenetically.



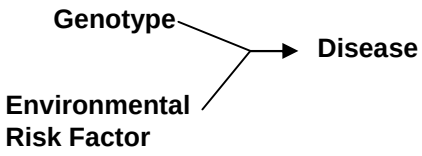
Model B: The genotype increases the effect of the environmental variable. People with only the genotype are unaffected.



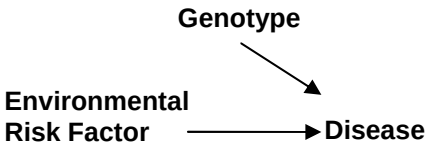
Model C: The genotype has an independent effect, which is exacerbated by the environmental variable. People without the high risk genotype are not affected.



Model D: Both the genotype and the exposure are required for disease risk. Neither has an independent effect



Model E: Both the genotype and the environmental variable have independent effects on the risk of disease. The overall effect is greater than the additive effects.



2.1.2.3 Gene-Environment Interactions on Disease Risk

Statisticians and epidemiologists may not regard the first model as an interaction effect; however the other four models match up with the recognised epidemiological definitions of gene-environment interactions, shown in table 2.1.

Table 2.1 Epidemiological Models of Gene-environment Interaction

Gene Variant	Environmental Risk	Model 1 (RR)	Model 2 (RR)	Model 3 (RR)	Model 4 (RR)
Absent	Absent	1.0	1.0	1.0	1.0
Present	Absent	1.0	1.0	Modest/High	Modest
Absent	Present	1.0	Modest/High	1.0	Modest
Present	Present	Very High	Very High	Very High	Very High
Biologically Plausible Model		D	B	C	E
Example of a condition		PKU	Podoconiosis	VP	COPD

The first type of gene-environment interaction to be considered is that in which both the environmental and the gene variant are necessary for the disease symptoms, either on its own will have no effect (Model 1/Model D). A well known example of this type of condition is Phenylketonuria (PKU), an autosomal recessive disorder that prevents the sufferer from being able to convert phenylalanine to tyrosine. Both the genetic variant and the presence of phenylalanine in the diet are necessary for the symptoms; if either is absent the person is asymptomatic. There is a small subgroup that are the exception to this rule and that is the children born to mothers with PKU, who are not themselves genetically predisposed to PKU, who have been exposed to high levels of phenylalanine in the blood due to their mothers enzyme deficiency. However, this can be prevented by the mother following an phenylalanine restricted diet [47].

A second type of model for a gene-environment interaction (Model 2/Model B) is when there is an established environmental risk, enhanced by the presence of particular genotypes, which, on their own confer no increase in risk. An example of this is podoconiosis, also know as non-filarial elephantiasis, which is a chronic and debilitating disorder endemic in regions where people are exposed to red clay soil derived from volcanic rock. The main symptoms are oedema of the foot and lower leg, caused by silica particles absorbed through the foot, causing an inflammatory reaction and destruction of the vessel lumen [48]. Only a small proportion of people exposed to red clay develop podoconiosis and that the disease tends to cluster in families [49]. Following a segregation analysis, they found that the most likely

explanation was that of an autosomal co-dominant gene, with age as a significant co-variate and wearing shoes having a protective effect [50]. This model fitted the data better than a model based entirely on environmental risks.

Another condition where the environmental risk is amplified by different genotypes is xeroderma pigmentosum (XP). Exposure to ultraviolet light is a risk factor for developing skin cancer. In combination with XP mutation, the risk becomes very high. Although not possible in practice, if an individual with the XP mutation avoided UV light altogether, they would have a similar skin cancer risk to the background risk [51].

In the third model (Model 3), a gene confers an increase in risk, yet the risk is enhanced or changed in some way depending on the environmental factors. An example of this is porphyria variegata, also known as variegate porphyria, (VP) an autosomal dominant disorder whose symptoms include abdominal pain, neuropsychiatric manifestations, sun sensitivity and blistering easily. Exposure to barbiturates, which is harmless for most people, for people with VP can lead to paralysis or even death [52]. The symptoms are caused by a partial deficiency of protoporphyrinogen oxidase (PPOX), PPOX is a mitochondrial enzyme involved in the haem biosynthetic pathway [53]. Mutations in the gene coding for this enzyme lead its deficiency and the resultant symptoms. The symptoms are worsened considerably on exposure to barbiturates.

Similarly a gene mutation might behave differently and develop new associated risks when exposed to different environmental variables. There are some genes involved in repair or oxidation pathways that might be a risk factor for a number of different cancers, including the CYP1A1 gene, with reported independent association with lung cancer risk [54]. Some studies have reported a relationship between coffee consumption and ovarian cancer risk [55, 56] and others have dismissed such claims completely [57]. However one study has documented a gene-environment interaction involving coffee and a variant of the CYP1A1 gene [58]. The study have found that having the gene variant alone does not increase risk of ovarian cancer, yet if people with this variant drink above median levels of caffeine, their risk of developing ovarian cancer is significantly higher.

Finally, if both a gene on its own and the environmental risk alone are associated with an increase in risk, but together the risk is significantly higher than the additive risk of both independent risks, this is a model of gene-environment interaction (Model 4). An example of this is Chronic Obstructive Pulmonary Disease (COPD), expected to be the fifth leading cause of disability by the year 2020 [59]. The most established environmental risk factor for COPD is smoking [60, 61]. There are other important environmental risk factors including childhood infections, air pollution and occupational exposure. However, there are also indications of a genetic component, one that may interact with the environmental variables [62]. For example, severe α_1-antitrypsin (AAT) deficiency, normally the result of a PI ZZ or PI Znull genotype, combined with smoking leads to more severe pulmonary impairment at an earlier age [63]. There are also reported variants of the CYP1A1 gene reported to interact in this way [64]. Further studies are needed to help clarify their finding as the segregation and linkage analysis approach used by Janus in 1985 assumes no interactions.

2.1.3 Variables

In the search for interaction effects, care must be taken when selecting each variable, whether environmental, genetic or outcome. The environmental exposure can be physical, chemical, biological, behavioural or a life event. When quantifying an environmental risk factor, it is important to consider how purely environmental a variable is. Usually gender and age are considered in the environmental exposure category, when both could also be considered genetically determined. Other factors that are classed as environmental may also be influenced by genetics, obesity, for example or even alcohol consumption. Thus it is important to investigate and think about each environmental exposure prior to analysis.

2.1.4 Studies and Analysis

For each new hypothesis or exploratory study it is also necessary to plan the most appropriate study and analysis method in advance. Often, the study to identify

interactions follows from a gene identification study or an environmental risk factor study that has shown unexpected results. More recently there have been studies primarily aimed at identifying interactions and methods adapted appropriately.

2.2 Search Strategy for Reviewing Literature

2.2.1 Search Tools

Initial searches were done using MEDLINE (National Library of Medicine), with the saved searches repeated in EMBASE (Excerpta Medica). After the results from the initial searches had been analysed, the references from relevant articles, or articles with a relevant section, were searched through and entered into an EndNote database. Once an article had identified a research group with a relevant interest, PubMed and Google Scholar were searched by author name for further articles from the same group. If an article was a basic introduction to a concept, a cited search was used to find any more recent articles that had cited the work. When back referencing from articles, Google scholar proved the fastest and most efficient way of finding the required article. When searching for specific articles, the other articles identified as most similar were also briefly read and assessed for relevance.

2.2.2 Basic search

There were two basic searches that were repeated for most of the subheadings. As the literature review was concerned with both genetic research in general and the more specific studies for gene-environment interactions, the first search looked for general papers on genetic studies. If any interaction studies are found from the basic searches, they were categorised as such and efforts were made in the search for gene-environment studies to ensure the same studies are identified by the criteria.

2.2.3 Inclusion/Exclusion criteria

For an article to be included, it needed to come from a credible scientific source. All types of paper were considered; whether reviews, commentary, specific studies or meta-analysis. They needed to focus on either a particular gene or genetic studies.

A study was excluded if the evidence was primarily lab based or methodological or if the primary study aim was not for humans. Studies with small sample sizes would only be used anecdotally if there were no other, more suitable, articles.

2.2.4 Search Methodology

In place of a second independent researcher repeating the search for inconsistencies, the entire search was discussed with the medical librarian, who crafted the search to maximise the results from the search engines involved. This was considered a suitable subsititute considering the timescale of repeated searches and the complexity of the subject matter.

There were common search terms used throughout, shown below. In each case the search was limited to humans and a variety of genetic and interaction synonyms were used.

An initial list of statistical methods was obtained through reading the “methods” sections of the papers that had been useful in defining interactions. This was followed by a search based on synonyms of “method” or “methodology,” before specific searches for each of the methods identified, using both their full names and acronyms. The initial descriptive paper for each method was tracked down, even if it was not directly relevant to gene-environment interactions. Papers comparing methodologies in high dimension data or for interactions were considered the most relevant. Possible bias of papers proving the superiority of a single method was taken into account and additional papers identified.

The basic search is shown below, with steps 5 and 6 only used when additional papers were needed, usually for specific examples where the environmental variable would be described as “BMI” or “eating carrots” without using the term “environment” or “exposure”

1. (1polymorphism$ or allel$ or genotyp$ or phenotype$ or isoform$ or mutation$ or gene or genes or genet$).mp. [mp=title, original title, abstract, name of substance word, subject heading word]
2. Limit 1 to humans
3. (environment$ or exposure).mp. [mp=title, original title, abstract, name of substance word, subject heading word]
4. 1 and 3 and interact$.mp. [mp=title, original title, abstract, name of substance word, subject heading word]
5. 1 and interact$.mp. [mp=title, original title, abstract, name of substance word, subject heading word]
6. 5 not 4

Abstracts were read and relevant papers selected, followed by citation searches using both back referencing, the citation tracker application and extracting relevant references from said selected papers.

2.3 Review of Statistical Methods

2.3.1 Introduction to Statistical Methods

There are a variety of statistical methods which have been developed to analyse the relationships between large numbers of genetic and environmental variables and disease status. Given the complexity and variety of gene environment interactions, there is no single, straightforward, statistical method that can accurately identify all possible types of interaction in all possible datasets. The following chapter discusses the strengths and weaknesses of the available statistical methods, including gene identification methods; epidemiological methods that can incorporate interaction terms; and methods that are designed to tackle a specific type of interaction between variables. It is important to first cover what is being measured, the data and the sources of the data in order to understand the methods themselves.

Significance and Estimation

When using conventional statistical/epidemiological methods to test for association between exposure and disease, there is a difference between significance testing and estimation; over the last fifteen years, there has been a move away from pure significance testing towards estimations or presentation of both [65]. Significance is testing whether or any differences in the data could have been produced by chance from variation during sampling and is normally in the form of a *p*-value (a value representing the likelihood of these results occurring by chance). *p*-values do not describe the magnitude of the relationship nor the statistical uncertainty surrounding the results [66]. Estimation involves quantitatively measuring the magnitude of an association producing a point estimate and a confidence interval around that estimate. This gives more information about the relationship between an exposure and an outcome and allows better assessments of its implications for interventions or further studies. In studies where non-statistically significant results are found, having a point estimate and confidence interval for an odds ratio instead of a *p*-value can help interpret the findings appropriately. For example, if the confidence interval is very wide, but only includes a value of close to 1.0 at one extremity of this interval, this

may indicate that the sample size was too small to yield statistically significant results.

Covariance

Statistical measures such as standard deviation and variance only work on data of a single dimension. As disease phenotype and gene-environment interactions incorporate multi-dimensions, the measure used to assess how much the dimensions vary from the mean, with respect to each other, is called covariance. Covariance is a statistical measure of the variance of two random variables, observed or measured in the same mean time period, that tend to vary together; when one of them is above its expected value, then the other variable tends to be above *its* expected value too. In this case the covariance between the two variables will be positive. The covariance would be negative when one variable is above and the other below their respective expected values.

Additive and Multiplicative Models

An additive model is one in which the effect of a risk factor or treatment will correspond to an absolute change in the outcome variable for all the members of the population under study. In a multiplicative model, the change will be proportional based on the previous values of the study population. For example an additive model will increase disease risk for every member exposed by 10%, so someone with 20% risk from other factors would now have 30% and someone with 60%, 70%. A multiplicative model increasing risk by 10% would increase the first risk from 20% to 22% and the second from 60% to 66%.

Accuracy Estimation Methods

It is important, especially when mining large data sets, to ensure that hypotheses suggested by the data are not being tested by the same data (sometimes referred to as Type III errors) [67]. To prevent this, different subsets of the data from those used for the model building or hypothesis generation can be used to validate findings and these are called accuracy estimation methods. A number of different exploratory methods being reviewed incorporate this approach.

The most commonly used method for estimating accuracy of results is cross validation, also known as rotation estimation. The data is randomly split into *x* approximately sized, mutually exclusive subsets. The method being validated is then trained (using data minus a subset) and tested (using the subset) *x* times and the accuracy estimated by dividing the number of correctly classified outcomes by the total number of outcomes [68]. The way cross validation is normally integrated into a study is shown in Figure 2.1. However, cross-validation may be wasteful of the data, with splitting the data leading to a smaller derivation set. In studies requiring a large sample size other methods should be considered.

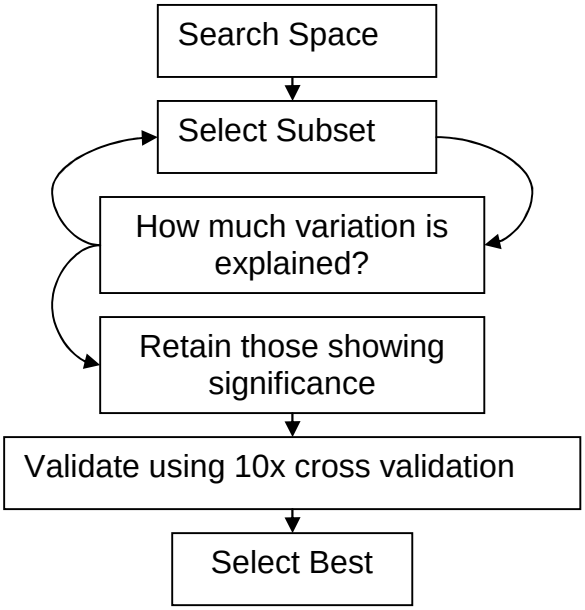


Figure 2.1 Cross Validation in Practice

One more efficient method in terms of sample size is bootstrapping [69], where the properties of an estimator are estimated by sampling from an approximate distribution. To generate an approximate distribution, the original dataset is randomly sampled and some members of the data are replaced by repetitions of other members. There is also "leave one out" validation, where one variable is removed at a time and the data tested accordingly, however, this can lead to high variation in estimates and is therefore unreliable [70].

2.3.2 Data

All the data is of the traditional DNA variety, using genetic code variations from the four bases (A, G, T and C). This corresponds to the real dataset under analysis but does not consider epigenetic or copy number variations. Aside from these more molecular based genetic markers, there are a number of quantitative genetic data types used for analysis. These include: Variable number of tandem repeats; microsatellite polymorphisms; Candidate SNPs; Haplotypes and Genome Wide Scans.

Variable Number of Tandem Repeats

Some studies identify variable numbers of tandem repeats, patterns of two or more nucleotides repeated directly adjacent to each other. Two nucleotides is a dinucleotide repeat, three is a trinucleotide repeat and a group of more than ten nucleotide repeats is considered a minisatellite.

Microsatellite Polymorphisms

Microsatellites are simple sequence repeats (SSQ) between 1-6 base pairs in length that, when polymorphic (ie vary between individuals) can be used as markers in genetic studies.

SNPs

Single Nucleotide Polymorphisms (SNPs) are DNA sequence variations in which a single nucleotide differs between individuals. Within a population, SNPs are assigned a minor allele frequency that describes the frequency of the least common allele at that locus. This can vary by population or ethnic group. Traditionally, sequence variants with a minor allele frequency greater than 1% were considered SNPs [71].

SNPs can lie anywhere on the genome: within non coding regions, coding regions or intergenic regions between genes. SNPs in non-coding or intergenic regions may still have an effect on phenotype as these regions can be involved in the regulation of transcription. Those in coding regions are divided into two groups, synonymous and non-synonymous, depending on whether the nucleotide change results in a different amino acid in the translated protein chain. The non-synonymous changes can result in

stop codon (a Nonsense, Ns, mutation) or a different amino acid (a missense mutation). NsSNPs, together with SNPs in regulatory regions, are believed to have the highest impact on phenotype [72].

SNPs are widely used in genetic epidemiological studies as markers of genetic variation, there are databases: GenBank, dbSNP and a number of disease studies that collate SNP data.

Haplotypes

In genetic analysis, a haplotype is a set of SNPs at multiple loci, transmitted together on one chromatid. For example, a first locus with the alleles G and C has the possible genotypes GG, GC and CC. A second locus has A and T, has the combinations AA, AT and TT. Therefore, across these two loci, there are nine possible genotypes. If a person is homozygous at one or both loci, for example genotype AGAC or AGAG, it can be deduced that the haplotypes are AA, GC and AA, GG respectively. Haplotypes are more complex when the person is heterozygous at both loci. There is collaboration among scientists worldwide working together to develop a haplotype map of the human genome. This is called the International HapMap Project (http://www.hapmap.org).

Genome Wide Scans

A genome wide scan (GWS) involves the genotyping of a large number of markers across the entire genome. The first genome wide association study (GWAS) was published in 2002 [73]. There have been many other studies since, including two studying Age Related Macular Degeneration (AMD) that complement each other and are therefore encouraging regarding the consistency of such a blanket approach [74, 75]. Although using such large numbers of genetic markers with traditional epidemiological statistical methods reduces the power for detecting interactions and increases the problems associated with multiple testing, the combination of GWS data and hypothesis driven research may aid advances in interaction research by having the data to test a prior hypothesis available [76, 77]. GWAS are themselves an effective hypthesis generating approach.

Environmental Data

The environmental variables in many disease studies are measured as part of the research, through either biological measurements (weight, blood chemistry); through questionnaires; from clinical records or observations taken by the researcher. There are a number of different methods for collating and measuring environmental variables for use in an epidemiological study. The methods of data collection and analysis are decided at the start of the study and take into account the most appropriate design for the study questions, time, cost and feasibility.

2.3.3 Statistical Methods for Gene Identification and Detection of Interactions

There are a number of different approaches attempting to describe gene-environment interactions. In some cases the idea of interactions is only explored when there is a prior hypothesis of interaction effects, either following a gene identification study or when biologically plausible.

There are also methods which attempt to visualise the data and integrate human skills into the data exploration process, through distorted overview displays and dense pixel displays [78] using human perception abilities to draw conclusions. This is intuitive, requires no understanding of complex mathematics or statistics and deals well with complex data. It is also possible to use clustering of parallel co-ordinates [79] or principal components [80]. Although, these methods may provide an invaluable method of generating a testable hypothesis for further study, they are not on their own a statistical method which can be consistently verified or compared and are therefore not useful in the context of this study.

2.3.4 Evaluation of Searching Methods

It is generally accepted that the influence of a particular gene is dependent on the context defined by other genes and by the environment. However, much of the literature, even when acknowledging this complexity, tends to focus on identifying

and characterising the effects of a single locus and the effects of the loci on variability. Such analysis is based on the assumption that interacting loci can be identified individually by their independent, marginal contributions to variability within a phenotype. This approach does not consider the possibility that the role of multi-locus functional genetic units is larger than that of single locus effects in influencing trait variation [81]. Single locus methods are only useful for studying single gene disorders or disorders caused by a single locus within a gene, but cannot detect complex patterns [82]. Single locus methods are not appropriate for studying complex disease as many genes may have marginal or non existent independent effects but contribute to a complex disease through their interactions with other genes [83].

2.3.5 Areas of Interest in Identifying Gene-Environment Interactions

Given the problems associated with large data sets and trying to identify interactions between variables within such large volumes of data, there are seven main areas in which the methods of analysis can be assessed and compared.

Dimensionality

The first problem in analysing data looking for gene-environment interactions is the dimensionality of interaction datasets. It is difficult to use many methods of analysis on data with so many variables, as the number of potential interactions increases exponentially as effects are added to the model. At such high dimensions, the data across numerous combinations of factors will become very sparse, with many cells in a contingency table containing little or no data: the curse of dimensionality. Any parameter estimates obtained would therefore have very large standard errors, which would increase the likelihood of type 1 errors [84].

Interactions

The second criteria for assessing the methods is how well they can detect interactions specifically, even for variables that have no, or marginal, main effects. A number of methods may detect interactions if they confer additional risk on top of the risk of the combined main effects, but would completely miss variables whose effect was dependent on interactions.

Power and the Effect of Sample Size

For every tested locus, there is the possibility of a type I error; therefore the power can be very low even when studying large data sets. Adjusting for multiple testing, using Bonferroni correction or False Discovery Rate (FDR), also leads to a decrease in power. Therefore, for a method to identify interactions successfully from a large data set, it must make some attempts to preserve power.

The problem of power is exacerbated by modest sample sizes, as considerably larger sample sizes are required to identify interactions than to detect main effects. If the numbers of events, or cases of a disease, are too few relative to the number of risk factors being tested, many methods are unable to accurately identify even the variables related to disease risk, ie the main effects, and will certainly not be able to identify interactions. This is quite closely related to the curse of dimensionality, yet would yield slightly different problems for parametric and non-parametric methods.

Marginal Effects

The fourth limitation of many methods in analysing genetic data is that many genes for complex common diseases have fairly small effect sizes. The ability of a method to identify risk factors where the relative risk is lower than 2 is an important criterion for effective analysis of genetic data.

Different Genetic Models

There are a number of different genetic models: dominant, partial dominance, dominant negative, recessive and co-dominant. There are also a number of ways that genetic factors behave differently than traditional environmental variables when related to disease risk. For example, genetic heterogeneity, where a single disorder or condition can be caused by different allelic or non -allelic mutations.

Computational Intensity

Another factor that can inhibit a statistical methodology is computational intensity. Although many modern computers can handle large data sets, the software may not translate well between systems and understanding the output can be time consuming and complicated. Minimising computational burden would be a benefit to a statistical method.

Applicability of Results

Finally, some methods deal so exclusively with theoretical statistics and the hypothesis space that the results cannot be translated to be of practical use. It is therefore important that any results gained from a method can be directly applied to a study population.

So in summary, the seven areas of interest to this review are:

1. **Dimensionality**
2. **Interactions**
3. **Power and the effect of sample size**
4. **Marginal Effects**
5. **Different genetic models; dominant, co-dominant etc.**
6. **Computational Intensity**
7. **Applicability of Results**

To these ends there is an assortment of statistical methods that fill a number of the above criteria and are worth investigation into their suitability for identifying gene-environment interactions in large data sets. The statistical methods under consideration in this review are summarised in Table 2.2:

Table 2.2 Summary of Statistical Methods for Interaction Analysis

Basic Gene Only Methods		Linkage Analysis
		Association Studies
Parametric Approaches	**Conditional Methods**	Multiple Regression
		Logistic Regression with Stepwise Selection
		ANOVA/ MANOVA
		Truncated Product of P-values
		Hotelling's T^2 test
		Multivariate Adaptive Regression Splines (MARS)
		Generalised Additive Models (GAM)
	Neural Network (NN) Based Methods	Feed Forward/ Back Propagation NN
		Parameter Decreasing Method (PDM)/ Parameter Increasing Method (PIM)
		Genetic Programming Optimised NN (GPNN)
		Grammatical Evolution NN (GENN)
Combined Approaches	**Two Step Approaches**	Set Association Approach
		Multiallelic Set Association (MSA)
Non Parametric Approaches	**Spectral Methods**	Singular Value Decomposition (SVD)
		Principal Components Analysis (PCA)
	Combinatorial Methods	Combinatorial Partitioning Method (CPM)
		Restricted Partitioning Method (RPM)
		Combinatorial Searching Method (CSM)
		Multifactor Dimensionality Reduction (MDR)
		Generalised MDR (GMDR)
		Odds Ratio MDR (ORMDR)
		Focused Interaction Testing Framework (FITF)
	Recursive Partitioning Methods	Patterning and Recursive Partitioning (PRP)
		Clustering and Recursive Partitioning (CRP)
		Random Forests

2.3.5.1 Basic, Gene-only Methods: Linkage and Association

Both linkage and association identify disease genes as being in the vicinity of a marker locus [85]. However there is a fundamental difference between a "necessary" disease allele with a penetrance between 90 and 95% and a "susceptibility" allele which confers an increase in risk but is neither necessary nor sufficient [86].

One approach to mapping disease genes involves scanning the entire genome for genes related to disease, in a hypothesis independent manner, by looking for regions that have been passed down through the generations largely intact from the founders of a population. Evidence for the ancestral segments looks for sections of DNA that are more commonly shared than would be expected from the relatedness of subjects, this difference is called "linkage disequilibrium"(LD). This parameter is also used to detect the presence of variants near or in disease genes [85]. If the presence of the allele is necessary for the disease to occur, the disease locus may be separate from the marker locus and in linkage disequilibrium with it, unless the association is complete.

Although predominantly used for case-control studies, LD can be observed in samples of affected sib-pairs that are independent from one another. The level of linkage disequilibrium is influenced by genetic linkage, mutation rate, recombination rate, random drift, non-random mating and population structure.

Association between genetic polymorphisms occurs when, due to genetic linkage, the association of the marker and disease alleles is non random. This occurs as a result of their proximity on the same chromosome and the fact that they are more likely to be inherited together.

Association studies can be direct or indirect [87] but are always based on the hypothesis that genetic variants affect disease susceptibility, even if they are neither necessary nor sufficient for disease to occur. Direct association studies catalogue all common variants and try and identify the disease susceptibility genes directly from the sample. Indirect studies use a smaller number of markers and incorporate LD measures to identify genes in the vicinity of a marker gene.

A large part of the rationale for searching for complex disease genes within families is the observation that having a parent with the disease confers an increase on risk in the offspring. These similarities between offspring and parents are attributable to the additive rather than the interactive effects of genes [88]. It could be argued that parents and offspring may share environmental exposures, however these are at such widely different times (childhood compared to adulthood) that the shared effects would not necessarily be expected to have the same effects. Statistically, a simple genotypic or allelic association with a dichotomous trait can be measured using a chi-squared test for significance.

It is extremely difficult for an association study to detect interaction effects as association approaches are designed to identify main effects. Interactions can affect power to detect genetic effects and limit the scope for replication in other studies [89].

2.3.5.2 Parametric Methods

Conditional Methods

Most traditional, conditional methods fall into the category called Generalised Linear Models (GLM), which use least squares regression to study the relationships between independent variables and outcome. This involves combining the model parameters, or regression co-efficients, to create a regression equation of the standard form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . \beta_n X_n + \varepsilon$$

The category of generalised linear models includes linear regression, logistic regression, log-linear regression, ANOVA and Poisson regression. Logistic regression is the most commonly used, especially for case control studies. However, all these methods have a great deal in common.

Linear Regression

Linear, or multiple, regression develops a model that uses a combination of values from continuous explanatory variables to predict the continuous value of a dependent variable.

Logistic Regression

Logistic regression relates a discrete outcome variable, for example case or control status, following a binomial distribution [90], to a set of explanatory variables, that can be discrete, continuous, binary or a selection of data types. The logistic function is used for such varied data as it can convert any value as an input value and use it to compute a value between 0 and 1. This prevents the calculation of impossible probabilities in cases where risk is combined or stratified in more complex analyses.

The actually methodology for logistic regression is different that the straightforward calculation in linear regression. The process is iterative, repeating cycles of calculation of maximum likelihood. It is possible to start with all the variables in the model and remove the least significant variables sequentially (backward selection), or start with an empty model and add in the most significant (forward selection). It is also possible to select the variables using a combination of forwards and backwards steps, called stepwise regression. Stepwise regression can have different, pre-determined starting points, depending on previous knowledge. It is also possible to start with variables known to have strong effects. After the first round of variable selection, the method can implement a stepwise procedure and work forwards and backwards to find the best model.

Analysis of Variance (ANOVA)

ANOVA is a branch of statistics originally devised by Fisher [91], and describes a collection of statistical methods where observed variance is partitioned into components using categorical explanatory variables. It is possible to use regression to fit an ANOVA model and use ANOVA to examine complex regression models, as it partitions the sum of squares variance into between group variance and within group variance.

As complex diseases have, by definition, several dependent variables, the ANOVA model most suited to investigation of the association between exposures and outcome is Multivariate ANOVA, or MANOVA. This technique attempts to identify the interactions among the independent variables and the association among dependent variables where there is more than one (correlated) dependent variable and the dependent variables cannot be combined. However, as each variable adds another degree of freedom, this technique is used in conjunction with a prior hypothesis.

Regarding gene-environment interactions, there are at least three problem areas in using ANOVA for analysis [92]. Firstly, there is always less power for detecting interactions than main effects as interactions are measured using sub-groups of the data and therefore smaller sample sizes and their detection is dependent on the size of their main effects. Secondly, estimates of main effects are inconsistent when interactions are present and this unreliability is increased in the presence of an unbalanced study design. Thirdly, as the gene frequencies are often unknown, their estimated effects are difficult to estimate making it difficult to apply results from an ANOVA study to other populations. Logistic regression, on the other hand, using contiuous explanatory variables, makes it possible to obtain a risk estimate directly from the regression co-efficient.

Other Conditional methods

There are a number of other statistical methods that rely on prior knowledge, a testable hypothesis or on good evidence for a loci or set of loci. They all have been developed to combine evidence from individual markers to test an overall null hypothesis that there is no association between the markers and the disease phenotype. In order to extend such a method to incorporate interaction effects would rely on specifying the potential interactions before analysis. One method that focuses on differences between populations, instead of within population is line cross analysis (LCA) [93]. There is the standard Hotelling's T^2 test [94], considered for genetic data by Xiong et al [95] or an adapted version of Hotelling's T^2 test to deal with haplotype data [96].

Other methods include multivariate adaptive regression splines (MARS) [97], truncated product of p-values [98], Generalised Additive Models and the global *U* statistic, which

combines univariate *U* statistics using generalised least squares [99]. These methods are not specifically designed to identify interactions, especially for variables with small or no main effect.

Some examples of practical uses where such methods have touched on interactions include using MARS to study multiple genetic contributions to hypertension [100, 101]. Hotelling's T^2 test is mainly used to study data on serum or microarray markers [102, 103].

Evaluation of Conditional Methods

Regarding the seven areas on which these methods are being assessed, logistic regression performs poorly in high dimensions, due to a high rate of false positives [104], which may be helped using stepwise selection. If the least absolute shrinkage and selection operator (LASSO), which minimises the residual sum of squares using penalty terms and structure detection, is applied, the co-efficients of the predictors with marginal effects is recoded as zero [105]. This allows a subset to be selected from the larger set for further analysis. As conditional methods rely on a prior hypothesis or candidate variables, they are not designed for mining large data sets.

In identifying interactions, logistic regression is not particularly effective unless the specific interaction is entered as an interaction effect, which would require a prior hypothesis. Using forward selection, only variables with a significant independent effect would be fitted and tested for interaction effects. This approach would miss any gene variations with interaction only effects. This could be overcome using backwards or stepwise selection, but then there will be an unmanageable number of degrees of freedom. The lasso shrinkage would also lose any interaction dependent risk factors and would therefore be inappropriate for detecting interactions. Backwards selection also does not work if the ratio of cases to possible predictors is too low and is therefore not practical with small sample sizes. Stepwise selection could be used but has a tendency to be over optimistic with results and, although technically possible, is not recommended for such high variables to case ratios [106].

To ensure sufficient power, conditional methods require a reasonable case: variable ratio, with estimates for the minimum number of events per variable ranging from 10

to 20. ANOVA in particular is very sensitive to noise or error and could therefore have trouble identifying marginal effects, although no conditional method would find a significant, yet small, effect when there were a high number of variables under study.

Also, the scale used to measure clinical variables may not correspond well to the scale on which genetic and environmental variables act. Transforming the scale, for example changing the scale to logarithmic, which is useful for many analytical approaches, assumes the absence of gene-environment interactions on the original scale. Such changes to the scale might actually make it more difficult to identify interactions [107].

As logistic regression models the relationship between predictors and disease risk for each individual, it cannot recognise subgroups with different gene to disease relationships. Therefore it cannot handle genetic heterogeneity well. The manner in which the data were entered would dictate if different dominance effects could be identified. The markers could be entered, with each variation as a positive or negative for a dummy variable, or they could be entered as haplotypes.

Given how high the degrees of freedom are when dealing with complex disease variables, none of these methods would be powerful enough to identify an interaction even in a large data set.

One advantage of both logistic regression and ANOVA is the ease with which widely available software can compute them and the large number of packages available which can be used to analyse data in this way. The greatest strength however, is that the results obtained from a conditional methods approach would be widely applicable to the population from which the sample was obtained.

Neural Networks

Artificial neural networks, often simply referred to as neural networks (NN), are non-linear statistical data modelling tools. They are not specifically a parametric or non parametric method as they provide large but not unlimited numbers of parameters. They were developed for tasks involving pattern recognition, signal filtering and data

classification [108]. They are based on the complex pattern of connections and processes used by our brains and are used in biology for analysis that is too complex and computationally intensive for traditional methods. Neural networks can be used to distinguish loci that are involved in an interaction, but have no main effect, from those that have no effect on the disease.

A neural network consists of an input layer, an output layer and one or more hidden layers. All the layers are composed of neurons (also known as units or nodes) and work in parallel with each other. The neurons of one layer are connected to the neurons of the next layer, with weights assigned to these connections. The weighting is the number of connections, for example in a case with 8 input neurons, 4 neurons on the hidden layer and a single output node, the weight would be (8*4) + (4*1) = 36. The number of input and output nodes is determined by the nature of the data, but the number of hidden layers, the parameters and configurations of the connections can be adjusted to optimise the analysis.

The process of developing a neural network involves splitting the data into a training set, used to calibrate the parameters of the network, and a testing set from which real results can be established. There are a number of different processes that take a network from a useless, basic premise with no set parameters to a useful data mining tool. However, only three are particularly suitable for the data type and study of gene-environment interactions: feed forward network (FFN), genetic programming optimised neural network (GPNN) and the parameter decreasing method.

Feed Forward Network (FFN)

The most commonly used type of neural network is the “feed forward” network (FFN), also known as the “back propagation” network (BPNN). This is the simplest neural network developed as the information only moves in one direction, forward, and there are no hidden loops or cycles in the network. In a feed forward structure, the information flows from the input layer, through the neurons of the hidden layers to the neurons of the output layer [109].

The earliest and most basic neural network is also referred to as a single layer perceptron, where the inputs are fed directly to the inputs via a selection of weights.

Each node calculates the sum of the input and the products of the weights and if it reaches a certain threshold, fires and assigns itself an activated value. The activated value is usually 1 and the deactivated value -1. A multi-layered perceptron consists of multiple layers and neurons, with each neuron in one layer having directed connections to the neurons of the next layer.

The appeal of neural networks is that they can learn from the training set of data and reduce the error (the difference between the calculated output and the expected output) through a series of repetitions. The most commonly used learning technique is called back propagation, where the error is fed back through the network to adjust the weights of each connection. This is repeated until the error is small and the network can be said to have learned the target function.

There are a few problems with this method, the main one being that the network may overfit the training data and fail to capture the true statistical process generating the data. Stopping the process early, after a pre-assigned number of repetitions can minimise this problem, as can ensuring there are a large enough number of cases in the training set.

Other types of neural network that are extensions of the feed forward network but not suitable for this study include: the Radial Basis Function (RBF) which works best in a small hypothesis space; the computationally intensive Kohonen self-organising methods; the Stochastic neural networks, most useful for studies in which some variables are unknown; and the modular neural networks, which separate the network into several small networks.

Feed forward neural networks have been used extensively in genetic epidemiology from basic testing of the method [110, 111] to gene identification studies for outcomes of interest [112-114]. Neural networks have been shown to reliably select SNP combinations that are associated with a multifactorial disease, childhood asthma [115]. However, neural networks have shown most promise in combination with other methods or variable selection procedures.

Parameter Decreasing Method (PDM)

A parameter decreasing method (PDM) can be used to select the important SNPs from a large sample of SNPs, before entering them into a neural network [115]. This is done by deleting one SNP and creating a model from the other SNPs, then replacing that SNP and deleting another one before making a model with the then remaining SNPs. This is repeated for all the SNPs and the models are compared. The missing SNP that correlates with the model with the lowest misclassification error is removed from the sample and the process repeated, deleting all the SNPs one by one. This is repeated until there is only one SNP left, the most important SNP. Then the other SNPs are put in order of importance. A model containing the 10 most important SNPs had the same prediction accuracy as a model containing all 25 SNPs in an experiment [116]. A permutation test can determine whether one of the selected SNPs is associated with the disease [117].

The PDM – neural network combination has been used to identify SNP combinations associated with childhood allergic asthma [115], and to select a set of ten important variables in postprandial lipaemia as a risk for cardiovascular disease [118].

Parameter Increasing Method (PIM)

The parameter increasing method (PIM) is the reverse of the PDM, starting with each single variable and selecting the best model before adding all the other possible SNPs individually and selecting the best model again [119]. PIM has been used in conjunction with neural networks to identify SNP combinations related to *H. pylori* susceptibility [119].

Genetic Programming Optimised Neural Network (GPNN)

One method that attempts to optimise the neural network architecture by using a machine learning approach called genetic programming (GP), is called genetic programming optimised neural network (GPNN) first proposed in 1991 [120]. The premise of GP is based on the principles of natural selection, with the fittest models and their offspring surviving to the next generation. As well as optimising the weights, GPNN employs a set of inputs selected from a larger set of predictors. The method was developed in 2003 with a view to identifying gene-gene interactions [121].

A sample of all the possible GPNN models is generated, each with a random selection of predictors taken from the larger set. The initial GPNN models may vary from one another in size. The models are assigned a fitness value proportional to its performance and resultant error rate. This fitness value informs the next set of models, so a predefined proportion of the best models are a subset of the next generation, sometimes referred to as their offspring. Other new models are produced by exchanging model parts between these models to make another subset of best models, the offspring of the best models. Then these subsets are grouped into a single group, the same size as the previous generation and replace the previous generation of models. The cycles then starts again to improve upon this generation, producing a set of models with even lower error values. This cycle continues to a pre-specified point, either when error reaches zero or negligible, or through a predefined number of generations. Then the model of the final generation with the lowest error is selected as the best model.

In a study looking for gene-gene interactions comparing traditional feed forward/back propagation neural networks (BPNN) with GPNN using artificial data, the GPNN showed improved prediction error and improved power. It can therefore be considered an improvement on BPNN when looking for gene-gene interactions [121]. BPNN has a tendency to overfit the data when non functional SNPs are present and therefore has a higher classification error than GPNN, which is more adaptable to new observations. GPNN has been shown to have high power at detecting relatively small effects (2-3% heritability) in data sets containing 2 or 3 locus interaction [122]. It also compares favourably, showing an increase in power, against logistic regression, [123] and stand alone GP [124]. GPNN has been used to identify a two locus interaction which confers an increase in risk for Parkinson's disease [125]. GPNN has been useful for selecting and modelling important predictors, but has not been tested for its efficiency in selecting the best predictors amongst a large number of variables.

Grammatical Evolution Neural Networks (GENN)

However, there are limitations to using the genetic programming algorithm: The implementation involves binary splits with each node attached to two nodes on the level below it. Although this was sufficient for small datasets, more complex datasets require more than two nodes. Similarly, there is a limit to the number of levels the

network can contain which may be problematic in more complex situations. It is also time consuming and restrictive to alter and recompile the source code for every alteration of GPNN.

In order to tackle these problems a similar model was proposed that uses the Grammatical Evolution (GE) learning algorithm [126] in place of the Genetic Programming method. This is similar to GPNN as it uses a machine learning approach to optimise the inputs from a pool of variables, weights and possible network architecture. The main advantages that the GE algorithm has over GP are based on flexibility. GE uses a linear genome and maps from the genome using grammatical rules similar to the DAN transcription rules to form mRNA. This is equivalent to comparing evolutionary changes at a chromosomal level, whereas GP is at the phenotypic level. GENN is therefore less computationally intensive than GPNN.

In comparisons using simulated epistatic genetic models, both GPNN and GENN performed better than the simple BPNN. Both GPNN and GENN reached the upper limit of power. It was suggested that simulated data is not the best test sample for the differences between the two methods as such artificial data do not generally contain much noise. Although not proven, this is the area in which GENN was designed to perform better than GPNN.

Neural Network Evaluation

Both the feed forward network and the parameter decreasing method can only handle a limited number of events per variable and are therefore not an efficient method for handling large data sets. However, neither GPNN nor GENN are affected by the curse of dimensionality as they only use a random sample of variables to build the first round of models and it is during the repetitions of the process that the most important variables are selected [127].

If there is an interaction present, it can be detected in a predictive sense by PDM. If one part of the interaction pair, or triplet, is removed it does influence the accuracy of the model predictions [115]. GPNN and GENN are the most powerful at detecting interactions between genes and the environment, although more work needs to be done to assess the limitations and parameters of these approaches and to compare

them. It is also not possible to isolate individual interactions from the larger predictive model.

When there is only a small sample, there is a neural network method called Radial Basis Function (RBF) that can be used [128]. The GPNN is designed specifically to work in a large hypothesis space but can be used on smaller samples of data. Although neural networks are successful at identifying small effects, as part of a group dynamic, in complex models [122], there is still work required in isolating these effects to look at them independently.

The neural network methods are effective in cases of genetic heterogeneity and in the presence of correlated markers. In power studies, GPNN has performed better than other similar methods. GPNN is not overly computationally intensive, and GENN is even less so. Neither GPNN nor GENN have open source code. If a practical stopping point is pre-assigned and over fitting is not allowed to happen. It is also a method where the results can be applied to the population; the figures are not a theoretical indicator but a practical one. However, such an application is likely to be as a predictive model not information on independent or two variable interaction effects.

Neural Networks are an interesting field with a very different approach to gene environment interactions and may be a useful comparison method. However the results are in the format of a predictive model, not easily divided into its constituent parts, it would therefore be difficult to compare to other methods attempting to identify specific interactions.

2.3.5.3 Combined Approaches

Two Step Approaches

There are some studies where a two step approach is used, with a non parametric method used to reduce the number of variables and then modelling this group of selected variables for interactions. For identifying gene-environment interactions, the first step would be applied to the large genetic data set and the environmental factors would be introduced at the second step. The second step can be a logistic regression

[129] using forward selection to create the model and backwards selection to remove the interactions that are not significant.

Set Association Approach

As a first stage of a two stage approach, the set association approach was developed by Hoh et al [130], and then improved to the current methodology a year later [131] as a non parametric method that could handle large numbers of predictor variables. It selects a subset of important markers, which can be either categorical or quantitative, from a large data set. It does this by evaluating sets of SNP markers at various positions in the genome, performing a simultaneous significance test on several sets of loci, whilst ensuring the type I error rate stays low.

The initial step of the set association approach is to determine a test statistic for each marker individually, which is a combination of measures of allelic association (AA) and deviation from the Hardy Weinberg equilibrium (HWE), with information included to try and minimise the effects of genotyping errors. The measure of association used is normally the χ^2 statistic, but other measures can be used. Although moderately high levels of deviation from HWE can indicate association with a risk locus, extremely high levels can indicate genotyping errors [132]. Therefore the statistics are "trimmed" to remove the top 1 percentile of Hardy-Weinberg Disequilibrium (HWD)

For the next step, a number of these informative SNPs from different genomic regions contribute to form a sum statistic of single marker statistics. The trait association statistic is chosen for each marker and the sets of statistics are summed. Finally, a permutation test is used to calculate significance. A limitation of the set association approach is that it deals only with bi-allelic loci as the χ^2 statistic would have different numbers of degrees of freedom for markers with different numbers of alleles. The set association approach also requires more research into the power and type I error rate.

Multiallelic Set Association (MSA)

MSA is an extension of the set association approach, which can be used on multi allelic markers [133]. MSA trims the markers using the same extreme HWD limitations as the bi-allelic set association on the higher values. However, the MSA also removes

a pre-specified (*d*) proportion of the smallest HWD *p*-values, along with their corresponding *p*-value for allelic association. This approach can theoretically be applied to any reasonable score function at each locus, allowing different data types to be combined and thus increasing the power of the analysis.

Set Association Evaluation

Regarding the curse of dimensionality, the set association approach reduces the large number of markers to a smaller set of important markers. However, a disadvantage is that interactions are only tested for between the markers that are selected as important. This would mean that important interactions with small independent effects might be missed, so the set association approach would not be a particularly strong method for identifying gene environment interactions.

Under a simulated multilocus inheritance model, Ott and Hoh found that a rather small number of case and control individuals can be sufficient to detect at least one disease locus of the three they simulated [134].

Despite being affected by genetic heterogeneity and correlation, the power of the set association approach in high dimensions is fairly high, and is better than if the Bonferonni correction or False Discovery Rate (FDR) procedures are used [135]. A sum statistic is the combined value of marker main effects and therefore performs better than methods that test each marker independently. For a meaningful power calculation the computer simulation should be carried out under a pre-specified, reasonable, multilocus inheritance model. The power can decrease considerably if prevalence of the disease gene increases or the heritability decreases. This can be explained by the oligogenic threshold model [136] as when prevalence is high and heritability is low, a small number of susceptibility alleles is sufficient for the disease to be expressed. However, in situations where neither interacting variable has a main effect, an exhaustive pairwise search of the genome was found to be more powerful than a two step selecting approach [137].

This method is also weak at detecting marginal effects and can be affected by genetic heterogeneity as the method tests the association between markers and disease for the whole sample. If there are different loci resulting in the same outcome, this will

decrease the association between each of the markers and the outcome and reduce the overall power. Correlated markers are also problematic as risk associated markers correlated with loci that do not influence disease will lead to an overrepresentation of these non susceptibility loci and thus reduce the power of the method. There has been an adjustment procedure developed to adjust the test statistic of a marker for correlation with other markers in the sum [135]. This negates the effects of correlation on power.

This method is easily implemented in a computer algorithm. There is open source software available in a programme called SunStat [138]. Results gained from the set association approach can be applied to the study population and can be fairly easily interpreted.

Examples of studies that have used the set association approach include looking for genetic variation relating to glucocorticoid excess in early onset Alzheimer's disease [139], gene-gene interactions in late onset Alzheimer's [140], susceptibility genes in head and neck cancers [141] and the occurrence of coronary artery restenosis following percutaneous transluminal coronary angioplasty (PTCA) [142].

2.3.5.4 Non-Parametric Approaches

Spectral Methods

Spectral methods are based on the analysis of eigenvalues, or singular values of a matrix, and attempt to solve differential equations using a series of known, smooth functions. They constitute a variety of methods that can be used in a number of applications including clustering, recognition and graph partitioning [143-146]. The most relevant spectral method for attempting to identify gene-environment interactions is singular value decomposition (SVD) which is applied in principal components analysis (PCA). SVD and PCA are common techniques for analysis of multivariate data and are now occasionally used in gene expression studies [145, 147].

Singular Value Decomposition/ Principal Components Analysis

SVD is a powerful technique dealing with sets of equations or matrices that are either singular or numerically very close to singular and is a factorisation procedure of rectangular, real or complex matrices. A number of methods other than PCA employ SVD; including fitting the data by least squares [148], computing the pseudo inverse [149] and matrix approximation [150]. SVD can both identify and offer solutions to the problems in a given matrix.

Taking a standard n×m matrix **A**, its first singular vector , normally called **v1**, is defined as the unit vector that stretches the most under the action of **A**, maximising ||**Av1**||**2.** In the subspace containing all the rows of **A**, the most stretched vector will also come from that subspace and therefore characterise it.

It is possible to construct trees using the basic tool from singular value decomposition (SVD) [151], however this is mainly used to develop phylogenetic trees, not statistical (classification, regression or decision) trees.

PCA is closely related to singular value decomposition; it is effectively applying SVD to the covariance matrix of the data. PCA is also known as the Karhunen-Loève transformation, the Hotelling Transformation or Proper orthogonal decomposition (POD) and is used as a tool in exploratory data analysis and for making predictive models.

PCA is defined in mathematics as an orthogonal linear transformation that transforms the data to a new co-ordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (the principal component), the second greatest variance on the second co-ordinate and so on. PCA reduces dimensionality by selecting and keeping the characteristics that contribute the most to the variance. Effectively PCA is rotating the co-ordinates so the transformed axes are aligned with the direction of maximum variance. PCA gives optimal results when a single source of information is corrupted with Gaussian noise and when the multidimensional scatter plot of the data forms a hyper elliptic shape [152].

There are a number of assumptions needed to be made in the process of PCA, so in order to obtain a result the applications become limited. These assumptions include: linearity of the data set and that the principal components are orthogonal with each other. There have been methods developed to work with PCA to avoid making these assumptions, kernel PCA and Independent Components Analysis (ICA), respectively [153, 154].There is also the assumption that large variance indicates importance. This would be the case if there was a high signal: noise ratio with the larger effects being an interesting dynamic and the smaller effects being noise. However, in genetic studies this is unlikely to be the case, as many genes only have marginal effects or work within a network of other genes and environmental variables. It is also a drawback that for PCA, variables under analysis need to be continuous.

PCA is used to try and identify and eliminate population stratification in genome wide association studies [155]. This works as an axis can be created that spans from Northwest to Southeast Europe and the PCs for each country are more genetically similar when physical proximity is closer [156].

Evaluation of Spectral Methods

PCA has been tried in a number of different genetic studies in an attempt to reduce dimensionality. Regarding gene-environment interactions, it may be possible to use PCA for some exposure variables. One example being sunlight exposure, and possible interactions with vitamin D receptor genes, as sunlight varies in a constant way from South-North and could be treated as a cline (a scale of continuous gradation). For SNP data, there could be false results from PCA where, from a sample of 20 SNPs, a group of correlated SNPs appears to have a bigger effect than the real effect of one SNP. As PCA does not identify the SNPs individually, it would be difficult to explain away confounding effects.

Identifying effects successfully in a study with a low ratio of cases to variables and the identification of marginal effects both depend on the variable being in trend (on a cline) with other variables. A cumulative, gradual difference may be detected where an inconsistent or isolated larger effect may be missed.

Compared to other popular gene identification techniques, PCA has been shown using simulations to be more powerful than other methods [157]. Computationally, PCA has to weigh up keeping the subspace that has largest variance against the computational intensity of analysing such a large space.

However, as PCA can only tell you the association between the phenotype and the SNPs included in the model as a whole, not the relationship with individual SNPs, it is impossible to extrapolate meaningful conclusions from results of a pure PCA. However, PCA can be used to eliminate SNPs that are in high LD from linkage analysis, thus reducing the dimensionality [158].

Combinatorial Approaches

Recently a number of approaches have been developed as tools specifically for detecting gene-gene and gene-environment interactions. These approaches give more insight into combinations and patterns for sets of genetic and/or environmental risk factors. They can be used for either binary or continuous outcome variables and were developed to explore high dimensional genotype space to try and predict the variation in quantitative traits in the general population

There are a number of related approaches that work on the basis that combining similar data into a single group reduces the dimensionality making the data more manageable and analysis of such data more efficient.

These methods include Combinatorial Partitioning Methods (CPM), Restricted Partition Method (RPM), Multi Dimension Reduction (MDR), Generalised Multi Dimensionality Reduction (GMDR) and the MDR extension Odds Ratio MDR (ORMDR). There is also a more generalised Combinatorial Approach which incorporates elements from the other methods.

Combinatorial Partitioning Method (CPM)

CPM was one of the first partitioning methods developed to try and analyse the large data sets being produced in genetic risk factor studies [159] and in response to the increasing burden and realised complexity of chronic, complex disease. The goal of CPM is to find a way to divide the multilocus genotypes into subgroups in such a way

that the overall trait variation can be explained by the characteristics of how the genes are grouped. There is no pre-specified genetic model required prior to analysis. CPM is used to study the effects of combinations of risk factors on a quantitative phenotype, and is therefore a suitable method for studying complex disease risk factors.

Traditionally, when testing if a locus has an effect on the phenotype, it is possible to use ANOVA as it tests the overall difference between the mean outcome values of different genotypes. It is difficult then, however, to determine the significance level without the problems of multiple testing. CPM however, determines the combinations of genotypes that influence the quantitative phenotype at the same time as defining groups of loci with similar phenotypic means.

There are three steps to combinatorial partitioning. The first step takes a subgroup of loci from the whole sample and combines the genotypes with the same outcome (case or control) into partitions. The set of partitions is evaluated, calculating the similarity within the partitions to the dissimilarity of partition means to assess how much variation can be accounted for. A group will be selected that predicts a pre-specified level of variance. The selection criteria can be the amount of disease explained by the identified group, or set, or the number of individuals in the identified group. A set with a small number of individuals can produce false positive results [159].

The second step is multi fold (usually ten fold) cross validation. The data is divided randomly into ten subgroups approximately equal in size. All the groups except one are used to estimate the mean of the genotype partitions of a set. The remaining one is used to compute the within partition sum of squares. Once the prediction error for this one group is calculated, the process is repeated for the other groups and then the average prediction error is calculated [160].

The third step is to identify the sets of loci that are most predictive of the variance and analyse what information they provide about loci combinations and relationships with disease risk. Selecting more than one set and comparing them increases information gain.

The main limitation of CPM is that the method is limited to studying pairs of variable loci (or other discrete variables) and cannot therefore fully capture all the epistatic or gene-environment effects. Extending to partitions of higher dimensions may be prohibitive in terms of the number of partitions to be examined. It is also unclear as to what capacity CPM has to detect additive effects of polymorphisms [161].

Examples of studies that have used CPM in interactions studies, though mainly gene-gene interaction, include a study of the genetic effects on plasma triglyceride levels [159] and the effect of metabolic pathway genes [162].

Restricted Partitioning Method (RPM)

RPM was designed by Culverhouse *et al* to try and reduce the burden of computational intensity of CPM, most of which can be considered unnecessary [163]. It is a form of recursive partitioning, similar to CPM but restricts the search, removing the partitions that only explain small amounts of variation.

In practice, the difference between RPM and CPM are the selection criteria, which consist of three steps. The first is a multiple comparison test to measure which groups have significantly different mean quantitative trait values. If all groups are equivalent, the process stops at this step. The second step is ranking pairs of genotype groups according to the difference between their genotypic means and merging the pair with the smallest difference into a single group. The algorithm then returns to the first step and repeats the process until the second step does not find significantly different means between the groups. If there are nine genotypes, as would be the case for two biallelic loci, the final partition will be found after no more than eight repeats of the process followed by an R^2 computation. For initial *n* genotypes, the algorithm will always stop after no more than *n-1* repeats.

Compared to CPM, this is much less computationally intensive: the same scenario which takes RPM 8 iterations and one R^2 calculation, would take CPM 21,146 R^2 calculations. This difference in computational intensity is even more pronounced for three way interactions with a maximum of 26 iterations for RPM to identify the most significant partition and 10^{21} partitions evaluated under CPM.

However, when studying large numbers of locus sets, each locus set needs several tests and each test needs a permutation test to calculate the p-value, so RPM can still be very computationally intensive. It also uses the Bonferonni correction to calculate the overall p-value from the individual p-values. The Bonferonni correction may be overly conservative, given the nature of the correlated locus sets.

RPM has been used in studies of cardiovascular disease risk factors [164], gene-gene interactions associated with autism [165] and behavioural reactions in different people following exposure to caffeine or nicotine [166, 167].

Combinatorial Searching Method (CSM)

The combinatorial searching method (CSM) was developed by Sha *et al* [168] and is similar to RPM in that it is an extension of CPM with the potential partitions being assessed for importance before analysis. Applying CSM to a data set filters all the possible sets of loci to retain the candidate locus sets for evaluation, using a new objective function based on cross-validation and partitioning. There are three steps to CSM as outlined in the Figure 2.2 below (Figure form Sha et al 2006).

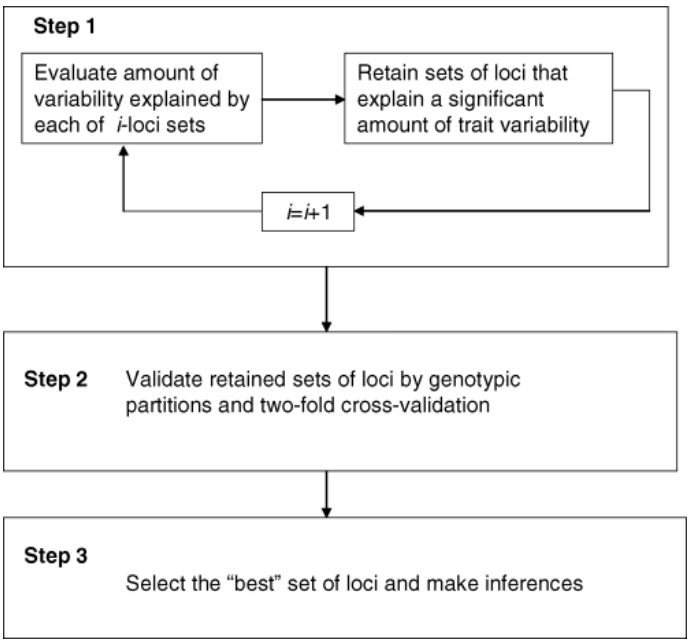


Figure 2.2 Combinatorial Searching Method

Step 1 involves searching every single locus set, two locus sets, three locus sets through to (pre-specified) *L* locus sets and retaining those that explain a significant amount of trait variability. The second step validates the locus sets that have been retained using two fold cross validation and assigns each locus set a value of the objective function. The higher the value, the more reliable and more predictable the locus set. Step 3 uses a permutation test to assess the statistical significance and calculate a p-value for the association between the locus set and the phenotype.

CSM has been used by Sha to study the variations of the angiotensin-converting enzyme (ACE) gene contributing to the ACE level and is being considered by a group studying genetic susceptibility to cancer [169].

Multifactor Dimensionality Reduction (MDR)

MDR can be used to analyse either genetic, environmental or a combination of genetic and environmental variables on a dichotomous outcome, such as case/ control status. It was first developed by Ritchie et al in 2001 in a study to identify high order interactions in sporadic breast cancer [170], however the theory and applications of the method are far reaching.

MDR is a variation of the combinatorial partitioning method, which is a method that involves collapsing high dimensionality data into a single dimension. In theory this allows interactions to be detected in relatively small sample sizes. The MDR algorithm looks for the strongest associations between variables and disease status. It only requires the input of two parameters; the numbers of variables to be selected at one time (N) and a threshold (T), guided by the goals of the study, as to what ratio of affected: unaffected distinguishes high risk genotypes from low risk. To find main effects, the N value can be set to 1.

MDR then incorporates a cross validation step. This paper uses ten fold cross validation which is the process where the data are divided into a training set (9/10ths of the data) and an independent testing set (the other 1/10).

One of the drawbacks of MDR is that it is necessary to pre-specify the number of interacting factors prior to analysis, so a three way interaction will go undetected in a

study into two way interactions. However, MDR can be adapted for use in case control, discordant sib-pairs and family based designs [84, 171, 172].

The majority of the work on interactions analysed using combinatorial methods has been done on gene-gene interactions. MDR has been used to identify multi locus interactions in a number of cancers including prostate cancer [173], bladder cancer [174] and sporadic breast cancer [170]. Other medical conditions where MDR has been used to identify gene-gene interactions include degenerative conditions such as familial amyloid polyneuropathy [175], Alzheimer's disease [176] and multiple sclerosis [177], common complex disease such as type 2 diabetes [178], asthma [179, 180], hypertension [181, 182], and myocardial infarction [183, 184], atrial fibrillation [185-188], autism [189, 190] and schizophrenia [191].

Generalised MDR (GMDR)

GMDR has a similar framework to MDR but has a wider application. Unlike CPM, RPM and MDR, GMDR allows adjustment for covariates, can handle both dichotomous and quantitative phenotypes. and can be applied to a number of different population based study designs [192]. GMDR uses the same data reduction strategy as the MDR method but changes the way the cells in the table are classified. Instead of describing the cells high risk or low risk based on their ratio of cases to controls, the GMDR uses a scoring method to classify cells.

The scoring method used in GMDR is a log-prospective likelihood of independent observations, with the predictor-variable vectors conditioned. The first stage in scoring the cells is to calculate Maximum Likelihood Estimations (MLEs) and scores for all individuals under the null hypothesis. Since the null hypothesis assumes no individual effects and no interaction effects, the score will be the same across all the cells. Then the cumulative score value is calculated for each cell. High risk cells have a score that is equal to or exceeds a pre-assigned threshold; low risk cells have a score below the threshold. The validity only depends on the availability of an appropriate statistic that can measure association between risk factors and the phenotype. Therefore, GMDR (like MDR) can be considered model free. GMDR can do everything MDR can do and more, including handling quantitative traits and

covariates. Compared to MDR, GMDR can improve prediction accuracy when the trait is influenced by covariates, even for complex and rare interaction models.

Although the GMDR approach can reduce some of the complications of MDR, it still uses high dimensionality computing. This is problematic when more than ten factors are used, and it has been suggested that combining this approach with knowledge about biological plausibility to select the factors may be more effective.

Odds Ratio MDR (ORMDR)

Instead of classifying multilocus genotypes into binary, high and low risk sub-groups based on case-control ratio, which can lead to a high false positive rate, ORMDR uses a variation on the Odds Ratio (OR) as a quantitative measure of disease risk [193]. This allows the genotype combinations to be ordered both by effect size, using the ORMDR measure, and by importance, using the confidence intervals.

ORMDR is the version of MDR that so far shows the most potential for identifying gene-environment interactions. It has successfully been used, alongside other methods, to identify a gene-gene interaction in Achilles tendinopathy [194].

Focused Interaction Testing Framework (FITF)

The Interaction Testing Framework (ITF) performs likelihood ratio tests in stages, performing joint tests of main and interaction effects conditional on lower order effects, so that the complexity of the test increases with the order of interaction considered. The joint tests of interaction are performed conditional on significant lower level effects [180].

Using ITF on its own in the presence of a large number of marginal effects, involves such a large number of tests that the type I error rate has to be adjusted and power is lost. Therefore the gene combinations are pre-screened using a goodness of fit χ^2 statistic that is dependent on association among the candidate genes in a pooled case control group. By controlling the false discovery rate in this way, the adjustments for multiple testing are not necessary.

The differences in performance between MDR and FITF depend on the genetic effects. MDR performs better when identifying a set of high risk multilocus genotypes for genes with little or no marginal effects. However, FITF performs better when the interactions involve additive, dominant or recessive genes. This difference is due to modelling assumptions, with MDR effectively the parametric method dependent on a susceptibility pattern and FITF as the non parametric method [180].

However, the testing on FITF used a sample of candidate genes found in literature, previous work and pathway genetics. It is unlikely that such an approach would perform so well in GWA without a prior hypothesis, where the number of comparisons, and therefore the number of tests, will be much greater [137].

FITF was used to identify a significant multilocus effect associating a group of three genes with childhood asthma [180].

Evaluation of Combinatorial Methods

The different combinatorial methods have different strengths and weaknesses in identifying gene-environment interactions. Without incorporating a filter or selection method prior to the combinatorial method, none of the approaches can handle very large data sets [186].

A very strong advantage of combinatorial methods is that they can all identify interactions without the variables having an independent main effect [163]. However, it is difficult to reach genome wide significance, for example - with the application of CPM to genes for coronary heart disease, the overall significance of the identified risk genes was 0.14 [159]. This is not considered a significant level,

Apart from the effects on power from genetic variability, combinatorial methods maintain their power well in the presence of errors, missing data or noise. The main advantage of using combinatorial methods is that they maintain their power to identify interactions when there is no main effect present [90]. The power of MDR for detecting gene-gene interactions actually increases when environmental variables are added to the model, as environmental variables may indicate subgroup differences.

Different genetic models can affect the power of combinatorial methods; with genetic heterogeneity and phenocopy (where a phenotype, under a certain environmental condition, is identical to a genotype determined phenotype) reducing the power and efficiency as different groups reaching the same outcome would decrease the variance of case: control ratio between partition groups. This decreases the prediction accuracy and the consistency of the model cross validation [160]. However, combinatorial methods are appropriate way to identify the dominance effects at the same time as identifying the risk locus. If there is a phenotype influenced by diallelic variation at locus A, a combinatorial method would identify locus A at the same time as separating the possible genotypes into genotypic partitions, for example: (AA, Aa) and (aa) in the case of dominance [159]. FITF has proved particularly effective at identifying genes with dominance, additive or recessive effects [76].

CPM, MDR and ORMDR are very computationally intensive, with RPM and GMDR designed to alleviate this problem to some degree. The results gained from a combinatorial approach can be applied to populations and the results are interpretable and useful in studies of disease risk.

Given the evaluation criteria, ORMDR is one of the most efficient methods currently available in the search for gene-environment interactions. Comparison of any novel method should be compared to ORMDR to justify any claims of success, or indeed failure.

Table 2.3 shows a summary of the different advantages or disadvantages of the methods assessed so far.

Table 2.3 Summary of the Strengths and Weaknesses of Different Methods

Feature	Success
Conditional Methods	
Dimensionality	Poor in high dimensions, high false positive rate
Interactions	Strong when interaction is pre-specified only
Power	Requires high case: variable ratio
Marginal Effects	Poor, made worse with transformation and loss of information
Genetic Models	Can be adapted if known in advance
Computational Intensity	Highly intensive
Applicability	Very Applicable, especially logistic regression
Neural Networks	
Dimensionality	Poor unless combined with a prior selection step
Interactions	Strong in predictive sense, hard to isolate individually
Power	Prior adaptation required
Marginal Effects	Cannot isolate individual effects
Genetic Models	Strong in a predictive sense
Computational Intensity	Intensity depends on adaptations, no open source code
Applicability	Applicable only as an entire model
Two Step Approaches	
Dimensionality	Reasonable if interacting variables have independent effects
Interactions	Only for interactions where variables have main effects
Power	Strong
Marginal Effects	Weak, especially weak interactions
Genetic Models	Affected by genetic heterogeneity
Computational Intensity	Relatively low computational intensity
Applicability	Applicable
Spectral Methods	
Dimensionality	Not appropriate for enough variable to assess
Interactions	Identification is dependent on trend not individual effects
Power	Powerful in correct circumstances
Marginal Effects	Cannot isolate individual effects
Genetic Models	Not possible to gauge, inappropriate
Computational Intensity	Not possible to gauge, inappropriate
Applicability	Not possible to gauge, inappropriate
Combined Approaches	
Dimensionality	Requires a prior filter stage
Interactions	Strong
Power	Strong
Marginal Effects	Strong
Genetic Models	Possible to draw conclusions on underlying model from results
Computational Intensity	Very computationally intensive
Applicability	Results interpretable

Recursive Partitioning (RP) Methods

Recursive partitioning methods are non parametric methods that divide the total dataset into smaller, more homogeneous subsets according to a set of predictor variables. This produces a regression or classification tree for continuous or categorical outcome variables respectively. Classification and regression trees were developed by Breiman, Freidman, Olshen and Stone in the early 80s [195]. The principal is that each object can be defined by its properties into a set of classes. The basic functions of a decision tree are to completely search an entire hypothesis space, select the best attributes and to decide if each branch can be justified. For the first division, the tree root, each attribute is statistically evaluated to see how well it fits the data. There are two different criteria for choosing the best attribute: ID3 trees, commonly used in machine learning, use information theory and gauge the best attribute as that which maximises the information gain ratio; Classification and Regression Trees (CART), more often used by statisticians, look for the decision that minimises classification error [195].

Similarly to neural networks, the tree components are referred to as "nodes," with the primary split the "root node" containing the total sample. The root node is split into two nodes which best improve the homogeneity of the case and control groups compared to the root node. Also, similarly to neural networks, cross validation is often used to create an optimal tree.

There are a number of advantages and disadvantages to recursive partitioning methods compared to other multivariate models. In some, but not all, cases there is an increase in accuracy using RP compared to other methods, [196, 197]. Despite performing similarly to logistic regression in predicting cognitive impairment, RP methods tend to be easier to explain and more intuitive, without involving any complex mathematics [198], which makes them a more useful tool in a clinical, as opposed to a statistical, setting. Sensitivity and specificity can be balanced during the selection process through prioritising and assigning misclassification penalties [199]. They do not make any implicit assumptions about the form of the underlying relationships between factors and outcome variables and can therefore identify synergistic interactions between factors and non linear relationships.

However, RP methods are not as effective when continuous variables are used [200] and are more susceptible to the problems of multiple testing and overfitting the data [199]. They are also perform poorly compared to logistic regression in identifying additive models [201] which is important for many clinical prediction models. However, these limitations mostly describe scenarios in which logistic regression can be used effectively.

There has also been work done combining logistic regression and recursive partitioning, which found that the combination was particularly effective at identifying at risk subgroups, with the RP methods identifying interactions that had previously gone undetected [202].

There are variations of recursive partitioning, including Cox linear recursive partitioning [199] which has been found to have better predictive accuracy than Cox linear modeling and simple recursive partitioning [200].

Recursive Partitioning in Combination

Combining RP with a different dimensionality reduction technique can reduce the number of terminal nodes and cross validated error rate [203]. Combining patterning (P) or Clustering (C) with RP, give acronyms PRP and CRP respectively.

PRP is an extension of CART that has been applied to genotype-phenotype association data, originally devised for viral genetics and later used for SNP association data. It assigns people to genotype groups based on their multi locus genotypes and then use this classification, or pattern, as a predictor in RP. MDR is effectively a special case of PRP where the tree is restricted to a single split and the misclassification error is used to measure impurity [204]. Therefore these two methods tend to give similar results, with only slight differences in accuracy estimates.

Compared to RP, both PRP and CRP reduced the error rate and PRP reduced both the false positive and false negative rates [203], with CRP reducing only the false positive.

Random Forests

The random forests approach uses a group of tree based models and identifies the class that is the mode of the classes from the different trees. The algorithm was developed by Breiman and Cutler [205] and combines models based on a random subset of the data to select important predictor variables. More important predictors will more successfully distinguish between cases and controls and will therefore be present in most of the trees and closer to the root node. Less important predictors will be present in less trees and nearer the terminal nodes [206].

There are two features that distinguish the random forest trees from those grown in a traditional deterministic manner: Firstly, the subsets used for growing the trees are generated using bootstrapping, repeating individuals within a sample to ensure the sample size is the same as the study sample. Secondly, the individuals not included in tree building are used to measure the prediction accuracy by the proportion of correctly and incorrectly classed individuals.

The majority of the large volume of research involving RP methods is in predicting patient specific probabilities of adverse outcomes or death. Such an approach can identify an interaction unrelated to main effects [202, 207, 208]. Recursive partitioning can be used for tissue classification using gene expression data [209] and was found to be more accurate in distinguishing distinct colon cancer tissues [210], although in some situations other methods perform better [211]. PRP was used to assign HIV medications to appropriate strains of the virus [203].

Evaluation of Recursive Partitioning Methods

RP methods perform well in high dimensions and are well designed to detect interactions between predictor variables. The random forest approach circumvents the dimensionality problem by selecting the most important predictors but does not have a cut off level at which a predictor is considered important [212]. However, the same study found that random forests could identify interactions where there were little or no main effects.

However this ability to identify interactions is affected by the sample size and the relative number of interactions present compared to the main effects. With a small

sample size, there are fewer learning and test samples and therefore greater uncertainty. In this situation a localised test can be used to balance the needs of validating the results and retaining the larger number of observations [213].

Recursive partitioning methods are used often in identifying marginal effect sizes, which was found to be justifiable using analytical and numerical arguments [214].

RP methods are able to identify genetic heterogeneity [212, 215] as different models are fitted to the data defined by early splits in the tree, effectively dividing the subpopulations into separate branches. Correlated markers can be a problem for random forests as the importance of risk SNPs can be underestimated. When disease risk is determined by haplotypes, RP can still identify candidate SNPs [214].

In large data sets, RP methods are less affected by the problems of multiple testing compared to more traditional methods as variables are identified by their relationship with the data instead of pre-defined thresholds. Combining RP with a patterning reduction technique (PRP) reduces the false negative rate and therefore increases the power of RP.

Given the cross validation and number of comparative splits, RP methods can be computationally intensive. This is reduced by incorporating a dimensionality reduction method prior to analysis. However the software is openly available and any results gained from RP methods can be applied to the population from which they are drawn. The use of RP methods in conjunction with other steps is an area of statistical development that has potential.

2.4 Summary of Findings

It has been suggested that, as many of the methods have different strengths and weaknesses, the best approach might be to use a number of different methods on the same data to try and identify the most important genes and understand the interaction patterns. Heidema et al suggest that the set association approach, MDR and random forests is a good combination [216]. However, the characteristics of the data set should be taken into account when deciding the most appropriate set of methods.

The two most successful current methods for identifying gene-environment interactions are stepwise regression and the Odds Ratio Multifactor Dimensionality Reduction (ORMDR) method. Both show successful results at identifying gene-gene interactions and are being used to investigate gene-environment interactions. However, the recursive partitioning methods show more potential for manipulation and are as yet relatively untested on gene environment interactions. Any changes made to ORMDR would be fine tuning an already good method, whereas the concepts of recursive partitioning have much more scope for further, more novel, development.

Recursive partitioning, therefore, forms the basis of the novel method developed by this project, in Chapter 5, to attempt to identify gene environment interactions in large data sets.

2.5 Conclusions

Methods that consider or exploit interactions between genes and the environment are of growing importance in the dissection of risk factors of complex disease. Even though methods have not yet been fully developed to deal with this complexity, it is essential that future genetic studies should gather information on behavioural and environmental factors. This will allow appropriate analysis to take place and ensure the data collection is not wasted.

The field of gene-environment interactions is an area of study both relevant and necessary to fully understand the nature of complex, chronic diseases. Development of a statistical tool that aids the identification of such interactions could have both clinical and public health benefits. Current methodologies are not wholly appropriate for the large volumes of data now available and the complexities of different potential disease models, but each has strength in a particular problem area. There is the potential to combine the methods or develop a new method using some of the same basic statistical theory

With techniques for molecular biology making such rapid advancements, it is becoming more common to analyse the entire genomes of study populations. In these gene identification studies, many environmental variables are also measured as potential confounders. Therefore data sets suitable for studying gene-environment interactions are becoming more common. The problems of power and dimensionality inherent in having so many variables are being tackled to some degree for gene identification studies, any resolution of which will also benefit interaction studies, though need further work. As a variety of studies of isolated populations are doing similar work on different populations, there may be differences in the genes identified depending on the environment of the population. If this happens to a number of genes, studies on possible interactions could become considered more important.

Chapter 3

Genetic and Environmental Influences on Colorectal Cancer

3.1 Introduction

3.1.1 Aims

The main aim of this review is to establish the research status of colorectal cancer. This identified potential genetic, environmental and interacting risk factors associated with disease risk, which informs the later practical application of the method.

3.1.2 Search Strategy

The colorectal cancer section used information from different sources including journals, reputable health information sites and health service reports. The search criteria were extended with clinical understanding of the condition, for example inclusion of the word “adenoma.” A more detailed version of the search criteria is shown below, again enhanced by back referencing, citation tracker and papers from other sections where relevant.

1. colorectal or colon or rectum or "large bowel" or "large intestine").mp
2. cancer or adenoma or carcinoma

Dietary variables:	Diet$ or food$ or nutriti$ or mineral$ (or specific term, e.g Retinol)
Genetic:	polymorphism$ or allel$ or genotyp$ or phenotype$ or isoform$ or mutation$ or gene or genes or genet$).mp.

3.2 Epidemiology of Colorectal Cancer

Incidence

Across the UK, colorectal cancer (CRC) is the third most common cancer in both men and women behind prostate and lung for men and breast and lung for women. The one exception was England in 2006, when there were slightly more colorectal cancers than lung cancers in women.

http://www.statistics.gov.uk/downloads/theme_health/MB1-37/MB1_37_2006.pdf
http://www.wales.nhs.uk/sites3/Documents/242/incpub2006%5F31Jan08.pdf
http://www.isdscotland.org/isd/
http://www.qub.ac.uk/research-centres/nicr/Data/OnlineStatistics/Colorectal/

There were approximately 36,700 cases of CRC in the UK in 2006, which is on average 100 new diagnoses every day.

Worldwide, the incidence rates of CRC vary by country, with developed countries having an incidence rate four times higher than developing countries [217]. At the population level there is a relationship with sex, with CRC more common in males than in females, with incidence in both sexes increasing as age increases. In fact, the majority of cancer occurs in older people [218]. Given the increasing age of the population in developed countries, this suggests not only a current public health problem but one whose importance may be set to increase.

Mortality

Despite an increase in incidence, there is a definite downward trend of deaths from colorectal cancer, even in recent years, as shown in table 3.1. Figure 3.1 shows the trend of deaths from colorectal cancer in the UK, the data from 2008 onwards is not yet available. It is worth noting the y-axis begins at 17 cases per 100,000 and that CRC is not close to being eradicated, as an initial reading may suggest.

Figures and graphs taken from the European Mortality Database:
http://data.euro.who.int/hfamdb/

Table 3.1 UK deaths from Colorectal Cancer

	2002	2003	2004	2005	2006	2007
SDR: Malignant neoplasm of colon, rectum and anus, per 100,000	19.13	18.91	18.62	18.31	17.91	17.74

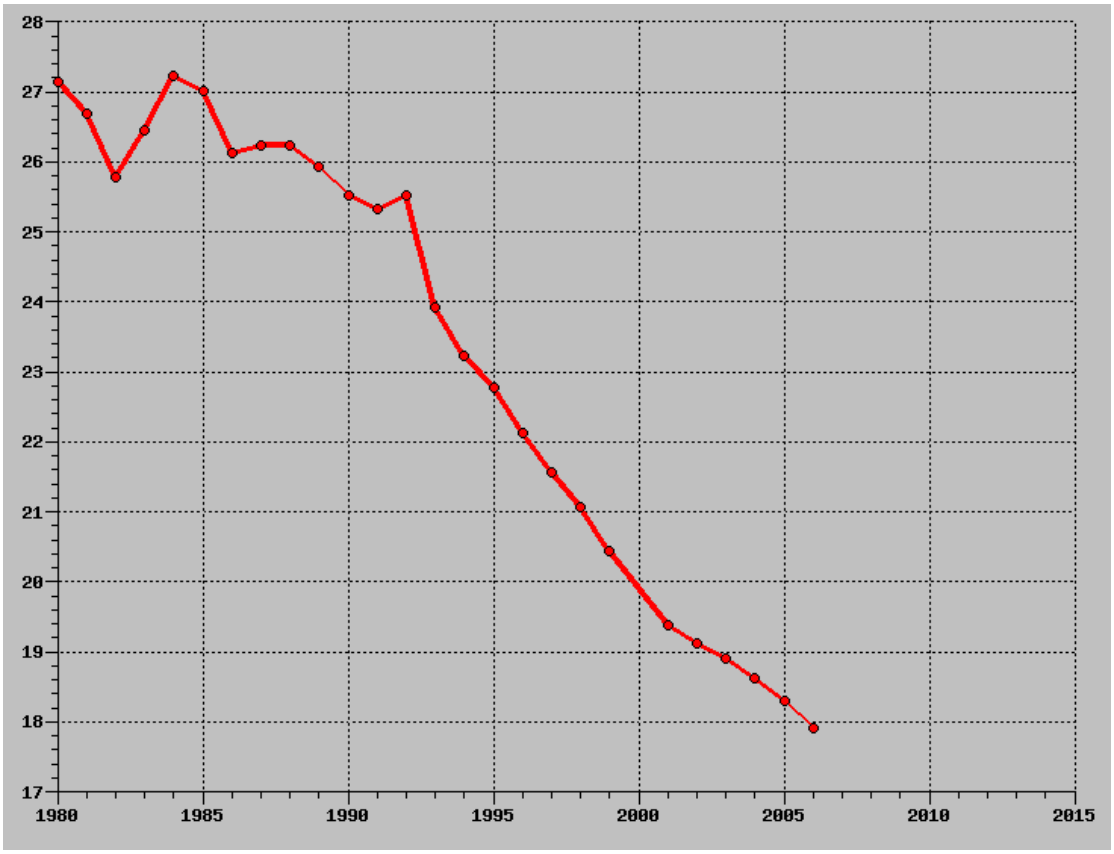


Figure 3.1 Mortality Rate from Colorectal Cancer from 1980 to Present Day

Genetic Screening and Prevention

There are a number of different approaches possible to try and reduce the mortality and morbidity of colorectal cancer, which can be broadly categorised into three groups: primary prevention, secondary prevention and chemoprevention.

Primary prevention involves identifying risk factors and promoting their avoidance in people's lifestyles so as to reduce colorectal cancer incidence [219]. The promotion of healthy diets, maintaining a healthy body weight and increased exercise would be considered primary prevention. Secondary prevention would involve screening people

for early signs of disease, in the case of CRC this would be the pre-neoplastic lesions and early stage tumours. Given the length of the latent period of tumorigenesis before the development of a symptomatic tumour, colorectal cancer is an ideal candidate for screening in this way. Chemoprevention is the use of drug compounds to reduce incidence of a disease, in the case of CRC this would involve non steroidal anti inflammatory drugs (NSAIDs) such as aspirin [220-224]. Fortifying a diet with calcium has also been shown to have a preventative effect [225] although other studies have found no effect [226] and some suggest that the effect is conditional on vitamin D status [227].

Colorectal Cancer Screening

The UK National Screening Committee has a set of criteria based on the condition, the test and the treatment proposed [228]. The condition should be an important health problem, be well understood, have incorporated practical primary prevention strategies and be understandable to those whom test positive. Colorectal cancer is therefore an ideal condition for screening. The tests should be simple, safe, precise, validated and acceptable to the general population with result thresholds agreed beforehand. There are a number of different possible screening mechanisms for CRC: Faecal Occult Blood Testing, colonoscopy, flexible sigmoidoscopies, double contrast barium enema and virtual colonoscopy. The suitability of these methods, especially their acceptability to the general population, does vary but a balance or combination can be used depending on the patient. The guidelines regarding treatment insist that a treatment is possible and that the policies regarding the treatment be well supported and agreed. The treatments for CRC do fit these criteria. Once these conditions are met, there are a number of conditions placed on the proposed treatment programme, finding the most suitable is an area under constant investigation [229-233].

Recently Scotland has introduced a nationwide screening programme using Guaiac Faecal Occult Blood Testing (GFOBT). The pilot involved three rounds of biennial GFOBT, between 2000 and 2007 for all people aged between 50 and 69 resident in Grampian, Tayside and Fife[234]. This pilot was based on evidence from other studies showing that regular FOBT can reduce mortalty from colorectal cancer [235, 236]. The pilot demonstrated that population-based colorectal cancer screening is feasible in

Scotland and should lead to a comparable reduction in disease-specific mortality; therefore a nationwide screening programme was introduced.

3.3 Clinical Characteristics of Colorectal Cancer

In general terms, cells following malignant transformation include the following characteristics [237]:

1. Self sufficiency in growth signalling, able to produce their own growth factors and self stimulate
2. Lack of sensitivity to agent that restrict growth
3. Evasion of apoptosis (programmed cell death)
4. Unlimited ability to replicate
5. Sustained angiogenesis, blood flow to tumour site
6. Ability to invade other tissues and metastasise

The combination of the rapid turnover of epithelial cells in the gastrointestinal tract and the hostile environment to which the cells are exposed has meant that gastrointestinal epithelium is an important tissue in cancer research. That colon cancer takes 10-15 years to develop and the progress through parallel histological and molecular changes has permitted extensive study of the development pathway. There are three main established, well recognised, characteristics: that there is a multipstep progression at the molecular level, that many inherited familial cancer syndromes correspond to key defects and the loss of genomic stability identified in tumours.

Colorectal carcinogenesis is a multi-step process involving global DNA methylation changes, hyperproliferation, adenoma formation and growth, specific somatic changes, and malignant transformation [238]. The development of colorectal cancer along these key stages has been standardised into a recognised process called the adenoma-adenocarcinoma sequence, a stepwise pattern of loss of methyl groups in DNA, mutational activation of oncogenes, for example K-ras, and inactivation of tumour suppressor genes, for example p53, [239], shown in Figure 3.2. There is epidemiological, clinicopathological and genetic evidence supporting the adenoma-adenocarcinoma sequence [240].

More generally, the development of tumours follows the "two-hit" principle proposed by Knudson [241]. The "first hit" is the disruption of the normal epithelium, via hyperproliferation, to form Aberrant Crypt Foci, which can turn into adenomatous or non-adenomatous polyps. Adenomatous polyps occur fairly regularly in the general population and often do not develop any further. However, if the "second hit" occurs, they undergo malignant transformation to become colorectal carcinomas. The development of colorectal cancer is a slow process, with the adenoma-adenocarcinoma sequence taking 10-15 years [242].

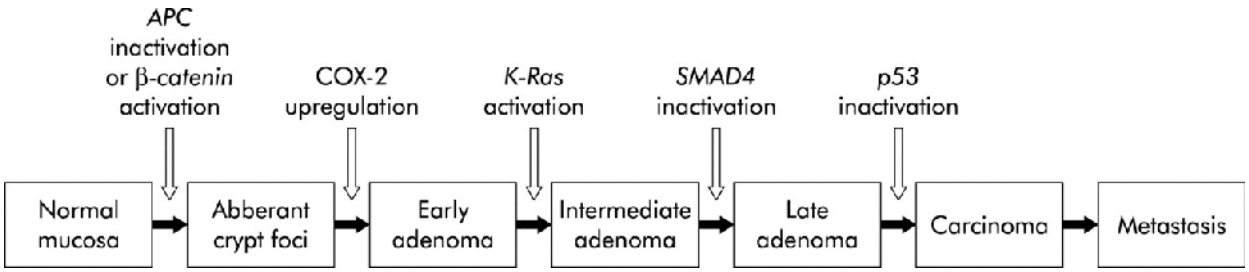


Figure 3.2 The Adenoma-adenocarcinoma Sequence, from Bronsens *et al*, 2005

The left side of the bowel is affected by cancer more often than the right: tumours in the sigmoid colon, rectosigmoid junction and the rectum together account more than half of all diagnoses. Figure 3.3 shows the percentage distribution of the different sites in the colorectal tract [243].

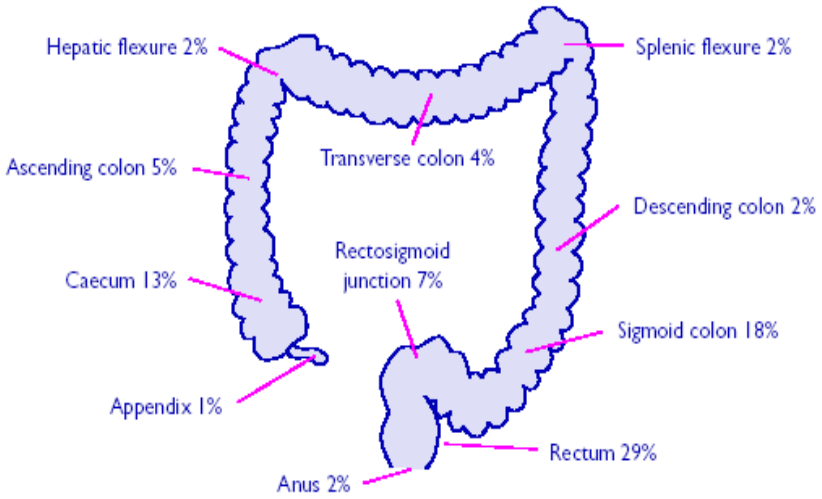


Figure 3.3: Percentage Distribution of Cases by Site within the Large Bowel, England 1997-2000 [243]

3.4 Environmental Variables Influencing Colorectal Cancer Risk

The hypothesis that colorectal cancer is influenced by environmental factors is supported by a number of different observations: the incidence of CRC is distinctly different in different countries [244, 245], migrant groups soon adapt to the risk estimates of their new country, often within a single generation of arrival [245, 246] and the sharp increase in the rate of colorectal cancer in Japan following the second world war [247].

A number of exposure variables are the same as those seen in other medical conditions, particularly cancer. Eating a healthy well balanced diet, not smoking or drinking and taking physical exercise a number of times a week are standard health recommendations, which will also help reduce your risk of CRC. For CRC specifically, it is plausible that dietary variables are a particularly important exposure variable as the large bowel has prolonged contact with digested food. In fact it has been suggested that following an optimal dietary approach, population wide, could reduce incidence of colorectal cancer by 70% [248].

BMI

Despite being a risk factor for so many health complications and disease, associations between BMI and colorectal cancer can usually be explained by accounting for dietary variables and physical activity [245].

Related Conditions

Inflammatory Bowel Disease (IBD) is a group of inflammatory conditions of the intestinal tract, the main two of which are Crohn's disease and ulcerative colitis (UC). The differences between these two conditions are the location and the nature of the inflammation, with Crohn's affecting the whole intestinal tract and UC being restricted to the colon and rectum.

Together with Familial Adenomatous Polyposis (FAP) and Hereditary Non Polyposis Colorectal Cancer (HNPCC), discussed later, ulcerative colitis (UC) is one of the top three high risk conditions for developing colorectal cancer with sufferers having a risk

an order of magnitude higher than those without it [249]. However, although FAP and HNPCC both have a well understood genetic basis, it is the chronic inflammation of UC that underlies the development of colorectal cancers [250]. People receiving treatment for UC are screened more regularly than the general public, so despite an increase in incidence, mortality rates are generally lower. Similarly, Crohn's disease increases the chance of developing colorectal cancer, especially those who develop Crohn's at a younger age, under 30 and patients with colitis (Crohn's inflammation in nature but presenting in the colon) [251].

A relationship between history of diabetes, and a higher fasting blood glucose, and risk of colorectal cancer has been found in women [252] but not in men. There is past evidence for a link, with some studies finding no difference between the sexes [253, 254] and some inconsistency regarding position of the tumours. A meta analysis fund that diabetes was a risk factor for colorectal cancer, with no significant differences between the sexes or position of tumour [255].

Another study found a significant decrease in incidence of hypertension, heart disease, stroke, chronic chest disease and chronic arthritis in people who developed colorectal cancer compared to controls [256]. They also found a significant association with haemorrhoids.

NSAIDs

Results from epidemiological, clinical and animal based studies have found some Non Steroidal Anti Inflammatory Drugs (NSAIDs), salicylate derivatives (including aspirin) and COX-2 inhibitors to have a preventative effect on CRC [257]. Even a low dose of aspirin has been shown to have a chemopreventative effect on adenomas in the colon [220].

There is both biological and epidemiological evidence for a protective role of NSAIDs in CRC risk. The painkilling mechanism of NSAIDs works by inhibiting both cyclooxygenase-1 (COX-1) and cyclooxygenase-2 (COX-2), catalytic enzymes involved in the synthesis of prostaglandins, through irreversible acetylation and competitive inhibition respectively [258]. COX-2 is an enzyme which is elevated in colorectal cancer tissues but not in normal colonic epithelial tissue [259, 260].

Epidemiological studies in a variety of populations and of different designs have found associations between NSAID intake and a decrease in incidence of colorectal neoplasms [261, 262]. However, randomised controls trials have had less success [263, 264].

One successful study split patients who had received treatment for CRC into an aspirin and placebo group. The aspirin group had lower levels of adenomas (17% to 27%) although there was no difference in adenoma size [223]. Interestingly a study of patients who had had adenomas removed found that low dose aspirin was associated with a reduced risk but higher dose aspirin was not [220]. Other studies have shown positive results after 20 years of continuous use [265-267].

Given the moderate effects of aspirin on CRC, it is not a suitable substitute for good screening, nor does the intention to treat analysis (1250 people for 10-20 years) indicate that population wide aspirin will reduce CRC incidence [261].

Smoking

Carcinogenic particles from tobacco could enter the cells in the colorectal mucosa either through the alimentary tract or the circulatory system. Nearly all recent research into the relationship between tobacco consumption and colorectal cancer has found a positive association between increase in consumption and increased risk [268]. Earlier studies, in the 1950s and 1960s, that found no associated risk, can be explained by the reduced timescale of smoking of the participants compared to more recent years. This association is also backed up by changes in trends in smoking between the sexes, with women becoming long term smokers later than men and their colorectal cancer incidence reflecting this. Chewing tobacco has been found to interact with particular genotypes to increase oral cancer risk [269].

Physical Activity

Physical activity has been one of the most consistent factors associated with a reduced risk of colon cancer. Early studies found that men with physical jobs had less colon cancer and more recent studies shows that increased amounts of exercise in leisure time can reduce risk [270]. Recent studies have found the effects on risk of rectal cancer to be equivalent to colon cancer [271]. This is backed up by cohort studies showing there is epidemiological evidence that men with high levels of either occupational or

recreational physical activity are at lower risk of developing colorectal cancer [245], an association that remains when diet and BMI are accounted for. However, research in Norway has found this association only to apply to males [252], though this could in part be explained by misclassification, with the physical activity level of housework being differently perceived.

Evidence has also been found to support an exercise based intervention with a control trial finding that middle aged or older people who took an hours exercise a day, six days a week had reduced colon crypt cell proliferation [272].

Deprivation

Comparing colorectal cancer incidence and survival against the Carstairs index, where 1 represents the least deprived category and 7 the most deprived, found that incidence varies little with deprivation but that survival increases with decreasing deprivation [245]. Men have a higher incidence than women across all seven deprivation categories. However, further research which adjusted for prognostic factors found the relationship between CRC and CRC associated death to be insignificant [273] but that death from other causes and co-morbidity were significantly higher in lower socioeconomic groups. One important factor is the lower acceptance to screening programmes in lower socioeconomic groups [274].

Female Hormones

A possible role of female hormone treatments in the progression of colorectal cancer was first suggested when an excess of colorectal cancer was found in nuns, who do not take contraception or fall pregnant [275]. When postmenopausal hormone treatments were less commonly used, the disease pattern of colorectal cancer varied between men and women, with more females than males developing cancer in the proximal colon, especially under 50 years of age. However, males over 65 have a similar excess compared to other cancers [276]. Males of all ages have a higher colorectal adenoma prevalence [277]. Therefore, without hormone supplementation, progression from adenoma to carcinoma in the proximal colon must be more common in women under 65 and may be related to female hormones. Biologically this is supported by mRNA for oestrogen and progesterone receptor proteins that has been found in the large

bowel [278], the expression of which decreases as carcinoma cells become more differentiated [279].

Over the last 25 years, the mortality from CRC has decreased slightly in men but much more drastically in women [280]. One possible explanation for this is the increased use of hormone replacement therapy (HRT) in postmenopausal women. Oestrogen could have an affect on CRC in a number of ways: by decreasing bile acid production, by decreasing the production of insulin-growth factor 1, by directly effecting the colorectal epithelium, or a combination of mechanisms early in the neoplastic process [281].

Most early studies found a relative risk of CRC and HRT either around or below zero [282]. A more recent meta analysis of 18 published observational studies [283] found there to be a 20% decrease of colorectal cancer between those who had never used HRT and those who had ever used it. This decrease was 34% for current users versus never users.

In studies of HRT use, other female only variables were identified that reduce colorectal cancer risk including number of children, with the risk decreasing for each additional child; breast feeding; the use of oral contraceptives; and BMI [284]. None of these factors were associated with HRT use.

Similarly, the oral contraceptive has been found to have a protective effect against CRC, probably due to the inhibitory effects of oestrogen on colon cancer cells [285]. A meta-analysis of 8 case control studies and 4 cohort studies found an 18% decrease in risk [286].

It is important to remember that the women receiving HRT are not a cross section of the general population but a group of women receiving treatment based on perceived need [277]. Any protective effect therefore may not extend to the general population. Given the decrease in mortality since the advent of HRT, it is unlikely that the medical problems for which HRT treatment is used are themselves providing some sort of protective effect and confounding the results. It is therefore more likely, assuming that behavioural factors such as social class have been adjusted for, that

case-control studies of HRT use may be underestimating the true protective effect conferred on the women who will benefit most from the treatment by comparing them to a lower risk control group.

It is also important to note that oral contraception has been associated with an increase in risk for a number of other cancers: breast, cervical and liver [285]. Therefore neither HRT nor the combined pill should be considered as a public health recommendation based on the potential reduction of colorectal cancer alone, considering the other risks.

3.5 Dietary Variables

Fibre

Traditionally known as roughage, an association between high levels of fibre consumption and low incidence of colorectal cancer was first studied scientifically by Burkitt in Africa, more than 35 years ago [287]. Since then fibre has been one of the most well researched, and most controversial, factors thought to affect colorectal cancer risk. Despite there being both biological and epidemiological evidence underlying the hypothesis, the limitations of different study designs have led to conflicting results across epidemiological, intervention and animal studies.

The biological evidence is based on the fact that fibre increases stool weight, which increases transit speed through the colon and reduces constipation, thus diluting the contents of the colon and stimulating the bacterial anaerobic fermentation [288]. The increased speed and fermentation reduce the contact between the intestine contents and the mucosa, leading to the production of short chain fatty acids, acetate, propionate and butyrate [289]. Butyrate reduces the pH of the colon and the production of secondary bile acids through primary bile acid binding.

Butyrate is also a major source of energy and helps reduce cell proliferation and induce apoptosis, which reduces the level of transformation of cells from colonic epithelium to carcinoma. It is possible that butyrate works by blocking the induced signalling events of IL6 and the IL6 receptor, an autocrine loop that promotes the development of many tumours [290]. It is also possible that butyrate modules the expression of glutathione S-transferases genes, thus enhancing toxicological defence in primary, adenoma and tumour human colon cells [291].

Epidemiologically, a strong associations was found in one of the largest prospective studies, a collaborative project of ten European countries, called the European Prospective Investigation of Cancer and Nutrition (EPIC) [292]. The collaborative study has an advantage over smaller studies, not just in sample size but in ensuring that homogenous eating habits do not increase measurement error to the level that it obscures all but the largest effects. In 2003, they found a significant association

between fibre consumption and reduced colon cancer risk, reported as both a hazard ratio between the highest and lowest fibre consumption quintile, and as a test for trend across all five quintiles. The Prostate, Lung, Colorectal and Ovarian Cancer Screening project team (PLCO) used sigmoidoscopies to screen and compared the dietary habits of the 33,971 people that were free of polyps in their large bowel to the eating habits of the 3,591 that has at least one adenoma in their colon [293]. They found that increased fibre from grains, cereals and fruits was associated with a decrease in CRC risk, but that there was no association for fibres from legumes or vegetables. A different study in Sweden found the benefits of fibre to come from fruit and vegetable consumption, with no protective effect from eating high level of cereal fibre [294].

However, a number of randomised clinical trials have failed to find any association between an increase in fibre consumption and adenoma recurrence. The Polyp Prevention Trial (PPT) tested a dietary intervention after four years [295] and then after a further four years follow up [296] finding that people could follow interventions successfully but that there was no effect on polyp recurrence. Other trials that show no effect include a study of wheat bran supplementation [297] and a high fibre, low fat intervention [298]. It has also been noted that the death rates from CRC are the same in vegetarians as non-vegetarians [299].

The sheer volume of confounding factors between people with a diet high in fibre and those with low levels can be make analysis difficult. One successful retrospective study adjusted for the higher fibre group being older, more likely to be female, better educated, exercised more, ate less red meat, lower levels of smoking, less alcohol consumption, more aspirin, higher consumption of folate and calcium. Pooled analysis of all the observational studies found the inverse association between fibre and colorectal cancer became insignificant when all other dietary variables were adjusted for [300]. Given that different sources of fibre may have different affects on risk, it is possible that fibre is in fact acting as a marker for an unmeasured substance that occurs jointly with fibre [293]. This substance could be glucosinolates [301], caratanoids, or beta-cryptoxanthin [302].

It is unlikely, given how large the sample sizes were, that the contradictory findings are the result of chance. However, treating dietary fibre as a variable that can be

accurately measured and characterised may be misleading and result in contradictory results. Dietary fibre is neither simple in structure nor simple to analyse [303]. Dietary fibre was first described in 1975 as the complex carbohydrates in the diet from plant sources that escape small bowel digestion and therefore reach the colon [304]. It was later defined as the plant cell walls that are resistant to digestive enzymes, however these vary greatly depending on species and cell type [305] and these different components may have differing effects on cancer risk, with some types of fibre even enhancing risk [306].

The intervention studies may have found no protective effect from additional fibre as the intervention level of fibre was too low [303] and there was no way to verify that people followed the recommended eating patters [296]. Similarly the nurses' study may have found no association as the nurses consumed low levels of fibre from cereals, and in the study by Mai et al [307], the 90th percentile of dietary fibre intake only consumed 18.2 g of fibre, well below the recommended amount. The same applies to the NIH-AARP Diet and Health Study, with the the 90th percentile eating only 15.9g of fibre a day [308]. Both the EPIC study and PLCO covered a much more varied range of eating habits and may therefore be comparing heterogeneous eating patterns including the beneficial fibres.

The timescale of the clinical trials may also be too short to be conclusive, with most lasting 3-4 years [309] and colorectal carcinogenesis in humans estimated to take 10-40 years to develop [310].

In conclusion, the results suggest that fibre, consumed at high enough levels, does have a protective effect but that further analysis subdividing the categories of fibre could show more consistent and verifiable results. Such work has been explored in animal models [311, 312] and although such data may be difficult to collect in humans, it may make comparisons between different studies more valid and useful.

Folate

Folate is micronutrient found abundantly in fruit and vegetables. It has been suggested that some of the inconsistencies between studies of fibre could be explained by the protective effects of fibre in European populations being confounded by folate intake,

whereas those in North America, where cereals are fortified with folic acid, are adjusted for possible confounding by folate [313]. However, increasing the sample size and adjusting for folate did not alter the association the EPIC study found between fibre and a reduction in risk [314], it did however suggest that the positive association between folate and reduced colorectal cancer incidence could be confounded by fibre intake [315].

Biologically, folate plays an important role in DNA synthesis and replication [316] with anti-folate agents used as chemotherapeutic treatments of cancer, as reduced folate inhibits cancer cell growth. Folate blocking drugs work as fast growing tissues, including cancers, require more folate for nucleotide synthesis and therefore up-regulate their folate receptors. Other possible mechanisms include altered DNA methylation or a relationship to the methylenetetrahydrofolate reductase (MTHFR) gene [317], which is involved in folate metabolism [318] although research does not indicate that these methods are likely in CRC progression [319]. However, there are also biologically plausible mechanisms by which a deficiency of folate could increase incidence of cancer, by reducing DNA repair efficiency. It is one vitamin where low levels can plausibly lead to an increase or a decrease of carcinogenesis.

A number of studies and meta analyses have shown an inverse relationship between folate intake and risk of developing CRC [320-322]. Although there is a protective effect of eating a folate rich diet, the degree of benefit seems to be greater for those that take a folate supplement [257]. However, there are also biological and epidemiological studies that show that increased folate can enhance the progression of tumours, increasing the risk of advanced lesions [323, 324]. It therefore seems likely that the timing of folate administration is important. Folate administration prior to the development of the first pre-neoplastic lesions can prevent initial tumour development, whereas increased folate after early lesions are established increases tumorigenesis [325].

Meat and Fat

One of the specific components of the western diet that has been proposed as a possible risk factors for CRC is increased intake of both animal fat and saturated fat [326]. A meta analysis of studies over the 1973-1999 period found that a high intake of red meat, especially processed meat, was associated with a moderate but significant

increase in CRC risk. Total meat consumption is not associated with risk but the risk is higher in regions where red meat composes a higher proportion of the diet. It was concluded that reducing red meat intake to 70g/week in these areas would decrease the risk of colorectal cancer for the population by 7-24% [327].

Other Dietary Variables

Other dietary variables that have been associated with CRC risk include calcium and vitamin D (both separately and in combination), alcohol, flavanoids and antioxidants.

Epidemiological studies have found an inverse association between consumption of dairy products and hypertension, stroke and colorectal cancer [328]. Increased consumption of calcium may have a protective effect by binding with bile and fatty acids and by direct inhibition of colonic epithelial-cell proliferation [257]. Pooled analysis of ten prospective studies found the protective effect of milk consumption was limited to tumours in the distal colon and rectum [329]. Most retrospective studies have also found an inverse correlation between calcium intake and CRC risk, though not always reaching significance. The main limitations of these studies is accurately measuring calcium intake and confounding from other dietary variables. One advantage to calcium supplementation, compared to other methods of chemoprevention, is that the preventative effect begins in a relatively short space of time, within a year of starting supplementation [225].

Alcohol has been associated with colorectal cancer risk both independently and as a negating factor for other protective dietary variables. The associated risk seems to be more consistent for proximal colon, distal colon and rectum [330]. Pooled analysis of a number of different studies found that a single determination of alcohol intake correlated with a slight elevation in CRC rate, an effect more pronounced for higher levels of alcohol consumption [331].

Despite being most widely known for their antioxidant activity, flavanoids may also use other mechanisms to provide health benefits. Flavanoids are found in plants and can be subdivided by chemistry into ten different subgroups. One category of flavanoid and three individual compounds have been associated with reduced risk of colorectal cancer [332]. This has been backed up by a randomised dietary intervention

trial which found that flavonols were associated with decreased risk of advanced adenoma recurrence, but not total flavanoid consumption [333]. However in a both meta analysis and a review other antioxidants have been shown to have either no effects or detrimental effects on the risk of different cancers of the gastrointestinal tract [334, 335].

Whatever the specific reasons, eating a diet rich in a variety of plant foods, including fruit, vegetables and wholegrains, remains the best option for reduction of colorectal cancer risk and for general good health.

3.6 Genetic Susceptibility to Colorectal Cancer

Similarly to other conditions with a genetic risk component, the inherited risk of colorectal cancer was initially identified by its tendency towards familial aggregation. Familial colorectal cancer is a public health issue due to its relatively high frequency, with 15-20% of colorectal cancers being familial [336]. Some twin studies have found that as many as 35% (95%CI of 10 - 48%) of cases of colorectal cancer have a genetic component [337]. The genetic risk factors involved in colorectal cancer are not yet fully understood and include genes with different dominance patterns, pathogenic mutations with low penetrance and gene-gene and gene-environment interactions.

Persistent genomic instability, which contributes to the accumulation of mutations in tumour suppressor genes and oncogenes, is essential for all colorectal cancers. The instability can arise through two separate distinct pathways those with microsatellite instability (MSI) and those that show chromosomal instability (CIN) [338], with both occurring after adenoma formation but before malignancy [339]. MSI mutations are recessive, with cancers occurring more frequently in the right hand side of the colon, having diploid DNA and behaving indolently. CIN mutations are dominant, the resulting cancers tend to be on the left hand side, show aneuploid DNA and behave aggressively [336]. Both types of instability can be characterised by associated mutations.

Cancer Syndromes

There are two autosomally inherited cancer syndromes that account for a significant minority of colorectal cancer cases and confer large increases in risk to those with such a family history: FAP and HNPCC.

The first, called Familial Adenomatous Polyposis (FAP), an autosomal dominant precancerous condition of the entire colorectal tract, characterised by the appearance of large numbers (>100) of adenomatous polyps. It is one of the most clearly understood and well studied of the inherited colorectal cancers with an incidence varying from 1 in 7,000 live births in Western countries to 1 in 22,000 in developing

countries [340]. FAP is associated with a number of other cancers, it is therefore important to follow people with FAP, even after a prophylactic colectomy [341].

A similar condition called Attenuated Familial Adenomatous Polyposis (AFAP) occurs later in life and in a slightly different area of the bowel (proximal colon as oppose to distal colon or rectum) [342]. There is also a condition called attenuated Polyposis which does not involve germline mutations in the APC gene, but in genes that interact with APC or mismatch repair genes which increase the risk of somatic APC mutation.

The second major condition is called Hereditary Non-Polyposis Colorectal Cancer (HNPCC), also known as Lynch syndrome, which is thought to account for 5-8% of colorectal tumours [336]. Lynch syndrome I is specifically familial colorectal cancer, Lynch syndrome II is other cancers of the GI tract or reproductive system. The average age of diagnosis of HNPCC is 11 years earlier than the average age of diagnosis for other colorectal cancers [343].

FAP and AFAP are the result of a CIN, germline, truncating mutation in the tumour suppressor gene APC, with the mutation on chromosome 5q, distal and proximal to 5' respectively. 90% of the people carrying this mutation will develop CRC by the time they are 45. A large number, around 40-60%, of the families diagnosed with HNPCC have pathogenic, MSI mutations in mismatch repair genes.

There is a rare autosomal recessive condition called Turcot's syndrome which is having both FAP and HNPCC which can lead to CRC and tumours in the brain and skin. Other than these two there are a number of other, less common syndromes which are associated with an increase in colorectal cancer risk, though these may be more strongly associated with other cancers: Peutz-Jeghers syndrome, Juvenile Polyposis, Cowden, Li-Fraumeni and Bloom syndrome. A summary of the genes involved is shown in table 3.2.

Table 3.2 Different Cancer Syndromes and their Associated Genes

Syndrome	Genes
Familial Adenoma Polyposis (FAP)	Adenoma Polyposis Coli (APC)
AFAP	
Hereditary Non-Polyposis Colorectal Cancer (HNPCC)/ Lynch	hMSH2
	hMSH6
	hMLH1
	PMS2
Li-Fraumeni	p53
Cowden	PTEN
Juvenile Polyposis	BMPR1A
	SMAD4(DCP4)
Peutz-Jeghers	STK11
Bloom	BLM
Familial GI stromal tumour	KIT
	PDGFRA

FAP - Adenomatous Polyposis Coli (APC)

The APC gene is on chromosome 5q21-22 and encodes a protein important for cell adhesion and signal transduction. The loss of APC is one of the first events in the chromosomal instability (CIN) pathway.

There have been more than 1000 mutations of the APC gene discovered [344] more than 300 of which are associated with disease. More than 98% of APC mutations are nonsense or frameshift, leading to the synthesis of a truncated protein which cannot suppress the cellular overgrowth. The most common occurs in 10% of FAP and is a deletion of 5 bases. A hypermutable tract in APC has been identified in the Ashkenazim population, but not in non Jewish populations, which both increases incidence and reduces age of onset for carriers [345]. In some cases, less obviously familial CRC can be caused by incompletely penetrant, comparatively rare missense mutations in the APC gene.

Biallelic inactivation of APC can also occur through other routes, Spirio et al found some families affected by attenuated polyposis had the loss of mutant germline APC accompanied by new somatic mutations in the remaining allele [346].

HNPCC

A large number of polymorphisms in the human mismatch repair (MMR) proteins MLH1, MSH2, MSH6, PMS1, and PMS2 have been found to co-segregate with HNPCC [347]. Over 90% of the polymorphisms discovered so far occur in hMLH1 (human mutL homolog 1) and hMSH2 (human mutS homolog 2) [348] as shown in the table below.

Table 3.3 MMR Gene Polymorphisms, adapted from Mitchell et al, 2002

Gene	Type of Mutation	Position (exon, nucleotide)	DNA Change
hMLH1	Nonsense	13, 1459, C-T	
		17, 1975, C-T	
		19, 2141, G-A	
		8, 676, C-T	
		2, 184, C-T	
	Missense	16, 1852, AA-GC	Lys618Ala
		4, 350, C-T	Thr117Met
		19, 2146, G-A	Val716Met
		4, 320, T-G	Ile107Arg
		2, 199, G-A	Gly67Arg
	Insertion	13, 1490, C	Frameshift from codon 497
	Deletion	16	3.5kb deletion
		16, 1846, AAG	Deletion lysine codon 616
	IVS deletions (d)/ insertions (i)	IVS*5, 4541 (d)	Out of frame deletion
		IVS*14, 1667 (d or i)	Allele silenced
hMSH2	Nonsense	7, 1216, C-T	
	Missense	6, 965, G-A	Gly322Asp
	Deletion	12, 1786, AAT	Deletion asparagines codon 596
		IVS*5, 942+3	In frame deletion, exon 5

Attenuated Polyposis

One of the main tumour promoting effects of the somatic mutations associated with attenuated polyposis is that they can lead to an over-activation of the Wnt (wingless) signalling pathway, up regulating expression of genes that promote cell growth. This pathway is shown in figure 3.4, and shows the functional role of APC and other molecules involved in signalling [349]:

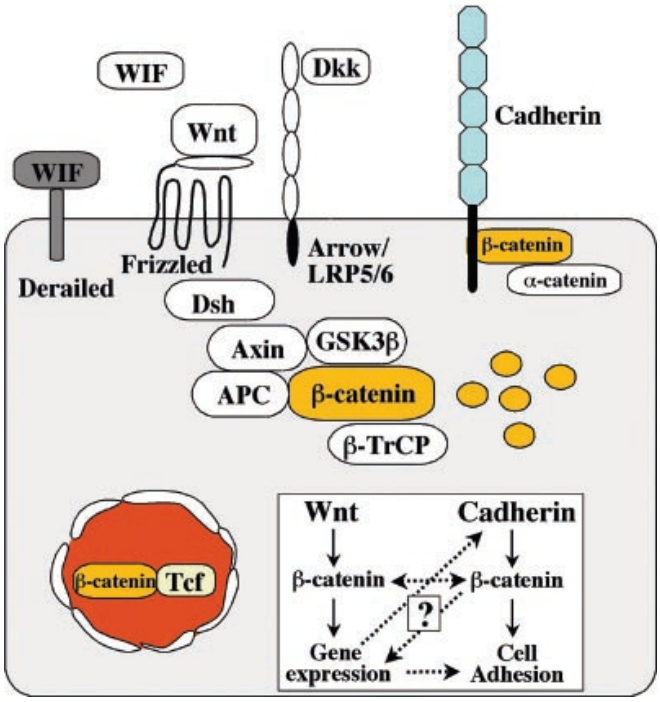


Figure 3.4 The Wnt Signalling Pathway

When the Wnt signal binds to the Frizzled compound, it also binds to AXIN, which leaves the β-catenin in its stable state. In the absence of Wnt signalling, the AXIN forms a compound with APC and GSK-3β, which phosphorylates β-catenin into a compound that is degraded in the proteasome. β-catenin, in its stable state, initiates cell proliferation, therefore its degradation halts cell proliferation. The mutated APC leads to an accumulation of β-catenin in the cell nucleus, leading to cell proliferation and colorectal cancer. β-catenin is also a key component of the cadherin complex, which controls cell-cell adhesion and influences cell migration [349]. Once the excess β-catenin translocates to the nucleus, it interacts with other transcription factors including the T-cell factor enhancing factor (TCF). TCF-4 is the predominant TCF expressed in colorectal epithelium [339]. Target genes up-regulated by TCF-4 include

CYCLIN D1, C-MYC, MATRILYSIN, C-JUN, FRA-1, urokinase-type plasminogen activator receptor and the peroxisome proliferators activator receptor.

The genes on the Wnt signalling pathway include: CTNNB1 (β-catenin gene, 3p22.1), BTRC (β-transducin repeat-containing protein gene, 10q24.3), ICAT (inhibitor of β-catenin and Tcf-4, 1p36.2), AXIN1, AXIN2 (17q23-q24), TCF4 and CDX-2. Other genes up-regulated by the Wnt pathway include two members of the CNN family of growth factors WISP1 and WISP3, and PPP2R1B a subunit of the PPA-2 tumour suppressor gene involved in β-catenin phosphorylation [350].

Mutations of β-catenin, much like mutations in APC, make β-catenin insensitive to the compound leading to its degradation. The majority of CTNNB1 mutations are missense mutations in a portion of exon 3 encoding for the GSK-3β phosphorylation consensus region of β-catenin, hindering the ability of GSK-3β to phosphorylate β-catenin [351]. The majority of CTNN1B mutations are very rare with the missense mutation N287S only found heterozygous in 0.6% of a control group [352].

Mutations in the AXIN1 gene include T1942C and G2063A, at 21% and 3% respectively in the general population [353]. There are other rare variants (P312T, R398H, L445M, D545E, G700S and R891Q) which have also been identified and may contribute to CRC risk [352].

The AXIN2 gene has been mapped to 17q23-q24, a region that shows frequent loss of heterozygosity in breast cancer, neuroblastoma and colorectal adenomas [354]. The AXIN2 gene codes for the Axin-related protein Axin2, which interacts with APC, GSK3 β and β-catenin to promotes β-catenin phosphorylation and subsequent proteasomal degradation [355]. However, the mutated Axin2 stabilises β-catenin and activates β-catenin/T-call factor signalling [354]. The deregulation of beta-catenin is an important event in the genesis of a number of malignancies.

Although the majority of human colorectal cancers have high levels of β-catenin or TCF transcription up regulated, due either to inactivation of the APC (tumour suppressor gene) or activating mutations of β-catenin, one group identified a CRC cell line with neither of these more common mutations. Instead they found a truncating

mutation, a 4bp deletion, in the CDX-2 gene [356], which combined with animal models, suggested that the CDX-2 gene may contribute to the tumour suppressor effects of APC. There is a G to A substitution in the Cdx-2 binding site, called 1e-G−1739A (rs11568820), which has been associated, the G allele at 81% and the A at 19%, following Hardy-Weinberg equilibrium.

Other genes related to β-catenin function include presenilin 1 (PS1), a gene normally associated with Alzheimer's disease, which regulates β-catenin stability by assisting its phosphorylation and subsequent degradation [357]. MYH is a mismatch repair gene, which has been implicated in multiple adenomas, including attenuated polyposis, and is a base excision repair gene that corrects oxidative DNA damage. Although not specific to the Wnt pathway, MYH/MUTYH increases the likelihood of G:C>T:A transversions in the gastrointestinal tract. Initially two Missense mutations were identified, Y165C and G382D, [358]. The genes involved in the Wnt signalling pathway leading to attenuated polyposis are shown in table 3.4:

Table 3.4 Genes Involved in the Wnt Signalling Pathway

Action	Genes
Up-regulation by TCF-4	CYCLIN D1, C-MYC, MATRILYSIN, C-JUN, FRA-1
β-catenin	CTNNB1
On Wnt pathway	BTRC, ICAT, AXIN1, AXIN2, TCF4, CDX-2
Up regulated by Wnt	WISP-1, WISP-3, PPP2R1B
β-catenin degradation	PS1
Related function	MYH

Other Syndromes

The remaining familial CRCs are composed of a large numbers of different syndromes involving polymorphisms in the tumour suppressor genes, mismatch repair and stability genes, oncogenes, metabolic and antioxidant genes and genes involved in different biological pathways.

Tumour Suppressor Genes

Tumour suppressor genes are often identified by the allelic imbalance within polyps. Other than APC, the most widely researched tumour suppressor gene in CRC is p53, mutations in the germ line of which lead to the rare, autosomal dominant Li-Fraumeni syndrome (LFS) [359]. This is susceptibility to a wide range of cancers, including breast, bone, brain and leukaemia.

Juvenile Polyposis syndrome (JPS) is an autosomal dominant syndrome characterized by a specific type of polyp in the entire gastrointestinal (GI) tract, particularly the stomach, small intestine, colon, and rectum [360]. The term "juvenile" refers to the nature of the polyp, not the age of the patient. There are two genes associated with JPS: BMPR1A and SMAD4 (also known as MADH4 and DPC4), which together account for approximately 20% of cases of JPS [361]. Mutations in BMPR1A have also been identified in association with Hereditary Mixed Polyposis Syndrome (HMPS), which is another autosomal dominant syndrome that can led to CRC but with a histological variety of polyps [362].

Peutz-Jeghers Syndrome (PJS), also known as Hereditary Intestinal Polyposis Syndrome (HIPS), affects the entire GI tract and is often diagnosed by areas of hyperpigmentation on the hands, feet and mouth [363]. Peutz-Jeghers can be caused by a variety of mutations in the STK11 gene (also known as LKB1), located on chromosome 19p13 which encode the serine/threonine kinase 11 protein [364], with most mutations leading to a truncated protein. However, there maybe other mutations which also lead to PJS [365] which could be due to a pathway relationship between PJS to another inherited tumour syndrome, Tuberous sclerosis complex (TSC) [366] which does not affect the colon or rectum.

There is a tumour suppressor gene PTEN, which removes phosphate groups from tyrosine, serine and threonine. Mutations in PTEN lead to a spectrum of disorders called PTEN hanartoma tumour syndromes (PHTS) which include Bannayan-Riley-Ruvalcaban syndrome (BRRS) and Cowden syndrome.

Mismatch Repair (MMR) Genes

Cancers with MMR gene defects carry tens of thousands of small insertions and deletions in short tandem repeats [367]. Other than Lynch syndrome and attenuated polyposis, variations a mismatch repair gene, BLM, can lead to a syndrome called Bloom, also known as Bloom-Torre-Machacek syndrome. This is an increased susceptibility to a broad spectrum of cancers including CRC.

Oncogenes

There are two main genes associated with familial GI stromal tumour: KIT and PDGFRA. One of the most prominent proto-oncogenes in CRC is a member of the RAS family of genes, K-RAS, with a small number of mutations in N-RAS [339]. Another gene often thought to be an oncogene, but is in fact a transcription factor with tumour suppressor characteristics, is TP53, coding for the p53 protein, germline mutation of which lead to Li-Fraumeni syndrome. There are a large number of mutations and modes of action of the p53 gene, which in its wildtype state recognises DNA damage and induce cell cycle arrest.

Polymorphisms in the Glutathione *S*-Transferase (GST) superfamily

The GSTs are a superfamily of dimeric phase II metabolic enzymes, which protect cellular macromolecules from damage by catalysing the conjugation of electrophilic molecules and products of oxidative stress with reduced glutathione [368]. They are involved in the metabolism of a number of environmental carcinogens, chemotherapeutic agents and endogenously derived reactive oxygen species.

In humans, there are five subfamilies: Alpha, Mu, Pi, Theta and Zeta [369], of which there are two Alpha classes and four Mu classes. There has also been a mitochondrial GST, GST Kappa, identified.

GSTM1 codes for the cytosolic enzyme GST- μ [370], with approximately 50% of Caucasians homozygous for a deletion in GSTM1 [371]. The GSTM1 null allele leads to an absence of GST- μ enzyme activity [372]. There has been conflicting results in studies into the risk associations with GSTM1 and colorectal cancer. Some studies found an association with increased colorectal cancer risk [373, 374] or no association [375-377], or differing association depending on the site of the tumour [378]. There is also

hypotheses that GSTM1 may interact with smoking or meat intake [379], where further studies would be useful.

There are two alleles of the GSTM3 gene: GSTM3*A and GSTM3*B. The GSTM3*B variant has a 6bp deletion in intron 6, which creates a recognition sequence for a Yin Yang 1 (YY1) transcription factor [380]. GSTM3 is linked to GSTM1, and it has been proposed that GSTM1*A confers an increase in susceptibility due to an association with GSTM3*B [381].

The Pi subclass, GSTP1, expressed in epithelial tissues has been found to be highly over expressed in colon cancer [382]. The GSTP1 gene is polymorphic, with nucleotide 313 of exon 5 changing from adenine to guanine resulting in an amino acid change at position 104 of the GSTP1 protein [383]. This change substantially diminishes CSTP1 enzyme activity [384]. The polymorphisms of GSTP1 include wild-type (AA) in 42 – 69% of the population, heterozygous (AB) in 35% and homozygous for the variant (BB) in approximately 105 [370].

The GSTT1 null allele, which abolishes GST θ enzyme activity [372] and has been associated with ulcerative colitis, has been found associated with increase in risk of colorectal cancer [375]. This supports the hypothesis that the GST T enzyme is involved in the detoxification of unidentified xenobiotics in the large bowel. This frequency of this allele in Europe is between 10-21% [379]. A meta-analysis also found that GSTT1 conferred an increased risk of colorectal cancer in Caucasians [385], and that the combination of the highest risk GSTM1and GSTT1 genes has an odds ration of more than 2 compared to the null genotypes. A different, more recent meta-analysis found both GSTM1 and GSTT1 null alleles to confer an increase in risk in Caucasian populations, but not in Chinese [386]. Another study found that the GSTT1 null genotype association with colorectal cancer was stronger in current smokers, suggesting a possible interaction [370], a later meta analysis found a similar, though still insignificant interaction with smoking [387].

Polymorphisms in the Matrix Metalloprotineases (MMPs)

Matrix Metalloprotease (MMP) are a group of zinc-dependent endopeptidases, which degrade products in the extracellular matrix (ECM). Such degradation in the tissue

surrounding a tumour is a critical process in cancer development and metastasis [388]. A number of MMPs have been associated with colorectal cancer, although there is speculation over whether higher levels of MMPs precede or follow the onset of cancers. The majority of the work has focused on the biological role of MMPs and their levels of gene expression in cancerous versus benign tissue.

The MMP family is comprised of a number of enzymes that share similar active domains and are numbered according to their order of discovery. An increased expression of MMPs 1, 2, 3, 7, 9, and 13 have been associated, in separate studies, with an increase in malignancy, microsatellite instability and a decrease in prognosis [389].

Other Metabolising Genes

Cytochrome P-450 CYP1A1 is expressed in the large bowel [390] and is involved in the metabolism of oestrogen and polycyclic aromatic hydrocarbons (PAHs), found in tobacco smoke and consumption of high levels of certain meats [268, 391]. There are three common polymorphisms and a fourth that has been detected in an African American cohort. The CYPAP1A1*2A (m1) and CYPAP1A1*2C (m2) are associated with higher levels of biomarkers for PAH exposure [392] but are not conclusively linked to an increase or decrease in risk. The CYPAP1A1*4 (m4) variant has only been studied once [393], when it was found to be associated with decreased risk of CRC, but this has not been verified in further work.

The alcohol dehydrogenase (ALDH2) gene has been associated with CRC in Japan, especially in heavy drinkers [394]. Microsomal Epoxide Hydrolase (mEH) is a protein which mediates the transport of bile acids and plays a central role in the metabolism of PAHs [395]. In smokers and those who regularly eat well done meat, there was an association between mEH polymorphism and CRC.

Methylenetetrahydrofolate reductase (MTHFR) is a vitamin B12 dependent enzyme that catalyses the conversion of 5, 10 methylenetetrahydrofolate, involved in purine and thymidine synthesis, to 5-methylenetetrahydrofolate, for methionine synthesis. This process also creates SAM, the universal methyl donor in humans involved in DNA methylation [396]. People with an identified polymorphism, MTHFR C677T

(rs1801133), which causes an amino acid change from an alanine to a valine, have ~30% of the enzyme activity of the CC wildtype, heterozygotes (CT) have ~65% [397]. A mutation in MTHFR could therefore influence both DNA synthesis and DNA methylation. This polymorphism has been found to reduce enzyme activity and has been investigated or an associated to CRC risk, with conflicting independent results [398-400]. There have also been studies that have identified differences in risk for dietary variables in a sample stratified for MTHFR genotype, suggesting a possible interaction role [401, 402]. There is another polymorphism A1298C, which also decreases enzyme activity. Three other polymorphisms have been identified: T1059C, T1317C and G1793A, though not studied in great detail [403-406].

Compounds phosphorylated by the phosphatidylinositol 3-kinases (PI3Ks) activate a wide number of downstream targets. The mechanisms of the PI3Ks may go some way to explaining the role of the tumour suppressor gene PTEN in tumour development. Polymorphisms in these genes, particularly PIK3R1, have been associated with colorectal cancer [407, 408].

N-Acetyltransferase Polymorphisms

N-Acetyltransferase (NAT) has two isozymes, NAT1 and NAT2, which are responsible for activating and deactivating aromatic and heterocyclic amine carcinogens. There are two polymorphic genes that code for NAT activity, NAT1 and NAT2, and one pseudogene, NATP [409]. An early observation, whilst using isoniazid to treat TB, was that the human population is divided into slow, intermediate and rapid acetylator phenotypes. There have been a number of conflicting studies regarding a possible association between the rapid NAT2 polymorphism and colorectal cancer [409, 410]. It is biologically plausible that rapid isoforms of NAT1 and NAT2 more readily activate certain carcinogens, thus putting the colon at more risk [411]. However, studies investigating the relationship between the NAT phenotypes and smoking or meat consumption have found indications of a possible interaction effect [412].

Anti- Inflammatory Pathway Genes

The idea that NSAIDs could prevent colorectal cancer is based on the relationship between chronic inflammation of the bowel and the associated increase in risk [249], the

assumption being that the reverse is true. Anti-inflammatories have also been shown to reduce adenomas in people with FAP [413].

The most widely recognised mode of action for aspirin and other NSAIDs is targeting the cyclooxygenase (COX) family which are important enzymes in the synthesis of prostaglandins, which have a role in inflammation. NSAIDs are all competitive inhibitors of COX, with aspirin covalently modifying the protein irreversibly [414]. There are two types of cyclooxygenase: COX1 and COX2. Aspirin inhibits COX1 and modifies COX2, which normally produces the pro-inflammatory metabolite prostaglandin E2 (PGE2), to produce lipoxins, which are anti inflammatory.

The mechanism by which inflammatory molecules influence CRC risk is still not fully understood. It has been suggested that the prostaglandins target the Wnt pathway, also the target of the mutated APC gene. In the Wnt pathway, NSAIDs act in a similar way to the action of wildtype APC, by reducing the cyclooxygenases, which in turn reduce the prostaglandins. In the presence of inflammation, the prostaglandins bind with a compound, activating a G-protein coupled receptor (Gαs) which binds to AXIN, having the same effects on β-catenin as the Wnt pathway [249].

It has also been suggested that COX2 inhibitors work by preventing the COX2 stimulation of tumour cell growth and formation of new blood vessels (neoangiogenesis) [415]. Polymorphisms of the COX family have been investigated, with an amino acid change Pro17Leu in COX1 and G(-765)C in COX2 both being associated with colorectal cancer risk. Polymorphisms in hepatocyte growth factor (HGF) and the ERK signalling pathway have also been associated with cox-2 activation [416].

Other modes of action include the up-regulation of the prostate apoptosis response 4 (Par-4) gene following NSAID treatment [417]. NSAIDs also antagonise NF-κB, a transcription factor found to be elevated in many cancers and with a role in inflammation [418]. There are certain metabolic enzyme genotypes that can modulate the chemopreventative effects of NSAIDs by impairing the metabolism, especially UGT1A6, UGT1A7 and CYP2C9 [419-421]. Other pro-inflammatory genetic polymorphisms associated with colorectal cancer risk include IL-10-592, an anti-

inflammatory cytokine with a role in modulating gastrointestinal tract inflammation [422].

Digestive Pathway Genes

For dietary variables associated with colorectal cancer, the different genes and resultant proteins involved in their metabolism may provide clues to different levels of susceptibility.

Vitamin D

Different polymorphisms in the vitamin D receptor (VDR) gene have been associated with a number of cancers and chronic diseases [423]. The CDX2 gene that may play a role in the action of APC, is normally related to the Vitamin D pathway.

Folate

Other than MTHFR, other genes involved in folate metabolism include methionine synthase reductase (MTRR A66G), cystathionine β-synthase (CBS exon 8, 68 base pair insertion) and thymidylate synthase (TS enhancer region and 3' untranslated region). Overexpression of TS is linked to resistance to TS-targeted chemotherapy drugs. Tandem repeats located in the TS enhancer region (TSER) gene have been shown to influence TS expression [400].

Low Penetrance Susceptibility Genes

There are a few genes with low penetrance that have been associated with colorectal cancer, these are: CASP8 [424], PUO5F1P1 , GREM1, SCG5,
8q24 [425-427].

Gene Environment Interactions

A number of different genes have been either found or hypothesised to interact with NSAIDs. For example, there is evidence of a possible interaction between aspirin and p53 mutations. A study comparing different compounds that work on the same pathway found that aspirin based G1 arrest and apoptosis (a mechanism that restrains cell proliferation) activated p53 and p21 in an ataxia-telangiectasia-mutated (ATM) kinase dependent way [428]. There is also evidence of an interaction between a variant

of the IL-10 gene and aspirin use [422] with carriers of the IL-10-592 allele being more likely to derive benefit from the anti-inflammatory actions of aspirin due to a lower production of their own anti-inflammatory interleukin-10.

An interaction between GST enzymes the toxicity of chemotherapeutic drugs has also been found [376]. GSTs are involved in the metabolism of polycyclic aromatic hydrocarbons (PAHs) for which there is inconsistent evidence for an association with CRC [370]. One study found no independent effects of GSTP1, M1 or T1 polymorphisms but did find an increased risk when interacting with the NAT2 slow genotype [429]. Significant interaction effects have been found between all three variants of the CYP1A1 gene and meat intake [392], between two of the variants and leafy vegetable consumption, but not with smoking.

There is no direct correlation between mEH and colorectal cancer, but there is a relationship within certain subgroups: smokers; meat consumption, especially those who consume well done meat; and certain GST genotypes [395]. Similarly, those with slow NAT polymorphisms are more likely to have an increase in risk as a result of smoke exposure and frequent consumption of red meat [412].

The association of the MTHFR gene and colorectal cancer has been proposed to be dependent on a number of different environmental variables. The protective effect of the 667T variant has been found to be enhanced by high levels of folate intake [399, 400, 430] an effect negated by high levels of alcohol [431]. The A2756G mutation has also been found to have a protective effect enhanced in people with a low alcohol intake [432]. It has been proposed that these effects could also be the result of other B vitamins on the same metabolic pathway. One study found that vitamin B2 was inversely associated with colorectal neoplasms, an effect more pronounced in those with the MTHFR TT genotype [402]. Low intake of folate, vitamin B12 and vitamin B6 have also been found to increase risk in those of the TT genotype [401].

Interactions for other genes on the folate pathway, other than the obvious potential interaction with folate, include methionine, alcohol consumption and vitamin B12 [400].

3.7 Conclusions

Colorectal cancer is a good example of a disease with both genetic and environmental risk factors, which potentially interact in a lot of ways. Combining the development of a novel method with the data from a colorectal cancer study could be an important first step for the field of interaction statistics and public health genetics.

Chapter 4

Introduction to the Data Used in Both the Simulation and Real Analysis

4.1 Introduction

Methodological development and evaluation were carried out using simulated data, before applying the new method to a real dataset consisting of colorectal cancer cases and controls. Artificial data allows us to dictate the outcomes and explore different complex underlying interaction models. In this instance the parameters are based on those found in the real dataset, to minimise the complications that features such as distribution may introduce in the later analysis. Therefore it is essential to describe and characterise the data from the colorectal cancer dataset before beginning the simulations.

4.2 Description of SOCCS data

The study of colorectal cancer in Scotland (SOCCS) is a large, prospective, epidemiological case control study in which large volumes of both dietary and genetic data have been collected for analysis. The methodology is explained in more detail as the Edinburgh arm of the COGENT study [433], including the supplementary tables.

4.2.1 Study Population

Cases were recruited by research staff based at clinical research centres across Scotland. Ascertainment bias was minimised by ensuring that recruitment occurred soon after diagnosis. Recruitment typically took place within 2-3 months of the initial diagnosis of adenocarcinoma of the large bowel.

The controls were identified at random within calendar, age, sex and area of residence restrictions from a population based register of people registered with a general practitioner in Scotland (the Community Health Index) and invited to participate. They were recruited in the same time period as the cases and matched, one control per case, for age (± 1 year), gender and area of residence. The matching was carried out to minimise the effects of gender, age and socioeconomic status and prioritise the identification of genetic differences. This study was designed with different aims to the ones in this thesis but the data is suitable for the analysis, however it makes it impossible to study the effects of age, gender or area of residence in terms of gene-environment interaction. The exclusion criteria included: death before ascertainment; being too ill to participate; if the cancer was a recurrence of colorectal cancer as opposed to the first presentation or if the patient was unable to give informed consent.

4.2.2 Preliminary Analysis of Data

The environmental exposures of interest include basic characteristics, such as BMI, dietary factors and other lifestyle measures, such as exercise and NSAID intake. The

case/control status of the simulated sample will be determined by a variety of different risk algorithms, testing different levels of confounding and attributable risk. Therefore an individual within the artificial dataset may be a case under a certain simulation and a control under another. This is necessary to ensure that the algorithm can detect gene-environment interactions in a range of plausible circumstances and the simulated sample is not biased towards what would be expected following basic analysis or reviewing the literature. Therefore it is more important to differentiate by biological factors, such as gender, than to identify differences between the cases and controls using basic, crude methods. Univariate analysis on the environmental variables and case outcome is carried out in detail in Chapter 8.

4.2.2.1 Basic Characteristics

Table 4.1 shows the distribution of age and anthropometric measurements by sex of the SOCCS population, with one outlier removed, which are then explained in more length individually.

Table 4.1 Description of Basic SOCCS characteristics

	Age (years)		Height (m)		Weight (kg)		BMI (kg/m^2)	
	M (n=2761)	F (n=2074)	M (n=2754)	F (n=2057)	M (n=2745)	F (n=2047)	M (n=2741)	F (n=2040)
Min	21	22	1.39	1.32	42.2	33.1	15.3	12.5
1st Quart	56	53	1.70	1.57	73.0	60.2	24.1	23.0
Median	65	62	1.75	1.61	81.6	66.7	26.4	25.6
Mean (sd)	63.1 (10.2)	61.2 (11.1)	1.75 (0.07)	1.62 (0.06)	82.6 (13.6)	69.0 (13.6)	26.9 (4.1)	26.4 (5.1)
3rd Quart	71	70	1.80	1.65	90.0	76.2	29.1	28.8
Max	83	101	2.13	2.06*	167.0	142.9	53.8	55.9

***after removal of outlier**

Age

The age of the study population ranges from 21 to 101 years and is approximately normally distributed with a mean of 62.3 years. Males and females did not differ significantly in age distribution. This is not what would be expected from a random sample or all recorded cases, where more cases would be found with increasing age. However, the SOCCS study wanted to focus on the genetic risk factors with the heritable factors being more associated with CRC at a younger age and therefore selected the study population accordingly.

Height, Weight and BMI

Preliminary analysis of height found one woman to be 2.565m (more than 8ft 4 inches tall), with a very low BMI (automatically calculated from the height and weight measures). These are unlikely values and suggest an error, either in measurement or recording of the data. Therefore the height variable for this participant has been recoded as a missing variable, along with BMI.

Once this variable was recoded as missing, the distribution of height for the whole population was as would be expected from a combined male and female population, shown in Figure 4.1. Approximately normally distributed but with two peaks (*)

representing the peaks of two normal distributions. When divided by gender, the distributions were approximately normal, shown in Figure 4.2.

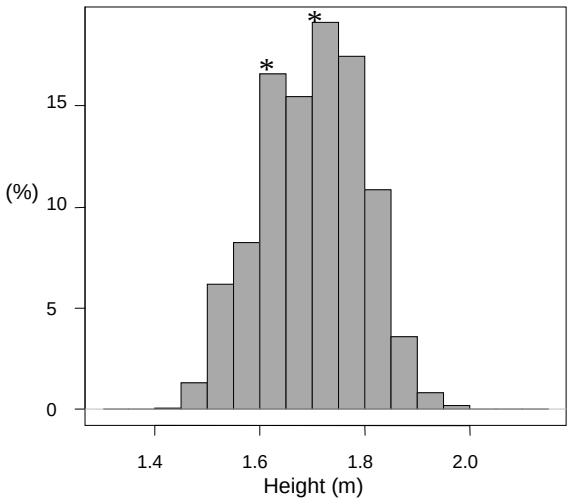


Figure 4.1 Histogram of Height for the Whole SOCCS Sample

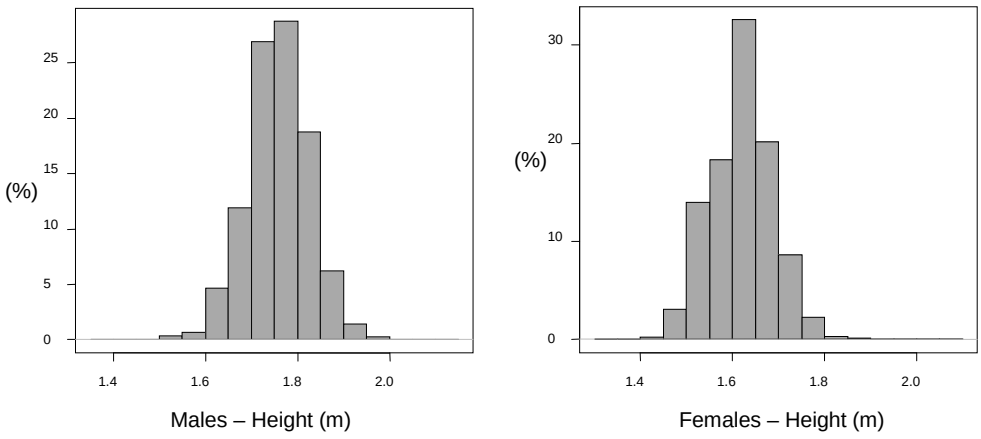


Figure 4.2 Height Separated by Gender

Separating the height by gender, results in two normally distributed populations with an average height of 1.75 and 1.62 metres, for males and females, respectively. Another area where there are differences dependent on gender is weight, with males tending to weigh more than females. However BMI is approximately the same for males as for females, with the females having a slightly larger range.

4.2.2.2 Dietary Variables

The dietary variables are all continuous measures calculated from a food frequency questionnaire in which the participants reported which foodstuffs they consumed and the approximate level of consumption. These data were then analysed by nutritionists to quantify the weights of nutrients, fibre and alcohol being consumed daily. The cases and controls were asked to complete the questionnaire retrospectively for the year prior to diagnosis or recruitment, respectively.

In describing the dietary variables, preliminary analysis showed the consumption of all the variables to be severely skewed. Although this is in some way expected, the degree of skew-ness required further investigation. There was a single participant, a control, for whom the consumption of every single dietary variable was the maximum amount registered, usually 4 or 5 times the next highest value. For example Retinol consumption had a median of 511µg and the second highest consumption of 12,465µg with this participant consuming 39,681µg. Other figures were similarly unrealistic, especially given that the participant had a BMI of 32 kg/m2, which although high is inconsistent with the recording of low levels of exercise and consumption of ten times the average number of calories than the rest of the sample. Therefore this person's data have been removed from the dataset.

Sources of vitamins and minerals in food and the recommended daily allowed (RDA), as considered in 2009, were found on the Food Standards Agency website:
http://www.eatwell.gov.uk/healthydiet/nutritionessentials/vitaminsandminerals

Recommendations for minimum recommended amount were found on the British Nutrition Foundation webpage:
http://www.nutrition.org.uk/home.asp?siteId=43&sectionId=s

Table 4.2 presents the preliminary analysis of the dietary habits of the SOCCS population as a whole. As case control status was allocated to people within a simulated population based on controlled and varied risk factors, it was not important at this point to divide the parameters by case-control status. In fact, it could have biased the method towards finding specific results from this particular population that

could be observed by basic analysis of separate risk factors without accounting for multiple testing. The solubility explains whether the nutrient is fat soluble and therefore does not need to be eaten every day to meet recommended requirements, or water soluble and not stored by the body for future use. The RDA is not applicable for nutrients that have sources other than dietary ones or for nutrients that have such a range of sources that it is unlikely someone could become deficient. RDA measures may be different for children or pregnant women, neither of which was included in the SOCCS study.

Table 4.2 Dietary variables from SOCCS Questionnaire

Variable	Min	1st Quart	Median	Mean	3rd Quart	Max	Solubility	RDA (adults, unless stated)
Vitamin D (µg)	0.08	2.55	3.90	4.82	5.82	71.13	Fat	NA
Retinol (µg)	22.0	319.0	511.0	697.6	814.0	12465.0	Fat	700 (M) 600 (F)
Calcium (mg)	124	840	1089	1158	1391	5092	NA	700
Thiamine (mg)	0.20	1.60	2.02	2.21	2.55	19.50	Water	1.0 (M) 0.8 (F)
Riboflavin (mg)	0.30	1.64	2.08	2.22	2.64	10.79	Water	1.3 (M) 1.1 (F)
Niacin (mg)	3.9	18.6	23.6	25.4	30.1	142.3	Water	17 (M) 13 (F)
Pantothenic Acid (mg)	1.11	5.26	6.67	8.23	8.75	68.99	Water	NA
Fibre (g)	2.9	15.9	20.7	22.4	26.9	98.8	NA	12-14 (FSA) 20-35 (FDA)*
Alcohol (g)	0.00	1.70	8.10	13.16	19.20	161.50	NA	NA
Vitamin B6 (mg)	0.39	2.20	2.79	2.98	3.50	15.28	Water	1.4 (M) 1.2 (F)
Biotin (µg)	8.0	38.9	48.7	51.7	60.2	275.6	Water	10 - 200
Vitamin B12 (µg)	0.30	4.90	6.90	8.19	9.90	80.00	Water	1.5
Folic Acid (µg)	45.0	256.0	322.0	343.2	400.0	1815.0	Water	200

*Different advisory bodies have different recommendations for fibre intake

As can be seen from the table, there is an extremely large range in consumption for most of the dietary variables, with at least a four-fold increase, rising to fifteen fold for retinol, between the 75th percentile and the maximum value. These maximum values did not all correspond to the same individual and are therefore less likely than the earlier high value to be a mistake in classficiation.

The same basic analysis was carried out on all the dietary variables, all of which were positively skewed but normal following transformation, except alcohol consumption which was re-classified as a binary variable. A summary of the importance or associated risks for each variable is given below, with Vitamin D illustrating the distribution.

Vitamin D

Vitamin D has an important role in maintenance of organ systems, regulating calcium and phosphorus in the blood and calcium absorption into the kidneys, which in turn enables normal mineralization of the skeleton. It also has a role in the immune system, promoting phagocytosis, anti-tumour activity and immuno-modulatory functions [434].

Vitamin D can be created by our skin on exposure to UVB sunlight. Vitamin D is found in oily fish and eggs. Other food sources include fortified foods such as margarine, breakfast cereals and powdered milk.

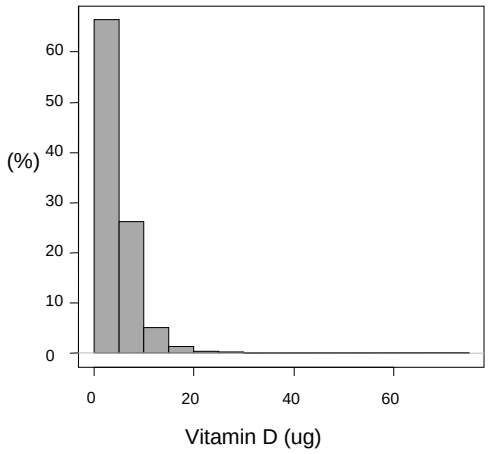


Figure 4.3 Distribution of Vitamin D

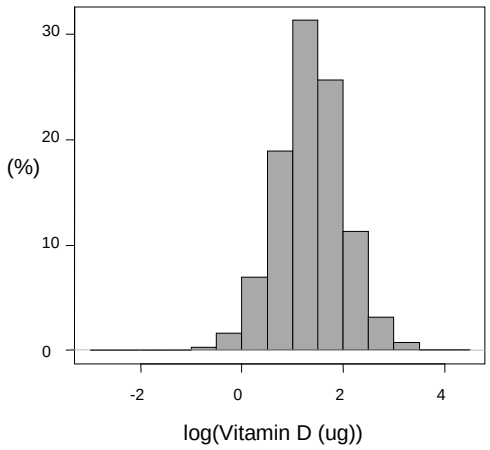


Figure 4.4 log Transformation of Vitamin D Distribution

In this population, the consumption of vitamin D, shown in figure 4.3, is very positively skewed with the mean (4.8) larger than the median (3.9) and a maximum consumption of 71.13μg. The transformation to the log(vitamin D consumption), however, is normally distributed, as shown in figure 4.4. Therefore, when generating artificial data that represents vitamin D consumption, the lognormal distribution was used.

Retinol

Retinol is also known as vitamin A, a fat soluble vitamin important for bone growth and vision, especially in low levels of light. It helps maintain healthy skin and mucosal layers and strengthens the immune system. Vitamin A deficiency in childhood can lead to xerophthalmia, a leading cause of childhood blindness in the third world (Humphrey 1992). Those suffering from Vitamin A deficiency tend to grow more slowly and suffer from more persistent or severe infections than those with normal vitamin A levels. It is also dangerous to take large doses of vitamin A supplement on top of a healthy balanced diet, as Vitamin A toxicity can lead to liver damage, a risk that increases with age [435].

Good sources of Retinol include cheese, eggs, oily fish, dairy products, fortified margarine and yoghurt. One of the richest sources of retinol is liver. In the study population retinol consumption is more drastically skewed than vitamin D, with a mean of 511.0 μg and a median of 697.6 μg. The consumed volumes are slightly below the RDA but not worryingly so.

Calcium

Calcium is essential for the normal growth and maintenance of bones and teeth, with long term deficiency leading to rickets [436] poor blood clotting, osteoporosis [437] and periodontal disease [438]. Calcium needs to be eaten regularly as it is water soluble and not well stored by the body.

Food sources for calcium include milk, cheese and other dairy foods, green leafy vegetables (except spinach), soya beans, tofu, and fish in which the bones are eaten, such as sardines and pilchards. There are also many fortified foods with added calcium including soya milk and soya produce, breakfast cereals and bread and other

baked goods made from fortified flour. Calcium in the SOCCS study population is not as skewed as many of the other variables, probably in part due to the large variety of possible sources, but still benefitted from transformation. Only 13.1% of the population eats less than the recommended 700mg a day.

Thiamine

Thiamine is also known as vitamin B1 and works with other B vitamins in metabolism and maintenance of nerves and muscle tissue. A deficiency of thiamine can result in a pandemic human deficiency disease called beriberi [439]. It is also a deficiency commonly seen in alcoholics, as alcohol and acetaldehyde have direct toxic effects on thiamine-related enzymes in the liver [440] .

Thiamine is found in most food types. Good sources of thiamine include pork, dairy products, vegetables, fruit and wholegrains. Some breakfast cereals are fortified with thiamine. The SOCCS population has a mean Thiamine consumption of 2.214 and a median of 2.02. Both the mean and median, and in fact the first quartile measure, are considerably more than the minimum RDA, which is 1mg for men and 0.8mg for women.

Riboflavin

Riboflavin is also known as vitamin B2, it is a micronutrient and is involved in cellular and metabolic processes including the metabolism of carbohydrates, fats, proteins and ketone bodies. There is evidence that riboflavin has a role in iron handling and a deficiency contributes to the aetiology of anaemia when iron intakes are low [441]. Biochemical signs of deficiency occur after only a few days of dietary deprivation.

Riboflavin is found in milk, eggs, fortified breakfast cereals, rice and mushrooms. Exposure to daylight can degrade riboflavin, so these items should be stored away from direct sunlight. However, the mean (2.217) and the median (2.08) are both higher than the RDA for riboflavin: 1.3mg for men and 1.1mg for women.

Niacin

Niacin is also known as vitamin B3 and is an essential human nutrient. It plays a precursor role in metabolism, DNA repair and production of steroid hormones by the adrenal gland. A prolonged deficiency of Niacin (or its precursor tryptophan) causes pellagra, a disease that is still endemic in parts of the world, including Africa and Mexico and within more vulnerable people in affluent societies (homeless people, alcoholics, those with eating disorders) [442].

Good dietary sources of niacin include beef, pork, chicken, wheat flour, maize flour, eggs and milk. The majority of the SOCCS population are consuming the minimum recommended daily amount of niacin (13mg for women, 17mg for men), with some taking considerably more.

Pantothenic Acid

Pantothenic Acid is also known as vitamin B5 and is necessary in the formation of coenzyme-A (Co-A). It is also an important metabolic vitamin critical in the metabolism and synthesis of carbohydrates, fats and proteins. A deficiency of pantothenic acid leads to general clinical malaise [443].

There is a large variety of sources of Pantothenic acid. Good sources include chicken, beef, potatoes, porridge, tomatoes, kidney, eggs, broccoli and whole grains such as brown rice and wholemeal bread. It is also found in many fortified breakfast cereals. More than 60% of the study population consumed between 5 and 10 mg of pantothenic acid per day. There are so many sources of pantothenic acid that there is no RDA.

Vitamin B6

Vitamin B6 is also known as pyridoxine and is a water soluble vitamin. It is a co-factor in amino acid metabolism and in the release of glucose from glycogen. It is important for maintaining blood sugar levels and for the formation of haemoglobin. A deficiency of vitamin B6 (along with low intake of folate) can lead to hyperhomocysteinemia and an increased risk of heart disease [444].

It is found in most food groups, especially meat, cereals and fortified breakfast cereals. With an RDA of 1.4mg for men and 1.2mg for women, the majority of people reported consuming at least this amount of vitamin B6.

Biotin

Biotin, also known as vitamin B7 or vitamin H, plays an important role in the citric acid cycle, the generation of biochemical energy during aerobic respiration. It is also important for cell growth, production of fatty acids and the metabolism of both fats and amino acids. Biotin helps maintain blood sugar levels and strengthens hair and nails [445].

Good dietary sources of biotin include meat, eggs and dried mixed fruits. However, deficiency is extremely rare, as intestinal bacteria generally produce an excess of the body's daily requirement. Therefore, some health boards do not even recommend a minimum daily amount. This was one of the least skewed of the dietary variables, with a mean of 51.7µg and median of 48.7µg, but still was more normally distributed after being transformed.

Vitamin B12

Vitamin B12 is important for the function of the brain and the central nervous system. It plays an important role in cell metabolism, especially in regards to DNA synthesis and regulation, but also in the metabolism of fatty acids and energy production. It is necessary to process folic acid. A deficiency of vitamin B12 can lead to anaemia and has been implicated in a spectrum of neuropsychiatric disorders [446].

Vitamin B_{12} is mainly found in meat products and certain algae such as seaweed. Good sources include meat, salmon, cod, milk, cheese, eggs, yeast extract, and some fortified breakfast cereals. The reported consumption of vitamin B12 varied from 0.3µg to 80.0µg, with a mean consumption of 8.2µg and a median of 6.9µg, in a positively skewed distribution. Despite being water soluble, vitamin B12 is stored by the body, in the liver and it takes years for a deficiency to develop. The RDA is only 1.5 µg a day, with almost the entire population reporting eating this much.

Folic Acid

Folic acid is also known as vitamin M and folacin. Folic acid and folate are the two forms of vitamin B9, a water soluble vitamin. Folate is necessary for producing and maintaining new cells as it is needed for synthesising the DNA bases thymine and purine. This is especially important when the body is undergoing a period of rapid cell division and growth, such as pregnancy [447]. A deficiency in folic acid during foetal development can cause neural tube defects and spina bifida in newborn babies [448].

Folate is found in small amounts in a large number of foods. Good sources include broccoli, Brussels sprouts, asparagus, peas, chickpeas, brown rice and fortified breakfast cereals. The distribution of the consumption of folic acid in the SOCCS population is only slightly positively skewed. The mean was 343.2µg and the median was 322.0µg. Both mean and median are above the RDA of 200µg a day.

Fibre

Fibre is not a vitamin or mineral, rather a class of materials that are continuous filaments or discrete extended segments. Dietary fibres, also commonly known as "roughage," are important for digestion, as the indigestible part of plant foods that move through the digestive system, absorbing water and making defecation easier. There are two types of dietary fibre, water soluble and insoluble. The soluble fibre undergoes a metabolic reaction, fermentation, producing gas and short chain fatty acids which themselves have considerable health benefits. Insoluble fibre attracts water and increases bulk and softness of the stool, making it easier to pass.

Sources of soluble fibre include oats, rye, beans, fruits and fruit juices, certain vegetables (broccoli, carrots), root vegetables and a seed husk called psyllium. Sources of insoluble fibre include wholegrains, bran, nuts and seeds, vegetables and the skins of some fruits, including tomatoes.

The British Nutrition Foundation recommends a minimum of 12-14g of fibre a day for a healthy adult, whereas the American Dietetic Association recommends a minimum of 20-35g a day for a healthy adult, depending on calorie intake. The median and mean reported fibre intakes of the SOCCS population were, at 20.7 and 22.4 respectively, both above the minimum requirement from both advisory bodies.

Total Energy Consumption

Energy requirements and consumption vary by gender, the following figures (4.5-4.8) show the female and the male energy consumption and the log transformation of the consumption respectively:

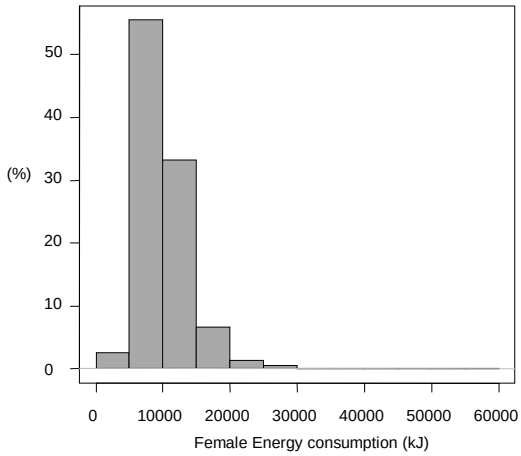


Figure 4.5 Male Energy Consumption

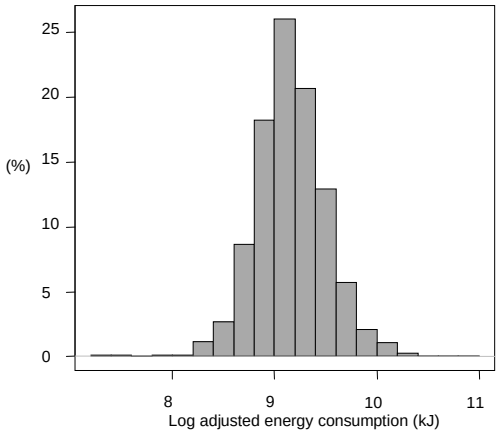


Figure 4.6 Male Energy Transformed

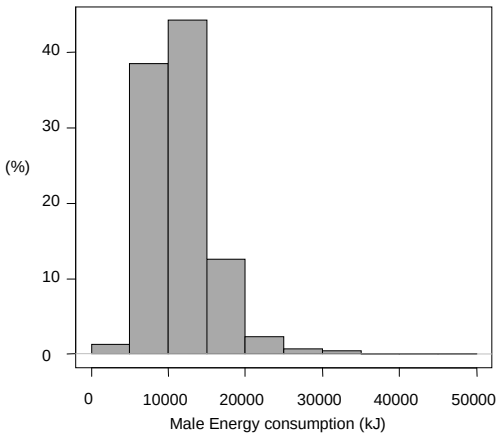


Figure 4.7 Female Energy Consumption

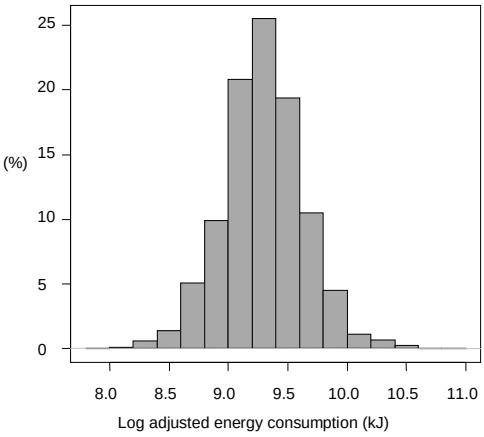


Figure 4.8 Female Energy Transformed

The distribution for both males and females is initially skewed, changing to normally distributed following log transformation.

Energy requirements vary from one individual to the next, depending on factors such as age, sex, body composition and physical activity level. It is therefore difficult to generalise about the study population. Using the mean BMI, median activity level and

age for each gender independently, table 4.3 shows the recommended and the real median levels of energy consumption, one calorie being equivalent to 4.1868kJ.

Table 4.3 SOCCS Energy Intake

	Recommended kcal	Recommended kJ	Real kcal	Real kJ (median/mean)
Males	1931	8085	2575	10780
Females	1747	7314	2241	9384

As the results show, the energy consumption is higher than would be recommended, although to assess overeating would require the individual recommended intake to be calculated and compared to the real intake.

Alcohol

Although technically, the chemical description of alcohol includes a number of compounds, in dietary terms alcohol is meant to mean ethanol. Unlike the other dietary variables it is the presence or increase of alcohol in the diet that is likely to be a risk factor as oppose to its absence. Alcohol is found in alcoholic beverages, of which there are three main categories – beers, wines and spirits. There are also trace amounts of alcohol in some recipes and some chocolate liqueurs.

One limitation of using a food frequency questionnaire for the year preceding diagnosis/recruitment is that it cannot differentiate between lifetime non drinkers and those abstaining due to past problems with alcohol: this is a common problem in the epidemiological studies of alcohol consumption. With the majority of the SOCCS population being retired, it is unlikely that their alcohol consumption in the questionnaire represents their lifetime habits. These limitations need to be taken into account during any analysis of alcohol (or any other dietary factor) consumption.

The distribution of alcohol consumption is different from all the other dietary variables. However, the data can be divided into groups based on frequency of alcohol consumption based on the risk categories and divisions in the Health Survey for Scotland 2003 (http://www.scotland.gov.uk/Resource/Doc/924/0019811.pdf). This

divides the group by those that drink less than three drinks a week compared to those who drink three or more drinks a week, as shown in table 4.4.

Table 4.4 Binary Alcohol Consumption by Gender

	Males (%)		**Females (%)**	
Number of alcoholic drinks per week	3 or more	Less than 3	3 or more	Less than 3
HSE 2003	42	58	26	74
SOCCS	46.5	53.5	31	69

4.2.2.3 Categorical Lifestyle Measures

There were a number of categorical environmental risk factors measured, a summary of which is shown below in table 4.5. The data was gathered using a questionnaire and therefore uses the patient's own definition of their physical activity level or smoking status. However, deprivation was assigned using postcode data.

Table 4.5 Description of Categorical Lifestyle Measures

Variable	**Category**	**Percentage of sample** (n)
NSAIDs	No Drugs	66.6 (3206)
	Mini Aspirin	16.9 (814)
	Other NSAID	12.1 (584)
	Normal Aspirin	3.2 (153)
	Mini Aspirin and other NSAID	0.7 (32)
	Normal Aspirin and other NSAID	0.5 (22)
Physical Activity (hours cycling or other sport)	0	55.9 (2595)
	1 - 3.5	25.6 (1189)
	3.5 - 7	11.7 (544)
	7+	6.8 (318)
Smoking	Non	43.3 (2074)
	Regular	38.6 (1867)
	Ex-regular	17.6 (852)
	Ex-occasional	0.9 (44)
Family History Risk	No	90.1 (4305)
	Yes	9.9 (471)
Deprivation	1	10.2 (452)
	2	20.2 (1002)
	3	24.6 (1290)
	4	25.5 (1133)
	5	10.6 (511)
	6	7.6 (318)
	7	2.3 (130)

4.2.2.4 Genetic Variables

The genetic variables in SOCCS are single nucleotide polymorphisms (SNPs), on the microarray Hap 550, which checks 550,000 human SNPs. The exact panel selected for this study were identified by reviewing the literature for genes with an association with colorectal cancer, found in Chapter 3.

The SNPs used in the simulations were selected using "Tag SNP Picker," this is an algorithm developed to select the best SNPs for a specified area, taking into account linkage disequilibrium with nearby variants to maximise efficiency and power [449].

This sample is just meant to represent a SNP panel of similar size to the one used for the real analysis. As the real analysis will inevitably happen later than the simulated analysis and benefit from any knowledge on limitations and strengths gained from it, knowledge of specific SNPs identified previously may bias the methodology towards these SNPs in particular. This is particularly problematic as the literature identifying specific SNPs contains examples from the same dataset as this study uses. It is therefore more appropriate to keep the panel more general and representative, not structured towards specific, expected SNPs.

4.3 Artificial Data

4.3.1 Generating the Data

A programme simulating all the data was constructed using R, the code is contained in Appendix 4.1. It was necessary to set a seed for the random number generation to ensure that the same data would be generated each time, if the entered parameters were the same, so the data analysed for chance effects were identical to that being analysed by the contrasting methods. Prior to variable simulation, the sample size, any risk effects and the number of populations being generated was entered. The populations were simulated sequentially, so the first was complete and saved before the simulation began on the second. The environmental and SNP data was generated separately before being combined into a data frame.

4.3.1.1 Environmental Variables

The first variables generated were those which differ between the sexes and gender, for example in generating the height for a male the following code was used:

```
height<-round(rnorm(SampSize/2, mean= 1.752, sd= 0.069),digits=2)
```

The normal distribution is used in this case as the male height was normally distributed; those variables that were normal following transformation were selected from the log normal distribution. The mean and standard deviation were also taken from the preliminary analysis of the SOCCS data. After initial analysis of the generated data, it became clear that very occasionally some figures were outside those realistically possible. Therefore a function, called “restrict” was developed for use in such cases. In such a case, the value was recoded as the maximum or minimum realistic value, for unrealistically small or large values respectively.

Any figures from the "Health Survey for Scotland" were given as a population mean and a standard error of the mean (SEM), which was translated to a standard deviation for the purpose of generating data.

4.3.1.2 Generating Genetic Variables

Following the systematic review of the literature to identify candidate genes, markers were chosen based on the information on the gene structure and linkage disequilibrium blocks available at the International HapMap Project (www.hapmap.org). The population of Utah residents with Northern and Western European ancestry from the CEPH collection was used as a reference collection as it most closely resembles the ancestry of the SOCCS population. Although there will be some differences in SNP frequencies between the reference population and the study population, the figures were used purely for simulation with the expectation that the method could be used on a number of different populations irrespective of SNP frequencies.

Genes and SNPs excluded

As the hypermutable tract in the APC gene identified in the Ashkenazi population has not been identified in non-Jewish populations, it was not included in simulations. Similarly, rare dominant mutations and those that more commonly cause other cancers were not simulated as people with other cancers were removed from the SOCCS study and it was therefore unlikely that the level of mutation in the study group will match the population frequency. These included K-RAS and TP53. Genes that would be impossible to identify using SNP data, for example the tandem repeats in the TSER gene, have not been included in the simulations. Known dominant polyposis syndromes, HNPCC or bi-alleleic MYH mutation carriers were excluded from the SOCCS analysis.

Included

As well as the candidate SNPs with not excluded, the simulations included genes in which the mutation rate increases as the cancer develops, despite the lack of

heritability. The mutations in juvenile polyposis - associated genes DCP4 and SMAD4 are found at 0% in adenomas but 35% in invasive carcinomas - suggesting a role in cancer progression, which would not be identified particularly well in a standard case-control study. As the full function is not yet known, it is possible there is an unidentified interaction effect. A full list of the polymorphisms used can be found in Appendix 4.2.

4.3.1.3 Assigning Case or Control Status

A number of different methods were investigated to assign case control status with differing success rates. Initially, the probability of being a case was estimated as the fitted value from a logistic regression model with a linear predictor capturing the relevant risk model, an example of which is shown below. In the example Beta1 refers to the regression co-efficient of the effect from not taking NSAIDs (the higher risk category) and Beta2 is the regression coefficient of the interaction between not taking NSAIDs and a risk allele.

$$\text{Risk of case} <- \alpha + (\beta_1 * \text{NSAIDs}) + (\beta_2 * \text{NSAIDs} * \text{allele}) + e$$

Therefore additional risk was added to individuals with particular traits. To ensure there were equivalent numbers of cases as controls, the whole simulation was run as a loop effectively filling the “case spaces” and “control spaces.” However, this was both time consuming and was producing data with an odds ratio very different from that assigned. Also, as the data will be case control, it was not good practice to be using a relative risk based estimate to assign risk. Generating the controls on a loop until there was enough was effectively drawing the controls from a 50:50 risk groups and the cases from 70:30 (if there was a 20% increase in risk). In practice this meant that there was too much statistical noise and that the desired effect was diluted.

Therefore a function was written that used the available information and basic algebra to assign case control status. In each simulation there will be a number of values or proportions that are pre-assigned and can be changed for different simulations, they

are also necessary to know for this function, and, during this description, have been assigned the following symbols:

OR = x Sample Size = S

Proportion exposed = y Proportion of cases = z

The a, b, c and d are as normally used in a traditional odds ratio, two by two table, with the defined features in bold:

	Outcome			
Exposure		Y	N	
	Y	a	b	**Sy**
	N	c	d	**S-Sy**
		Sz	**S-Sz**	**S**

Using these equations to find b, c and d in terms of a found that:

$b = Sy - a$ $c = Sz - a$ $d = S - Sz - Sy + a$

Then, substituting these relationships into the sample size expression, a + b + c + d = S, we obtain:

$$(x - 1)\,a^2 + (Sy + Sz - Sxy - Sxz - S)a + S^2xyz = 0$$

$\Rightarrow Aa^2 + Ba + c = 0$

Where:

$A = x - 1$ $B = (1 - x) * Sy + (1 - x) * Sz - S$ $C = S^2\,xyz$

As these are all known values, they will at this point in the function have a numerical form and be entered into the quadratic root solver ($Root = -B \pm \sqrt{(B^2 - 4AC)} / 2A$). This will give two roots, the positive solution is then selected. This method allows the case control ratio to be determined in advance and is more appropriate for a case control study where an Odds Ratio would be used to assess any effects. The code for this function is included in Appendix 4.1.

4.3.2 Simulation Designs

There are three main stages to the simulation study: the first involves the different parameters within the method and how to maximise the power; there are data variations obvious prior to analysis, such as sample size or case-control ratio; and finally there is the underlying data model, including inheritance and interaction model. This simulation allows data to be simulated that represents each of these situations. These areas are covered in more detail in Chapters 5 and 6.

4.4 Discussion

The Scottish Diet Action Plan (SDAP) [450] states that the diet of Scotland is so poor that if current trends continue, diet will remain a contributing factor in the poor health of Scotland relative to the rest of the UK and Western Europe. Such a poor diet increases the toll of unnecessary premature death, long-term illness and dental ill-health within Scotland.

However, the nutrient intake of the majority of the SOCCS population falls within the healthy range for the majority of variables. One reason for this difference is that the dietary variables are based on self reported consumption from a food frequency questionnaire, which may be subject to recall bias or be affected by what people are aware they should be eating, thus underestimating the unhealthy foods and overstating the healthy.

A shortcoming of using artificial data is the generation of each variable independently. For example, there would be no correlation between the level of calorie consumption and the likliehood of being assigned a high BMI. The patterns of association between risk factors within a real data set are absent from the generated set. Trying to assess and define every small relationship the variables may have with one another would be extremely difficult and time consuming. For the main aim of this project, although such associations may be desirable, they do not necessarily have to mimic those in real life. It is likely that, by chance, there will be spurious associations between variables and if such cases stand out for that reason, hopefully further analysis will elucidate this, much as would happen in real life. For example, even if the artificial data set had people with lower BMIs eating more than their heavier counterparts, it is the effect of this association on the analysis that is important not the specifics of the relationship. In fact, an unrealistic relationship between variables would suggest that the data had not been generated with expected results in mind.

However, the risk estimates for different dietary variables may have been influenced by risk factors falling into such groupings. People who follow a healthy diet may have

better outcomes for a number of different reasons and combination of reasons yet the majority of risk factor studies approach these risk factors in a univariate fashion. In fact only fibre and folate are being studied, in some places, as having an association. Current statistical methods and standard sample sizes do not allow such complexity to be taken into account for every study, yet it is worth bearing in mind when doing simulation studies that in fact the main effect may really be the sum of associated minor effects.

Chapter 5

The Development of a Novel Method to Identify Gene Environment Interactions

The Mixed Tree Method

5.1 Introduction

As discussed in Chapter 2, the size of many data sets compiled for disease-risk factor studies are too large for the convenient use of traditional statistical methods, making it extremely difficult to identify interactions between genetic and environmental factors. For this purpose it is important to either develop new methodologies or use well recognised methods in different ways to search through and break down the data. The term for such searching of so many possibilities is "data mining." Although the term is sometimes used pejoratively to refer to methods of manipulating or repeating statistics to get a desired result from a standard data set, which is more accurately described as data dredging, data mining can be a justified, investigative endeavour, laying no claim to a significance threshold, for such large volumes of data.

5.1.1 Aims

The aim of this section is to optimise the power of the Mixed Tree Method for a variety of different data scenarios. Success will be considered by the computational intensity, successful identification of the interaction and the identification of false positives. This will allow an efficient design, both informative and enabling further simulation studies, whilst controlling the computational burden. Small scale simulation studies are used to investigate features of the method and guide methodological development.

5.1.2 Data mining

Data mining is the extraction of patterns and relationships from large data sets. Ten years ago, the improvements in technology in many different scientific fields meant the volumes of data available were rapidly increasing requiring a new generation of tools for analysis. This field of development was called knowledge discovery in databases (KDD) [451].

There are three phases to data mining: exploration, model building and deployment. The exploration phase involves preparing the data for analysis and selecting features to bring the number of variables to a manageable level. The model building and validation phase extracts the knowledge from the data and determines the best possible way to describe the patterns within the data. The deployment phase uses this knowledge to create estimates for outcomes outside the dataset and to make predictions of future outcomes.

Data mining is used in a number of different areas including business: marketing [452], tailoring advertisements to groups, analysis; government work: security and identification of dangerous people, trends in voting and areas of influence, identifying money laundering [453, 454]; and scientific research: bioinformatics, physics, genetics [455]. It should be noted, however, that data mining in medical research comes with a host of ethical and practical considerations including: privacy of data and patient confidentiality; data ownership and potential lawsuits; and the explanation of potential risks and benefits [456, 457].

There are a number of different methods that can be considered to be forms of data mining, depending on how they are used. For example, regression is commonly used without a prior hypothesis on which variables might be having an effect on the outcome or how the effect comes about. One popular method of data mining is tree based; the currently used forms of tree based searching are described in section 2.3.2 but in order to develop a new form of using these methods it is important to start from their basic form and understand how they can be used.

5.1.3 Tree Methods and Applications

> "Our philosophy in data analysis is to look at the data from a number of different viewpoints. Tree structured regression offers an interesting alternative for looking at regression type problems. It has sometimes given clues to data structure not apparent from a linear regression analysis. Like any tool, its greatest benefit lies in its intelligent and sensible application."
>
> - Breiman, Freidman, Olshen, Stone

Background

The most widely used implementation of decision trees, for statisticians, is CART, which stands for Classification And Regression Trees. This is a recursive partitioning method developed by Breiman, Freidman, Olshen and Stone in the early 1980s [195]. When using CART, a hypothesis space is completely searched considering all the possible values, for all the variables, before selecting the attribute that best divides each set into distinct groups based on the selected outcome. CART only uses binary splits. CART considers the optimal split to be the one with the largest decrease in node impurity (Δi). In bioinformatic circles, a number of slightly different algorithms, ID3, C4.5 or C5.0, are used. These algorithms use a measure of uncertainty called information entropy and choose the split with the greatest information gain.

Classification Trees and Regression Trees

Classification trees are used for categorical outcomes, or discrete target variables (e.g. case/control status in an epidemiological context). The aim is to retain "purity" of the nodes, where an ideal model would produce nodes that solely contain either cases or controls, not a mixture of both. The effectiveness of a split is based on the homogeneity of the subsamples following the split. CART can employ a number of different splitting rules based on different measures of classification accuracy. Three commonly used splitting rules are the gini splitting criteria, twoing and entropy, descriptions of which are found in Table 5.1. In some cases the resulting tree is fairly insensitive to which measure of purity is used and gives a fairly consistent result across the different measures. The goodness of fit for classification trees is measured using misclassification rates.

In Table 5.1, $\mathbf{p}i$ is the relative frequency (Observed numbers in a class / total observations) of class i at a node (t) where a split is performed, also called the relative frequency of defectors. The left and right hand sides of a split are referred to as **L** and **R** respectively.

Table 5.1 Commonly used Splitting Criteria for Partitioning Methods

Purity Measure	Description	Uses
Gini (IG)	$\Sigma p_i(1-p_i)$	Most efficient when p=0 or 1
Twoing	$(PLPR/4)[(\Sigma\|p(i\|tL) - p(i\|tR)\|]^2$	Useful when 50:50 split required
Entropy	$-\Sigma p_i log p_i$	Most efficient when p = 0.5

The gini index is an impurity measure based on estimated class probabilities, which is maximised when all the observations are equally distributed across the classes. Gini impurity is calculated by adding together the range of probabilities of a series of objects being selected multiplied by their risk of misclassification. The Gini impurity measure is zero when everything is correctly classified. The gini splitting criteria attempts to identify the largest homogeneous group from the data and separate it from the remaining data in order to lower impurity [458].

Twoing attempts a more even separation than the gini index. Twoing denotes the variables as classes. At each node, the classes are separated into two superclasses and the decrease in node impurity (Δi) is calculated as a two class problem, finding the split to maximise Δi. The advantages to twoing are the strategic splitting where initially apparently dissimilar classes can be grouped together by having more "important" characteristics in common before being separated further down the tree. Twoing does not ignore any characteristic, but separates the populations that are too small to affect the split differently (usually differentiating them from the more significant by putting them in parenthesis). This makes both the position and relative importance of a variable in a tree easy to understand.

The entropy based splitting criteria is based on normalised information gain, a non symmetric measure of the differences between two probability distributions [459]. An advantage of the entropy based trees is that they are heuristic. Based on Occam's razor, entropy splitting methods have a preference for simple trees over larger trees, though not necessarily the smallest.

There are other splitting criteria that vary only slightly from the main ones discussed. Class probability is based on the same equation as the gini criteria but focuses on the

probability structure of a tree, not the classification structure or prediction success. Cross probability splitting attempts to divide the data based on probabilities of response and assigns class based on probability trees. It is theoretically possible for class probability to employ the twoing or entropy algorithm in place of the gini.

Selecting the optimal tree

It is important to bear in mind that the building of an accurate tree relies upon the data set being representative; otherwise the results will only be directly applicable to the sample. It is important to verify and validate the resulting trees, which can be done in a number of ways. For searching methods, the most common method for identifying the optimal model is cross validation. Cross validation requires the partitioning of a data set into subsets at random and performing the searching method on one of the subsets (the training set) and using the other set (the testing set) to validate the results. Normally a numbers of rounds of cross validation are performed and the results collated and averaged across the rounds. Ten fold cross validation, used in MDR and described in section 2.3.2, would split the data into ten sets, using the subset of nine of these to train the model and the remaining subset to test it for each subset in turn. Other cross validation methods include “repeated random sub sampling validation”, where the dataset is randomly split each time into a training and a testing set and “leave one out cross validation” where repeatedly and methodically a single observation is removed from the data set to be used as the testing set.

More specifically to tree methods, identifying the best tree depends on the balance between accuracy, computational burden, bias, variance and manageable data. Trees are usually pruned to show the most useful data. Pruning can be done after a large tree is grown and then pruned back to a manageable size or can be determined before the tree is grown, although this may be difficult to know in advance. Pruning is done by sequentially collapsing the nodes that have the smallest difference in purity and involves a trade off between bias and variance. Cost complexity pruning adds a complexity cost per terminal node and the cost complexity measure is calculated by adding the complexity cost of the tree to the misclassification cost penalties. Such costs are pre-determined and are based on the requirements of each analysis. If accuracy is so important that it is preferable for every variable to have a node

containing itself, there will be no misclassification; this would be achieved by setting the complexity cost very low and the misclassification cost high.

Weaknesses of Tree Methods

The weaknesses of tree based methodologies are that they are unable to capture a linear relationship, the best they can do is a step function approximation of a linear relationship. It would therefore not be the most efficient way to model a dataset, if the main underlying structure in the dataset is linear. However, this is not the primary aim of exploring large data sets to identify possible gene environment interactions.

5.2 Mixed Tree Method (MTM)

5.2.1 Rationale behind the Mixed Tree Method

One of the important factors in developing this approach is the inherent difference between the epidemiological history of environmental variables and that of genetic variables. Environmental risk factors are well studied and fairly well understood, with large volumes of evidence supporting association and complementary biological mechanisms. It is also unlikely that anything completely unprecedented will be discovered regarding the risk of an environmental variable, as an environmental factor which has never, in any study, been shown to confer risk, will not normally have been measured and recorded for analysis That is not to say that studies of interactions add nothing to environmental risk factor studies, but that they more likely add to the understanding of the risk factor than to its discovery. Therefore, it would be inefficient to mine environmental variables in the same way as the genetic data, looking for completely new risk factors.

A second important issue is that studies into the risk of environmental variables tend to have considerably larger odds ratios for the independent effects. If all the environmental and genetic variables were included in a single statistical model, the more minor environmental risk factors would be considered more important than the most important genetic risk factors. For example, in a decision tree, the environmental variables would dominate the earlier branches and disrupt the later branches containing the genetic factors, making it possible to miss an interaction effect between the two.

Similarly, even main effects for polymorphisms are likely to be a magnitude higher than for interaction effects, so it might be expected that independent genetic effects would dominate the upper branches of the individual trees. There are currently no methods available that remove such variables in order to treat them differently and allow the study of second order effects in general.

Therefore any novel method must accommodate the following:

- Different approach to environmental and genetic variables
- Larger effect sizes of environmental variables
- Genetic main effects vs. interaction effects

5.2.2 Outline of Method

The data are analysed sequentially, using separate data sets that each contain a single environmental variable and all the genetic variables. Once the process is completed for the first environmental variable, it moves on to the second, and then the third and subsequent ones until all the environmental variables have been used in one complete analysis. Missing values are removed from the single analysis for which there is no environmental data, although any variable with a large proportion of missing data will be looked at in more detail to try and elucidate if there is a non-random reason for the missing values. If testing a more specific hypothesis, it is possible to run the method just for one environmental variable, or a smaller set of SNPs, and study the results in more depth.

Step 1

For continuous variables, the data set is forced to produce a single split tree using only the environmental variable, the cut-off value for the grouping is then isolated and two data sub groups are produced, those below or equal to the cut-off value and those above it, similar to the first step of a CART approach. In the case of binary variables, the data is just split into the two variables present. For categorical variables with more than two levels, dummy variables are created for each level and the single split divides up the data based on them. For example, NSAIDs can be a number of different drugs and combinations of the two, so each drug is entered as a dummy variable with both the individual and combined presence of a drug coded as true.

Within each analysis, any genetic effects dominating the subtrees will be treated as having both potential main effects and interaction effects; the method will then be re-run with them removed to minimise their effects on other variables. Once a genetic

main effect has been found to have a significant main effect, at the end of one analysis, it will be removed from the data for the further analysis. It can also be planted as a root in its own analysis to look for gene-gene interactions.

Step 2

Each subgroup from step 1 is then used to produce a separate tree, effectively the same as two branches of the same tree after one split, using the genetic variables. The results from the two subtrees are then compared to one another and the most important variables are selected for the next step. Variations developing the selection algorithm to maximise the selection of important trees is explored in this chapter. The variables are considered important if they are at the top of either sub tree or if there is a large difference between the sub trees regarding the position of the variable.

Step 3

Once the best method of selecting important variables has been identified through preliminary analysis, the important variables from the third step are entered into a logistic regression model. Depending on their selection criteria they will be entered as an independent effect, an interaction effect, or both. After the regression step, the final results are compiled. A summary of all three steps is shown in Figure 5.1.

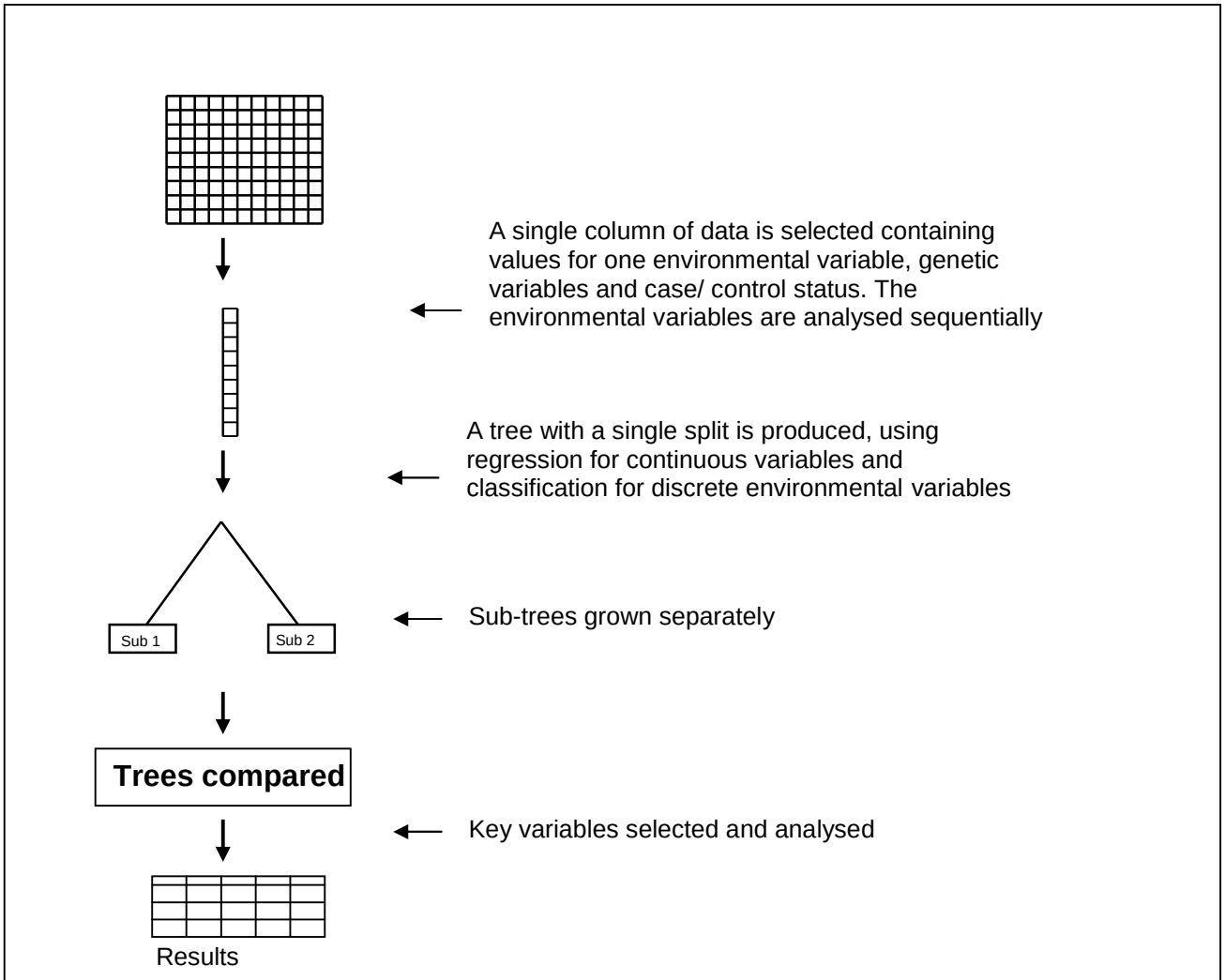


Fig 5.1 Graphic Depiction of the Basic Mixed Tree Method

This is the basic structure, the more specific elements of the method still to be assessed and decided include: the particular method of tree growth and tree control; splitting criteria; and identification and analysis of potential interactions. Using small simulation analysis to develop the method and determine the specfics of each step is summarised in this chapter.

5.2.3 Computational Intensity

Initial simulations showed that the MTM is reasonably computationally intensive, particularly in the context of simulation studies with a large number of repeated analyses. This was explored using a number of different data generation methods and operating systems, none of which reduced the problem to a manageable level.

Without a technical solution, it was necessary to look more closely at the different elements of the study, especially the data input variables, to identify the time consuming elements. The areas requiring such study included the number and format of environmental variables being entered into the analysis. It was therefore a reasonable judgement to carry out two analyses, one containing an interaction effect and one with no determined effects for comparison.

5.3 Method Development

The method was implemented using R (http://www.r-project.org/) incorporating the rpart function [460] and the random forest function [461] and a help guide on recursive partitioning in R [462]. However, the majority of the method was written for this thesis, including selection of the data for each analysis, the loops for implementing the functions and the settings chosen. More importantly the selection algorithm and analysis was developed and perfected drawing on the results from the practice analysis.

5.3.1 Data Preparation

The initial steps of an MTM analysis involve preparing the data, identifying which data set is to be used, removing missing values and specifying which environmental variable will provide the initial split or "root." As this was a simulation study, there were no issues of missing data, however in a real analysis univariate anaylsis of the variables would be carried out prior to analysis and the missing values treated according to the results of such analysis.

In the following sections, in cases where each individual result has to be examined, a smaller dataset extracted from the main simulation data was used that contained information on NSAID intake and the simulated SNPs associated with genes from the Wnt pathway only. This subset contained 63 SNPs from 14 different genes. In this smaller test run, only 100 datasets were run each time. In the larger analyses, 175 SNPs and 1000 replications were used.

Occasional limitations during the data generation step led to a reduction in the expected identification rate, purely as an artefact of the data. In these instances the same analysis was also run using an optimum version of regression, where only the variables involved in the risk model were entered into the analysis model. This was purely as a comparison to show that the MTM was analysing imperfect data and to

indicate the results that could be expected from a “perfect” performance and was therefore more appropriate for gauging the success of the method.

5.3.2 Presentation of Results

As the method develops different measures of success are used to measure performance, depending on which steps are finalised. Throughout, one of the main measures of success will be the identification of the pre-determined effect, either main effect or interaction effect. This will be referred to as the “target SNP,” in the case of a main SNP effect, “target interaction” for the target SNP and its interacting environmental variable and “interacting SNP” for the SNP involved in the interaction but identified separately.

To determine the most effective way to identify interactions or main effects, different features of the results were considered. The different modes of SNP selection are recorded as they may give clues to the underlying model of the interaction.

The most obvious candidates are the SNPs at the top of each sub-tree. Theoretically, being at the top, or very near the top, of both trees would suggest a potential independent effect, although possibly an interaction effect as well, these variables are refered to as “Ind” variables. Being at the top of one tree but not important in the other would more likely suggest an interaction effect. SNPs identified in this way are referred to as “Top” variables. Although an Ind variable occupies a top position, it is the relative positions between the two subtrees that determines classification. The number of positions under consideration to define these labels is explored in section 5.3.3.3.

Another feature that may identify a SNP as being important is its association with movement between subtrees, which requires measurement of the difference in position – with position referring to the ranking of variable by importance following the random forest analysis. These are referred to as “Move” variables. SNPs exhibiting such positional differences can be identified by a ranking score which

measures the difference in position between the two subtrees whilst according more weight to having a high position in one tree:

$$\textbf{Score} = (\text{pos in tree 1} - \text{pos in tree 2})^2 / (\text{smaller position})^2$$

The following table gives a summary of the different features of the SNPs following the random forest stage and how they will henceforth be described. In a number of situations the SNP at the top of one subtree would also be the one exhibiting the most movement, in which case it is classed as Top:

Table 5.2 Key for Describing Variables Selected Following the Random Forest Step of the MTM

Variable name	Identification Method
Independent Type 1 (Ind1)	Top of both subtrees after random forest step
Independent Type 2 (Ind2)	Top of one sub tree and position 2-5 in the other
Top variable 1 (Top1)	Variable top of one sub tree, below position 5 in the other
Top variable 2 (Top2)	Variable second in sub tree, below position 5 in the other
Move variable 1-n (move1)	Variable occupying a top 5 position with greatest difference between sub trees
Move variable 1-n (move2)	Variable occupying a top 5-10 position with greatest difference between sub trees

The initial analysis measures the success of different splitting criteria, by studying in detail how often the interacting SNP is identified using the different splitting criteria and in which format.

When attempting to identify the potential interacting SNPs, success at identifying a target SNP with a main effect will quantify the results defined as Ind1 and Ind2. Identification of the interaction will measure the relative success of the different result features to assess the most effective way for the method to select the variables to enter into the regression step.

5.3.3 Implementation of the Mixed Tree Method

With the basic structure of the mixed tree method decided, it was necessary to determine the specifics of each step. For the sub-tree stage both the growing and comparison of the sub-trees need pre-specified limits, the recursive partitioning stage needs a choice of splitting criteria and finally, a determination of how to best select potential interactions from the data.

5.3.3.1 Tree Growing

Tree size is not limited during the growing process, so the trees tend to be large and overfitted. In order to recognize the important genes in a large tree, the trees can be reduced in size or large numbers of trees generated and then averaged.

There are two methods that can be used to reduce tree size: pruning and shrinking [463]. Pruning determines importance using a cost-complexity measure and recursively snips off the least important branches of a tree. The pruning is done by automatically selecting the complexity parameter associated with the smallest cross-validated error. Pruned trees have some advantages over un-pruned trees: they are smaller, avoid over-fitting and are better at dealing with statistical noise.

Instead of pruning a tree towards the main root node, the nodes can be shrunk towards their direct, nearest, parent node. Whereas pruning reduces the tree as a whole with the relative importance of each branch compared to all others, shrinking is more specific to the node involved. Shrinking uses a convex combination parameter; a fitted probability at a node, a convex combination of the probability of the parent and the frequency of the offspring nodes [464]. This is effectively bagging, combining the output from bootstrapped samples, each branch individually.

However, producing a single tree from all the data, even using pruning or shrinking to make the results more accessible, is susceptible to a large amount of statistical noise and influenced by random effects. It is not possible to differentiate between the random differences between subtrees with no effects and those that differ because of a real main or interaction effect.

Therefore the option of integrating other tree approaches at this stage was explored. As discussed in chapter 2, the random forest method creates a number of trees from the same dataset and then averages across the tree. Therefore using a random forest approach to generate the subtrees reduces the statistical noise. The advantages to this approach include balancing the error in unbalanced data sets and generating a measure of importance for the variables in determining classification, the splitting criteria. To decide on a method, both a measure of decrease in accuracy and a measure of

decrease in node impurity, known as the Gini index, were tried. Twoing and entropy were not compatible with the random forest R function or preferable in theory. Twoing is more informative looking at individual trees, but such information gain would be negated by the combinatorial approach of random forests. Also, attempts by twoing to create similarly sized groups would lead to a similarity across the forest trees, instead of creating the forest average tree from the most information available. Similarly, entropy has a preference for simple trees and the resultant forest would not be created from trees with the variation and information of node impurity or Gini.

5.3.3.2 Splitting Criteria

Which splitting criteria would be more successful in the mixed tree method was not clear theoretically or from the literature. Therefore examining a number of small, practice simulations using only the steps described thus far, was the best way to decide between them. In this case it was important to be able to study the details of every result to get a clear picture of the results, information that would be impractical to study from 1000, or even 100 simulations. Therefore ten datasets of ten populations were analysed using first the accuracy measure, then the Gini criterion. It would have been possible to develop a large number of trees and select ten at random, however, the use of different set seeds allows random variation whilst allowing repeated and more in depth analysis on specific data.

Ten datasets of ten were produced with an OR of 2 for a SNP-environment interaction. The results of different selection criteria using a basic, single interaction with no main effects model are shown in table 5.3. Detection is described as the proportion of identification under the different selection options described in Table 5.2. The measure of which splitting criteria is more successful is based on a combination of identification and the method of selection. For example, if the results indicated that it was necessary to consider all top three (Top3) positions, in both subtrees, to achieve a 90% identification rate, that is taking six variables into account, five of which will be false positives. However if it is possible to achieve an 80% identification rate using just the SNPs at the top of the tree, that involves only 2

variables – up to one false positive. In this scenario the latter would be considered more successful.

Table 5.3 Detection Results by Splitting Criteria

Set Seed	Splitting Criteria	Detection Rate (ID method)	More Successful
29	Gini	100% (Top) or 90% (Move3 + Ind1)	Gini
	Accuracy	90% (Top), 1 null result	
17	Gini	90% (Top) or 100% (Top + Move1)	Gini
	Accuracy	90% (Top), 1 null result	
88	Gini	100% (Top) or 100%(Move1 + Ind3)	Gini
	Accuracy	90% (Top) or 100% (Top3)	
52	Gini	100% (Top) or 100% (Move1 + Ind1)	Gini
	Accuracy	90% (Top) or 100% (Top2)	
34	Gini	90% (Top and Move1)	Gini
	Accuracy	80% (Top2)	
111	Gini	90% (Top) or 100% (Top2) or 100% (Top + Move2)	Gini
	Accuracy	80% (Top + Move 3)	
9	Gini	90% (Top) or 100% (Top3)	Gini
	Accuracy	100% (Top3) or 90% (Top + Move3)	
68	Gini	90% (Top2) or 90% (Top + Move1)	Gini
	Accuracy	90% (Top2)	
94	Gini	90% (Top)	Accuracy
	Accuracy	100% (Top)	
40	Gini	100% (Top) or 100%(Move2)	Gini
	Accuracy	100% (Top)	

From these results, the Gini index performs better in the majority of cases than accuracy in identifying a basic interaction. Simply selecting the variables from the top of each sub tree would identify the interaction 92% of the time, rising to 95% if the variable with the greatest movement was also identified (assuming it was a different variable, which was usually not the case). Using accuracy, the interaction was identified 84% of the time, with no improvement if movement was included in the selection criteria.

Although different data models may slightly change the selection criteria, these are the two most basic measures and provide a good representation by which to compare the two indices. In all the following analyses, the Gini criterion was used.

5.3.3.3 Identification of Potential Risk Factors

With the tree growing and splitting criterion determined, the next stage was how best to decide which SNPs to select from the trees to maximise the chance of identifying any underlying effects. Considering that the selection of a single SNP could involve entering both the SNP and its potential interaction into a regression formula, it is necessary to minimise the number of variables being entered into the regression formula as far as possible. It is also important that the method is capable of identifying independent SNP effects as well as interaction effects, or a combination of both depending on the underlying interaction model. For a hypothesis generating method, all additional information gain is beneficial.

Data containing independent effects

The first step in developing the method to identify independent effects was to establish the best way to select them and then evaluate them. To establish the selection criteria, a dataset was generated with the parameters shown in Table 5.4:

Table 5.4 Data Parameters – Selection Criteria

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
NSAIDs	1	0.6664	63	1000	0.5	Dominant
Target SNP	2.0	0.6634				
Interaction	1	0.4421				

Focusing on the selection stages, prior to the regression step, table 5.5 shows how often an independent effect is detected using three different criteria: the first column shows the rate of identification if a variable is considered to have a main effect only when it is at the top of both sub-trees, described earlier as Ind1; the second when

selecting variables at the top of one sub-tree and positions 1-3 in the other; with the third column considering position 1-5 alongside a top position, which was earlier described as Ind2; the fourth column considers top 3 positions in both trees simultaneously; and similarly, both top 5 in the final column.

Table 5.5 Effect of the Sub-tree Position on Identification Rate

Positions	1 & 1	1 & (1-3)	1 & (1-5)	(1-3) & (1-3)	(1-5) & (1-5)
Identification (%)	43	64	76	96	97
False detection (%)	2	6	6	31	105

From these results, there is a large improvement in performance without too large an increase in false detection when considering positions 1-5 alongside one variable in the top position. However, considering positions 1-3 or 1-5 in both sub trees simultaneously leads to a substantial increase in the false detection.

Therefore, a reasonable balance between specificity and sensitivity in identifying an independent effect includes variables at the top of one sub tree and in position 1-5 in the other tree. The increase in successful identification is greater without an equivalent increase in false detection. There is still one last methodological step after this one to eliminate false positives. The following table shows the results when incorporating this final stage, at different significance thresholds, thus adding a quantitative element to the results.

Table 5.6 Results from Identification of Independent Genetic Effects

Significance Threshold	SNP Effect Identified (%)	False SNP only result (%)	False Interaction result (%)	False Environmental result (%)
0.01	100	20	16	4
0.001	100	7	4	0
0.0005	99	6	1	0
0.0001	96	2	0	0

Data containing only an interaction effect

For the second simulation, the data contained no independent effect; only people with both the risk SNP and the environmental factor were at increased risk. In this case the SNPs at the top of each sub-tree were selected (Top1) as the earlier work on splitting criteria found this to be the strongest selection method.

However, the best parameters and positions involved in selecting the Move variables were explored further. Initially, the SNP that showed the most movement and occupying position 2-5 in one sub-tree were selected, the sub-tree position was then extended to 2-9 to gauge any improvement. Table 5.7 shows the parameters of the simulation, Table 5.8 the results at significance levels 0.01 and 0.001.

Table 5.7 Data Parameters- Interaction Effect

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
NSAIDs	1	0.6664	63	1000	0.5	Dominant
Target SNP	1	0.6634				
Interaction	2.0	0.4421				

Table 5.8 Identification of Interaction when Interaction OR = 2

Sub-tree positions considered	Significance Threshold	Interaction Identified (%)	SNP Effect Identified (%)	False SNP result (%)	False Interaction result (%)
2-5	0.01	54	100	3	10
	0.001	29	93	0	2
2-9	0.01	54	96	15	3
	0.001	30	91	0	2

Therefore, increasing the Move positions considered did not dramatically increase the identification of the interactions, for this model. Increasing to an ever greater number of positions is unlikely to have much more of an effect as the denominator of the Move calculation is based on the position.

Data containing both main and interaction effects

Using the results from the independent and interaction analyses, it was then necessary to check that neither dominated the results when used in combination at the expense of the other. Therefore data were generated based on a model in which the gene had an independent effect, which was amplified by the environmental factor, the parameters shown below in table 5.9, the results in table 5.10.

Table5.9 Data Parameters- Main and Interaction Effects

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
NSAIDs	1	0.6664	63	1000	0.5	Dominant
Target SNP	2.0	0.6634				
Interaction	2.5	0.4421				

Table 5.10 Identification of interaction when SNP effect OR = 2, interaction = 2.5

Significance Threshold	Gene Effect Identified	Interaction Identified	NSAIDs Identified	False Gene result	False Interaction result
0.01	100	94	19	1	13
0.001	100	81	8	1	0
0.0005	100	77	5	1	0
0.0001	100	65	1	0	0

Even at low significance thresholds, the mixed tree method was extremely successful at identifying the gene effect and moderately effective at identifying the interaction. Considering the interaction was only increasing the OR by 0.5 and there would have been some statistical noise from the data generation, this was a good result. Further, in depth analysis of the underlying interaction model can be found in Chapter 6.

5.4 Analysis of Parameters under Investigator Control

There are a small number of parameters under investigator control (or nearly so), the selection of which can affect the results. The environmental variables being entered for analysis, given that the Mixed Tree Method (MTM) is effectively analysing them sequentially, could be entered either as a group or individually. The format of the environmental variable can also be changed, especially in cases where there are well recognised thresholds or very small numbers in a categorical grouping.

In order to optimise the results in subsequent simulations and future practical application, analyses were run to measure the effect of these parameters. The aim was to construct the method with maximum power, with respect to the parameters under investigator control. However, in some instances, the computational intensity in trying to run so many datasets, each so large in nature, on a standard, personal computer was problematic and time consuming.

Initial simulations were run using different numbers of environmental variables, to measure computational intensity and the feasibility of the other simulations. The second parameter was the format of the environmental variables, especially when there was an option to re-classify from continuous to binary.

5.4.1 Number of Environmental Variables

The number of environmental variables analysed at one time was primarily an issue affecting the simulation studies, where the required number of replications increased the computational burden to an unmanageable level. The most obvious candidate for reducing the time of the analysis without losing too much information was the number of environmental or “root” variables being analysed at one time. The MTM effectively analyses them separately, so measuring successful identification for a variable with an effect does not necessarily need all the variables to be entered into the analysis simultaneously. It would depend on what the computational benefits are for different assortments of variables.

A number of simulations were run comparing the time taken to run different numbers of datasets containing different numbers of environmental variables. More detail can be found in Appendix 5.1 but prior to any adaptations it took more than 300 hours to run analysis on 1000 datasets each containing 21 environmental variables.

There were three factors that influenced run time: number of "root" variables; number of replications; and the number of datasets generated and being saved in the system memory. Once the false error rate had been established there was no analytical advantage to running large number of variables alongside the target variable for each run. As long as there was one variable with an effect and one that had no set effect with which to compare, it was possible to gain a full picture of what the results would be in such a situation without adding too much time to the simulation.

5.4.2 Format of Environmental Variables

There are situations where prior knowledge may dictate an obvious way to divide a categorical or continuous variable into binary or discrete groupings, for example the dietary variables have a recommended level of consumption. Simulations were run to test the effect of variable type on identification rate.

Categorical Environmental Variables

To analyse the effect of recoding categorical environmental variables, duplicate analyses were run using NSAIDs as the split variable in both the categorical and re-classified format. The recoding is shown in table 5.11 and the parameters for the test simulation in table 5.12. In this instance, an optimised regression, as described in section 5.3.1, was run for comparison to identify to where in the method the difference could be attributed.

Table 5.11 Recoded Class of NSAIDs from Categorical to Binary

Variable	Discrete classes (%)	New classes (%)
NSAIDs	None (66.6)	0 (66.6)
	mini aspirin (16.9)	1 (33.4)
	standard aspirin (12.)	
	other NSAID (3.2)	
	mini + other (0.7)	
	standard + other (0.5)	

Table 5.12 Simulation Parameters – Data Type

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
NSAIDs/ NSAIDs bin	1	0.6664	175	1000	0.5	Dominant
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

The results from this simulation, showing true and false positive identification across a spectrum of significance thresholds, are shown in Table 5.13. Table 5.14 shows the logistic regression results for an analysis of only the target variables.

Table 5.13 Percentage of Datasets Identifying Interaction, NSAIDs as Discrete Compared to Binary Variable

	Significance Threshold			
	0.01	0.001	0.0005	0.0001
NSAIDs in Discrete Classes				
Target Int	34.2	13.6	9.9	5.0
Target SNP	91.6 (574/658)	74.4 (608/864)	68.5 (586/901)	50.0 (450/950)
Neither Int or SNP	8.4	25.6	31.5	50.0
Other Int	9.8	1.7	0.5	0.1
Other SNP	2.3	0.6	0.3	0.1
When Recoded as Binary Variables				
Target Int	54.2	26.1	19.4	10.5
Target SNP	97.0 (428/458)	91.1 (650/739)	87.4 (680/806)	77.2 (667/895)
Neither Int or SNP	3.0	8.9	12.6	22.8
Other Int	22.1	4.9	2.4	0.3
Other SNP	3.1	0.8	0.6	0.2

Table 5.14 Percentage of Datasets Identifying Interaction, Difference in Discrete and Binary Classification Using an Optimised Version of Logistic Regression

	Significance Threshold			
	0.01	0.001	0.0005	0.0001
NSAIDs in Discrete Classes				
Target Int	28.4	9.9	7.3	3.6
Target SNP	83.1	58.8	50.8	33.6
Neither Int or SNP	16.3	40.9	48.8	66.2
When Recoded as Binary Variable				
Target Int	54.8	25.8	20.7	10.8
Target SNP	98.3	92.0	88.0	77.4
Neither Int or SNP	1.7	7.9	11.9	22.5

These results show that MTM identifies interactions with binary variables more successfully than discrete classes, which could be in part explained by the risk model being based on an effect which is essentially binary, that of an elevated risk from not taking any NSAID. There is a greater chance that the categorical format would misclassify the variables in some cases and therefore have poorer results. In the majority of successful datasets for the discrete variable, it was splitting in effectively the same way as when it was recoded as binary.

Using the categorical variables to split was least successful at identifying the target gene when the subgroups were most different in size. For example, if the data had split based on the consumption of any non-aspirin NSAID or not, only 3.3% of the population would be in one tree and 96.7% in the other. This suggests that binary recoding may be more valuable in circumstances where there is a logical grouping of small variables into a cohesive group; as such small groups may be wrongly selected more often as different, just through chance. It is worth noting though that some of this effect could be due to the binary nature of the risk model and if this is a concern, the analysis can be repeated using the categorical groupings.

Therefore in further analysis, when there is an obvious divide – both logically, based on the nature and prior knowledge of the variable, and numerically, giving close to 50:50 split – recoding will be used. In some circumstances these criteria are better met

by recoding the variable into new ranked categories, for example: smoking in table 5.15 which shows, for the four categorical variables, the recoding which should maximise their results and will be used in Chapters 6 and 8:

Table 5.15 Recoding of Categorical Variables into Binary Classes

Variable	Discrete classes	Proportion (%)	Binary classes	Proportion (%)
smoking	current	27	2	27
	ex regular	4	1	32
	ex occasional	28		
	non	41	0	41
exercise	none	55.9	0	55.9
	0-3.5 hrs/day	25.6	1	44.1
	3.5 – 5 hrs/day	11.7		
	5+ hrs/day	6.8		
NSAIDs	none	66.6	0	66.6
	mini aspirin	16.9	1	33.4
	standard aspirin	12.1		
	other NSAID	3.2		
	mini + other	0.7		
	standard + other	0.5		
deprivation	1	9	0	77
	2	21		
	3	26		
	4	23		
	5	11	1	23
	6	7		
	7	3		

Continuous Variables

Compared to binary and ranked variables, continuously distributed variables had the lowest identification rate and highest computational intensity during the preliminary runs. On further investigation it was found that was largely down to the settings for the partitioning step of the method. For the other variable types, the difference in complexity required between resultant subsets to justify a split was set very low in order to force the method to split when there were no real effects. However this meant that for continuous variables, the split favoured extreme outliers and very small

resultant subsets on one side of the tree. This was fixed by setting the cost complexity parameter higher and the minimum resultant sub-tree size to 20% of the initial sample.

There are a number of different approaches for turning quantitative traits into binary covariates. The most commonly used, especially in the context of nutritional variables, are recommended thresholds. These however may not always give the most information possible, especially if they divide the data into very uneven subgroups. Given that the nature of tree based methods requires a split, the ideal divide would be both data driven in nature and result in sub groups of similar size.

To examine the effect of recoding a continuous variable as binary the following parameters were used, with an OR of 1.5 for being overweight, defined as having a BMI of greater than or equal to 25, (and having the SNP) and 2.5 for being obese, BMI greater than or equal to 30, (and having the SNP). The overweight category in the table includes those who are obese as the risk is additive, to have a BMI above 30 is also to have one above 25. The frequencies are approximate as the weight and height are generated on a spectrum before the cut-off limits are applied. Table 5.16 shows the simulation parameters, 5.17 the results. The analysis is only run twice using the terms overweight or underweight, the obesity grouping is mentioned in the parameter table as it confers additional risk during the data generation stage.

This was also examined more thoroughly using the optimised regression, as dicussed in section 5.3.1. Table 5.18 shows the optimised logistic regression results with those from the MTM in brackets for comparison.

Table 5.16 Simulation Parameters – Continuous versus Binary Classification for BMI

Risk Effects Present (b = binary)	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
BMI **Overweight (b)** ***Obese (b)** **Underweight (b)**	1	 ~0.562 ~0.258 ~0.09	175	1000	0.5	Dominant
Target SNP	1	0.6634				
Interaction **Overweight (b)** ***Obese (b)** **Underweight (b)**	 1.5 2.5 1	 ~0.373 ~0.171 ~0.06				

Table 5.17 Percentage of Datasets Identifying Interaction for BMI as a Continuous Variable and Recoded as Binary

	Significance Threshold			
	0.01	0.001	0.0005	0.0001
BMI as a continuous variable				
Target Int	56.1	27.9	22.1	13.4
Target SNP	24.7	10.0	7.4	3.2
Neither Int or SNP	43.9	72.1	77.9	86.6
Other Int	10.4	2.3	1.8	0.6
Other SNP	28.3	9.0	6.2	2.4
When Recoded as Binary (overweight) Variable				
Target Int	81.6	56.3	48.0	31.5
Target SNP	1.3	0.1	0	0
Neither Int or SNP	17.9	43.6	52.0	68.5
Other Int	15.8	4.0	2.4	0.7
Other SNP	60.2	21.4	14.7	6.6

Table 5.18 Percentage of Datasets Identifying Interaction with BMI from Optimised Regression versus MTM

	Significance Threshold			
	0.01	0.001	0.0005	0.0001
BMI as a continuous variable				
Interaction (Int) found	57.3 (56.1)	28.1 (27.9)	12.2 (22.1)	13.3 (13.4)
Gene found	25.2 (24.7)	10.5 (10.0)	7.8 (7.4)	3.3 (3.2)
Neither Int or gene found	42.7 (43.9)	71.9 (72.1)	77.9 (77.8)	86.6 (86.7)
When Recoded as Binary (overweight) Variable				
Interaction (Int) found	81.9 (81.6)	57.3 (56.3)	49.1 (48.0)	32.2 (31.5)
Gene found	31.7 (1.3)	11.7 (0.1)	9.1 (0)	3.6 (0)
Neither Int or gene found	17.9 (18.1)	43.6 (42.7)	52.0 (50.9)	68.5 (67.8)

In some instances the MTM performed better than the optimised regression. This can be explained by the presence of other variables, reducing the residual variation attributed to the environmental main effect and random statistical noise. From these results it can be seen that almost the entire improvement in identification using the binary classification can be attributed to the regression step of the method.

This variable also appears to be more successful when recoded as binary but again the risk model has been optimised towards the binary variable as risk was assigned at a higher level above a cut-off of 25 during the generation step. Recoding would be less accurate when increase in risk does not line up identically to established parameters (e.g. BMI thresholds of 25 and 30 conferring an associated risk). Therefore, a number of small simulations were run with an associated increase in risk for a BMI above 23-33, but keeping the 25 cut-off for analysis. This will help distinguish between the preference for binary variables and the problem of matching the risk model to the analysis. The results are shown in Figure 5.2:

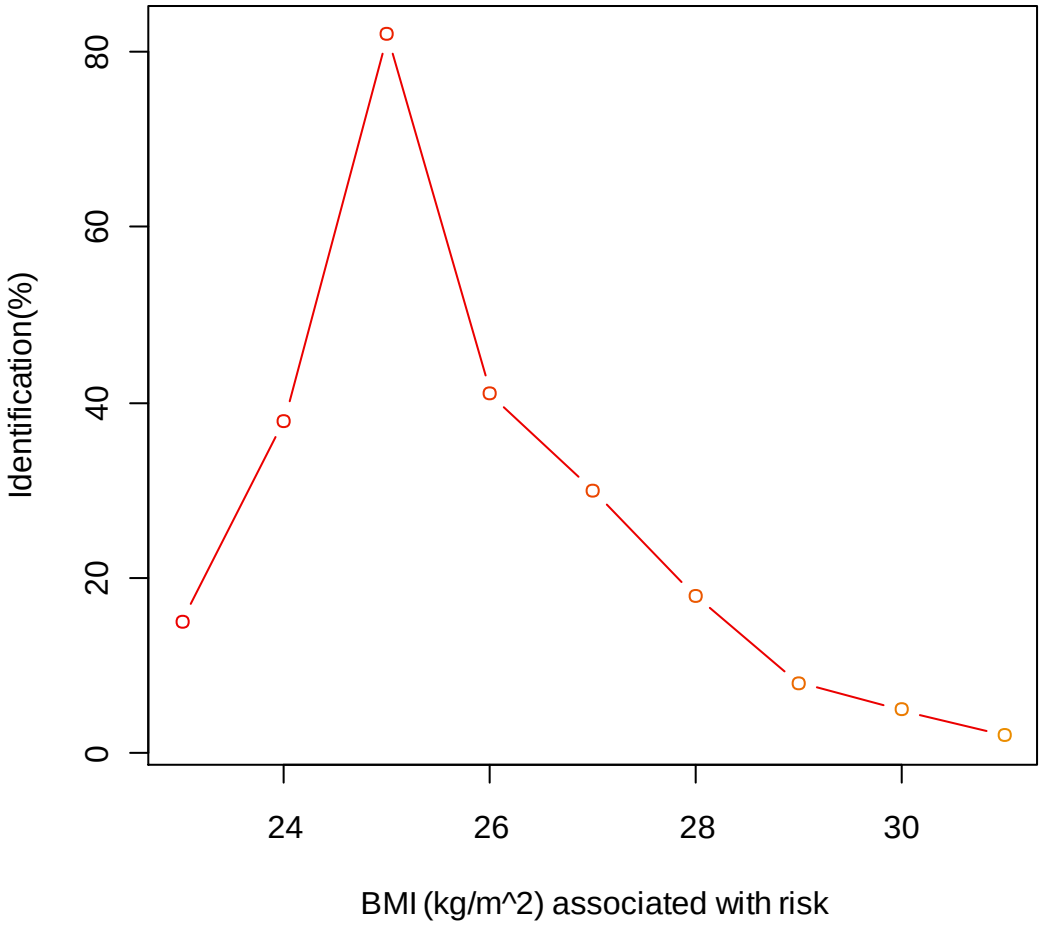


Figure 5.2 Association between BMI Associated Risk and Predefined Cut-off at 25

It does therefore not seem beneficial to split the data into pre-specified binary or ranked groups prior to the whole MTM analysis. However, prior to the regression stage regrouping the variables based on the cut-off point from the earlier recursive split might improve the performance of this stage of the analysis. To assess if this was the case, the same analysis was run again using the continuous variable, but before the final regression step the variables were re-grouped based on their earlier partitioning threshold. The results from this are shown in Table 5.19.

Table 5.19 Identification when the Binary Classification for the Regression Step is Dictated by the Earlier Recursive Split

	Significance Threshold			
	0.01	0.001	0.0005	0.0001
Interaction (Int) found (%)	65.2	38.8	31.1	18.8
Gene found (%)	96.7	90.1	88.1	80.4
Neither Int or gene found (%)	3.3	9.8	11.8	19.5

Comparing these results to the earlier table, this technique was more successful than either the regression or MTM on a continuous variable. The proportion of datasets that did not detect either the interaction or the SNP was even lower than when the variable was recoded as binary. As the initial advantage from binary recoding was primarily explained away by the more effective regression step, recoding only at the final stage using a data driven boundary keeps the mixed tree method more flexible in dealing with different cut off points, especially when the data has not been generated using identical limits, as it was here.

5.4.3 Summary of Mixed Tree Method

In order to maintain effectiveness but reduce computational time the simulations in the next chapter were run with only the two environmental variables, NSAIDs and smoking, to assess the levels of successful identification and the false positive identification respectively. Variables originally in discrete classes were regrouped to form binary variables with a ratio as close to 50:50 as possible. Continuous variables were analysed in their original format, with some limitation defined to ensure reasonable group sizes, until the final regression step where they are entered as binary.

Using a data driven split for the variables prior to regression improves the applicability of regression to the analysis. As hypothesis generating, a more quantitative way to compare the interactions across the different environmental variables is useful and a logistic regression step is well suited to this purpose following the variable selection.

Therefore after some basic analysis the final specifications of the method include the following list:

- Random forests instead of pruning for the initial sub-tree summary
- Using the Gini index as splitting criteria
- Identifying independent effects through their high position in both sub-trees
- Selecting potential interactions from their tree position (top of one sub tree) and movement using a score criteria that assesses variables in a top 5 position
- A final logistic regression step
- A single environmental variable involved in an interaction and a single variable not involved is adequate for future simulations
- The MTM has greater power analysing binary variables than categorical, and the regression step is more effective when continuous variables are recoded using the initial data-driven dichotomosation.

Taking these decisions into account, a depiction of the final method is shown in Figure 5.3. In chapter 6 the MTM is used on replicated datasets to ensure consistency.

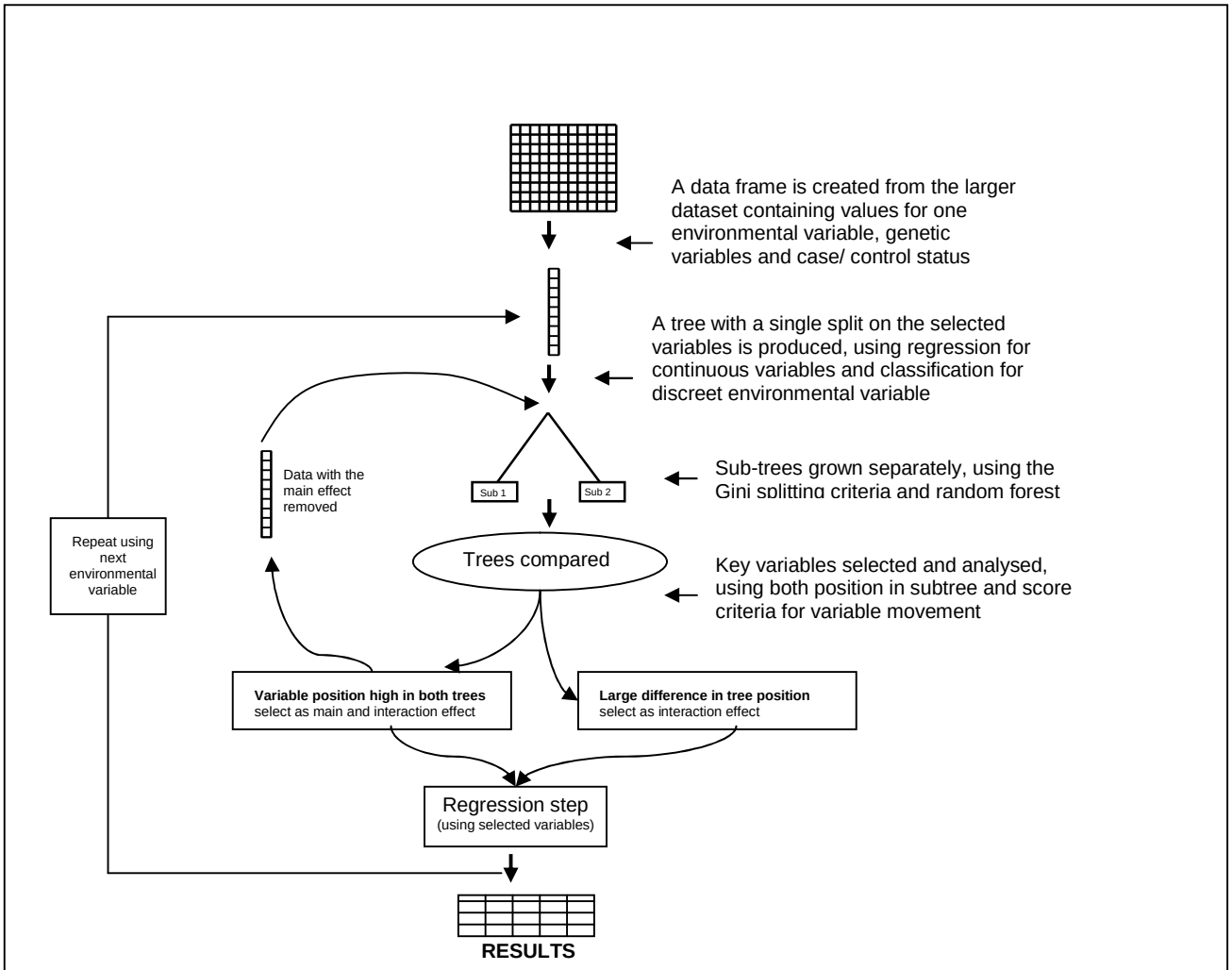


Figure 5.3 Mixed Tree Method incorporating Improvements

5.5 Discussion

The preliminary runs of the method showed some unexpected results. The sensitivity of the method and degree of positive results yielded when trying to define the splitting criteria was reassuring. However, the high level of false positive results found during the sub tree comparison is something that will have to be considered very carefully. In effectively forcing the method to choose the best variables, even from a pool of variables with marginal or no effects, a large false positive rate is inevitable. Although steps were taken to minimise this through implementation of the random forest and the logistic regression steps, the underlying theory of selecting the most important variables in relative, not threshold based, terms may result in low specificity and mean the method is better suited to particular study designs or situations. The following simulations focus on a case-control design, as that is the format of the real data analysed in Chapter 8.

5.5.1 Similarities and Differences with other methods

The classification procedure CART is forced to find the split point for a single variable but not used for the entire tree generation. The theory behind CART stays the same but the use changes slightly. The trees are grown using Random Forests, however it is how the resultant trees compare to trees grown from data on the other side of the initial split that is important, not so much the trees themselves. Logistic regression is used as an estimation stage following the hypothesis generated by the data mining.

What makes MTM different is that it treats the environmental variables and the genetic variables differently and exploits their inherent differences to best identify interactions between them. Other methods which combine genetic and environmental data, as if they are of one type, will find it hard to remove the larger and more dominant effects of the environmental variables. Also, the way potential effects are detected using this method gives more information about their mode of action.

5.5.2 Conclusions

The mixed tree method was highly successful at identifying interactions and main effects, maintaining a high level of success as the significance threshold became smaller and the level of false positive identification decreased. The small simulation has shown that the identification for main effects was improved, without a large increase in false positive results, when taking sub tree positions 1-5 into account instead of just 1-1, or 1-3 (assuming position 1 in the other tree). However, identifying the variables involved in an interaction with no main effects was not improved by increasing the catchments.

Preliminary results indicated that the success of the mixed tree method will vary depending on the underlying model of interaction and may be more limited in the absence of genetic main effects. There was a high level of false positive identification, however, which is why it was necessary to add a regression step at the end. Although logistic regression alone would not reduce the false positives without any correction for multiple testing, following the MTM a much smaller selection of variables is entered into the regression and minimises this problem. High false positive rates are more acceptable, at a practical level, in hypothesis generation than they would be if the method was intended to be definitive. It was however a shortcoming of the method, which will need to be considered during each future simulation.

Overall, the results reported in this chapter were promising and indicated a successful role for this method in hypothesis generation. The parameters and limitations of the method are more fully explored in Chapter 6.

Chapter 6

Testing the Parameters of the Mixed Tree Method Using Simulated Data

6.1 Introduction

6.1.1 Aims

The central aim of this chapter is to assess the effectiveness of the mixed tree method in identifying gene-environment interactions in data with a variety of known parameters. Following on from the results in Chapter 5, where methodological and technical aspects of the method were assessed, this section uses the finalised method to analyse data containing different dimensions or containing different features of interest.

6.1.2 Simulations and Parameters

The data under analysis contains the simulated list of environmental and genetic variables, identified from the literature and described in more detail in Chapter 4. The analysis is run sequentially using a list of environmental variables as the initial split or "root" variable. Construction of the artificial data allows variation in the nature of the data and the interaction present. As in Chapter 5, the parameters of each analysis are presented in a table, an example of such a table, for data with no main effect but an interaction effect is shown below:

Table 6.1 Example of Table Showing Simulation Parameters

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
NSAIDs	1	0.6664	175	1000	0.5	Dominant
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

The majority of the simulations use these basic parameters with variation in one area approached at a time, the table will show in bold the variable under study and the different values assigned to that variable in different simulations. The risk factor

frequency refers to the prevalence within the simulated population. The effect size is dictated by reversing the calculation of an Odds Ratio from a contingency table, knowing the OR in advance and filling in the probability of entering each of the cells in the table accordingly, as described in Section 4.3.1.3. All the other variables are generated straightforwardly using the probability of falling into each category; therefore there is some statistical noise between the datasets and the desired estimates. More detail on the data generation algorithm can be found in Chapter 4.

Throughout the chapter is concerned with interaction identification; however the parameters under investigation fall into three main categories:

- Changes in known data characteristics
- Changes in underlying effects
- Combinations of known and underlying effects.

These analyses are approached in order of increasing complexity to ensure the information gain can be correctly attributed to the correct variable and any problems are understood and adapted for in later analyses. Similarly to chapter 5, in cases where imperfect data may be affecting the results, an optimised version of regression, containing only the target variables, is used for comparison.

Due to the nature of the SOCCS dataset the interaction considered are all passive interactions, based on the presence or absence of a SNP variant. Future work includes data on gene expression and active interaction models, but this has not been examined in this study.

Data Characteristics

These are the features of the data that can be seen prior to analysis. These include the sample size, number of SNPs being analysed and the allele frequency of these SNPs. It is important to assess how effective the MTM is for these characteristics before undertaking more compex analyses.

Underlying Effects

Underlying effects are the risk associations and interactions that are found from the analysis of a normal dataset, but can be generated deliberately for simulated data. This includes the effect size of any risk estimates, both main and interaction, and the more complex features of interaction model and inheritance model. The interaction model is the combination of main and interaction effects present and the inheritance model refers to the recessive, dominant and co-dominant nature of inheritance.

Combinations

The last section looks at how the effects of some of the parameters influence the results of others. The combinations include the variability of the following: effect size and sample size together; and the possible combinations of allele frequency, inheritance model and interaction model.

6.1.3 Presentation of Results

The measures of success used throughout will be the proportion of the datasets where the interaction is correctly identified and the proportion of false positive results. Given the increase in computational intensity with increasing number of environmental variables, only two are used for these analyses: NSAID intake and smoking. NSAID intake attempts to identify the interaction (SNP*NSAIDs) and measures the other identified SNPs or interactions alongside the real interaction.

Smoking has no real effects and provides a comparative measure of false positive identification, for both interactions involving the target SNP and the false identification of interactions between smoking and a different SNP. Only the latter is a true measure of false positive identification as the identification of an interaction involving the target SNP with variables other than that intended may add information to the type of interaction or suggest a main effect.

6.1.4 Simulation Limitations and Decisions

The simulated dataset will not include any missing data. When real data is being analysed, the nature and potential reasons behind any missing data will dictate how it is handled. In the context of a simulation study, especially when comparing different methodologies on identical data, simulated missing-ness would be an entire sub-topic comparing the different imputation methods for each different method. Analysis on real data will examine the specifics of any missing data before a decision is made on the most appropriate way to handle it.

Throughout the simulation, the exposure to the environmental variable is approximately 50%. The difference in unknown exposure is investigated through different allele frequencies. To examine the effect of a range of environmental exposures, especially those at the extreme ends of the spectrum, would require adjustment of the preferred splitting criteria. Given that the real data is more easily grouped into similarly sized samples, it is more important to maximise the identification power for this scenario. However, the environmental variable and how it is grouped is somewhat under investigator control and can be dictated prior to analysis to ensure groups of similar sizes. For continuous variables, it is possible to dictate the minimum proportion desired in each group to negate the effect of outliers.

The focus in these simulations will be on a two way interaction, which is most appropriate for the comparison methods and the sample sizes involved in the real data set. Future work would include more complex interaction models. Similarly, the focus will be on a case-control design, similar to the SOCCS study design from which the real data is drawn. Different case-control ratios are explored, giving some insight into the ability of the MTM to handle this feature of other study designs.

6.1.5 Analysis

Once the optimum method and data input approach was established, the analysis most similar to that of real data was undertaken. The first stage is the analysis of the most basic characteristics, in the sense that they are known prior to commencement of

analysis. This is followed by the underlying effects and the potential interactions between them. The results include a measure of false positive identification and the successful identification rate.

6.2 Analysis of Simulated Data

This section details the different parameters of the data that can be seen prior to analysis but cannot be manipulated or changed. These parameters include the sample or population size, the number of SNPs from the candidate panel, the allele frequency of these SNPs and the case-control ratio.

6.2.1 Sample Size

In order to determine the effect that the sample size has on the levels of successful identification and false positives, simulations were run with the parameters shown in Table 6.2:

Table 6.2 Simulation Parameters – Sample Size

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	**Sample Size**	Proportion Cases	Inheritance Model
Smoking	1	0.510	175	**200, 500, 1000, 2000, 3000**	0.5	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

There were memory constraints in analysing 1000 datasets containing 2000 or 3000 people, on a standard computer. It is important for comparison with other methods and the adaptation of the method for other analysis to solve this instead of using specialist equipment. Therefore, this analysis was done by running 500 datasets twice or 250 datasets four times, using different set seeds (29, 17, 88, 63). As using different set seeds can sometimes result in correlations between streams of pseudo-random numbers generated [465], these four numbers were used to generate 1000 random numbers and these numbers plotted on a scatter plot to look for any structure. No such correlations were observed. This is shown in more detail in Appendix 6.1. The results from the simulation are shown below in table 6.3.

Table 6.3 Percentage of Datasets Identifying Interaction at Different Sample Sizes

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Sample Size	**200**	**500**	**1000**	**2000**	**3000**	**200**	**500**	**1000**	**2000**	**3000**	**200**	**500**	**1000**	**2000**	**3000**	**200**	**500**	**1000**	**2000**	**3000**
Target Int	1.9	8.4	54.3	90.1	98.9	0.6	2.2	26.0	72.9	93.0	0.2	1.3	19.6	65.2	90.0	0.2	0.5	10.7	49.4	81.8
Target SNP	7.2	28.1	97.2	100	100	3.4	21.2	91.0	100	100	2.2	18.2	87.6	99.9	100	0.8	11.6	77.0	99.8	100
Neither Int or SNP	92.7	71.9	2.8	0	0	96.3	78.8	8.7	0	0	97.6	81.7	12.3	0.1	0	99.0	88.3	23.0	0.1	0
Other Int	24.9	26.9	21.0	20.7	20.9	3.2	4.7	4.4	5.2	5.2	1.9	3.1	1.9	3.4	2.8	0.2	1.1	0.2	1.1	1.1
Other SNP	17.6	8.2	3.1	1.9	1.3	3.7	1.3	0.7	0.2	0	1.6	0.9	0.3	0	0	0	0	0	0	0
Splitting on environmental variable with no effect (smoking)																				
Target SNP	1.6	10.0	66.4	97.4	99.9	0.5	4.4	44.7	91.7	99.3	0.5	3.4	37.5	88.2	98.7	0	2.2	24.0	80.1	97.6
Target SNP and smoking	0	0.5	1.0	1.7	1.5	0	0	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0
Other Int	25.4	28.5	25.3	25.4	26.2	2.2	3.0	7.2	4.8	5.4	0.9	2.2	3.9	3.2	3.4	0	0.5	1.2	0.9	0.9
Other SNP	19.1	13.9	16.5	17.7	19.9	4.2	2.1	5.4	3.7	4.5	2.4	1.7	3.0	2.0	3.1	0.7	0.3	1.1	0.8	0.9

Int refers to the interaction, with Target Int being an interaction between NSAIDs and the expected (Target) SNP. These descriptions are used throughout.

There are very few surprises regarding the effect of sample size on identification. The low level of successful identification in small sample sizes (200 – 500 people) is to be expected with such a high significance threshold. It would not be recommended to use a complex, multistage analysis on such a small sample size. The power of the MTM, like many other statistical methods, increases with sample size.

6.2.2 Number of SNPs

For the majority of statistical methods reviewed, the number of SNPs included in the analysis was a limiting factor for successful identification, even at numbers much smaller than the number of SNPs in a GWAS. To measure if this is also the case for MTM the following data was generated.

Table 6.4 Simulation Parameters – SNP numbers

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	**Number of SNPs**	Sample Size	Proportion Cases	Inheritance Model
Smoking	1	0.510	**50, 100, 175, 300**	1000	0.5	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

As with the sample size simulations, there were computational limitations in generating 1000 datasets, each with 300 SNPs, therefore two runs of 500 were generated using different set seeds (29 and 17) before being combined in the table. There were no large discrepancies between the two sets of 500 prior to combination.

During initial test simulations it was found that at a very low number of candidate SNPs (10 or less), there were a number of simulations that were producing unusual results due to the probability of being in the "top 5," being artificially high Once these methodological restrictions had been recognised and characterised, the results from the main analysis are shown in Table 6.5 below.

Table 6.5 Percentage of Datasets Identifying Interaction as SNP Numbers Increase

	Significance Threshold															
	0.01				0.001				0.0005				0.0001			
Splitting on Target Environmental Variable (NSAIDs)																
Sample Size	**50**	**100**	**175**	**300**	**50**	**100**	**175**	**300**	**50**	**100**	**175**	**300**	**50**	**100**	**175**	**300**
Target Int	59.3	54.8	54.3	56.4	31.1	30.9	26.0	28.2	24.4	24.1	19.6	21.6	12.7	12.7	10.7	11.7
Target SNP	97.6	97.6	97.2	98.3	91.2	91.3	91.0	92.9	88.4	88.4	87.6	89.7	80.4	78.4	77.0	78.3
Neither Int or SNP	2.4	2.8	2.8	1.7	8.7	8.7	8.7	7.1	11.4	11.6	12.3	10.3	19.5	21.4	23.0	21.7
Other Int	13.2	17.3	21.0	19.7	1.6	3.4	4.4	4.5	0.8	2.1	1.9	2.3	0.4	0.6	0.2	0.1
Other SNP	2.1	2.2	3.1	2.0	0.4	0.3	0.7	0.4	0.2	0.2	0.3	0.1	0	0	0	0
Splitting on environmental variable with no effect (smoking)																
Target SNP	75.3	73.6	66.4	68.9	48.3	49.9	44.7	48.6	41.0	41.8	37.5	41.1	25.5	26.6	24.0	26.4
Target SNP and smoking	1.2	1.3	1.0	1.1	0.1	0.1	0	0.1	0	0.1	0	0	0	0.1	0	0
Other Int	17.1	19.4	25.3	24.2	2.4	3.4	7.2	6.1	1.5	1.5	3.9	3.5	0.1	0.7	1.2	0.9
Other SNP	8.0	13.9	16.5	14.8	1.7	3.1	5.4	4.0	0.8	2.3	3.0	2.0	0.2	0.7	1.1	0.5

The results from differing SNP numbers do show some interesting results, in particular it was interesting that the number of SNPs under analysis at one time does not adversely affect the successful identification of the interaction nor increase the volume of false positives, unless very small numbers are involved. When there is an interaction present between a SNP and environmental variables, neither with a main effect, the SNP is often selected as a main effect – both for the variable it interacts with and other variables that have no main or interacting effects. This may in part be due to the absence of any other main effects, so the slight increase in associated risk for the proportion of carriers that are experiencing the interaction effect confers the effect to the group as a whole. The SNP has a higher sensitivity and specificity than its interaction term. In terms of false positive identification, the rate is lower in the presence of real effects and does not increase as the number of SNPs (and therefore the potential number of interactions) increases.

Although current genetic screening can contain much larger numbers of genes than 300, it is reassuring that for the numbers analysed the successful identification is consistent as the number of SNPs increases yet the number of false positives does not increase. Although it is not possible to extrapolate from such low numbers to millions of SNPs, this is a good result compared to other approaches.

It may also suggest that for large numbers, a method that effectively ranks the results against one another may be better suited than one with absolute significance thresholds, although this is unlikely to be the MTM specifically, without adaptation. This method has been designed to work on a selected panel of SNPs, which the literature review suggests is around 200 SNPs for colorectal cancer. Therefore, these results are reassuring that the MTM may be suitable for such analysis.

6.2.3 Allele Frequency

To assess how effective the MTM is at identifying risk alleles and their interactions, it is necessary to compare the identification rate when the target allele is at a variety of frequencies within the study population: therefore a number of simulations were run when the target allele is at frequencies 0.1, 0.25, 0.5, 0.75 and 0.9 under a dominant

inheritance model. The genotype frequencies are assumed to be in Hardy Weinberg equilibrium based on the allele frequency. The parameters for this simulation are in table 6.6, the results in Tables 6.7.

It is worth noting that at very low frequencies, the odds ratio associated with the interaction effect will be closer to 1 than in the higher frequencies. It is difficult to accurately simulate an OR of 2.5 with only 0.1% in the exposed group.

Table 6.6 Simulation Parameters – Allele Frequency

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency					Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Smoking	1	0.510					175	1000	0.5	Dominant
NSAIDs	1	0.6664								
Target SNP	1	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**				
Interaction	2.5	**0.07**	**0.17**	**0.34**	**0.51**	**0.60**				

Table 6.7 Percentage of Datasets Identifying Interaction at Different Allele Frequencies under a Dominant Inheritance Model

Significance Threshold																				
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	35.9	66.9	38.5	2.5	0.1	16.6	38.4	15.7	0.5	0.1	12.9	31.0	11.8	0.3	0.1	5.4	19.0	5.9	0	0
Target SNP	57.7	97.7	89.9	10.4	0	57.4	96.4	76.6	4.5	0	56.9	94.9	70.0	3.7	0	53.1	91.1	54.6	1.4	0
Neither Int or SNP	42.3	2.3	10.1	89.0	99.9	42.6	3.6	23.0	95.4	99.9	43.0	5.1	29.6	96.2	99.9	46.9	8.8	44.8	98.6	100
Other Int	25.2	22.6	21.8	27.4	4.6	4.6	4.6	5.6	6.1	6.1	2.9	2.8	3.1	3.4	3.3	1.1	0.6	0.9	1.0	0.4
Other SNP	17.7	1.7	3.9	32.3	35.4	5.2	0.4	0.9	9.9	10.3	2.5	0.3	0.8	6.6	5.8	0.3	0	0.2	1.7	1.2
Splitting on environmental variable with no effect (smoking)																				
Target SNP	15.3	70.7	53.1	2.1	0	13.7	58.6	31.3	1.1	0	13.2	53.4	24.5	0.7	0	10.7	39.6	13.6	0.3	0
Target SNP and smoking	0.5	1.4	0.9	0.4	0	0.1	0.2	0.1	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0
Other Int	29.4	24.7	22.9	27.7	26.2	5.7	5.3	5.3	6.3	5.7	3.2	2.6	2.4	4.3	1.3	1.3	0.5	0.4	0.7	0.5
Other SNP	28.5	17.8	17.1	36.9	37.6	7.2	5.1	4.9	11.2	10.1	4.7	3.0	2.8	5.9	1.0	1.0	0.6	0.7	1.5	1.0

In order to fully interpret these results it is important to take the risk factor in context of "exposure." This would be the proportion of the population that are exposed to the risk allele: in the case of a dominant inheritance model this would be people carrying either one or two copies of the allele, in a recessive inheritance model the exposed group are only those carrying two copies as a single copy would not confer an increase in risk, or exposure. As the model is dominant and the genotype frequencies following Hardy Weinberg equilibrium, the exposure rates are higher than they appear, table 6.8 shows how the allele frequency relates to exposure rate in this case:

Table 6.8 Relationship between Allele Frequency and Exposure

Allele Frequency	0.1	0.25	0.5	0.75	0.9
Proportion Exposed	0.19	0.4375	0.75	0.9375	0.99

Taking the exposure rate into account, the level of identification being most successful at an allele frequency of 0.25 is unsurprising – it is the point at which the level of exposure is closest to 0.5. As the allele frequency and exposure increase past this point identification rate decreases and the level of false positives increases. The trend of identification is very similar to the rate at which the target SNP is identified following different environmental splits, where there is no main or interaction effect present (in this case smoking).

6.2.4 Case-control Ratio

In order to use the MTM in situations where the case-control ratio is not perfectly matched, a simulation was done using the parameters shown in Table 6.9. This will be important if the method is to be applied to data gathered from a cohort study, where the proportion of controls is normally higher than cases. The results are shown in the Table below, 6.10.

Table 6.9 Simulation Parameters – Case-Control Ratio

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	**Proportion Cases**	Inheritance Model
Smoking	1	0.510	175	1000	**0.25, 0.5, 0.75**	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

Table 6.10 Percentage of Datasets Identifying Interaction for Different Case-Control Ratios

Significance Threshold	0.01			0.001			0.0005			0.0001		
Splitting on Target Environmental Variable (NSAIDs)												
Cases (%)	**25**	**50**	**75**	**25**	**50**	**75**	**25**	**50**	**75**	**25**	**50**	**75**
Target Int	13.5	54.3	23.2	4.3	26.0	9.6	2.9	19.6	7.0	1.6	10.7	2.3
Target SNP	54.8	97.6	56.8	40.0	91.0	40.1	34.0	87.6	33.4	18.9	77.0	22.7
Neither Int or SNP	44.9	2.8	42.9	59.7	8.7	58.5	65.8	12.3	65.5	80.9	23.0	77.1
Other Int	21.6	21.0	20.2	4.3	4.4	4.1	2.4	1.9	2.2	0.7	0.2	0.7
Other SNP	15.6	3.1	14.6	4.6	0.7	4.1	2.9	0.3	2.4	0.7	0	0.3
Splitting on environmental variable with no effect (smoking)												
Target SNP	21.9	66.4	18.4	12.3	44.7	7.8	10.5	37.5	6.2	4.6	24.0	2.6
Target SNP and smoking	0.3	1.0	1.3	0	0	0.1	0	0	0	0	0	0
Other Int	26.8	25.3	25.8	5.1	7.2	5.9	3.0	3.9	3.6	0.8	1.2	1.1
Other SNP	26.8	16.5	25.9	7.3	5.4	7.0	4.6	3.0	4.4	1.1	1.1	1.4

These results suggest that a 50:50 split, or close to it, gives the best possible level of identification, this is not a surprising result. It is reassuring that identification of the SNP is still fairly high in the lower ratios as this suggests some flexibility towards cohort data.

6.3 Underlying Data Structure

There are some features of the data that are not obvious and are only understood after analysis; they are in fact the reason to do analysis in the first place. These variables include the presence of a main SNP effect, the size of the interaction effect or association, the inheritance model of the SNP and the underlying interaction model. There is also the interplay between these factors to consider.

6.3.1 Effect Size

The size of the underlying effect is in the form of an Odds Ratio (OR), the odds of being a case in the exposed group to the odds of it occurring in the unexposed. Therefore, the variability of identification for different effect sizes may be influenced by the sample size. Following the basic simulation of differing effect size, combinations of effect size and sample size will also be analysed. To study effect size, an OR of between 1.5 and 3.5 is assigned to the interactions term using the parameters shown in table 6.11 The results are shown in table 6.12.

Table 6.11 Simulation Parameters – Effect Size

Risk Effects Present	**Effect Size (OR)**	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Smoking	1	0.510	175	1000	0.5	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	**1.5, 2.0, 2.5, 3.0, 3.5**	0.4421				

Table 6.12 Percentage of Datasets Identifying Interaction across Different Effect Sizes

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Effect Size (OR)	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**
Target Int	4.3	26.9	54.3	76.2	87.6	1.4	9.3	26.0	48.4	68.1	0.9	7.3	19.6	40.3	59.3	0.3	3.6	10.7	24.8	43.6
Target SNP	18.8	75.1	97.2	99.8	100	7.5	57.6	91.0	99.3	99.8	5.5	50.4	87.6	98.7	99.8	3.0	33.5	77.0	95.6	99.6
Neither Int or SNP	80.6	24.6	2.8	0.2	0	92.1	42.0	8.7	0.7	0.2	94.3	49.1	12.3	1.3	0.2	97.0	66.2	23.0	4.4	0.4
Other Int	26.1	23.7	21.0	19.7	19.3	4.9	4.9	4.4	5.4	4.1	2.7	2.7	1.9	2.6	2.0	0.4	0.2	0.2	0.5	0.4
Other SNP	27.7	8.8	3.1	1.7	1.3	8.6	3.2	0.7	0.4	0.2	5.1	2.5	0.3	0.2	0.2	1.2	0.8	0	0.1	0
Splitting on environmental variable with no effect (smoking)																				
Target SNP	6.0	32.4	66.4	89.8	96.5	3.0	17.0	44.7	72.3	88.4	2.1	13.7	37.5	64.5	83.1	0.6	7.9	24.0	49.4	71.4
Target SNP and smoking	0.8	1.0	1.0	1.5	1.6	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
Other Int	24.3	27.8	25.3	26.4	25.8	6.3	6.6	7.2	5.6	6.9	3.5	3.9	3.9	3.4	3.9	0.7	0.8	0.9	1.1	1.2
Other SNP	33.6	23.3	16.5	14.3	17.1	10.8	7.5	5.4	3.8	4.9	6.4	4.4	3.0	2.7	3.1	1.4	0.8	0.5	0.8	0.7

These results show a number of interesting, if not unexpected, points. They repeat the pattern seen earlier of less false positives in the presence of real effects. As the effect size increased, the level of identification increased and the level of false positives decreased. This was true for the false positives suggesting an interaction or main SNP effect with the NSAIDs and for false positive SNPs identified under the smoking variable. However, the false positive rate of interactions between smoking and other SNPs was nearly constant across the different effect groupings for the target interaction.

As would be expected, as the effect size increased, so did the likelihood of identifying the target interaction, especially if both interactions and SNPs were considered. The identification held consistently across the significance thresholds, whereas most of the false positives did not.

Unexpectedly, the target SNP was identified at a very high level in the smoking analysis. This suggests that in the absence of main or interaction effects, the SNP which has an association when involved in an interaction is still more strongly associated than other variables when splitting on other environmental variables.

6.3.2 Effect Size for Independent Effects

In order to measure how effective the MTM was at identifying an independent SNP effect, a simulation was run with the parameters in table 6.13. Table 6.14 compares this result to the results when an equivalent interaction term is generated, but no main effect.

Table 6.13 Simulation Parameters – Independent Effect

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
NSAIDs	1	0.6664	175	1000	0.5	Dominant
Target SNP	**1.5, 2.0, 2.5, 3.0, 3.5**	0.6634				
Interaction	1	0.4421				

Table 6.14 Percentage of Datasets Identifying Interaction across Different Effect Sizes for a Main Effect versus an Interacting Effect

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Main SNP Effect																				
Effect Size	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**
Target SNP	20.7	80.3	99.0	99.7	100	9.0	59.6	92.6	99.3	99.8	7.1	52.7	89.4	98.9	99.8	3.8	35.3	77.8	96.0	99.6
Target Int (False +ive)	1.7	2.0	2.1	1.9	1.9	0.2	0.2	0.2	0.3	0.2	0	0.2	0.1	0.1	0.2	0	0.1	0	0	0.1
Neither Int or SNP	78.4	19.0	0.7	0.1	0	90.9	40.3	7.4	0.6	0.2	92.9	47.2	10.6	1.1	0.2	96.2	64.7	22.2	4.0	0.4
Other Int	25.3	23.9	26.0	26.6	25.4	6.4	5.8	5.4	6.5	6.0	4.2	3.5	3.4	3.8	3.8	1.6	1.5	1.2	1.1	0.4
Other SNP	27.9	20.7	27.0	29.5	32.0	7.7	5.5	7.6	8.8	9.8	4.3	3.5	4.8	5.4	5.5	1.1	0.9	1.0	1.2	1.3
Interaction Effect																				
Target Int	4.3	26.9	54.3	76.2	87.6	1.4	9.3	26.0	48.4	68.1	0.9	7.3	19.6	40.3	59.3	0.3	3.6	10.7	24.8	43.6
Target SNP	18.8	75.1	97.2	99.8	100	7.5	57.6	91.0	99.3	99.8	5.5	50.4	87.6	98.7	99.8	3.0	33.5	77.0	95.6	99.6
Neither Int or SNP	80.6	24.6	2.8	0.2	0	92.1	42.0	8.7	0.7	0.2	94.3	49.1	12.3	1.3	0.2	97.0	66.2	23.0	4.4	0.4
Other Int	26.1	23.7	21.0	19.7	19.3	4.9	4.9	4.4	5.4	4.1	2.7	2.7	1.9	2.6	2.0	0.4	0.2	0.2	0.5	0.4
Other SNP	27.7	8.8	3.1	1.7	1.3	8.6	3.2	0.7	0.4	0.2	5.1	2.5	0.3	0.2	0.2	1.2	0.8	0	0.1	0

These results show that the MTM can successfully identify a SNP main effect at a level of 99% at an effect size 2.5, continuing to find the effect across an increasingly strict significance threshold. The presence of a main SNP effect is only identified as an interaction with the splitting variable at a very low level, which is not affected by the identification of the main SNP effect. The false positive rate of SNP identification is also unaffected by the presence and identification of an independent SNP effect.

Conversely, the presence of an interaction effect identifies the interacting SNP as a main effect at a high level, increasing as identification of the interaction increases. This analysis shows that the presence of main SNP affects do not adversely affect the level of false positive identification and removing them from the data and repeating the analysis could expect a similar level of false positives to that of a data set with no main effects at all.

6.3.3 Effect Size by Sample Size

To assess the difference in identification rate across different combinations of sample size and effect size, simulations were run on sample containing 500 and 2000 people to compare against the results for 1000, for effects sizes 1.5, 2.0, 2.5, 3.0 and 3.5. The results, at significance level 0.01 are shown in Figure 6.1, the results are recorded in more detail in Appendix 6.2.

These results show that the identification increases as the sample size increases and as the effect size increases. For all effect size and sample size combinations, the interacting SNP is identified more often on its own than as an interaction.

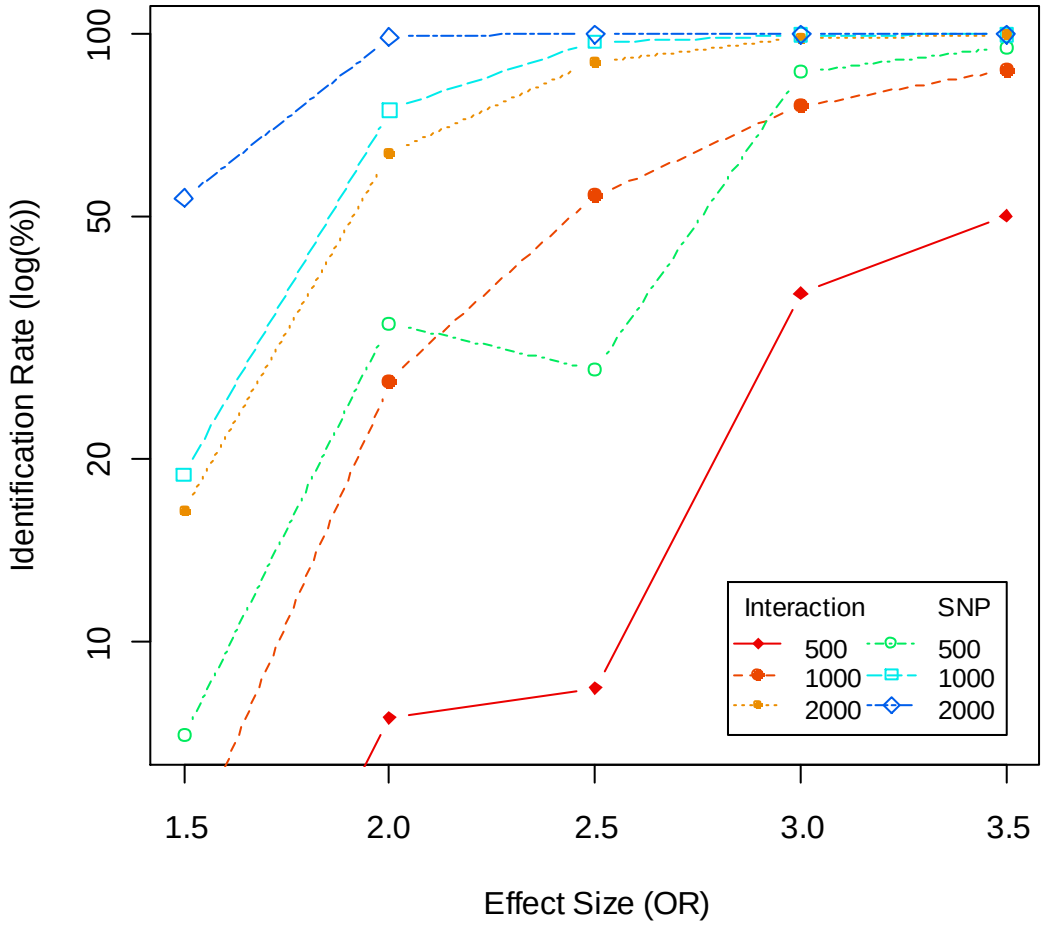


Figure 6.1 **Level of Identification at Different Effect and Sample Sizes**

6.3.4 Inheritance Model and Allele Frequency

Having looked at the effect of allele frequency under a dominant inheritance model (table 6.7), the same allele frequencies were used to assess the successful identification of the recessive and co-dominant models. The co-dominant model is slightly different to analyse than the dominant and recessive, in having two different risk levels: an increase in risk if heterozygous for the risk allele and a further increase if homozygous. The simulations to test the effect of allele frequency in a co-dominant model have an odds ratio of 1.5 for the heterozygotes, relative to the homozygous wildtype, rising to 2.5 for those homozygous for the risk allele.

The parameters of these simulations are shown in table 6.15, the results for recessive and co-dominant inheritance models can be found in tables6.16 and 6.17, respectively:

Table 6.15 Simulation Parameters – Allele Frequency and Inheritance Model

Risk Effects Present	Effect Size (OR)	**Risk Factor Frequency**					Number of SNPs	Sample Size	Proportion Cases	**Inheritance Model**
Smoking	1	0.510					175	1000	0.5	**Recessive**
NSAIDs	1	0.6664								
Target SNP	1	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**				
Interaction	2.5	**0.07**	**0.17**	**0.34**	**0.51**	**0.60**				
Smoking	1	0.510					175	1000	0.5	**Co-Dominant**
NSAIDs	1	0.6664								
Target SNP	1	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**				
Interaction	2.5	**0.07**	**0.17**	**0.34**	**0.51**	**0.60**				

Table 6.16 Percentage of Datasets Identifying Interaction at Different Allele Frequencies, under a Recessive Inheritance Model

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	0	1.3	37.2	68.7	42.3	0	0.1	14.7	38.7	20.5	0	0.1	11.0	30.9	17.1	0	0	5.3	20.1	9.0
Target SNP	0	6.3	88.8	98.5	72.2	0	3.1	73.7	97.4	72.0	0	2.4	66.0	96.4	70.9	0	1.1	51.2	91.5	65.0
Neither Int or SNP	100	93.5	10.9	1.5	27.4	100	96.9	26.1	2.6	28.1	100	97.6	33.5	3.5	29.2	100	98.9	48.3	8.3	35.1
Other Int	31.0	27.2	19.8	21.1	22.9	6.2	6.5	4.6	4.0	4.8	3.4	3.3	2.4	2.0	2.6	0.6	0.8	0.8	0.5	0.5
Other SNP	36.5	33.3	3.6	2.1	11.0	9.5	9.4	1.2	0.5	2.9	5.5	5.9	0.8	0.2	1.8	1.3	1.3	0.3	0	0.3
Splitting on environmental variable with no effect (smoking)																				
Target SNP	0	1.9	52.3	70.8	19.3	0	0.8	30.3	59.4	17.8	0	0.5	23.9	52.9	17.2	0	0	13.2	38.2	15.2
Target SNP and smoking	0.1	0.1	0.3	1.1	0.7	0	0	0	0	0.1	0	0	0	0	0.1	0	0	0	0	0
Other Int	29.6	25.4	26.6	23.9	27.7	6.5	5.6	4.8	5.0	6.7	3.8	3.1	2.8	2.5	4.4	1.2	1.2	0.5	0.7	1.2
Other SNP	34.1	32.5	16.5	17.4	28.1	8.1	9.5	4.7	5.4	9.9	4.6	4.8	2.4	3.2	4.7	1.2	1.2	0.7	1.1	1.0

Table 6.17 Percentage of Datasets Identifying Interaction at Different Allele Frequencies, under a Co-Dominant Inheritance Model

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	16.3	41.0	31.4	0.2	0	7.5	20.4	13.4	0	0	5.7	16.1	10.6	0	0	2.2	7.9	4.8	0	0
Target SNP	27.1	74.6	70.9	1.7	0.1	27.1	72.4	62.1	0.7	0	26.8	70.1	56.0	0.4	0	25.3	60.0	40.0	0.2	0
Neither Int or SNP	72.9	25.4	29.0	98.3	99.9	72.9	27.5	37.5	99.3	100	73.2	29.8	43.6	99.6	100	74.5	39.7	59.8	99.8	100
Other Int	24.9	22.0	22.7	27.6	25.9	4.4	4.5	4.3	5.9	5.5	2.5	2.7	1.7	4.1	2.9	0.5	0.5	0.5	0.6	0.8
Other SNP	25.5	10.6	12.1	34.9	33.3	6.3	3.2	3.9	10.7	10.0	3.6	1.8	2.1	5.8	5.5	0.7	0.6	0.3	1.0	1.3
Splitting on environmental variable with no effect (smoking)																				
Target SNP	6.3	32.3	33.3	1.2	0	6.1	27.4	19.9	0.6	0	5.7	24.4	15.6	0.6	0	5.2	16.7	8.1	0	0
Target SNP and smoking	0.8	1.1	0.9	0.5	0	0.1	0.2	0.2	0.1	0	0	0.1	0.1	0.1	0	0	0	0	0	0
Other Int	29.2	30.5	26.6	29.5	29.4	6.4	6.1	5.7	6.5	6.2	4.1	3.2	3.3	3.3	3.5	0.6	0.8	1.0	1.0	0.7
Other SNP	30.5	25.7	26.6	35.2	35.8	9.0	7.9	5.8	9.8	11.3	5.1	5.0	3.4	5.1	6.4	1.2	1.8	0.8	0.8	1.3

Dominant and Recessive Inheritance Models

Although different in practice, the risk of a dominant or recessive allele is effectively contributing to the same variable in the mixed tree model – exposure to the "at risk" genotype. The exposed group will either be those with at least one copy of the gene or only those with both but theoretically the analysis is the same, comparing exposure status to outcome.

The identification of the interaction is lowest when the ratio of exposed and non-exposed is largest. In the dominance model, an allele frequency of 90% translates to 99% of the sample being in the exposed group. Similarly, a person in the recessive model an allele frequency of 10% has only a 1% chance of being exposed. So few exposed makes it very difficult to identify effects, even when actively looking for them using traditional methods. For example, even in an optimised regression model containing only the target SNP, the NSAIDs and their interaction, the level of identification was low. Table 6.18 show how this regression model compares to the mixed tree method results using all 175 SNPs when the target SNP has a frequency of 0.9 and dominant behaviour (therefore 99% of the population is exposed).

Table 6.18 Percentage of Datasets Identifying Interaction in an Optimum Regression and MTM for Low Exposure Frequencies

	Significance Threshold							
	0.01		0.001		0.0005		0.0001	
	Reg.	MTM	Reg.	MTM	Reg.	MTM	Reg.	MTM
Target Int	1.6	0.1	0.2	0	0	0.1	0	0
Target SNP	1.7	0	0.1	0	0	0	0	0
Neither Int or SNP	96.9	99.9	99.7	99.9	100	99.9	100	100

The analysis is most successful when the exposed group contain approximately half of the sample. For a dominant model this is most similar to an allele frequency of 25%, translating to 43.75% of the population carrying the risk allele as either a heterozygote or homozygote. In the recessive model, an allele frequency of 75% translates to 56.25% of the population being exposed. This goes some way to explaining the difference in results from the two models.

The Co-Dominant Model

The trends of identification for the co-dominant inheritance model are similar to the dominant model, this makes sense as the exposure group is the same just with varying levels of risk associated. The MTM does not estimate the size of risk, so cannot differentiate well between the co-dominant and dominant models.

The results are probably poorer for the co-dominant model because of the difficulty of assigning risk to small proportions of the sample. In effect the sample is being split in three before risk is being assigned, so the chance that one of the fractions will be too small to effectively represent the desired OR is higher.

Conclusions on Inheritance Model

For all three models, the most easily identified SNP has an exposure frequency as close to 0.5 as possible. Therefore, where possible, a literature search for information on inheritance model prior to analysis would help identify the best SNPs for each particular analysis (between 50-75% for recessive and 25-50% for dominant).

Similarly to the results from different SNP numbers, the target SNP is identified in the analysis containing other environmental variables and at a higher rate than the interaction effect. Again, the false error rate decreases when a real effect is present and is not affected in any other way by the inheritance model of the target SNP.

6.3.5 Interaction Model

To recap the definitions of interaction type from section 2.1.2.2:

Type A: Having certain genotypes increases the likelihood of being exposed to the environmental factor, which in turn increases the risk of disease.

Type B: The environmental factor has a main effect, increased in the presence of certain genotypes which themselves have no main effects

Type C: The genotype has a main effect, enhanced by the environmental factor which has no independent effect

Type D: Neither have a main effect, but together the gene and environmental factor confer an increase in risk

Type E: Both have main effect, the combination of which is higher than their additive combined risk

Data were generated for each of the interaction models above, with OR of 1.5 for the first line/main effects and 2.5 for the second, cumulative risk estimates. The simulation parameters are shown in Table 6.19, the results in table 6.20.

Table 6.19 Simulation Parameters – Interaction Model

Risk Effects Present	**Effect Size (OR)**					Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Interaction Type	**A ***	**B**	**C**	**D**	**E**					
Smoking	**1**	**1**	**1**	**1**	**1**	0.510	175	1000	0.5	Dominant
NSAIDs	**2.5**	**1.5**	**1**	**1**	**1.5**	0.6664				
Target SNP	**1.5**	**1**	**1.5**	**1**	**1.5**	0.6634				
Interaction	**1**	**2.5**	**2.5**	**2.5**	**2.5**	0.4421				

* The associated risk for the target SNP in interaction model A describes the increase on risk of taking NSAIDs, not an increased risk directly of case status.

Table 6.20 Percentage of Datasets Identifying Interaction for Different Interaction Models

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Interaction Model	**A**	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
Target Int	0.2	80.4	78.1	54.3	61.1	0	55.4	52.4	26.0	33.9	0	48.1	45.3	19.6	28.0	0	33.2	28.6	10.7	16.5
Target SNP	0.3	99.3	99.4	97.2	96.7	0	97.5	97.5	91.0	92.1	0	96.2	96.3	87.6	88.0	0	92.0	91.8	77.0	80.1
Neither Int or SNP	99.6	0.6	0.6	2.8	3.2	100	2.4	2.4	8.7	7.8	100	3.6	3.6	12.3	11.8	100	7.8	7.9	23.0	19.8
Other Int	26.4	17.2	16.4	21.0	18.7	5.7	4.0	4.1	4.4	4.0	3.3	2.0	1.8	1.9	1.4	0.8	0.2	0.1	0.2	0.3
Other SNP	35.6	2.5	2.0	3.1	2.3	8.8	0.5	0.4	0.7	0.7	4.7	0.2	0.2	0.3	0.4	0.5	0.1	0	0	0.2
Splitting on environmental variable with no effect (smoking)																				
Target SNP	18.5	74.4	77.4	66.4	65.2	9.1	56.9	59.6	44.7	47.0	7.0	49.8	54.0	37.5	40.3	2.6	33.2	38.0	24.0	27.0
Target SNP and smoking	1.0	1.6	1.4	1.0	1.2	0.1	0	0.1	0	0	0.1	0	0	0	0	0	0	0	0	0
Other Int	26.5	26.4	26.0	25.3	25.4	5.8	5.7	5.6	7.2	6.5	3.4	3.1	3.3	3.9	1.1	1.1	1.0	0.7	1.2	1.0
Other SNP	27.2	15.4	13.8	16.5	15.4	8.2	4.1	3.7	5.4	4.5	4.9	2.8	2.0	3.0	0.9	0.9	0.5	0.6	1.1	0.7

Conclusions on Interaction Type

As shown in the table, the interaction least detected – by quite some margin – was Type A, and the best Types B and C. Taking SNP identification into account as an identification, all four types other than A had very high levels of identification. As seen in all previous runs, in the presence of a real, identified effect, the level of false positive identification of both interactions and SNPs was lower.

It is debateable whether type A is an interaction at all, and not just a two step causal pathway, which might explain why interactions are not so easy to identify using the Mixed Tree Method (MTM). The important step that is not working well under this model is the initial splitting of sub-groups. When the initial sample is split based on the environmental variable the SNP types are also divided into non-equal groups, with more of the risk SNP genotypes in the groups positive for the risk factor. After this splitting step however, there is no actual difference between the risk SNP and other SNPs within the subgroup – the risk SNP only affects the likelihood of being in the subgroup in the first place. Therefore the poor results are not entirely unexpected.

For the other four interaction types the results are reassuring. The relatively high identification in Types B and C, especially B, is down to the effectiveness of the environmental split and the variation in SNP allele frequencies within the groups. In interaction Type D, the lower power may, in part be explained by the MTM less consistently splitting the groups into similarly sizes sub-groups, as there is no main environmental effect to make such a clear division. With larger sample sizes, the identification rates were higher. Chapter 7 will compare these results to those obtained using different analyitical methods to establish whether 70-80% identification can be considered high in such a context.

6.3.6 Interaction Model, Allele Frequency and Inheritance Model

As the allele frequency and inheritance model have a combined effect that can be explained by the level of exposure, analysis combining interaction model and allele frequency will suffice in suggesting the trend by exposure. Table 6.21 shows the parameters for these simulations:

Table 6.21 Simulation Parameters – Allele Frequency and Interaction Model

Risk Effects Present	**Effect Size (OR)**					**Risk Factor Frequency**	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Interaction Type	**A ***	**B**	**C**	**D**	**E**					
Smoking	**1**	**1**	**1**	**1**	**1**	0.510	175	1000	0.5	Dominant
NSAIDs	**2.5**	**1.5**	**1**	**1**	**1.5**	0.6664				
Target SNP	**1.5**	**1**	**1.5**	**1**	**1.5**	**0.1, 0.25, 0.5, 0.75, 0.9**				
Interaction	**1**	**2.5**	**2.5**	**2.5**	**2.5**	0.4421				

The results are displayed in Figure 6.2 and shown in table 6.22, for a significance level of 0.01. Multiple tables covering the results in more detail can be found in Appendix 6.3. Interaction Model A was almost wholly unsuccessful, with the results similar to what would be expected by chance, it has therefore not been included in the graph.

As an interaction can be identified either as an interaction effect or as a main SNP effect for a single environmental variable, figure 6.2 displays the results as both looking for interaction effects only and looking for the proportion of time the vairable is identified as either an interaction or a main SNP effect for that enironmental variable.

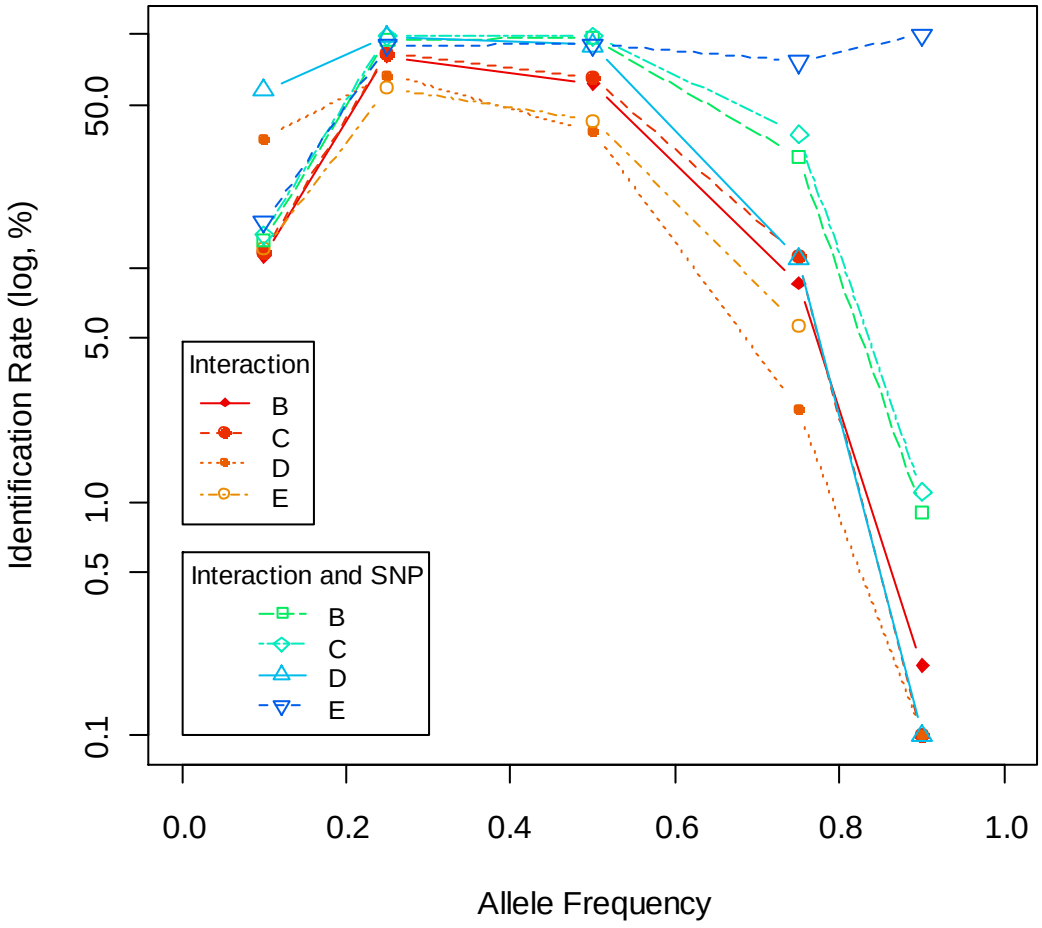


Figure 6.2 Interaction Model and Allele Frequency

Table 6.22 Percentage of Datasets Identifying Interaction for Different Interaction Models and Allele Frequencies

	Interaction Model																			
	B					**C**					**D**					**E**				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	11.2	80.8	61.8	8.6	0.2	11.6	83.3	65.7	11.2	0.1	16.3	41.0	31.4	0.2	0	12.1	59.2	42.9	5.7	0
Target SNP	13.2	94.8	96.6	29.0	0.7	13.9	98.6	98.2	35.7	1.0	27.1	74.6	70.9	1.7	0.1	15.8	89.9	90.6	23.5	0
Neither Int or SNP	86.8	5.2	3.4	70.1	99.1	86.1	1.4	1.6	62.6	98.9	72.9	25.4	29.0	98.3	99.9	84.2	10.1	9.3	75.8	100
Other Int	26.1	17.0	18.9	24.8	27.1	17.7	11.9	12.7	10.6	17.0	24.9	22.0	22.7	27.6	25.9	28.5	20.1	19.6	23.6	28.5
Other SNP	31.7	3.5	3.4	21.2	37.9	18.0	1.4	2.7	10.4	20.5	25.5	10.6	12.1	34.9	33.3	29.8	4.3	4.8	21.8	38.0
Splitting on environmental variable with no effect (smoking)																				
Target SNP	0.8	48.9	63.0	8.4	0.2	1.2	64.8	73.2	11.6	0.1	6.3	32.3	33.3	1.2	0	2.0	47.2	55.1	7.1	0.2
Target SNP and smoking	0.1	0.9	1.6	1.0	0.3	0.1	0.5	1.1	0.9	0.4	0.8	1.1	0.9	0.5	0	0.2	1.6	1.3	0.6	0.2
Other Int	30.4	25.7	28.3	28.0	30.1	19.3	18.0	20.0	18.3	19.8	29.2	30.5	26.6	29.5	29.4	30.0	25.2	25.3	29.0	28.6
Other SNP	34.1	21.9	18.3	32.2	35.1	21.2	12.1	10.9	15.3	18.6	30.5	25.7	26.6	35.2	35.8	31.9	22.1	17.2	31.5	37.4

There were no great differences between interaction models when measuring the differences over different allele frequencies. The same general pattern applied for all the different interaction types, with the exception of type A. For all the interaction types, the power of identification was higher as the exposure neared 50% (in the dominant inheritance model used this is best represented by an allele frequency of 0.25).

At the lower frequencies, so an 0.1 allele frequency corresponding to 19% exposure, interaction type D, where neither the gene nor environmental variable have a main effect but combined have an interaction effect, performed slightly better than the other interaction types. This could be an artefact from the data generation as effectively only one risk estimate is being calculated and added to a single at risk group. Splitting the data effectively into two different at risk groups prior to adding the risk effect is less efficient at the lower frequencies as the at risk in each group will be very small.

Surprisingly, the type D interactions had the lowest rate of the target SNP being identified when the root split was smoking. Unsurprisingly, type C, where there was a genetic main effect compounded by the environmental variable, had the highest rate.

Like all the other simulations, the level of false positive identification for interaction models B-D decreased when real effects were detected for both interaction and main SNP effects.

6.4 Ranking of Results

One of the main issues is how to move from these results to a method of identifying potential interaction in real data. As the analysis is carried out on a pre-selected SNP panel, then undergoes a number of parameter decreasing stages before being entered into a regression model, it would be statistical bad practice to take the p-value estimate as true and conclusive, without any correction or context.

Therefore, in order to identify the most plausible hypothesis from these results it is necessary to develop a hierarchy of evidence and rank them accordingly. The following criteria are a summation of the results from this chapter and the theoretical basis of how this method will be applied to real data in a biological context.

Strong Evidence of a Main Effect

- Identified as an important variable in both subtrees of a single analysis and across the majority of environmental variables
- In most cases a p-value for the independent effect below 0.01

Strong Evidence of an Interaction Effect

- Identified as an interaction term following the logistic regression step
- Highly significant p-value, below 0.01 or 0.05 depending on sample size and other p-values obtained
- Biological Plausibility
- Replicated in Other Work

Intermediate Evidence of an Interaction

- Identified as an interaction with a borderline effect

OR

- Identified as a main SNP, which was significant and only found at the top of one sub-tree for a single environmental variable
- Biologically Plausible

Counter Evidence of an Interaction

- If further analysis reveals the genotype selected to be the heterozygote, whilst grouping the homozygotes together – this is not biologically plausible
- Identified more than once throughout the full analysis, at low significance levels
- No prior evidence base

6.5 Discussion

From the analyses presented in this chapter there were a number of common observations: the proportion of false positive results decreased in the presence of real interaction effects, decreasing as the level of successful identification increased; the real effects were more likely to be identified consistently across the different significance thresholds, especially if the odds ratio is high or the sample size large; the SNP involved in an interaction with one environmental variable was still identified frequently as an independent effect for other variables. This last point may be an artefact from the lack of any other marginal effects but may suggest a good way to identify potential interactions: look for important SNPs across all variables then look at the interaction identification rates for that SNP. However, further work would be required to analyse combinations of effects before such an approach could be validated.

The number of SNPs did not affect either the identification of real interactions or of false positives, but did require slightly different settings for very low SNP numbers, which can be easily adjusted. As regression is the significance determining step, it was not surprising that the significance level, effect size and sample size had interplay with similar parameters to regression – improving at higher sample and effect sizes. In fact at the lower effect sizes (OR =1.5) the identification was ten times higher with a population of 2000 than 500. The larger sample sizes showed increased identification but not a higher false positive rate.

The model of inheritance was only relevant in dictating the exposure vs. non exposure groups. In real analysis, where the inheritance model is known, the SNP selected should split the sample as close to 50% exposed as possible. Therefore, if the gene itself is known to have a dominant inheritance pattern, a SNP of a lower frequency, approximately 25-50% would be preferable to that of a higher frequency. Where unknown, the 50:50 SNP frequencies would be recommended.

Four out of five interaction types were consistently identified; with Type A interactions not being picked up. It is debatable whether this kind of interdependence

can even be classed as an interaction, it does not fit the criteria for interaction with regards to statistical theory. Other interaction types were identified well (power between 0.75 and 0.95) with the interaction more often being identified as a SNP for all interaction types, especially at higher sample sizes – this suggests that it is the numbers in the sub groups limiting identification when applicable.

The data from these simulations are based on a case-control study: an interesting area of future research would compare results across different study designs, particularly founder populations, where the people under study have common ancestry. In cases where the population contains a high level of relatedness, there would be stretches of DNA inherited together and therefore SNPs associated with one another. This would have to be considered during the SNP selection process but would not in fact inhibit the identification of a key SNP, especially if the method was being used to generate hypotheses – in which circumstances position of SNPs would be seen and the explanation become clear. The random forest procedure (as currently implemented in R) allows settings for relatedness between the variables. With careful planning, there is no methodological reason the mixed tree method would not work on founder populations – however, it is an area for further research.

It is important, before starting analysis by MTM, to look carefully at the dataset and establish the best way of entering the data for each of the environmental variables. In a number of cases, particularly those with unusual or extreme measures, this will involve some preliminary data preparation. However, the method has flexibility and can be adapted to overcome a number of problems.

An important question is: how generalisable are the results gained from the MTM? They are hypothesis generating, so not directly applicable to a population and particularly not to individuals within that population. However, used within a larger framework and investigation they could prove to be a crucial ingredient to a new arm of gene-environment research.

At high sample sizes the results were very good, with close to 100% identification (depending on which measure is used) with no corresponding increase in false positives. These results are particularly reassuring given that the sample analysed in

Chapter 8 contains a similar number of participants (1216). However, MTM, like most statistical methods, would not perform well on a small sample size with a large number of potential explanatory variables. The results on a sample of only 200, were poor and it would not be a recommended sample size for MTM analysis. Between 500 and 1000 participants in total would be the recommended sample size, with the number of cases and controls roughly equivalent.

Chapter 7

Alternative Analytical Methods: How Their Performance Compares to the Mixed Tree Method

7.1 Introduction

To give a wider context to the results from the mixed tree method analysis the same data were analysed using other statistical methods that have shown some success in identifying interactions. These methods were discussed in more detail in Chapter 2, where the methods considered most successful were logistic regression, regression using stepwise selection, Multifactor Dimensionality Reduction (MDR) and the extension to MDR to give a quantitative measure of association, Odds Ratio MDR (ORMDR). Neural Networks (NN) provide a useful diagnostic tool for complex diseases where interactions are present but are not directly comparable as they do not select individual interacting effects and have a tendency to overfit the data.

7.1.1 Preparing and recoding the data

The datasets from the previous chapter required no adaptations for analysis using logistic regression or regression using stepwise selection. There is no missing data in the simulated datasets and in situations where transformation would aid the analysis, this will be accommodated. The environmental variables were analysed sequentially – analysing NSAIDs then smoking to represent a "real" effect and a false identification rate, respectively. Analysing the interactions between environmental variables is not a main objective for this study and would complicate the regression analysis, giving it an unfair disadvantage.

A number of the covariates were dichotomised prior to modelling analysis, which allows a more direct comparison with the MTM and MDR, which either performed better or was restricted to binary covariates, respectively. Furthermore, it has been claimed that, from the clinical point of view, binary covariates might be preferred because they offer a simple risk classification into high versus low, assist in making treatment recommendations, and in setting diagnostic criteria [466]. On the other hand, dichotomisation results in loss of information and power, if a linear rather than threshold association pertains, and non-linear relationships such as U shape associations will not be detected [467, 468]. This must be considered, especially for any

analysis identifying an interaction between a SNP and a continuous variable. However in this dataset all the variables were normally distributed, either naturally or following logarithmic transformation, so such problems were minimised to some degree.

In order to use the MDR/ORMDR function, all the variables (both genotypic and phenotypic) under analysis had to be parameterised as “0”, “1” or “2”. Therefore, in simulating binary data, the levels were replaced with “0” and “1” but the same set seed used as before, so the actual data content remained unchanged. Continuous variables were generated as such and categorised based on commonly recognised limits, such as overweight having a BMI greater then 25 or above/below recommended retinol intake. It was necessary to generate and re-classify to ensure the data was the same as the MTM simulations, generating into two groups would have used the random number generator differently and there may have been differences introduced between the datasets by chance.

The genetic variables were recoded so that “0” referred to the homozygous combinations considered to be the wild type, or reference allele. “1” was the heterozygote and “2” was the homozygous mutant. The allele selected as the reference allele was dictated by its status as such in the HapMap database.

7.2 Methods

In order to assess how well the MTM compares to other statistical methods in the search for gene-environment interactions, the data simulations were repeated and analysed using the other methods. The results, both successful identification and false positive identification, are compared quantitatively and other factors, such as computational intensity and practicality, discussed. In each case the success is measured by the identification of the interaction and the false positive results, unless otherwise stated.

7.2.1 Logistic regression

The generalised linear model function from R [469] was used with the setting for binomial regression and forward selection. The case status was the binomial outcome variable with the explanatory variables including all the SNPs, an environmental variable and the possible gene-environment interactions. The model was such that gene-gene interactions were not considered, to minimise the number of variables where possible, although this in itself would be unusual for such analysis.

7.2.2 Logistic Regression using Stepwise Selection

Stepwise regression was carried out using the step function from the MASS library in R, with the stepAIC command using the parameter: direction = “both”. AIC (Akaike Information Criterion) is a goodness of fit measure that favours smaller residual error in the model, but penalises larger models with equally good fit and helps avoiding overfitting. Similarly, the outcome was case control status and each environmental variable analysed separately, alongside all the SNPs and potential interactions.

7.2.3 MDR

Once the data had been generated and recoded accordingly, the data was analysed using the ORMDR function in R [470], the first step of which is the MDR analysis. As MDR chooses the best combination of risk factors, normally genotypes, and then classifies these combinations into low or high risk sub groups, the measure of success used is different. It looks for the combination with the lowest misclassification rate from all the possible combinations, using cross validation and training set to identify the combination and a test set to calculate the prediction error. Therefore success is measured by identification rate and misclassification rate.

The specific parameters used include: ten fold cross validation; the cross validated sets being randomised; the maximum number of genes in combination set to two. In some cases, when the numbers of cases and controls is not equivalent, it is necessary to balance the proportion of cases and controls expected. This makes MDR robust to class imbalance [471]. The simulated data has balanced data, it is not necessary to adjust the threshold used for assigning high and low risk labels.

7.2.4 ORMDR

ORMDR is effectively an extension of MDR that calculates risk estimates and confidence intervals for these estimates. Although the ORMDR odds ratio estimator is different, technically, from the standard estimator – it can be applied in the same way with significant results having a confidence interval that does not include the value 1. As well as classifying groups as high or low risk, the confidence intervals supplied by ORMDR indicate the significance and size of this classification [193].

Similarly to the Mixed Tree Method, the quantitative measure of significance allows the results to be evaluated at different significance thresholds. This is a better direct comparison than just selecting the best interaction, as basic MDR does.

7.3 Results

The literature suggested that large numbers of SNPs within a given dataset would represent the most obvious challenge for the above methods. Whereas effect size, sample size and allele frequency will affect accuracy, a large number of variables is outwith the recommendations for use. For this reason, success over varying SNP numbers was the first parameter examined; the other parameters are shown alongside in table 7.1.

Table 7.1 Simulation Parameters – SNP Numbers

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	**Number of SNPs**	Sample Size	Proportion Cases	Inheritance Model
Smoking	1	0.510	**50, 100, 175, 300**	1000	0.5	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

The results are presented in the same way, with the same measures of success as Chapter 6: the target interaction involves the environmental variable and SNP that have been assigned an interacting effect. A false positive interaction for an environmental variable involves a different SNP, a false positive SNP main effect includes any SNP except the target SNP.

7.3.1 Logistic Regression

The standard number of SNPs for comparing the other parameters, 175, was not well tolerated using logistic regression with the number of parameters preventing convergence. The results from analysis on 50 and 100 SNPs are shown below, with a comparison to the results from Chapter 6 using the MTM at a significance level of 0.01. It is worth noting that the recommended Event Per Variable (EPV) ratio is usually considered to be 10:1, with some estimates as high as 20:1, so these analysis are often outside the recommendations for using logistic regression.

Table 7.2 Percentage of Datasets Identifying Interaction at low SNP numbers for Logistic Regression versus MTM

Number of SNPs	**50**	**50**	**100**	**100**
Method	Regression	MTM	Regression	MTM
Splitting on Target Environmental Variable (NSAIDs)				
Interaction (Int) found	53.5	59.3	40.3	54.8
SNP found	98.2	97.6	97.7	97.6
Neither Int or SNP found	1.8	2.4	2.2	2.8
NSAIDS/ other SNP	81.9	13.2	304.8	17.3
False SNP	65.3	2.1	172.8	2.2
Splitting on environmental variable with no effect (smoking)				
SNP still identified	77.2	75.3	77.9	73.6
SNP/smoking interaction	1.4	1.2	9.0	1.3
Both SNP and Env wrong	94.4	17.1	578.1	19.4
False SNP found	63.2	8.0	162.5	13.9

As can be seen, the sensitivity of both methods to detect the interaction is similar for across both samples. However, the level of false positive identification for both the false interaction and the false main SNP effects is considerably higher for logistic regression and increases with the number of SNPs under analysis. With regard to computational intensity, logistic regression was the faster method for 50 SNPs, but considerably slower and prone to computational problems regarding convergence when the data contained 100 SNPs. To counter the increase in computational problems, slight adaptations were made to the regression parameters, specifically an increase in the number of iterations.

When the sample contained 175 SNPs, logistic regression was time consuming and generated a large number of errors. The glm algorithm did not converge for any of the datasets, and in some cases the probabilities 0 and 1 occurred (for one dataset every single SNP had an interaction with a significance level of 0). The glm settings were adapted, considering up to 10,000 iterations and increasing the positive convergence tolerance to 0.01, which despite the increased computation, left one dataset unable to produce any results.

The results do suggest, however, that 175 SNPs and their interaction with the environmental variable has a very high false positive rate. For example, after

removing the most obvious problematic results gave, for 1000 simulations, a false detection rate per simulation for interaction involving NSAIDs of 2780.8%. It was also impossible, computationally, to run 1000 populations in one analysis – it was necessary to run the simulation in four sets of 250 replications.

Although not shown, at 175 SNPs the detection rate for the interaction dropped to 17.9% for regression. Therefore, for the standard analysis of 175 SNPs, logistic regression alone does not work as effectively nor as easily, both computationally and in terms of power, as the Mixed Tree Method. It is also not possible to compare logistic regression to the range of other parameters compared in Chapter 6, as the standard number of SNPs for the other simulations is 175.

In conclusion, logistic regression is not an appropriate comparison method for this study, nor an appropriate analytical method for this type of problem as:

- The false positive rate is too high
- Non convergence

7.3.2 Logistic Regression Using Stepwise Selection

7.3.2.1 Number of SNPs

The same parameters were used in this simulation as for logistic regression, as shown in Table 7.3. Similarly to logistic regression, the computational time was extensive with a single run (NSAIDs and SNPs) containing 175 SNPs taking 110 hours, as opposed to approximately 8 hours for MTM. The results are shown below in table 7.3 for a significance level of 0.01.

Table 7.3 Percentage of Datasets Identifying Interaction for Different SNP Numbers, Comparing Stepwise Regression and MTM

	Stepwise Regression				MTM			
Number of SNPs	**50**	**100**	**175**	**300**	**50**	**100**	**175**	**300**
Splitting on Target Environmental Variable (NSAIDs)								
Target Int	58.0	57.8	55.8	58.7	59.3	54.8	54.3	56.4
Target SNP	97.9	98.0	97.7	98.5	97.6	97.6	97.2	98.3
Neither Int or SNP	2.1	2.0	2.3	1.5	2.4	2.8	2.8	1.7
Other Int	10.9	27.4	75.4	71.9	13.2	17.3	21.0	19.7
Other SNP	82.9	206.4	475.7	471.9	2.1	2.2	3.1	2.0
Splitting on environmental variable with no effect (smoking)								
Target SNP	89.4	90.2	88.4	89.7	75.3	73.6	66.4	68.9
Target SNP and smoking	0	0.2	0.2	0.4	1.2	1.3	1.0	1.1
Other Int	1.9	3.9	13.0	11.3	17.1	19.4	25.3	24.2
Other SNP	61.8	150.9	359.0	345.7	8.0	13.9	16.5	14.8

As the results show, the level of identification is very similar for stepwise regression as MTM, with a similar preference in detecting SNP main effects over interaction effects. However, the false positive identification rate is considerably higher for stepwise regression compared to the MTM, increasing as the number of SNPs increases. However, the false positive SNP identification is similar for 175 SNPs and 300, which is an improvement on logistic regression which increased in a more linear fashion. Given the computational time and false positive rate, there is no advantage to using stepwise regression in place of the MTM as the SNP numbers increase.

7.3.2.2 Sample Size

Considering the computational burden of the previous simulations, each set of parameters for simulations examining sample size and subsequent variables were run

100 times instead of 1000. The features of the simulations used to evaluate the identification at different sample sizes are shown in table 7.4:

Table 7.4 Simulation parameters - Sample Size

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	**Sample Size**	Proportion Cases	Inheritance Model
Smoking	1	0.510	175	**200, 500, 1000, 2000, 3000**	0.5	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

The populations containing only 200 and 500 members were too small to run stepwise selection effectively. Outside the recommended case to explanatory variable ratio (10:1), the probabilities 0 and 1 occurred for a large number of the simulations. The results for 1000, 2000 and 3000, at a significance level of 0.01, are shown in the table below, with the MTM results rounded up to facilitate comparison across the different number of replications. For ease of comparison, all of the remaining analyses are presented in this way, showing the results at the significance threshold of 0.01.

Table 7.5 Percentage of Datasets Identifying Interaction for Different Sample Sizes, Comparing Stepwise Regression and MTM

	Stepwise Regression			MTM		
Sample size	**1000**	**2000**	**3000**	**1000**	**2000**	**3000**
Splitting on Target Environmental Variable (NSAIDs)						
Target Int	56	90	98	54	90	99
Target SNP	98	100	100	97	100	100
Neither Int or SNP	2	0	0	3	0	0
Other Int	75	61	39	21	21	21
Other SNP	476	352	297	3	2	1
Splitting on environmental variable with no effect (smoking)						
Target SNP	88	100	100	66	97	100
Target SNP and smoking	0	1	0	1	2	2
Other Int	13	13	4	25	25	26
Other SNP	359	248	220	17	18	20

For both stepwise regression and the mixed tree method, identification increased as the sample size increased, with both methods performing equally well. The false positive rate was considerably higher using stepwise regression compared to MTM, with the exception of the identification of interacting effects with both variables incorrect. This suggests the MTM has a preference for interaction effects, as was intended. The SNP was more often identified than the interaction for both methods, across all sample sizes.

7.3.2.3 Allele Frequency

In order to compare the level of identification across different allele frequencies, a similar simulation to that in chapter 6 was run, except for 100 runs in place of 1000. The parameters are shown below and the results in table 7.7.

Table 7.6 Simulation Parameters – Allele Frequency

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency					Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Smoking	1	0.510					175	1000	0.5	Dominant
NSAIDs	1	0.6664								
Target SNP	1	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**				
Interaction	2.5	**0.07**	**0.17**	**0.34**	**0.51**	**0.60**				

Table 7.7 Percentage of Datasets Identifying Interactions for Different Allele Frequencies, Comparing Stepwise Regression and MTM

	Stepwise Regression					**MTM**				
Allele Frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Splitting on Target Environmental Variable (NSAIDs)										
Target Int	25	65	42	3	1	36	67	39	3	0
Target SNP	88	100	89	22	2	58	98	90	10	0
Neither Int or SNP	12	0	11	78	97	42	2	10	89	100
Other Int	56	67	66	81	74	25	23	22	27	5
Other SNP	463	423	467	484	478	18	2	4	32	35
Splitting on environmental variable with no effect (smoking)										
Target SNP	77	94	80	13	4	15	71	53	2	0
Target SNP and smoking	0	0	0	0	0	1	1	1	0	0
Other Int	13	14	13	16	14	29	25	23	28	26
Other SNP	385	347	334	378	356	29	18	17	37	38

As the results show, both methods were most effective in detecting the target interaction, for a dominant inheritance model, when the allele frequency was approximately 25%. The results for successful identification were similar in the two approaches, but the stepwise regression had a much higher rate of false positive identification, for both interaction and SNP main effects. The degree of false positive identification for stepwise regression did not appear to be influenced by the presence

or identification of real effects, unlike MTM where the level of false positive identification decreased when a real effect was identified. As with the sample size comparison, stepwise regression had a much higher level of identification, both real and false, for SNPs compared to interaction effects. MTM had a higher false positive rate of interaction identification than main effects.

7.3.2.4 Case Control Ratio

To assess the difference in results between stepwise regression and MTM for different case-control ratios, a simulation of the following parameters was run.

Table 7.8 Simulation Parameters – Case-control Ratio

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency	Number of SNPs	Sample Size	**Proportion Cases**	Inheritance Model
Smoking	1	0.510	175	1000	**0.25, 0.5, 0.75**	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	2.5	0.4421				

Table 7.9 Percentage of Datasets Identifying Interaction for Different Case-control Ratios, Comparing Stepwise Regression and MTM

	Stepwise Regression			MTM		
Proportion Cases	**0.25**	**0.5**	**0.75**	**0.25**	**0.5**	**0.75**
Splitting on Target Environmental Variable (NSAIDs)						
Target Int	23	56	27	14	54	23
Target SNP	76	98	73	55	98	57
Neither Int or SNP	24	2	27	45	3	43
Other Int	100	75	69	22	21	20
Other SNP	562	476	493	16	3	15
Splitting on environmental variable with no effect (smoking)						
Target SNP	56	88	53	21.9	66.4	18.4
Target SNP and smoking	1	0	2	0.3	1.0	1.3
Other Int	18	13	9	26.8	25.3	25.8
Other SNP	415	359	354	26.8	16.5	25.9

Both the stepwise regression and the MTM performed best when the proportion of cases and controls was similar. The false positive rate was higher for both methods when the sample was unbalanced towards either cases or controls.

7.3.2.5 Effect Size

Using the same parameters as chapter 6, shown again in table 7.10, a number of different populations were simulated with the effect size of the interactions varying from 1.5-3.5 (OR). The results are shown in table 7.11.

Table 7.10 Simulation Parameters – Effect Size

Risk Effects Present	**Effect Size (OR)**	Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Smoking	1	0.510	175	1000	0.5	Dominant
NSAIDs	1	0.6664				
Target SNP	1	0.6634				
Interaction	**1.5, 2.0, 2.5, 3.0, 3.5**	0.4421				

Table 7.11 Percentage of Datasets Identifying Interaction for Different Effect Sizes, Comparing Stepwise Regression and MTM

	Stepwise Regression					MTM				
Effect Size	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**
Splitting on Target Environmental Variable (NSAIDs)										
Target Int	2	27	56	76	84	4	27	54	76	88
Target SNP	35	83	98	100	100	19	75	97	100	100
Neither Int or SNP	65	17	2	0	0	81	25	3	0	0
Other Int	46	62	75	62	68	26	24	21	20	19
Other SNP	442	440	476	453	457	28	9	3	2	1
Splitting on environmental variable with no effect (smoking)										
Target SNP	23	71	88	97	99	6	32	66	90	97
Target SNP and smoking	0	0	0	1	1	1	1	1	2	2
Other Int	7	3	13	8	11	24	28	25	26	26
Other SNP	334	333	359	343	323	34	23	17	14	17

Unsurprisingly, the level of successful identification increased as the effect size increased, the false positive rate was unaffected. Successful identification for the interaction effect was very similar to that of the MTM, but again the false positive rate was considerably higher. The interacting SNP was identified more often in the smoking analysis for stepwise regression.

7.3.2.6 Inheritance Model and Allele Frequency

In another repeat of the simulation in Chapter 6, the parameters shown below, data was simulated to represent populations with a different allele frequency for the risk (target) allele under different inheritance models. The results from a dominant model can be found in Table 7.7. The results for the recessive and co-dominant model are in Tables 7.13 and 7.14 respectively.

Table 7.12 Simulation Parameters – Allele Frequency and Inheritance Model

Risk Effects Present	Effect Size (OR)	Risk Factor Frequency					Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Smoking	1	0.510					175	1000	0.5	**Recessive**
NSAIDs	1	0.6664								
Target SNP	1	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**				
Interaction	2.5	**0.07**	**0.17**	**0.33**	**0.5**	**0.59**				
Smoking	1	0.510					175	1000	0.5	**Co-Dominant**
NSAIDs	1	0.6664								
Target SNP	1	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**				
Interaction	2.5	**0.07**	**0.17**	**0.33**	**0.5**	**0.59**				

Table 7.13 Percentage of Datasets Identifying Interaction for Different Allele Frequencies under a Recessive Inheritance Model

	Stepwise Regression					**MTM**				
Allele Frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Splitting on Target Environmental Variable (NSAIDs)										
Target Int	0	22	27	5	51	0	1	37	69	42
Target SNP	13	87	79	100	96	0	6	89	99	72
Neither Int or SNP	87	13	21	0	4	100	94	11	2	27
Other Int	18	56	42	67	81	31	27	20	21	23
Other SNP	380	499	385	444	487	37	33	4	2	11
Splitting on environmental variable with no effect (smoking)										
Target SNP	1	78	68	96	89	0	2	52	71	19
Target SNP and smoking	0	0	0	1	0	0	0	0	1	1
Other Int	9	24	18	10	12	30	25	27	24	28
Other SNP	380	427	342	334	363	34	33	17	17	28

Table 7.14 Percentage of Datasets Identifying Interaction for Different Allele Frequencies under a Co-Dominant Inheritance Model

	Stepwise Regression					MTM				
Allele Frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Splitting on Target Environmental Variable (NSAIDs)										
Target Int	15	39	30	1	0	16	41	31	0	0
Target SNP	71	91	88	12	7	27	75	71	2	0
Neither Int or SNP	29	9	12	88	93	73	25	29	98	100
Other Int	18	52	54	56	13	25	22	23	28	26
Other SNP	391	447	458	489	380	26	11	12	35	3
Splitting on environmental variable with no effect (smoking)										
Target SNP	63	87	76	7	5	6	32	33	1	0
Target SNP and smoking	0	0	0	0	0	1	1	1	1	0
Other Int	8	21	9	10	20	29	31	27	30	29
Other SNP	367	353	320	400	363	31	26	27	35	36

As with the other simulations, the identification results were very similar in character to those found using the MTM. The closer the proportion exposed to the interaction became to 0.5, the better the results. The false positive rate was much higher for stepwise regression than MTM, as was the computational time.

7.3.2.7 Interaction Model

The five models of interaction described and simulated in Chapter 6 were repeated using the parameters in the table below. Although the labelling of type A as an interaction is debateable, the simulation was carried out in case the stepwise regression recognised it as such. The results are shown in Table 7.16.

Table 7.15 Simulation Parameters – Interaction type

Risk Effects Present	**Effect Size (OR)**					Risk Factor Frequency	Number of SNPs	Sample Size	Proportion Cases	Inheritance Model
Interaction Type	**A ***	**B**	**C**	**D**	**E**					
Smoking	**1**	**1**	**1**	**1**	**1**	0.510	175	1000	0.5	Dominant
NSAIDs	**2.5**	**1.5**	**1**	**1**	**1.5**	0.6664				
Target SNP	**1.5**	**1**	**1.5**	**1**	**1.5**	0.6634				
Interaction	**1**	**2.5**	**2.5**	**2.5**	**2.5**	0.4421				

* The associated risk for the target SNP in interaction model A describes the increase on risk of taking NSAIDs, not an increased risk directly of case status

Table 7.16 Percentage of Datasets Identifying Interaction for Interaction Models A-E, Comparing Stepwise Regression and MTM

	Stepwise Regression					**MTM**				
Interaction Model	**A**	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
Splitting on Target Environmental Variable (NSAIDs)										
Target Int	0	78	78	56	60	0	80	78	54	61
Target SNP	29	99	100	98	99	0	99	99	97	97
Neither Int or SNP	71	1	0	2	1	100	1	1	3	3
Other Int	0	89	11	75	71	26	17	16	21	19
Other SNP	76	516	99	476	514	36	3	2	3	2
Splitting on environmental variable with no effect (smoking)										
Target SNP	31	90	94	88	91	19	74	77	66	65
Target SNP and smoking	1	1	0	0	0	1	2	1	1	1
Other Int	0	16	2	13	17	27	26	26	25	25
Other SNP	85	397	77	359	396	27	15	14	17	15

The results for interaction type A were better for stepwise regression than MTM, although still worse than any of the other interaction models. Interestingly, the false

positive rate and the computational time were both much lower for the type A model, this contrasts with MTM, for which type A had by far the highest false positive rate.

In terms of successful identification, the other four models were fairly similar between stepwise and MTM, but the false positive rate was higher for stepwise regression. The SNP was identified at a higher rate in the smoking stepwise analysis compared to the MTM.

The within stepwise comparisons show that the identification was highest, the false positive identification rate lowest and computational time much reduced (approximately 20% of the time taken compared to other models) for interaction model C. This is the model containing an independent genetic effect, which is enhanced by the environmental effect.

7.3.3 MDR and ORMDR

In order to use the MDR and ORMDR in a comparable way to the MTM and stepwise regression, a dataset was created containing case status, NSAID intake, smoking and the first 50 SNPs, all categorised and recoded in the appropriate way. This follows the instructions in R for using the ORMDR function [472]. The literature suggests that from this data, MDR should be able to identify the strongest interaction effect and quantitatively measure the strength of the interaction using the ORMDR extension.

However, the initial analysis using MDR and ORMDR produced consistently wrong but similar results across the datasets, identifying interactions with an OR which, when calculated by hand, showed a smaller OR than that of the target interaction in the same dataset. Such a calculation was necessary as the ORMDR is not an equivalent measure to OR. The identified interactions are shown below in Table 7.17, the specific names of the SNPs are not relevant as they do not relate to the real SNP in any way other than distribution.

Table 7.17 ORMDR results

Interaction Identified	Number of identifications
smoking and SNPx	26 / 1000
smoking and SNPy	54 / 1000
smoking and SNPz	10 / 1000

Looking closely at the results of the first dataset, both the identified and target interactions showed a very significant interaction with the outcome using a χ^2 test (p values of 1.06×10^{-142} and 4.49×10^{-70} respectively). ORMDR selects the former based on significance and not the size of the effect. Correspondence with the author of the ORMDR function in R confirmed this to be the case.

One other complication in this case occurred for the dominant inheritance model of the simulated data, with both those heterozygous and homozygous for the risk allele at similar risk. When divided into separate groups, the groups were inevitably smaller and the corresponding CIs wider. As this was the selection criterion, such an inheritance model was at a disadvantage. Considering that one limitation of the MDR method is the requirement for prior knowledge regarding continuous variables and cut-off points, such grouping could reduce the power to detect effects due to the smaller group size.

Another interesting feature is that it was possible to re-order the columns within the dataset to get different results, an indication that the method or the computing function being used were not performing accurately. Correspondence with the author of the ORMDR function concluded that as these simulations use a risk model that confers an increase in risk, not an absolute relationship, such results were an artefact of there being no “real” effects present. Therefore, the model used throughout this thesis was not an appropriate test bed for MDR or any related functions. Any potential bugs in the programme are now being investigated by the author as such differing results suggest there is a computational problem.

Given that ORMDR is based on a biological causal pathway model of inheritance and the data in this case were generated to confer an increase in risk, it was not an entirely

suitable method for analysis. To prioritise variables based on their precision over their effect size would be ill advised for the type of interaction effect under consideration in this thesis.

7.4 Discussion

The problem of dimensionality was not solved by using logistic regression, where the problem of high levels of false positive identification could be seen most clearly. Even restricting the potential interactions to those between an environmental variable and all possible SNPs (i.e. not considering SNP-SNP interactions), so that each additional SNP only increased the potential for interactions linearly, not exponentially, still generated models containing too many variables per observation. However, logistic regression is a very good way to quantify the risk and significance of any findings following a data selection method.

Using stepwise selection for variable selection in logistic regression improved the performance and provided a good comparison method. The computational burden was high and the results showed a higher false positive rate and a preference for identifying independent effects. However, if known in advance that the interaction model contained a genetic main effect, it would be useful to use both stepwise regression and MTM and to compare the results. Stepwise regression was the most effective comparison method and had very good levels of identification.

MDR and ORMDR are designed to use a slightly different selection criterion, the significance of the results rather than the size. It is based on a biological causative model and more suited to use in pathway genetics where such effects are more likely to be absolute. Had there been no computational problems, such a comparison may have been useful, but using currently available software there is no appropriate way to test the effectiveness.

Each method analysed has been developed for a slightly different purpose and with different data requirements or criteria. Logistic regression has been developed primarily for environmental variables but has been implemented in many genetic studies without adaptation. Using the stepwise variable selection function can be used to reduce the dimensionality problem but has a very high level of false positive identification, and is more suited to analyses where more information is known in advance or a lower event to variable ratio. MDR was developed to find SNP-SNP

interactions, although claims are made regarding its adaptability for gene-environment interaction detection as well. None of these methods were designed for the data characteristics or for the interaction type being looked at in this study and cannot easily be compared to MTM in this respect.

However, stepwise logistic regression is directly comparable and provides strong evidence that MTM maintains the power, found in other methods, to identify interactions whilst decreasing the false positive identification rate. This is a reassuring result, particularly for increasing SNP numbers which the previous chapter showed was a strength of the MTM.

Chapter 8

MTM Analysis on Colorectal Cancer Data

8.1 Introduction

The initial literature review identified a selection of SNPs from their associated genes that could be simulated to realistically represent a sample of candidate SNPs. However, during the development and testing stages of the method, the field of SNP analysis gained considerable momentum. It was not practical to keep going back and adding to the original panel during method development, as the simulations would then be slightly different from one another and as an artificially simulated panel, it would not have made much difference to the outcome. However prior to analysis of real data it is essential to review more current, SNP based work, to identify a representative panel of SNPs for real Mixed Tree Analysis. It was also necessary to adjust the original list, ensuring all candidate SNPs were present on the SNP chip used in the SOCCS dataset. The same search strategy was used as in Chapter 2, with more focus on SNPs in place of genes and a particular search on the results found from GWAS.

8.1.1 SNPs Associated with High Risk Genes or Disorders

A number of common, low penetrance genetic variants acting in either an additive or multiplicative manner have been associated with familial and sporadic colorectal cancer. Some of the SNPs have been identified by their association with genes considered to influence colorectal cancer risk, some more recently through GWAS.

Although previously mismatch repair genes were not included in the list of SNPs used in the simulation studies as the SOCCS dataset excludes those with other forms of cancer, the most consistent association was found to be with rs6983267, which is in the mismatch repair gene MYC. This SNP has also been found associated with prostate cancer risk [425, 427, 473]. Alongside rs6983267, another SNP on MYC is present on the genotyping panel used in the SOCCS study: rs7014346. A similar gene to that identied in the initial review, SMAD4, that is involved with TGFβ signalling, SMAD7, contains three SNPs more recently found to have an association with CRC. SMAD7 is also involved in Wnt signalling. One gene considered in Chapter 2 due to

an association with Peutz-Jeghers Syndrome (PJS), STK11 (also known as LKB1) has more recently been associated with the Wnt pathway [474], combining two areas of interest identified in the earlier literature review (genes associated with syndromes and genes on the Wnt pathway).

Two well characterised cancer genes were added to the initial list; CRAC1 is a high penetrance susceptibility gene and CDH1 is an invasion suppressor. A number of signalling genes have been identified, including: POU2AF1, a B cell specific transcriptional co-activator; Eukaryotic translation initiation factor 3 (EIF3H); Gremlin 1 (GREM1) which is an antagonist of the bone morphogenetic protein family. Bone Morphogenetic Protein (BMP) 2 and 4 are both involved in signalling and apoptosis. RHPN2 is a Rho binding protein. These genes and their associated SNPs are shown in Table 8.1. SNPs highlighted in bold are present on the SOCCS SNP panel and were available for analysis in this study.

Table 8.1 SNPs Associated with CRC Identified by their Genetic Association

Gene	SNPs	References
MYC	**rs6983267**, rs10505477, **rs7014346**, rs10090154	425-427, 473, 475-479
SMAD7	**rs4939827, rs12953717, rs4464148**	480, 477, 479
POU2AF1	**rs3802842**	481, 482
CRAC1/GREM1	**rs4779584, rs10318**	483
EIF3H	**rs16892766**	484
VDR	**rs1544410**	485
HFE	**rs1800562**	486
MLH1	**rs1799977**	487
BMP4	**rs4444235**	433, 481, 488
CDH1	**rs9929218**	
RHPN2	**rs10411210, rs7259371**	
BMP2	**rs961253**	

8.1.2 SNPs Identified from GWAS

A number of genes have been identified or corroborated by GWAS in which a large number of markers across the entire genome are studied. With the false positive

identification rate a problem with such large numbers of variables, consistency of findings across different studies and combining study populations can enhance the validity of results. There are 10 SNPs, including rs6983267, which have been consistently identified [489], eight of which are mentioned in table 8.1: rs16892766, rs3802842, rs4779584, rs4939827, rs4444235, rs9929218, rs10411210 and rs961253. The other, rs10795668, is not situated within a particular gene. Zanke et al combined a number of different studies and identified the 8q24 locus (covering SNP rs6983267).

A number of SNPs have been identified from GWAS that are not associated with a known gene. These are shown in Table 8.2, with those available for analysis on the SOCCS dataset in bold, as before.

Table 8.2 SNPs Identified from GWAS

SNPs	References
rs12701937, rs6038071, **rs11014993, rs10795668**	[490]
rs355527, **rs4951291, rs4951039**	[433]

8.1.3 SNPs Interacting with Environmental Variables

A number of interactions have been identified between certain genes and environmental variables and, less commonly, SNP-SNP interactions have been identified.

Three genes that have been found to interact with smoking and/or alcohol are: XRCC1, XRCC3 and XPD, which are all DNA repair genes that are known to increase cell turnover and mutations rate. Whether there is a true XPD-smoking interaction is still somewhat debatable [491, 492]. The Growth Hormone 1 gene (GH1) has a genotype associated with a decrease in risk of colorectal cancer, the same

sample also found an association with drinking alcohol and a possible interaction between the genetic and environmental factors.

Activating Transcription Factor 3 (ATF3) can be a tumour suppressor in colorectal cancer. Tolfenamic Acid (TA) is a NSAID with anti-tumour and pro-apoptotic associations. A study by Lee et al described TA stimulation of ATF3 expression, which then induces apoptosis providing evidence of a biologically plausible interaction [493]. This is mediated through the phosphorylation of ATF2 – both of which contribute SNPs to the new panel. Similarly, expression of XRCC3 has been found to be susceptible to aspirin at a biological level [494]. PPAR has also been found to have a role connected to APC and NSAIDs [495].

There are a few other unusual findings that may suggest interactions. Haiman et al found that the rs6983267 association with CRC, although replicated across most populations, did not confer a significant increase in risk for those of African American or Japanese decent [425]. However a slightly different SNP at the same locus, rs7014346 at 8q24.21 has been found to have an association with CRC risk in African Americans [496]. This could in some way be explained by the different frequencies of the SNPs within these populations, the GG/TT genotypes being at a frequency of 0.679/0.057 in African Americans as opposed to 0.212/0.239 in Caucasians. The Odds Ratio estimates are similar but the significance levels differ, so it is possible this is a statistical artefact based on different allele frequencies. The CYP2E1 polymorphism was only found to confer an increase in risk for Caucasians [497]. They also found that the tissue type affected the association between genetic variants and the cancer development. Some studies found differences in risk for colon and rectal cancers associated with different genotypes [479].

Genes that have been implicated in an interaction are shown in table 8.3, for which candidate SNPs have been identified using HapMap, as before, and those available for this analysis are shown in bold.

Table 8.3 SNPs Involved in Interactions

Gene	SNPs	References
Interactions with smoking and alcohol		
XRCC1	rs2682556, rs25487, rs213334, rs3213282, **rs2854501, rs1799778, rs2854496, rs3213255**	[498]
XRCC3*	rs861537, **rs861531, rs861539**	
XPD (ERCC2)	**rs13181, rs17799787**, rs3916874, rs238416, rs50872, rs238404, **rs50871**, rs1799786, **rs1618536**	
GH1	rs4080076	[499]
Interactions with NSAIDs		
ATF3	**rs1195472, rs2137424, rs4543871, rs1877474**, rs11576473, rs10494952, **rs1567710**, rs3125295, **rs9430097**, rs17019438, **rs3125296**, rs10746435, rs10735510, **rs3125289**	[493]
ATF2	rs212349, rs1153685, rs7578569, **rs11888507, rs7566401, rs13388308**	
XRCC3*	rs861537, **rs861531, rs861539**	[494]
PPAR	rs1040436, **rs2016520, rs1053049, rs2038086**	[500]

8.1.4 SNP – SNP interactions

Abuli et al found an interaction between rs6983267 and rs9929218 [489]. By taking the ten most commonly identified SNPs, all of which are included in the table above, and running a pairwise interaction analysis, only the one interaction effect was found after adjustment for multiple testing.

8.1.5 Conclusions from the Literature

There are ever increasing numbers of SNPs being identified as being associated with CRC, and the presence of consistent findings in different datasets is a good indicator of a true association. The majority of these loci are within genes that have a biologically plausible association with CRC. The evidence of possible interactions is

still sparse but has for the most part come from analysing a single environmental risk factor against several genetic factors, not from analyses of multiple environmental and genetic variables. This is reassuring as it corresponds to the approach, and the reasoning behind the approach, taken in MTM analysis.

8.2 SOCCS Data

The SOCCS data are from a case-control study in which control subjects were matched by age, gender and area of residence in Scotland to CRC cases. A full description of the environmental variables recorded in the study can be found in Chapter 4. The mixed tree analysis is run on a subset of the whole sample, 1216 people (461 cases and 755 controls), who had data available for the environmental variables and the SNPs required, the univariate analysis is carried out on both samples to identify any biases introduced through using the smaller sample.

The main bias to be considered is survivor bias; however the samples were drawn from the participants at the same time as the environmental data was collected, which should reduce survivor bias by some degree. In the case of degraded samples, it would be impossible to resample from those who did not survive the interval between initial and follow up sampling. Both the larger and smaller samples would be prone to recall and sampling biases, inherent in the retrospective case control design. However, analysing both the larger and smaller samples individually for environmental effects explores to some extent, any bias introduced through selection of the smaller sample.

8.2.1 Data Preparation and Quality Control

Individuals with poor quality data can be identified by the proportion of missing data. In the case of environmental variables this could suggest difficulties in the questionnaire procedure or follow up. In the case of genotype data, this could result from a low quality or degraded DNA sample, suggesting other SNPs from that sample may be unreliable. It is also necessary to identify any SNPs with a higher than normal proportion of missing values and remove these from the analysis where appropriate.

There was a single person in the study population who had missing data for 59 out of 222 variables; they were removed prior to the analysis. In the panel of SNPs used in this analysis there was a single SNP (rs9430097) where 46/1216 (3.78%) of the sample had a missing value. As this SNP was one of nine in the same gene, the loss in

information incurred from eliminating it was minimal. There were only four other SNPs containing any missing data at levels of 1.48%, 0.82%, 0.57% and 0.49%, levels unlikely to suggest significant problems. As the proportion of missing data was so low and evenly spread between cases and controls, it was considered safe to leave them in the analysis with the missing values replaced by the most common genotype.

Missing environmental data can be handled slightly differently, given the nature of the statistical method. Individuals were only removed from the analysis when they had missing data on the environmental variable under study but were included for all other analysis. Had there been significant volumes of missing data, such an approach may have led to the analysis being carried out on effectively different study populations, with the risk of introducing bias. For example, if there was an environmental exposure only measured during follow up, those able to give that information would naturally be those who had survived long enough to be measured. However, in this case, such missing data was so sparse that any effects would be minimal.

8.3 Univariate Analysis

In order to fully interpret any interaction results, it is necessary to carry out univariate analysis on the environmental variables. This provides context to allow some understanding of the underlying model of interaction. The population who filled in environmental data and food frequency questionnaires is larger that the sample who have both environmental data and have genotyping data. The univariate analysis are run on both the sample with environmental data (approximately 4,800 people depending on the variable) and the sample with both environmental and SNP data (1,216 people). This allows interpretation of the variable effects for the specific sample analysed and for interactions in a wider context.

8.3.1 Methods

As seen in Chapter 4, the environmental variables are of a number of different types. There are binary, ordinal and categorical variables, and continuous variables distributed normally, log-normally or non-normal even after transformation. Different descriptive and comparative methods are appropriate for each type, as described below.

Binary Variables

As described more fully in Chapter 4, the only binary variable is gender, although for basic analytical purposes alcohol is treated as above and below a recommended threshold value. The distribution of NSAID use, smoking status and exercise were re-classified as binary variables for the mixed tree analysis. As alcohol was not normally distributed, for either the larger or smaller samples, even following logarithmic transformation, the binary counts of above and below a recommended intake were used for analysis. In terms of NSAID intake, frequent use was defined as an intake of at least 4 days per week for at least 1 month. This was a SOCCS definition, not one specific to this study.

As gender was matched when selecting the controls, it therefore can not be analysed as an independent risk factor in this case. Gender remains in the dataset, however, as there are other variables that differ by gender. For these binary variables, an Odds Ratio estimate was used to assess association with outcome. For NSAID intake, the binary variable is analysed using the Pearson chi-squared (χ^2) statistic, with the null hypothesis that there is no difference between the groups.

Categorical and Ordinal Variables

The categorical variables are NSAID intake and family history risk. Although NSAID intake has an element of ranked increases these increases do not correspond to the level in a linear way. The ordinal variables include smoking, level of physical activity and deprivation level, some of which may be better analysed in a binary format. As both exercise and deprivation increase in a natural order, the χ^2 test for trend can be used to assess a relationship.

For this gene-environment analysis the family history risk variable was removed as it cannot be considered independent from genetic effects. In simulations, the mixed tree method had low success in identifying the type of interaction that would follow the sequence: genes influencing family history, which in turn influences risk of disease. As it is debatable whether this model is even an interaction and given the low success rate, analysis of the family history is not appropriate for this study.

Normally Distributed Variables

There were two variables that were normally distributed: Age and BMI. Table 8.4 shows the mean and standard deviation for both the case and control samples. BMI for both cases and controls was normally distributed with a similar standard deviation; therefore a t-test could be used to compare between the means of the two groups. The cases and controls were age-matched, so there was no need for further analysis on the age variable. However, it is worth noting that the sample in the analysis is younger than the sample as a whole. The younger cases are more likely to have their case status influenced by genetic effects, as the environmental effects are more cumulative and the influence from the environment increases with age. If the age variable had not been matched between cases and controls and was under study, such selection may have introduced bias towards the genetic effects. It is possible that variables such as

exercise or alcohol intake may be associated with age, so it is necessary to be aware of potential sampling bias.

Table 8.4 Comparison of Cases and Controls for Normally Distributed Variables

Variable	Environmental Data (n=4837)		Environmental and SNP data (n=1216)	
	Mean of Cases (sd)	Mean of Controls (sd)	Mean of Cases (sd)	Mean of Controls (sd)
Age (years)	62.0 (10.7)	62.4 (10.5)	50.3 (5.54)	51.16 (5.96)
BMI (kg/ m^2)	26.6 (4.42)	26.8 (4.63)	26.9 (5.04)	27.2 (5.29)

Variables Normally Distributed Following Log Transformation

Almost all the dietary variables were positively skewed and responded well to log transformation. Therefore the median and inter-quartile ranges, shown inside brackets, were used to describe the distribution and the differences between cases and controls. For the univariate analysis the dietary variables were transformed and then analysed using a t-test

8.3.2 Univariate Results

Table 8.5 shows the case-control comparisons for the categorical and ranked variables, using either Pearson's χ^2 or the χ^2 test for trend. The *p*-values for binary categorisation are shown in brackets. The sample containing environmental data and the subset with both environmental and SNP data were analysed separately as such analyses have slightly different aims (relevance to population versus relevance to the interactions results).

Table 8.5 Comparison of Cases and Controls in Both Samples, for Categorical, Ordinal and Binary Re-coded Variables

Sample		Environmental Data (maximum n = 4837)			Environmental and SNP data (maximum n = 1216)		
Variable	**Grouping**	**Cases (%)**	**Controls**	***p* value (binary)**	**Cases**	**Controls**	***p* value (binary)**
Frequent NSAID intake	None	1449 (70.5)	1757 (64.8)		378 (82.0)	567 (75.3)	
	Mini Aspirin	311 (15.1)	503 (18.2)	5.115×10^{-6}	18 (3.9)	46 (6.1)	0.0786
	Other NSAID	220 (10.7)	364 (13.2)		56 (12.1)	118 (15.7)	
	Aspirin	47 (2.3)	106 (3.8)	(7.05×10^{-7})	3 (0.7)	14 (1.9)	(6.40×10^{-3})
	Other NSAID + mini Aspirin	14 (0.7)	18 (0.7)		3 (0.7)	4 (0.5)	
	Other NSAID + Aspirin	13 (0.6)	9 (0.3)		3 (0.7)	4 (0.5)	
Smoking	Non	874 (42.4)	1200 (43.2)		228 (42.2)	361 (53.4)	
	Regular	818 (39.7)	1049 (37.8)	0.044	119 (22.0)	203 (37.6)	0.430
	Ex-regular	343 (16.6)	509 (18.3)	(0.179)	190 (35.2)	111 (20.6)	(0.578)
	Ex-occasional	26 (1.3)	18 (0.6)		3 (0.6)	1 (0.2)	
Level of Physical Exercise (hours per day) (n=2061-98)	0	1139 (58.0)	1456 (54.3)		223 (49.7)	340 (45.9)	
	0 – 3.5	486 (24.8)	703 (26.2)	0.029	143 (31.8)	240 (32.4)	0.062
	3.5 – 7	205 (10.4)	339 (12.6)	(0.577)	48 (10.7)	116 (15.7)	(0.213)
	7+	133 (6.8)	185 (6.9)		35 (7.8)	44 (5.9)	
CARSTAIRS Deprivation Index	1	194 (9.4)	258 (9.3)		40 (8.7)	61 (8.1)	
	2	434 (21.1)	568 (20.5)	0.936	99 (21.5)	139 (18.4)	0.509
	3	532 (25.8)	758 (36.8)		118 (25.7)	212 (28.1)	
	4	488 (23.7)	645 (23.2)	(0.799)	105 (22.8)	167 (22.1)	(0.417)
	5	218 (10.6)	293 (10.6)		51 (11.1)	86 (11.4)	
	6	140 (6.8)	178 (6.4)		29 (6.3)	67 (8.9)	
	7	54 (2.6)	76 (2.7)		18 (3.9)	23 (3.0)	

This shows that the binary version of NSAID intake is associated with a significant difference in risk, with a decreased risk for those taking any form of NSAID compared with none. The risk of smoking is only borderline significant, with a very small increase in risk for smokers compared to non smokers in a binary format in the larger sample. Also in the larger sample there is a significant decrease in risk as exercise increases from 0 hours to 0-3.5 hours and then to 3.5-7 hours. The 7+ hours do not fit the trend but this could be in part down to the smaller sample sizes, as could the absence of such a significant effect in the smaller sample size. There is no relationship between deprivation and case status in this data. The differences between the larger and smaller samples are mainly evident in the lower *p*-values for the smaller sample, as would be expected.

The cases and controls have a mean BMI of 26.6 and 26.8 respectively. The analysis of association between case status and BMI gives a *p*-value of 0.3895, which showed that there was no significant difference between cases and controls with respect to BMI. Similarly, the OR for consuming above recommended levels of alcohol compared to below was 0.9979, very close to 1, suggesting no effect.

Table 8.6 shows the results for the univariate analysis of the dietary variables, the significance estimated by log transformation, followed by a t-test. It also shows how the picture of independent effects changes by using the smaller sample for the analysis compared to the larger sample.

Table 8.6 Case and Control Comparison of Dietary Variables, Based on log-transformed Data

	Whole group environmental Data (maximum n = 4837)			Subset with environmental and SNP data (maximum n = 1216)		
Variable	**Median Cases (mean log) (max n=2061)**	**Median Controls (mean log) (max n=2776)**	**Significance**	**Median Cases (mean log) (max n=461)**	**Median Controls (mean log) (max n=755)**	**Significance**
Vitamin D (ug)	3.94 (1.36)	3.88 (1.35)	0.527	4.70 (1.35)	4.62 (1.30)	1.73×10^{-3}
Retinol (ug)	549 (6.33)	478 (6.20)	1.78×10^{-9}	730 (6.21)	655 (6.16)	6.19×10^{-2}
Calcium (mg)	1114 (7.00)	1074 (6.96)	1.60×10^{-4}	1215 (7.03)	1173 (6.96)	0.449
Thiamine (mg)	2.04 (0.73)	2.01 (0.70)	1.38×10^{-2}	2.33 (0.74)	2.31 (0.73)	0.220
Riboflavin (mg)	2.12 (7.46)	2.065 (7.11)	1.67×10^{-3}	2.31 (0.77)	2.26 (0.74)	0.892
Niacin (mg)	23.6 (3.17)	23.6 (3.15)	0.142	27.2 (3.23)	27.0 (3.23)	0.954
Pantothenic Acid (mg)	6.78 (1.98)	6.575 (1.94)	1.82×10^{-2}	8.61 (2.00)	8.92 (2.00)	0.778
Vitamin B6 (mg)	2.79 (1.03)	2.79 (1.01)	3.22×10^{-2}	3.08 (1.06)	3.07 (1.05)	0.633
Biotin (ug)	49.1 (3.89)	48.3 (3.87)	0.1254	52.8 (3.90)	52.2 (3.89)	0.354
Vitamin B12 (ug)	7.0 (1.95)	6.8 (1.92)	0.1289	7.95 (1.93)	8.05 (1.90)	0.445
Folic Acid (ug)	324.0 (5.78)	321.0 (5.76)	5.06×10^{-2}	353 (5.81)	351 (5.79)	0.900
Fibre (g)	20.7 (3.03)	20.6 (3.02)	0.4378	22.9 (3.05)	23.0 (3.05)	0.900
Energy (kJ) - Males	11240 (9.33)	10580 (9.26)	1.837×10^{-7}	12743 (9.39)	11780 (9.31)	0.218
Energy (kJ) - Females	9556 (9.18)	9239 (9.15)	6.10×10^{-2}	10891 (9.23)	10469 (9.20)	0.261

The first point of interest is the different significance values between the two groups; in most cases you would expect a lower level of significance for the smaller sample size. Vitamin D is the only exception to this trend, with a higher significance in the smaller sample, which may be in part attributable to the slightly more skewed sample, simply down to chance or an artefact of the younger age of the sample and trends in eating habits.

It was is intriguing to notice that the cases reported greater intake of almost every measured variable, except Niacin and vitamin B6, which are equivalent. The controls reported greater exercise levels, although not significantly, yet had a slightly higher BMI overall, although again this was not significant. Although this could be due to chance, it is counter intuitive and may suggest either weight loss due to illness in the case group or recall bias, with the cases having more accurate recall when filling out the food frequency questionnaire. It also suggested that it would be easier to identify a dietary variable that conferred an increase in risk than one with a preventative effect.

Similarly, the sample selected for genotpypic data have higher rates of consumption than the overall sample. These people were selected for the first phase of genotyping by being in the youngest 10% of the overall sample. This was to enhance the risk attributable to genetic effects, as cumulative environmental risk increases with age.

In the larger sample the most significant independent, nutritional effects were retinol, calcium and riboflavin consumption, in each case with the cases reporting significantly greater intake than the controls. The difference in total energy consumption between cases and controls was also highly significant, but only in males. There were some borderline effects: thiamine, pantothenic acid and vitamin B6. In the smaller sample, the only significant effects were for vitamin D and fibre, however even borderline effects inform the understanding of the results following the mixed tree analysis.

8.4 Mixed Tree Analysis

The environmental variables were analysed sequentially using the mixed tree method and the partitioning method best suited to the variable type. Any independent results found consistently across the different variables were removed and analysed separately to identify any SNP-SNP interactions. The positively skewed variables were log-transformed prior to partitioning.

Any individual with missing environmental data was only removed from the analyses relevant to that specific variable. The missing SNP data was replaced with the mode of the rest of the data for that variable, using the "na.roughfix" command within the random forest function. This was suitable as the data were missing at fairly equal proportions between cases and control and at very low frequencies.

8.4.1 Mixed Tree Results

The results for categorical and ordinal variables, entered as binary variables, found with a significance threshold of below 0.05, are shown in Table 8.7. The significance value given here is only relevant in the context of comparison within the method and building an evidence base for further work, due to the multiple selection and testing procdures the significance thresholds are not themselves applicable. The results from initial MTM analysis on dietary variables are shown in table 8.8. Taken in context with the results from the earlier univariate analysis on the environmental variables, these results will inform the decision on which potential interactions require further investigation. The significance threshold of 0.05 is not in itself definitive, given the selection stages prior to this analysis, but is intended to add to a body of evidence of potential interactions.

Table 8.7 MTM Results for Categorical, Ranked and Binary Variables

Environmental Variable	SNPs or Interactions Identified	Estimate (SE)	Significance (p-value)	
NSAIDs	rs2481952	0.33 (0.09)	6.44×10^{-4}	*
	rs1799977*NSAIDs	0.74 (0.22)	9.53×10^{-4}	
Smoking	rs4918766*smoking	-0.66 (0.18)	2.17×10^{-4}	
Exercise	rs50871	0.40 (0.13)	1.32×10^{-3}	
	rs50871*exercise	-0.46 (0.17)	8.09×10^{-3}	
Deprivation	rs2481952	0.33 (0.10)	5.56×10^{-4}	*
	rs6761246*deprivation	0.25 (0.10)	1.31×10^{-2}	
BMI	rs2481952	0.25 (0.11)	3.45×10^{-2}	*
BMI overweight	rs10318	0.75 (0.34)	2.81×10^{-2}	*

Table 8.8 MTM Results for Dietary Variables

Dietary Variable	log split	SNPs or Interactions Identified	Estimate (SE)	Significance (p-value)	
Vitamin D (µg)	1.21	none			*
Retinol (µg)	5.25	rs10733118	0.25 (0.12)	3.09×10^{-2}	*
		rs10733118*Retinol	0.24 (0.11)	3.95×10^{-2}	*
Calcium (mg)	7.31	none			*
Thiamine (mg)	1.14	rs2481952	0.58 (0.17)	1.75×10^{-3}	*
Riboflavin (mg)	0.42	rs2481952	0.45 (0.19)	1.39×10^{-2}	*
Niacin (mg)	3.13	none			*
Pantothenic Acid (mg)	1.64	rs12988520	0.71 (0.34)	1.81×10^{-2}	
Vitamin B6 (mg)	1.21	rs2481952	0.58 (0.23)	1.29×10^{-2}	*
Biotin (µg)	3.41	rs706716	3.31 (1.12)	5.12×10^{-3}	*
		rs706716 * Biotin	-0.84 (0.29)	5.69×10^{-3}	*
Vitamin B12 (µg)	1.64	none			*
Folic Acid (µg)	5.46	none			*
Fibre (g)	2.73	none			*
Energy (kJ) - Males	9.62	none			
Energy (kJ) - Females	9.47	rs4148325	7.72 (3.36)	3.13×10^{-2}	*
		rs4148325*energy	-0.82 (0.36)	3.49×10^{-2}	

*rs2481952 identified as one of the most important variables prior to the regression step of the MTM, irrespective of later significance findings

A large number of the analyses (18 /22) identified the SNP rs2481952 as a main effect (marked with a * in Tables 8.7 and 8.8). None of the variables identified an interaction effect with this SNP, which suggested that for this dataset this SNP does have an independent effect of the other variables. This SNP was removed from the dataset and entered as a root variable to look for any potential SNP-SNP interactions. The results from the SNP-SNP analysis are shown in Table 8.9, for both splitting rs2481952 by combining CC and CT and for combining TT and CT, thus representing the presence of a C allele and a T allele respectively.

Table 8.9 SNPs with a Potential Interaction with rs2481952

rs2481952 split	SNPs or Interactions Identified	Estimate (SE)	Significance (*p*-value)
T allele	rs4779584	0.35 (0.18)	2.77×10^{-3}
	rs2481952 * rs50871	0.41 (0.19)	3.28×10^{-2}
C allele	rs4464148	-0.30 (0.10)	4.44×10^{-3}

Once rs2481952 was removed from the analysis, the MTM returned the results listed in Tables 8.10 and 8.11 and these were classified by strength of evidence, as described in Section 6.4. At this stage the strongest evidence of an interaction is represented by the identification of the interaction effect with a p-value below 0.01, followed by the identification of the potential interacting effect below 0.01. Those identified as interaction effects are considered to have stronger evidence of interaction than when the SNP is identified independently but for only one environmental variable. Following this brief classification, biological plausibility and similar findings in other work are then assessed.

Table 8.10 MTM Results for Categorical, Ordinal and Binary Variables Once rs2481952 Removed

Environmental Variable	SNPs or Interactions Identified	Estimate (SE)	Significance (p-value)	Strength
NSAIDs	rs4464148	-0.35 (0.10)	4.79×10^{-4}	Strong
	rs4464148 * NSAIDs	0.69 (0.24)	6.27×10^{-3}	
	rs1799977*NSAIDs	0.72 (0.22)	1.35×10^{-3}	Strong
Smoking	rs4086116*smoking	0.72 (0.21)	7.34×10^{-4}	Strong
Exercise	rs50871	0.41 (0.13)	1.32×10^{-3}	Strong
	rs50871*exercise	-0.46 (0.17)	7.69×10^{-3}	
BMI overweight	rs10318	0.77 (0.34)	2.54×10^{-2}	Weak
Alcohol	rs12953717	0.57 (0.27)	3.37×10^{-2}	Weak
Deprivation	None			
BMI	None			

Table 8.11 MTM Results for Dietary Variables Once rs2481952 Removed

Dietary Variable	SNPs or Interactions Identified	Estimate (SE)	Significance (p-value)	Strength
Folic Acid (ug)	rs706716	5.02 (1.65)	2.31×10^{-3}	Strong
	rs706716 * Folic Acid	-0.86 (0.28)	2.50×10^{-3}	
Biotin (ug)	rs706716	2.92 (1.13)	9.49×10^{-3}	Strong
	rs706716 * Biotin	-0.74 (0.29)	1.03×10^{-2}	
Energy (kJ) - Females	rs4148325	7.87 (3.36)	1.92×10^{-2}	Strong/ Intermediate †
	rs4148325*energy	-0.84 (0.36)	2.15×10^{-2}	
Niacin (mg)	rs4148325	1.76 (0.78)	2.41×10^{-2}	Intermediate
	rs4148325*Niacin	-0.49 (0.24)	3.92×10^{-2}	
Retinol (ug)	rs10733118	-1.63 (0.76)	3.14×10^{-2}	Intermediate
	rs10733118*Retinol	0.25 (0.12)	3.70×10^{-2}	
Riboflavin (mg)	rs6761246	0.54 (0.20)	6.24×10^{-3}	Intermediate
Thiamine (mg)	rs6761246	0.46 (0.18)	9.22×10^{-3}	Intermediate
Vitamin B6 (mg)	rs12988520	0.59 (0.25)	2.02×10^{-2}	Weak
Vitamin D (ug)	none			
Calcium (mg)	none			
Pantothenic Acid (mg)	none			
Vitamin B12 (ug)	none			
Fibre (g)	none			
Energy (kJ) - Males	none			

† Considering the smaller sample size, the less significant *p*-value does not necessarily represent less evidence of an interaction

8.4.2 Summary of Results

A possible main effect was identified for one SNP, rs2481952 within the gene CDX2, which may also have been involved with SNP-SNP interactions. NSAIDs had a potential interaction effect with rs4464148 (within SMAD7) and rs199977 (within MLH1), with SMAD7 also having a potential relationship with alcohol. There was a cluster of effects, both independent and interactions, identified on the gene UGT1A6, including possible interactions with vitamin B6, riboflavin, thiamine, niacin and female energy intake. With respect to smoking, which itself was not significantly

associated with CRC case status, there was a potential interaction with rs4086116, without the equivalent SNP having an independent effect. rs4086116 is found in the genetic region coding for CYP2C9. Both biotin and folic acid identified rs706716, on PIK3R1, as having an independent effect and as an interacting variable, suggestive of a marginal genotype effect modified by the dietary variable. The two SNPs within GRAC1/GREM1, rs4779584 and rs10318, were identified by rs2481952 and being overweight respectively. Similarly, the models identified an effect of rs50871 within XPD, as a main effect, interacting with exercise and interacting with rs2481952. Results that were only borderline significant included retinol and rs10733118 (within MTR), the latter as both an independent effect and as a component of the relevant interaction. Alcohol consumption had a borderline interaction with rs12953717, which is also found in the coding region for SMAD7.

8.5 Discussion of Results

It is worth noting that the p-values given in the results are purely in context and not to be considered conclusive, given the number of pre-selection stages inherent in the method. A p-value that would normally be considered highly significant is used in the MTM results as contributing towards evidence there might be an interacting effect suitable for further study.

CDX 2

The SNP consistently found to have an independent main effect (rs2481952) is located in the genetic region coding for the intestine restricted transcription factor *caudaltype homeobox2* (CDX2) gene. The CDX2 gene, as other CDX genes, has homologues across a wide range of species, indicating that it has been evolutionarily conserved. The CDX genes play a crucial role in the intestinal tract development, differentiation and maintenance [501]. Research has shown that many of the genes necessary for normal development also have important roles in the process of carcinogenesis [502]; this is due, in part, to the similarities in cell growth and reproduction in the two processes.

Regarding the role of CDX2 in CRC, there have been a number of different proposed mechanisms: including effects on the Wnt pathway, an APC target, the Vitamin D pathway, and a possible target for inactivating mutations in MMR genes [356].

The earliest studies of this role were based on the levels of CDX2 expression, finding that the expression levels were lower in adenomatous polyps and cancers compared to normal mucosa, negatively correlating with the degree of dysplasia found [503]. Interestingly, CDX2 expression was absent in most colonic adenocarcinomas, except for a subset of colon cancer with high levels of microsatellite instability. CDX2 expression has been found in other intestinal tumours and it has been suggested as a marker in the differential diagnosis of primary versus metastatic adenocarcinomas with unknown origins elsewhere is the body [504]. Similarly, such expression could reflect the metastatic condition of the CRC patients [505].

However, the effect on CRC risk and biological role for this gene are not yet well documented. There are numerous hypotheses including: APC/beta-catenin signalling, Wnt pathway, Vitamin D pathway, microsatellite instability, relationship with inflammatory bowel disease (IBD) and insulin production. However, recent research suggests that CDX2 has a role in sporadic CRC, but is unrelated to any inflammatory influences [506]. The majority of associations assume that CDX2 has a role in the transcription of other genes, including the VDR gene [507]. CDX2 has been found to directly regulate the *Multidrug Resistance 1* (MDR1) gene by binding to regions in the promoter region [501]. Interestingly, one study found co-occupancy of CDX2 and TCF4 across short genomic regions and evidence that implicates CDX2 in directly binding intestinal cells [508]. One of these regions spans the SNP rs6983267, within the MYC enhancer region, which is the SNP most often associated with CRC risk.

The other SNPs associated with rs2481952 identified as SNP-SNP interactions were: rs4779584, which is on the same gene as the SNP found to interact with being overweight (rs10318); rs50871, which was also found to potentially interact with exercise; and rs4464148, which was also found during the NSAIDs analysis. The latter two may indicate either a complex association pathway or simply an increased risk, occurring by chance, for these two SNPs. The first SNP, however, along with having the most significant effect, is on the same region as another associated SNP (rs10318) and might therefore indicate a real SNP-SNP-overweight relationship. However, there is not enough evidence to generate a specific hypothesis in the role of CDX2 in CRC other than to keep these potential interactions in mind during further study.

SMAD 7 and NSAIDs

An interesting finding was that alcohol, NSAIDs and the main gene effect were all identified with SNPs in the SMAD7 (mothers against decapentaplegic homologue 7) region of chromosome 18, a gene involved in inflammation and the modification of both transforming growth factor β (TGF-β) and Wnt signalling pathways. The association between NSAIDs and CRC risk was particularly strong for those with the CC genotype at rs4464148, a significance level of 1.87×10^{-3}, which is extremely high given the sample size of only 116 people. The cases were significantly less likely to have taken an NSAID regularly, only 3/52 having done so, compared to 18/64 in the

control group. The heterozygotes had a slightly smaller and less significant association (*p*-value of 1.02 $x10^{-2}$) and the people homozygous for the TT allele had no association at all between NSAID intake and CRC risk. This apparent dose response relationship suggests a co-dominant inheritance model and increases the likelihood that these results have an underlying biological mechanism.

SMAD7 is a gene which has often been identified as having a relationship with colorectal cancer, although usually associated with colitis [509]. All three SMAD7 SNPs included in this study have been identified in a previous study as being associated with adenomas and cancers [480], with similar results being found again, for two for the three SNPs, in a later study [510]. Another study found similar effects but restricted to women, which could be associated with the different rates of inflammatory disease between the genders [511]. This analysis identified rs4464148 as having both a main effect and an interaction with NSAIDs, which is similar to findings in another study where rs4939827 was found to have different risk estimates based on stratification by NSAID status [509].

There are other roles of SMAD7, potentially associated with CRC, which have been studied in other biological areas of research, particularly regarding its role in inflammation. SMAD7 promotes the anti-inflammatory action of the TGF-β pathway [512], by binding to the receptor complexes, thus blocking the downstream signalling events. The deficiency of TGF-β has been found to lead to increased inflammation [513]. SMAD7 has also been associated with the progression of CRC, with SMAD7 amplification being a selected event during progression of colorectal tumors [514].

The rs1799977 SNP did not demonstrate a main effect but did appear to contribute to a significant interaction with NSAID use. rs1799977 is found in the β-catenin gene (CTNNB1) which is an activating compound for the Wnt pathway [515]. This SNP has been previously identified as having a low level, but significant, effect on CRC risk [516] and is also associated with breast and prostate cancers. It is found in MLH1, the mutations of which are among the most frequent causes of Hereditary Non-Polyposis Colon Cancer (HNPCC) [517].

Overall, the role that NSAIDs play in reducing colorectal cancer risk is likely to be associated with the inflammation, through modification of either the TGF-β or Wnt pathways. This is both biologically plausible and repeatedly found throughout the literature. The results of this study combined with the results from previous work suggest that the hypothesis of NSAIDs interacting with SMAD7 or CTNNB1 is the most likely and a good area for further study.

UGT1A6 and B vitamins

The SNPs that were implicated in possible interactions with Riboflavin/vitamin B2 and Thiamine/vitamin B1 (rs6761246), Niacin/vitamin B3 (rs4148325), and vitamin B6 (rs12988520) are found in the coding region for the same gene: Uridinediphosphate Glucuronosyl Transferase 1A6 (UGT1A6). This gene is more normally associated with the protective effects of NSAIDs, particularly aspirin [420]. However such a cluster of results for the B vitamins all on the same gene does suggest a potential area of interest for new work.

A study comparing micronutrient intake with five cancers, including CRC, in women observed no association between individual B vitamins and CRC [518]. A different study found that increased plasma concentrations of both vitamin B6 and Riboflavin were associated with decreased CRC risk [519]. However, this study found no interactions between genetic polymorphisms and B vitamins. The gene MTHFR has been found to interact, in a preventative manner against CRC, with high levels of folate and vitamin B6 [520]. This was not replicated in this study but does suggest a possible, if not yet fully understood, protective association with high levels of B vitamins.

The SNP that may interact with niacin (rs4148325) was also identified as a potential interacting variable for female energy intake. The same SNP was also identified for female energy intake, both at fairly low levels of significance, which suggests any results should be treated with caution and may be simply be an artefact of the data or a very low significance level main effect. It is also possible that UGT1A6 is an important gene with strong variations in risk based on SNP combinations. However, the lack of corroborating work and currently understood biological mechanism make it difficult to draw any conclusions from these results.

CYP2C9 and Smoking

There is a possible interaction, with a highly significant effect, between smoking and rs4086116, which is found in the gene CYP2C9. No main effect was found for the SNP, and neither was smoking associated with CRC in this study, suggesting an interaction model where neither factor has a main effect. A more detailed examination of the results shows that the subgroup that smokes has a very significant association between rs4086116 and CRC, whereas the non smokers have no association.

A previous study found that the odds of CRC decreased as the level of the CYP2C9 enzyme decreased and a different SNP variant was found to be associated with decreased CRC risk but increased polyp risk [521]. However, other studies have found no evidence of a relationship between genetic variations in CYP2C9 and tobacco-related cancers. Kaur-Knudsen et al, did not detect any association between CYP2C9 genotype and CRC [522], and previous research into an interaction found that CYP2C9 activity was not affected by smoking [523].

As CYP2C9 is involved in the bioactivation and detoxification of polycyclic aromatic hydrocarbons (PAHs) derived from tobacco smoke, the interaction hypothesis is biologically plausible. There have been studies analysing the different amino acid change in CYP2C9 in relation to smoking status, but only identifying main effects for both variables [524]. It is possible that the variation in smoke exposure in the different studies may be confounding the ability to identify the effect of CYP2C9. The possible interaction between this gene and smoking should be investigated further, both biologically and statistically.

PIK3R1, Folic Acid and Biotin

The phosphoinositide-3 kinase (PI3K) pathway is involved in cancer cell growth, survival and resistance to chemotherapeutic agents [525]. The results from this study show a main effect of PIK3R1, a gene on the PI3K pathway, associated with an increase in CRC risk. However, when this SNP (rs706716) interacted with folate or biotin, the risk decreased significantly. As folic acid/ folate are involved in DNA synthesis and methylation and a known target for chemotherapy, this result should not be dismissed. However, there is no direct evidence of a relationship between folate and this SNP in currently available research.

GREM1 and Obesity

GREM1 (gremlin 1), also called CRAC1 or HMPS, was found to have a potential interaction with being overweight and CRC status. The effect size was relatively large but not particularly statistically significant. There has been research suggesting that GREM1 is an important candidate gene for Total Body Lean Mass (TBLM) [157] and therefore such an association is not entirely implausible. Although the result is not enough to be conclusive, in this case, there could be scope for further research on GREM1 within CRC and obesity research. The other marker SNP on GREM1 (rs4779584) was found to potentially interact with the main SNP identified earlier, on CDX2.

XPD and Exercise

One SNP on XPD, rs50871, was identified as having an independent effect with increased CRC risk, decreased by an interacting effect with increased exercise. Those with the CC genotype have a negative association between CRC and exercise with a *p*-value of 7.13×10^{-3}, which is surprisingly strong given the small sample size (297 people with CC genotype). Such an effect was not replicated for the other genotypes.

XPD (xeroderma pigmentosum complementary group D) is a DNA repair gene, also known as ERCC2 (excision repair cross-complementing rodent repair deficiency, complementation group 2). This has previously been associated with an increase in CRC relapse [526] and a target for chemotherapeutic agents [527], and a potential modifier of alcohol based risk [498]. The association between alcohol and CRC only applies for the carriers of a particular codon, that would result from the amino acid change to the CC genotype, the same subgroup as this study found as potentially interacting with exercise. The similarity in mechanisms suggests that the CC genotype, or the Glu/Glu codon confers increased risk, a risk which may in some way be modified by lifestyle factors.

Borderline Results

The possible main and interaction effect of rs10733118 with retinol, vitamin A, is only borderline significant. rs10733118 is found in MTR (Methionine Synthase) a gene involved in folate metabolism. However, further investigation showed that both the homozygous genotypes had no significant difference for retinol consumption

between cases and controls, with the heterozygous having only a borderline difference. This is not suggestive of a real effect with a realistic inheritance model and therefore is quite likely to be a false positive result.

Pantothenic Acid identified a mildly significant main effect for rs12988520, on MMP2. The low level of significance and the lack of prior evidence and a biologically plausible hypothesis suggest this is likely to be a false positive result.

8.6 Summary

From the perspective of the MTM methodology, there are a number of promising findings from the analysis of the SOCCS data, primarily the ease with which the method could be applied and the low false positive rate. Although, obviously, it is not possible to determine exactly which are true or false positive results, the number of total results and the consistency of results suggest that the false positive rate is lower than for the simulated studies. This can be explained by the lower levels of statistical noise, especially when there would be an association between variables but they were simulated independently, but the results were better than were expected.

A number of the results are candidates for future work, many having biological plausibility or consistency of findings. There are too many variables, and too few events per variable, to allow a direct comparison of the results that may be gained through entering all the potential interaction into logistic regression for comparison. However, this use of regression following selection is useful in giving a quantitative estimate of significance, but further statistical research will be required to dictate the appropriate significance thresholds for such methods.

Chapter 9

Discussion

9.1 Conclusions on Research Question

The central tenet of this thesis is that enabling the identification of gene-environment interactions will expand and complement current epidemiological research on the aetiology of common complex disease. The research question has had to keep pace with other scientific advances, especially the increased technology and availability of data, with such advances moving much more quickly than development of methodological and statistical approaches required for such analysis.

The Mixed Tree Method takes a first step in resolving this problem, with more success than comparable methods. Features that suggest wide applicability of MTM include a lower computational burden and the ability to handle different formats of data than other current methodolgies. MTM also compared favourably in terms of successful gene-environment interaction identification for the majority of data models investigated. Although future work is required to determine the success on GWAS, the results from the MTM are promising for larger numbers of variables. Even if not fully applicable in the current format, providing evidence towards a ranking based approach to results in place of significance or estimation, for large datasets, may be a useful outcome.

9.1.1 Literature Review

The first aim of the thesis was to assess the methods currently used to identify gene-environment interactions. Reviewing the literature identified a number of methods with different strengths and weaknesses. A number of these methods had already been combined with other algorithms or extended to try and handle greater number of variables, with differing success. However, tree based methods cover a range of techniques, easily combined with one another and with very little prior research looking at interactions. The review selected other analytical approaches that could be used as a comparison for any new method developed and a potential niche for method development identified.

9.1.2 Development of Novel Method

The second aim was the development of a novel method. There is no method currently available that treats the genetic variables and environmental risks in different, or weighted, ways to accommodate the differences in their behaviour and effect sizes. Along with this feature, the specific steps of the method were assessed through small simulation studies in order to maximise results in the later work. This stage was also critical for ensuring all the written code was doing exactly as intended at each step.

9.1.3 Evaluation of Method

The third aim was to adapt the method to different data and evaluate the success of the method on data under different risk models. In essence there are four areas that need to be considered for using the MTM, analysing the results and applying any results to the general population.

(i) Data Input – Are the data used in the analysis appropriate, complete and accurate?

(ii) Study Designs – For which study designs is this an appropriate analytical tool?

(iii) Underlying Data Structure – which data characteristics have the greatest power for interaction identification?

(iv) Application – How well can any results be interpreted and applied to the general population?

The strength and limitations of this method, with respect to these four aspects are considered below.

Data Inputs

As statistical analysis is trying to identify particular patterns and effects within the dataset, any inaccuracies or missing data will have detrimental effects, if analysis is

possible at all. It is therefore necessary to run some preliminary tests on the data, remove outliers or variables with large volumes of missing data. If SNP choice is available, then selection of the allele frequencies most suitable to the potential inheritance model is essential. Missing data should be well characterised and any imputations or removals dictated by these features on a case by case basis.

With regards to variable type, environmental variables can be binary, categorical, ranked or continuous. However binary groups that are very different in size could lead to a number of problems when the trees are generated and adaptations should be made to minimise these problems. The majority of variables under investigation were continuous and therefore the grouping was done by a data driven split, calculated by the partitioning step. Therefore it is possible to adjust the parameters so that the loss of information from different group sizes is minimised.

As an environmental exposure rate of 50% is unrealistic, such a decision on parameters would need to take into account the sample size and the potential loss of information before deciding how far from 50:50 the grouping could afford to be. Adaptation was achieved straighforwardly by incorporating simple parameters into the code to define the data type in advance and treat it accordingly. Further details of these changes can be found in Chapter 5. It should be remembered, however, that prior to analysis it is necessary to define the variable type; binary, categorical, ordinal or continuous, for each of the environmental variables, as the partitioning step handles them slightly differently.

One further consideration regarding the quality of data is the SNP selection stage of any such analysis. As the genes and SNPs selected for this study had all previously been associated with CRC, there is ascertainment bias present and previously undiscovered SNPs are less likely to be included in such analysis. However, striking a balance between ascertainment bias of a SNP panel and the false positive rate for large numbers of variables is not a problem specific to this method or analysis. In fact, as this method is more easily comparable to current methodologies that analyse data with small variable to case ratios, as oppose to GWAS, the results of analyses using large numbers of variables are reassuring.

Study Designs

Although designed and tested exclusively on case-control data, there is no reason that, as long as the case control ratio is not too large, the method could not be used on cohort data, especially in cases of nested case control studies. As genetic data do not change over the life course, assigning a retrospective genetic dimension to a previous cohort study would also be suitable. Although epigenetics would be a natural home to studies of gene-environmenta interactions, the data gathering is still in its infancy and would require methodological adaptations.

Without further research it is impossible to gauge the effect of a founder population on the success of the MTM. However, the SNPs coding for the same gene in the real dataset did not tend to be identified together in any analysis despite an increased likelihood that they would be inherited together in the family based controls. Although in no way definitive, this is reassuring that common ancestry may not provide an insurmountable obstacle for future work on such datasets.

Underlying Data Structure

The simulation studies had a combination of expected and surprising results. It was understandable that the identification of a statistically significant association was highest when the proportion that were exposed and unexposed were roughly equivalent. In terms of SNP exposure, this equivalency was dictated by the allele frequency and underlying inheritance model. For environmental variables the power could sometimes be increased by re-grouping the variables. As would be expected, the rate of identification increased with effect size and a real interaction effect was identified more consistently across an increasingly stringent scale of significance thresholds. For example, an effect with an Odds Ratio of 2 is detected, as either interaction or SNP main effect, in 75% of cases.

It is debatable whether a factor that is on the causal pathway can be considered an interaction at all. Assuming that it can be, the MTM shows very poor results at detecting it. For all the other interaction models, the results are fairly promising, with the slight differences in some way explainable by the data generation method in this instance.

However, the most unexpected and promising result is the relatively low false positive rate, compared to the most equivalent methods, in the face of increasing SNP numbers. However, the largest SNP panel simulated contained 300 SNPs, so the boundaries of this advantage are yet to be fully determined, especially given that GWAS may identify several hundred thousand SNPs. Although a small step in the right direction, this result must be kept in context. It is possible that the more general principal of ranking results against each other to select the most significant candidates for further work may prove more valuable than the direct application of MTM itself.

Application

The results gained from MTM analysis cannot be directly applied to the general population. As a data mining technique, even with a significance measuring final step, any results provide a theoretical basis for further study and not a fully objective measure of association. It would also be necessary to study data from studies not case-control in design as the recruitment of such studies prevents them being wholly representative of the wider population.

In theory, any data containing both environmental variables and a relatively large sample of SNPs can be analysed using the MTM. As a data mining tool, it is more appropriate to use this technique when there is little prior information. In situations where a complex disease has been studied by different groups at the GWAS level, MTM is a useful way of analysing the different resultant SNPs identified, as the final stage of a meta-analysis. MTM gives more information than reducing the number of identified SNPs by means of adjustment for multiple testing.

It is also useful in cases where a risk has been identified in an initial GWAS or case control study, yet replication cannot reproduce the results. The MTM could be used with any environmental variables that are markedly different between the populations to add more information and possibly explain the differences.

9.1.4 Comparative Methods

Current methodologies evaluated both from the literature and experimentally, have different strengths and weaknesses. Having too many explanatory variables per outcome event in regression is a well recognised problem [39, 104]. In this particular scenario under study, with large numbers of variables and unknown interactions, logistic regression did not perform very well at identifying interactions. Using the stepwise selection procedure in logistic regression improved the identification rate, and proved to be the most successful comparable method, yet the false positive rate continued to be prohibitively high. However, the quantitative nature of the regression results, both for estimation of the associated risk and significance, provide a very good summary method and way of identifying the most promising associations from a smaller dataset. It is therefore a good last step following a parameter selection method and was used as such in the mixed tree method.

The MDR and ORMDR approaches to identifying interactions were not fully compatible with the data type in this study. The risk effects did not have absolute penetrance and there were marginal effects present. The data had been simulated to imitate an increase in risk, whereas the MDR method is most effective in situations with complete penetrance.

9.1.5 Analysis of Real Data

Prior to the analysis of real data in chapter 8, there were two main unknown questions: how the method would handle such data and how easy it would be to interpret the results? Compared to the simulated data, the overall identification rate for real data was lower than the expected false positive rate, probably as a result of decreased statistical noise and suggesting a relatively low false positive rate. One drawback of the simulation was the presumed independence and separate simulation of each variable, when in real data there are underlying associations between the variables. For example, in the simulation it was entirely possible for one person to eat the most of one nutrient and the least of another, whereas the real data more likely

represents a spectrum of unhealthy-healthy, with many of the similar risk factors forming similar population groupings.

The results showed a very likely independent main effect of rs2481952, interpretable by the importance across a number of environmental variables. Had one of the variables shown a strong interaction effect with this variable then it may have proven to be an interaction effect, with the perceived main effect a artefact of the interaction model. However in this case, it was clearly a main effect, a result that corresponds to similar findings throughout the literature, and warrants further analysis.

A number of the identified SNPs showed independent and interaction effects, in some cases with effect estimates in opposite directions, suggesting an increase in risk associated with the SNP but reduced by interaction with an environmental variable. The fact that such information was so easily interpretable is reassuring for further analysis using this approach.

9.2 Summary

In general, given the number of variables under consideration, in order for a traditional method to maintain power, extremely large sample sizes are required, which are expensive and time consuming. The more complex methods also tend to be very computationally intensive and take time to both set up and run analyses. Finally, the applicability of results from the study to the general population is a problem often encountered in eidemiology. As well as the inherent problems in sample selection prior to a study, the figures and estimated produced by many methods does not easily convert to an interpretable relationship or risk to the general population.

The importance of identifying gene-environment interactions is becoming more relevant as different study populations yield varying results from genome wide association studies (GWAS). More and more potential results are being produced; some which will turn out to be false, some which will stand up to repeated analysis and some which need an understanding of the environmental contex to be meaningful.

9.3 Public Health Implications

Many scientific endeavours are concerned solely with the increase in knowledge of a particular area, whilst refusing to be drawn on the potential benefits from such knowledge. However, epidemiological work, both conventional and genetic, usually has a basis in improving or understanding ill health at a population level, even if the target population is relatively small.

The benefits in identifying gene-environment interactions are likely to be most obvious with regard to drug targets. Being able to classify a subgroup that responds particularly well, or particularly badly, to a drug therapy informs future treatment options and prescribing decisions. The other benefits include an increased understanding of the disease and the ability to compare data with different findings to assess if an underlying interaction effect could be confounding the results. Given that NSAIDs are protective against CRC, yet can have other gastrointestinal side effects, it is useful to be able to make a more specific risk estimate with a basis in personal genetics and drug interactions, before advising daily NSAID intake.

Another area of research where MTM analysis may be fairly straightforward and provide useful insights is the study of maternal effects. It is relatively easy to follow up a woman during the nine months of gestation and there is the possibility of interactions between environmental or lifestyle factors during pregnancy and the genes of either the mother, or of the resultant offspring.

The identification of gene environment interactions will also help the general understanding of complex, chronic medical conditions. There are already recognised risks surrounding lifestyle and family history and the identification of specific risks is simply a more detailed examination of these risks. The population impact of this increased knowledge would be dependent on the specific findings, how applicable they were to the general population, how easy any sort of intervention would be and a whole host of other variables. In the case of colorectal cancer risk in Scotland, a number of the identified genes were also risk factors for other cancers, so the information gained has a wide reaching impact.

One of the most useful applications of the MTM is for risk factors that are only found in a proportion of the studies looking for them. Explaining the difference in terms of interaction, where possible, would mean that conflicting results actually added information to another study instead of negating its results.

This would be particularly pertinent given the large differences between the countries and populations being used to assemble genetic databases. Current approaches either study the within-population effects and then compare the results to see if they match those found in other populations or combine the samples into a larger population, where the effects of interaction may actually hide the real effects if not considered properly. For example, in Asian populations the consumption of dairy products tends to be lower which would mean a gene-interaction effect dependent on lactose would not be found within such a population. Combining this population with a European sample, not Asian in descent, would decrease the estimated Odds Ratio of the effect of the gene, had it not already been found to interact. This meta-analysis would effectively lose useful information. Entering such meta-analyses with the additional hypothesis from an MTM analysis could alert researchers to these complications.

9.4 Further Work

The results from different analyses of simulated data demonstrate the different strengths and weaknesses of different approaches to the problem of gene-environment interactions in large data sets. There are still areas where more work needs to be done to gauge success under different conditions. There are some results which have proved beyond the current capabilities of the MTM and there are analyses where the MTM approach could prove useful.

Method Testing

In order to understand more fully both the uses and limitations of the MTM, simulation studies could be undertaken to explore more complex underlying data models, including relatedness between subjects, related variables and sequential casual pathways (referred to earlier as Type A interactions). It would also be interesting to create samples with a number of different interactions, some with the same gene involved in different ways with different environmental variables. There is increasing use of genotyping large numbers of people within founder populations, so it is especially important to check the effects of relatedness on the MTM and adapt it accordingly. Finally, studies containing a model of genetic heterogeneity would be an interesting addition.

Other testing could include a simulation of larger scale, genome wide association study to study the scalability of the MTM. This would require substantially greater computational time, but is a necessary piece of work, considering the pace of development in molecular genetics.

Potential Adaptations

The method is already fairly adaptable; however, a more detailed analysis of the necessary significance thresholds would be beneficial. The issue of multiple testing is a potential area of criticism, when analysing such large numbers of variables in a single analysis. Although not problematic in the data mining steps of the analysis, when making statistical estimates on the significance threshold of any findings, it could be considered statistically invalid to use a threshold as high as 0.05. To confirm

a more appropriate step, or an accurate threshold may require further work and statistical development.

Potential Analyses

Ideally the next step would be to use the MTM on a similar dataset, for a different complex disease. This might include Type 2 diabetes, or an autoimmune condition which may have a genetic predisposition and an environmental trigger.

From the results on the real data, it would be particularly interesting to follow up the findings on NSAIDs and the main SNP effect on risk of CRC. It would also be useful to combine a number of the B vitamin results into a general measure of B vitamin intake, and repeat the analysis. This would be particularly interesting as the majority of analysis on dietary variables have been carried out in a univariate fashion.

9.5 Conclusions

Gene environment interactions are an important part of understanding complex disease, yet notoriously difficult to identify. The data type and characteristics of the interaction itself can vary greatly and there are only a few statistical tools capable of analysing even a few of these possibilities. The development of the mixed tree method and development of the corresponding software package is a useful and effective way to detect interactions between genetic and environmental risk factors for complex disease. This method can be used to analyse interactions between any number of environmental variables and up to 300 SNPs in populations of between 200 and 3000 people.

The Mixed Tree Method performed very well in particular circumstances and less well in others. The most obvious examples of variation in performance were found with the different underlying interaction models, but the variation across different effect and sample sizes should also be a consideration in any further analysis. The other results point to a general rule, that exposure of around 50% has the greatest power for identifying interaction effects.

An important finding however, was the consistency of both the identification rate and false positive rate as the number of SNPs increased. No other method has performed as well on similar data. Stepwise regression was unable to handle an increasing number of variables, even when the data was manipulated in a similar way to MTM regarding the environmental variables. None of the other methods were directly comparable. The handling and success of the real data analysis proved that the MTM is able to cope with real data, in some ways better than the simulated data, and a small number of potential interactions were identified for further analysis.

In summary, gene-environment interactions are an important part of the puzzle in complex disease aetiology and the least studied. There is a need to develop methods to identify interactions and none of the current methodologies are wholly appropriate. The Mixed Tree Method provides a new way of approaching this analysis, a way that showed considerable success in both simulated and real data analysis.

References

1. Lilienfield DES, P D. Foundations of Epidemiology. New York: Oxford University Press; 1994.

2. Kupper LLH, M D. Interaction in Epidemiological Studies. American Journal of Epidemiology 1978;108(6):447 - 53.

3. Blot WJ, Day NE. Synergism and Interaction: Are They Equivalent? American Journal of Epidemiology 1979;110(1):99 - 100.

4. Rothman KG, S; Walker, AM. Concepts of Interaction. American Journal of Epidemiology 1980;112(4):467-70.

5. Plato. The Republic; 360BC.

6. Wedekind CS, T; Bettens, F; Paepke, A J. MHC-dependent mate preferences in humans. JSTOR 1995;260(1359):245 - 9.

7. Brewer G. Annotation: Human ecology, an expanding role for the human geneticist. American Journal of Human Genetics 1971;23:92-4.

8. Omenn GS. Prospects for Pharmacogenetics and Ecogenetics in the New Millenium. Drug Metabolism and Disposition 2001;29(4):611 - 4.

9. Faustman EO, GS. Risk assessment, in *Casarett and Doull's Toxiclogy:The Basic Science of Poisons*. 5th ed. New York: McGraw-Hill; 1996.

10. Omenn GS. Putting environmental risks in a public health context Public Health 1996;111:514 - 6.

11. Kraft P, Hunter D. Integrating epidemiology and genetic association: the challenge of gene-environment interaction. Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences 2005;360(1460):1609-16.

12. Muin J Khoury RD, Marta Gwinn, Mary Lou Lindegren, and Paula Yoon. Do We Need Genomic Research for the Prevention of Common Diseases with Environmental Causes? American Journal of Epidemiology 2005;161(9):799 - 805.

13. Caraco Y. Genes and the Response to Drugs. New England Journal of Medicine 2004;351(27):2867 - 9.

14. Milano GAL, W; Formento, J.L.; Largillier, R; Campone, M; Chamorey, E; Francoual, M; Ferrero, J.M. HER2 genetic polymorphism and pharmacodynamics of trastuzumab-based treatment in breast cancer patients. Journal of Clinical Oncology 2005;23(16s):501.

15. DOH DoH. An organisation with a memory. In; 2000:xiv, 92p.

16. LIPID TLTIwPiIDSG. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. New England Journal of Medicine 1998;339:1349 - 57.

17. Niino M FT, Yabe I, Kikuchi S, Sasaki H, Tashiro K. Vitamin D receptor gene polymorphism in multiple sclerosis and the association with HLA class II alleles. Journal of Neurological Science 2000;177(1):65 - 71.

18. ACC/SCN. Controlling Iron deficiency. State of the Art Series. Nutrition Policy Discussion Paper no. 9. In; 1991.

19. Miller LH. Impact of malaria on genetic polymorphisms and genetic diseases in Africans and African-Americans. Proc Natl Acad Sci USA 1997;91:2415-9.

20. Connor RI PW, Sheridan KE, Koup RA. Macrophages and CD4+ T lymphocytes from two multiply exposed, uninfected individuals resist infection with primary non-syncytium-inducing isolates of human immunodeficiency virus type 1. Journal of Virology 1996;70(12):8758 - 64.

21. Wang F-S. Current status and prospects of studies on human genetic alleles associated with hepatitis B virus infection. World Journal of Gastroenterology 2003;9(4):641-4.

22. Gardner MJ, Hall N, Fung E, et al. Genome sequence of the human malaria parasite Plasmodium falciparum.[see comment]. Nature 2002;419(6906):498-511.

23. Winkelmann BRH, J ;Kraus, W E ;Merlini, P ;Keavney, B ;Grant, P J ;Muhlestein, J B ;Granger,C B. Genetics of coronary heart disease: Current knowledge and research principles. American Heart Journal 2000;140(4):S11 - S26.

24. Phillips CG, KJ; Poole, C. Lead editorial: The need for greater perspective and innovation in epidemiology. Epidemiological Perspectives and Innovations 2004;1(1):1-4.

25. Skrabanek P. Has risk-factor epidemiology outlived its usefulness? American Journal of Epidemiology 1993;138(11):1016.

26. Buchanan AV, Weiss K, M FS. Dissecting complex disease: the quest for the Philosopher's Stone? International Journal of Epidemiology 2006;35:562 - 71.

27. Taubes G. Epidemiology faces its limits. Science 1995;269:164-9.

28. Smith GD, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. The Lancet 2005;366(9495):1484-98.

29. Rose G. The strategy of preventative medicine. Oxford: Oxford University Press; 1992.

30. Rose G. Sick individuals and sick populations. International Journal of Epidemiology 1985;14:32 - 8.

31. Committee UNS. The UK National Screening Committee's Criteria for appraising the viability, effectiveness and appropriateness of a screening programme.

32. Willett W. Balancing life-style and genomics research for disease prevention. Science 2002;296:695-8.

33. Ioannidis JPA. Commentary: Grading the credibility of molecular evidence for complex disease. International Journal of Epidemiology 2006;35:572-7.

34. Howe GH, T; Hislop, TG; Iscovich, JM; Yuan, JM; Katsouyanni, K; Lubin, F; Marubini, E; Modan, B; Rohan, T; Toniolo, P; Shunzhang. Dietary Factors and Risk of Breast Cancer: Combined Analysis of 12 Case-Control Studies. Journal of the National Cancer Institute 1990;82(7):561-9.

35. Katsouyanni KT, D; Boyle, P; Xirouchaki, E; Trichopoulo, A; Lisseos, B; Vasilaros, S; Macmahon, B. Diet and breast cancer: A case-control study in Greece. International Journal of Cancer 1986;38(6):815-20.

36. van Gils CH, Peeters PHM, Bueno-de-Mesquita HB, et al. Consumption of vegetables and fruits and risk of breast cancer.[see comment]. JAMA 2005;293(2):183-93.

37. Weed DL. Commentary: Rethinking epidemiology. International Journal of Epidemiology 2005;35:583 - 6.

38. Nichter MN, M; Vuckovic, N; Quintero, G; Ritenbaugh, C. Smoking experimentation and initiation : qualitative and quantitative findings among adolescent girls. Tobacco Control 1997;6:285-95.

39. Concato J, Feinstein A, Holford T. The risk of determining risk with multivariable models. Annals of Internal Medicine 1993;118(3):201 - 10.

40. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 2010;86(1):6-22.

41. Baranzini SE, Galwey NW, Wang J, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Human Molecular Genetics 2009;18(11):2078-90.

42. Ottman R. Theoretical Epidemiology. Gene-Environment Interaction: Definitions and Study Designs. Preventative Medicine 1996;25:764-70.

43. Swerdlow DM, ED; Rodriguez, M; Tejada, E; Ocampo, C; Espejo, L; Barrett, TJ; Petzelt, J; Bean, NH; Seminario, L. Severe life-threatening cholera associated with blood group O in Peru: implications for the Latin American epidemic. Journal of Infectious Disease 1994;170(2):468-72.

44. Ahlbom A, Alfredsson L. Interaction: A word with two meanings creates confusion. European Journal of Epidemiology 2005;20:563-4.

45. Hermisson JW, GP. The population genetic theory of hidden variation and genetic robustness. Genetics 2004;168:2271-84.

46. Kroymann JM-O, T. Epistasis and balanced polymorphism influencing complex trait variation. Nature 2005;435:95-8.

47. Lee PJ, Ridout D, Walter JH, Cockburn F. Maternal phenylketonuria: report from the United Kingdom Registry 1978-97.[see comment]. Archives of Disease in Childhood 2005;90(2):143-6.

48. Price EW. The site of lymphatic blockage in endemic (non-filarial) elephantiasis of the lower leg. Journal of Tropical Medicine and Hygiene 1977;80:230-7.

49. Price EW. Endemic elephantiasis of the lower legs in Rwanda and Burundi. Tropical Geographical Medicine 1976;28:286-90.

50. Davey GG, E; Adeyemo, A; Rotimi, C; Newport, M; Desta, K. Podoconiosis: a tropical model for gene-environment interactions? Transactions of the Royal Society of Tropical Medicine and Hygiene 2007;101:91-6.

51. Hunter DJ. Gene-environment interactions in human diseases. Nature Reviews Genetics 2005;6:287-98.

52. Deybach JdV, H; Nordmann, Y. The Inherited Enzymatic Defect in Porphyria Variegata. Human Genetics 1981;58:425-8.

53. Deybach JC, Puy H, Robreau AM, et al. Mutations in the protoporphyrinogen oxidase gene in patients with variegate porphyria. Human Molecular Genetics 1996;5(3):407-10.

54. San Jose CC, A; Benitez, J; Carrillo, J A; Jimenez, M; Gervasin, G. CYP1A1 gene polymorphisms increase lung cancer risk in a high-incidence region of Spain: a case control study. BMC Cancer 2010;10(463).

55. Kuper HT-E, L; Harlow, BL; Cramer, DW. Population Based Study of Coffee, Alcohol and Tobacco Use and Risk of Ovarian Cancer. International Journal of Cancer 2000;88:313-8.

56. Stensvold IJ, BK. Coffee and cancer: a prospective study of 43,000 Norwegian men and women. Cancer causes and control 1994;5:401-8.

57. Larsson SC, Wolk A. Coffee consumption is not associated with ovarian cancer incidence. Cancer Epidemiology, Biomarkers & Prevention 2005;14(9):2273-4.

58. Terry KT-E, L; Garner, EO; Vitonis, AF; Cramer, DW. Interaction between CYP1A1 Polymorphic Variants and Dietary Exposures Influencing Ovarian Cancer Risk. Cancer Epidemiology, Biomarkers & Prevention 2003;12:187-90.

59. Murray CL, AD. The Global Burden of Disease. Cambridge: Oxford Univesity Press; 1996.

60. States SGotU. The health consequences of smoking: chronic obstructive lung disease. In: Services DoHaH, ed. Washington DC; 1984.

61. Stang PL, E; Silberham, C; Kempel, A; Keating, ET. The prevalence of COPD. Using Smoking Rates to Estimate Disease Frequency in the General Population. Chest 2000;117:354S-9S.

62. Sandford AJ, Silverman EK. Chronic obstructive pulmonary disease. 1: Susceptibility factors for COPD the genotype-environment interaction. Thorax 2002;57(8):736-41.

63. Janus EP, NT; Carrell, RW. Smoking, lung function, and α1-antitrypsin deficiency. Lancet 1985;1(8421):152-4.

64. Song NT, W; Xing, D; Lin, D. CYP 1A1 polymorphism and risk of lung cancer in relation to tobacco smoking: a case-control study in China. Carcinogenesis 2001;22(1):11 - 6.

65. Walter SD. Methods of Reporting Statistical Results from Medical Research Studies. American Journal of Epidemiology 1995;141(10):896 - 906.

66. Wolfe RC, G. Communicating the uncertainty in research findings:confidence intervals. J Sci Med Sport 2004;7(2):138 - 43.

67. Mosteller F. A k-sample slippage test for an extreme population. Annals of Mathemetics and Statistics 1948;19:58 - 65.

68. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Fourteenth International Joint Conference on Artificial Intelligence 1995; San Mateo; 1995. p. 1137 - 43.

69. Efron BT, R. An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability): Chapman & Hall; 1993.

70. Efron B. Estimating the error rate of a prediction rule improvement on cross validation. Journal of the American Statistical Association 1983;78(382):316 - 30.

71. Brookes AJ. The essence of SNPs. Gene 1999;234(2):177 - 86.

72. Vasily RP, B; Shamil, S. Human non-synonymous SNPs: server and survey. Computational Biology 2002;30(17):3894 - 900.

73. Ozaki KO, Y; Iida, A; Sekine, A; Yamada, R; Tsunoda, T; Sato, H; Sato, H; Hori, M; Nakamura, Y; Tanaka, T. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nature Genetics 2002;32(4):650 - 4.

74. Edwards AOR, R; Abel, K J; Manning, A; Panhuysen, C; Farrer, L A Complement factor H polymorphism and age-related macular degeneration. Science 2005;308:421 - 4.

75. Haines JLH, M A; Schmidt, S; Scott, W K; Olson, L M; Gallins, P; Spencer, K L; Kwan, S Y; Noureddine, M; Gilbert, J R; Schnetz-Boutaud, N; Agarwal, A; Postel, E A; Pericak-Vance, M A. Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. Science 2005;308:419 - 21.

76. Thomas DC. Are we ready for Genome-wide Association Studies? Cancer Epidemiology Biomarkers & Prevention 2006;15(4):595 - 8.

77. Bourgain C, Genin E, Cox N, Clerget-Darpoux F. Are genome-wide association studies all that we need to dissect the genetic component of complex disease. European Journal of Human Genetics 2007;15:260 - 3.

78. Klein DA. Visual Exploration of Large Data Sets. Communications of the ACM 2001;44(8):39 - 44.

79. Johansson JC, M; Jern, M. 3-Dimensional Display for Clustered Multi-Relational Parallel Coordinates. In: Ninth International Conference on Information Visualisation 2005; Sweden; 2005.

80. Thomas AC, N J. Graphical Modeling of the Joint Distribution of Alleles at Associated Loci. American Journal of Human Genetics 2004;74:1088 - 101.

81. Templeton A. Epistasis and the evolutionary process. Oxford: Oxford University Press; 2000.

82. Moore JHW, S M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays 2005;27(6):637 - 46.

83. Culverhouse RS, B K; Lin, J; Reich, T. A perspective on epistasis: limits of models displaying no main effect. American Journal of Human Genetics 2002;70:416 - 71.

84. Hahn LR, MD; Moore, JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 2003;19(3):376-82.

85. Hodge SE. Linkage Analysis versus Association Analysis: Distinguishing between Two Models That Explain Disease-Marker Associations. American Journal of Human Genetics 1993;53:367 - 84.

86. Greenberg DA. Linkage analysis of "necessary" disease loci versus "susceptibility" loci. American Journal of Human Genetics 1993;52(1):135 - 43.

87. Collins FS, Guyer MS, Chakravati A. Variations on a theme:cataloging human DNA sequence variation. Science 1997;278:1580 - 1.

88. Colhoun H, McKeigue P, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet 2003;361:865-72.

89. Gauderman WJ. Sample Size Requirements for Association Studies of Gene-Gene Interaction. American Journal of Epidemiology 2002;155(5):478 - 84.

90. Ritchie MH, LW; Moore, JH. Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy and Genetic Heterogeneity. Genetic Epidemiology 2003;24:150-7.

91. Fisher RA. The correlations between relatives on the supposition of Mendelian inheritance. In: Trans R Soc Edinburgh; 1918; Edinburgh; 1918. p. 399 - 433.

92. Demuth JPW, M J. Experimental Methods for Measuring Gene Interactions. Annu Rev Ecol Evol Syst 2006;37:289 - 316.

93. Lynch MW, B. Genetics and Analysis of Quantitative Traits. Sunderland: Sinauer Assoc; 1997.

94. Hotelling H. The generalisation of Student's ratio. Annals of Mathemetics and Statistics 1931;2:360 - 78.

95. Xiong MZ, J; Boerwinkle, E. Generalised T2 test for genome association studies. American Journal of Human Genetics 2002;70 1257 - 68.

96. Fan RK, M. Genome asssociation studies of complex diseaseby case-control design. American Journal of Human Genetics 2003;72:850 - 68.

97. Freidman JH. Multivariate Adaptive Regression Splines. The Annals of Statistics 1991;19:1 - 67.

98. Dudbridge FG, A; Koeleman, B. Detecting multiple associations in genome-wide studies Human Genetics 2006;2:310 - 7.

99. Schaid DM, SK; Hebbring, SJ; Cunnigham, JM; Thibodeau, SN. Nonparametric tests of association of multiple genes with human disease. American Journal of Human Genetics 2005;76:780 - 93.

100. Ge DZ, H; Huang, Y; Treiber, F A; Harshfield, G A; Snieder, H; Dong, Y. Multilocus Analyses of Renin–Angiotensin–Aldosterone System Gene Variants on Blood Pressure at Rest and During Behavioral Stress in Young Normotensive Subjects. Hypertension 2007;49:107 - 12.

101. Gu DS, S; Ge, D; Chen, S; Huang, J; Li, B; Chen, R; Qiang, B. Association Study With 33 Single-Nucleotide Polymorphisms in 11 Candidate Genes for Hypertension in Chinese. Hypertension 2006;47:1147 - 54.

102. Lu YL, P-Y; Xiao, P; Deng, H-W. Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. Bioinformatics 2005;21(14):3105 - 13.

103. Wallace CN, S; Braund, P; Zhang, F; Tobin, M; Falchi, M; Ahmadi, K; Dobson, R; Marcano, A; Hajat, C. Genome-wide Association Study Identifies Genes for Biomarkers of Cardiovascular Disease: Serum Urate and Dyslipidemia. American Journal of Human Genetics 2008;82(1):139 - 49.

104. Peduzzi PC, J; Kemper, E; Holford, TR; Feinstein, AR. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. Journal of Clinical Epidemiology 1996;49(12):1371 - 9.

105. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 1996;58:267 - 88.

106. Kirkwood BRS, J A C. Essential Medical Statistics. 2nd ed. Oxford: Blackwell Publishing; 2003.

107. Falconer DM, TFC. Introduction to Quantitative Genetics: Harlow: Longman; 1996.

108. Welstead S. Neural Networks and Fuzzy Logic Application in C/C++. New York: John Wiley and Sons; 1994.

109. Lucek PRO, J. Neural Network Analysis of Complex Traits. Genetic Epidemiology 1997;14:1101 - 6.

110. Bhat A, Lucek PR, Ott J. Analysis of complex traits using neural networks. Genetic Epidemiology 1999;17(Suppl 1):S503 - S7.

111. Ott J. Neural Networks and Disease Association Studies. American Journal of Medical Genetics (Neuropsychiatric Genetics) 2001;105:60 - 1.

112. Curtis DN, B; Sham, P C. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. Annals of Human Genetics 2001;65(1):95 - 107.

113. Sherriff A; Ott J. Applications of neural networks for gene finding. Advances in Genetics 2001;42:287 - 98.

114. Li WH, F; Falk, C T. Design of artificial neural network and its applications to the analysis of alcoholism data. Genetic Epidemiology 1999;17(Suppl 1):S223 - S8.

115. Tomita YT, S; Hasegawa, Y; Suzuki, Y; Shirakawa, T; Kobayashi, T; Honda, H. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma BMC Bioinformatics 2004;5(120).

116. Tomita YT, S; Suzuki, Y; Shirakawa, T; Kobayashi, T; Honda, H. Artificial Neural Network Model for Prediction of Childhood Asthma Using Single Nucleotide Polymorphism Data. Genome Informatics 2003;14:593 - 4.

117. North BVC, D; Cassell, P G; Hitman, G A; Sham ,P C. Assessing Optimal Neural Network Architecture for Identifying Disease-associated Multi-marker Genotypes using a Permutation Test, and Application to Calpain 10 Polymorphisms Associated with Diabetes. Annals of Human Genetics 2003;67(4):348 - 56.

118. Valavanis IKM, S G; Grimaldi, K A; Nikita, K S. Analysis of Postprandial Lipemia as a Cardiovascular Disease Risk Factor using Genetic and Clinical Information: An Artificial Neural Network Perspective. IEEE 2008.

119. Mutoh HH, N; Tajima, K; Kobayashi, T; Honda, H. Exhaustive exploring using Artificial Neural Network for identification of SNPs combination related to *Heliobacter pylori* infection susceptibility. Chem-Bio Informatics Journal 2005;5(2):15 - 26.

120. Koza JRR, J P. Genetic generation of both the weights and architecture for aneural network. In: Neural Networks - Seattle International Joint Conference; 1991; Seattle, WA, USA; 1991. p. 397 - 404.

121. Ritchie MW, BC; Parker, JS; Hahn, LW; Moore, JH. Optimization of neural network architecture using genetic programming improves detection and modelling of gene-gene interactions in studies of human disease. BMC Bioinformatics 2003;4(28).

122. Motsinger AMD, S M; Hahn, L W; Ritchie, M D. Comparison of Neural Network Optimization Approaches for Studies of Human Genetics. In: Applications of Evolutionary Computing. Heidelberg: Springer Berlin; 2006:103 - 14.

123. Ritchie MDC, C S; Moore, J H. Genetic Programming Neural Networks as a Bioinformatics Tool for Human Genetics. LECTURE NOTES IN COMPUTER SCIENCE 2004;Sect. 438 - 48.

124. Bush WS, Montsinger AA, Dudek SM, Ritchie MD. Can neural network constriants in GP provide power to detect genes associated with human disease. Lecture notes in Computer Science 2005;3449:44 - 53.

125. Motsinger AAL, S L; Mellick, G; Ritchie, M D. GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics 2006;7(39).

126. O'Neill MR, C. Grammatical evolution. IEEE Transactions on Evolutionary Computations 2001;5:349 - 57.

127. Hoh JO, J. Mathematical multi-locus approaches to localizing complex human trait genes. Nature Reviews Genetics 2003;4:701 - 9.

128. Lippmann RP. Pattern classification using neural networks. Communications Magazine, IEEE 1989;27(1):47 - 50, 9 - 64.

129. Ott JH, J. Statistical multilocus methods for disequilibrium analysis in complex traits. Human Mutation 2001;17:285 - 8.

130. Hoh JW, A; Zee, R; Cheng, S; Reynolds, R; Lindpainter, K; Ott, J. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Annals of Human Genetics 2000;64:413 - 7.

131. Hoh JW, A; Ott, J. Trimming, Weighting and Grouping SNPs in Human Case-Control Association Studies. Genome Research 2001;11:2115 - 9.

132. Xu JT, A; Little, J; Bleecker, E R; Meyers, D A. Positive results in association studiesare associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? Human Genetics 2002;111:573 - 4.

133. Kim SZ, K; Sun, F. Detecting susceptibility genes in case-control studies using set association. BMC Genetics 2003;4(Suppl. 1):1 - 15.

134. Ott JH, J. Set Association Analysis of SNP Case-Control and Microarray Data. Journal of Computational Biology 2003;10(3-4):569 - 74.

135. Wille AH, J; Ott, J. Sum Statistics for the Joint Detection of Mutiple Disease Loci in Case-Control Assocation Studies With SNP MArkers. Genetic Epidemiology 2003;25:350 - 9.

136. Abecasis GR, Cookson WOC, Cardon LR. The Power to Detect Linkage Disequilibrium with Quantitative Traits in Selected Samples. American Journal of Human Genetics 2001;68:1463 - 74.

137. Evans DMM, J; Morris, A P; Cardon, L R. Two-Stage Two-Locus Models in Genome-Wide Association. PLoS Genetics 2006;2(9):1424 - 32.

138. Hoh J. Sumstat. In. Yale University, New Haven; 2008.

139. de Quervain DJFP, R; Wollmer, M A; Grimaldi, L M E; Tsolaki, M; Streffer, J R; Hock, C; Nitsch, R M; Mohajeri, M H; Papassotiropoulos, A. Glucocorticoid-related genetic susceptibility for Alzheimer's disease. Human Molecular Genetics 2004;13(1):47 - 52.

140. Thornton-Wells TAM, J H; Martin, E R; Pericak-Vance, M A; Haines, J L. Confronting Complexity in Late-Onset Alzheimer Disease: Application of Two-Stage Analysis Approach Addressing Heterogeneity and Epistasis. Genetic Epidemiology 2008;32:187 - 203.

141. Danoy PM, S; Dessen, P; Pignat, C; Boulet, T; Monet, M; Bouchardy, C; Lathrop, M; Sarasin, A; Benhamou, S. Variants in DNA double-strand break repair and DNA damage-response genes and susceptibility to lung and head and neck cancers. International Journal of Cancer 2008;123:457 - 63.

142. Zee RYLH, J; Cheng, S; Reynolds, R; Grow, M A; Silbergleit, A; Walker, K; Steiner, L; Zangenberg, G; Fernandez-Ortiz, A; Macaya, C; Pintor, E; Fernandez-Cruz, A; Ott, J; Lindpaintner, K. Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. The Pharmacogenomics Journal 2002;2:197 - 202.

143. Alon N, Kahale N. A spectral technique for coloring random 3-colorable graphs In: press A, editor. The twenty sixth annual ACM symposium on Theory of computing; 1994; 1994. p. 346-55.

144. Drineas PF, A; Kannan, R; Vempala, S; Vinay, V. Clustering large graphs via the singular value decomposition. In: Tenth Annual ACM-SIAM Symposium on Discrete Algorithms; 1999; 1999. p. 219 - 99.

145. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000;97(18):10101 - 6.

146. Yeung KYR, W L. Principal component analysis for clustering gene expression data. Bioinformatics 2001;17(9):763 - 74.

147. Wall MER, A; Rocha, L M. Singular value decomposition and principal component analysis. In: Berrar DPD, W; Granzow, M, ed. A Practical Approach to Microarray Data Analysis. Kluwer: Norwell; 2003:91 - 109.

148. Zhang YL, Q; Dai, G; Zhang, H. A New Recursive Least-Squares Identification Algorithm Based On Singular Value Decomposition. In: The 33rd Conference on Decision and Control; 1994; Florida; 1994.

149. Golub GK, W. Calculating the Singular Values and Pseudo-Inverse of a Matrix. JSTOR 1965;2(2):205 - 24.

150. Gilula Z. Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables. Biometrika 1979;66(2):339 - 44.

151. Eriksson N. Tree construction using singular value decomposition. Cambridge: Cambridge University Press; 2005.

152. Lennon MM, G; Mouchot, M C; Hubert-Moy, L. Independent Component Analysis as a tool for the dimensionality reduction and the representation of hyperspectral images. IEEE 2001;6:2893 - 5.

153. Comon P. Independent component analysis, A new concept? Signal processing 1994;36:287 - 314.

154. Mika SS, B; Smola, A; Müller, K R; Scholz, M; Rätsch, G. Kernel PCA and de-noising in feature spaces. In: Advances in neural information processing systems II; 1998; Cambridge, MA: MIT press; 1998. p. 536 - 42.

155. Price AP, NJ; Plenge, RM; Weinblatt, ME; Shadick, NA; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 2006;38(8):904 - 9.

156. Heath SC, Gut IG, Brennan P, et al. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 2008;16(12):1413-29.

157. Wang KA, D. A Principal Components Regression Approach to Multilocus Genetic Association Studies. Genetic Epidemiology 2008;32:108 - 18.

158. Allen-Brady K, Miller J, Matsunami N, et al. A high-density SNP genome-wide linkage scan in a large autism extended pedigree. Mol Psychiatry 2008.

159. Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Research 2001;11(3):458-70.

160. Moore JH, Lamb JM, Brown NJ, Vaughan DE. A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. Clinical Genetics 2002;62(1):74-9.

161. Tahri-Daizadeh NT, D A; Nicaud, V; Manuel, N; Cambien, F; Tiret, L. Automated Detection of Informative Combined Effects in Genetic Association Studies of Complex Traits. Genome Research 2003;13:1952 - 60.

162. Kristensen VNT, A; Geisler, J; Faldaas, A; Grenaker, G I; Lingjaerde, O C; Fjeldstad; Yakhini, Z; Lonning, P E; Borresen-Dale, A L. Multilocus analysis of SNP and metabolic data with a given pathway. BMC Genomics 2006;7(5):1 - 14.

163. Culverhouse RK, T; Shannon, W. Detecting epistatic interaction contributing to quantitative traits. Genetic Epidemiology 2004;27:141-52.

164. Payseur BAC, A G; Hixson, J; Boerwinkle, E; Sing, C F; . Contrasting multi-site genotypic distributions among discordant quantitative phenotypes: the APOA1/C3/A4/A5 gene cluster and cardiovascular disease risk factors. Genetic Epidemiology 2006;30(6):508 - 18.

165. Coutinho AM, Sousa I, Martins M, et al. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Human Genetics 2007;121(2):243 - 56.

166. Childs E, Hohoff C, Deckert J, Xu K, Badner J, de Wit H. Association between ADORA2A and DRD2 Polymorphisms and Caffeine-Induced Anxiety. In: Neuropsychopharmacology; 2008.

167. Lou X-YC, G-B; Yan, L; Ma, J Z; Zhu, J; Elston, R C; Li, M D A Generalized Combinatorial Approach for Detecting Gene-by-Gene and Gene-by-Environment Interactions with Application to Nicotine Dependence. American Journal of Human Genetics 2007;80(6):1125 - 37.

168. Sha QZ, X; Zuo, Y; Cooper, R; Zhang, S. A Combinatorial Searching Method for Detecting a Set of Interacting Loci Associated with Complex Traits. Annals of Human Genetics 2006;70:677 - 92.

169. Maxwell CAM, V; Sole, X; Gomez, L; Hernandez, P; Urruticoechea, A; Pujana, M A. Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. Molecular Cancer 2008;7(4):1 - 10.

170. Ritchie MDH, L H; Roodi, N; Bailey, L; Dupont, W; Parl, F; Moore, J. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer The American Journal of Human Genetics 2001;69(1):138 - 47.

171. Hahn LM, JH. Ideal discrimination of discrete clinicla endpoints using multilocus genotypes. In Silico Biology 2004;4:183-94.

172. Martin ER, MD; Hahn, LW; Kang, S; Moore, JH. A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. Genetic Epidemiology 2006;30:111-23.

173. Xu JL, J; Wiklund, F; Sun, J; Lindmark, F; Hsu, F C; Dimitrov, L; Chang, B; Turner, A R; Liu, W; Adami, H O; Suh, E; Moore, J H; Zheng, S L; Isaacs, W B; Trent, J M; Grönberg, H. The Interaction of Four Genes in the Inflammation Pathway Significantly Predicts Prostate Cancer Risk. Cancer Epidemiology Biomarkers & Prevention 2005;14:2563 - 8.

174. Andrew AS, Nelson HH, Kelsey KT, et al. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. Carcinogenesis 2006;27(5):1030-7.

175. Soares ML, Coelho T, Sousa A, et al. Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: complexity in a single-gene disease. Human Molecular Genetics 2005;14(4):543-53.

176. Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH. A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. Genetic Epidemiology 2006;30(2):111-23.

177. Brassat D, Motsinger AA, Caillier SJ, et al. Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. Genes Immun 2006;7(4):310-5.

178. Cho YM, Ritchie MD, Moore JH, et al. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia 2004;47(3):549 - 54.

179. Chan IHS, Leung TF, Tang NLS, et al. Gene-gene interactions for asthma and plasma total IgE concentration in Chinese children. Journal of Allergy and Clinical Immunology 2006;117(1):127-33.

180. Millstein JC, D V; Gilliland, F D; Gauderman, W J. A Testing Framework for Identifying Susceptibility Genes in the Presence of Epistasis. American Journal of Human Genetics 2005;78:15 - 27.

181. Sanada H, Yatabe J, Midorikawa S, et al. Single-Nucleotide Polymorphisms for Diagnosis of Salt-Sensitive Hypertension. Clin Chem 2006;52(3):352-60.

182. Williams SM, Ritchie MD, Phillips Iii JA, et al. Multilocus Analysis of Hypertension: A Hierarchical Approach. Human Heredity 2004;57(1):28-38.

183. Coffey CS, Hebert PR, Ritchie MD, et al. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene Interactions on risk of myocardial infarction: The importance of model validation. BMC Bioinformatics 2004;5(49):1 - 10.

184. Mannila MNE, P; Ericsson, C; Hamsten, A; Silveira, A. Epistatic and pleiotropic effects of polymorphisms in the fibrinogen and coagulation factor XIII genes on plasma fibrinogen concentration, fibrin gel structure and risk of myocardial infarction. Thrombosis and Haemostasis 2006;95(3):420 - 7.

185. Tsai C-T, Lai L-P, Lin J-L, et al. Renin-Angiotensin System Gene Polymorphisms and Atrial Fibrillation. Circulation 2004;109(13):1640-6.

186. Moore JHG, J C; Tsai, C T; Chiang, F T; Holden, T; Barney, N; White, B C. A flexible computational framework for detecting characterizing and interupting statistical patterns of epistasis in genetic studies of human disease suscepitibility. Journal of Theoretical Biology 2006;241(2):252 - 61.

187. Motsinger AA, Ritchie MD. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genetic Epidemiology 2006;30(6):546-55.

188. Asselbergs F, Moore J, van den Berg M, et al. A role for CETP TaqIB polymorphism in determining susceptibility to atrial fibrillation: a nested case control study. BMC Medical Genetics 2006;7(1):39.

189. Ma DQ, Whitehead PL, Menold MM, et al. Identification of Significant Association and Gene-Gene Interaction of GABA Receptor Subunit Genes in Autism. The American Journal of Human Genetics 2005;77(3):377-88.

190. Ashley-Koch AE, Mei H, Jaworski J, et al. An Analysis Paradigm for Investigating Multi-locus Effects in Complex Disease: Examination of Three GABAA Receptor Subunit Genes on 15q11-q13 as Risk Factors for Autistic Disorder. Annals of Human Genetics 2006;70(3):281-92.

191. Qin S, Zhao X, Pan Y, et al. An association study of the N-methyl-D-aspartate receptor NR1 subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray. Eur J Hum Genet 2005;13(7):807-14.

192. Lou X-Y, Chen G-B, Yan L, et al. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. American Journal of Human Genetics 2007;80(6):1125-37.

193. Chung Y, Lee SY, Elston RC, Park T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics 2007;23(1):71-6.

194. Raleigh SM, van der Merwe L, Ribbans WJ, Smith RKW, Schwellnus MP, Collins M. Variants within the MMP3 gene are associated with Achilles tendinopathy: possible interaction with the COL5A1 gene. British Journal of Sports Medicine 2009;43(7):514-20.

195. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Boca Raton, London, New York, Washington DC: Chapman & Hall / CRC; 1984.

196. Kattan MWH, K R; Beck, J R. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. Computers and Biomedical Research 1998;31(5):363 - 73.

197. Kattan MW. Comparison of Cox regression with other methods for determining prediction models and nomograms. Journal of Urology 2003;170(6 Pt 2):S 6 - 9.

198. James KEW, R F; Kraemer, H C. Repeated split sample validation to assess logistic regression and recursive partitioning: an application to the prediction of cognitive impairment. Statistics in Medicine 2005;24:3019 - 35.

199. Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. Journal of Chronic Disease 1984;37(9 - 10):721 - 31.

200. Lee JWU, S H; Lee, J B; Mun, J; Cho, H. Scoring and staging systems using cox linear regression modeling and recursive partitioning. Methods of Informatics in Medicine 2006;45(1):37 - 43.

201. Austin PC. A comparison of regression trees, logistic regression, generalised additive models, and multivariate adaptive regression splines for predicting AMI mortality. Statistics in Medicine 2007;26:2937 - 57.

202. Nelson LMB, D A; Longstreth Jnr, W T; Shi, H. Recursive partitioning for the identification of disease risk subgroups: a case control study of subarachnoid hemorrhage. Journal of Clinical Epidemiology 1998;51:199 - 209.

203. Foulkes ASD, V; Hertogs, K. Combining genotype groups and recursive partitioning: an application to human immunodeficiency virus type 1 genetics data. Applied Statistics 2004;53(2):311 - 23.

204. Bastone L, Reilly M, Rader DJ, Foulkes AS. MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. Human Heredity 2004;58(2):82-92.

205. Breiman L. Random Forests. Machine Learning 2001;45:5 - 32.

206. Bureau A, Dupuis J, Falls K, et al. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology 2005;28:171 - 82.

207. Sauerbrei WM, H; Prompeler, H J. Differentiation of benign and malignant breast tumors by logistic regression and a classification tree using Doppler flow signals. Methods of Information in Medicine 1998;37:226 - 34.

208. Gansky SA. Dental data mining: potential pitfalls and practical issues Advances in Dental Research 2003;17 109 - 14.

209. Zhang HPS, B. Recursive Partitioning in the Health Sciences New York: Springer; 1999.

210. Zhang HPY, C-Y; Singer, B; Xiong, M. Recursive partitioning for tumour classification with gene expression microarray data. Proceedings of the National Academy of Sciences 2001;98(12):6730 - 5.

211. Romualdi CC, S; Campagna, D; Celegato, B; Cannata, N; Toppo, S; Valle, G; Lanfranchi, G. Pattern recognition in gene expression profiling using DNA array: a comparitive study of different statistical methods applied to cancer classification. Human Molecular Genetics 2003;12(8):823 - 36.

212. Lunetta KLH, L B; Segal, J; Van Eerdewegh, P. Screening large scale assocation study data: exploiting interactions using random forests. BMC Genetics 2004;5(32).

213. Zhang HPH, T; Bracken, M B. A TREE-BASED METHOD OF ANALYSIS FOR PROSPECTIVE STUDIES. Statistics in Medicine 1996;15(1):37 - 49.

214. Zaykin DVY, S S. Large recursive partitioning analysis of complex disease pharmacogenetic studies. II. Statistical considerations. Pharmacogenetics 2005;6(1):77 - 89.

215. Province MAS, W D; Rao, D C. Classification methods for confronting heterogeneity. Advances in Genetics 2001;42:273 - 86.

216. Heidema AGB, J M A; Nagelkerke, N; Mariman, E C M; van der A, D L; Feskens, E J M. The challenge for genetic epidemiologists: how to analyse large numbers of SNPs in relation to complex disease. BMC Genetics 2006;7(23):1 - 15.

217. Parkin DMW, S L; Ferlay, J; Teppo, L; Thomas, D B. Cancer Incidence in five continents vol. VIII. IARC Scientific Publication 2002;155.

218. Vainio H, Miller AB. Primary and Secondary Prevention in Colorectal Cancer. Acta Oncologica 2003;42:809-15.

219. Martinez ME. Primary Prevention of Colorectal Cancer: Lifestyle, Nutrition, Exercise. In: Tumor Prevention and Genetics III. Berlin Heidelberg: Springer; 2005:177 - 211.

220. Baron JA, Cole BF, Sandler RS, et al. A Randomized Trial of Aspirin to Prevent Colorectal Adenomas. N Engl J Med 2003;348(10):891-9.

221. Robert Benamouzig, Jacques Deyra, Antoine Martin, et al. Daily soluble aspirin and prevention of colorectal adenoma recurrence: one-year results of the APACC trial1 1 The authors thank the women and men who participated in the study, P. E. Douziech for coordination of treatments, and the hospital pharmacists for preparation of the treatments in the trial centers. Gastroenterology 2003;125(2):328-36.

222. Benamouzig R, Deyra J, Martin A, et al. Daily soluble aspirin and prevention of colorectal adenoma recurrence: one-year results of the APACC trial1 1 The authors thank the women and men who participated in the study, P. E. Douziech for coordination of treatments, and the hospital pharmacists for preparation of the treatments in the trial centers. Gastroenterology 2003;125(2):328-36.

223. Sandler RS, Halabi S, Baron JA, et al. A Randomized Trial of Aspirin to Prevent Colorectal Adenomas in Patients with Previous Colorectal Cancer. N Engl J Med 2003;348(10):883-90.

224. Ridker PM, Cook NR, Lee IM, et al. A Randomized Trial of Low-Dose Aspirin in the Primary Prevention of Cardiovascular Disease in Women. N Engl J Med 2005;352(13):1293-304.

225. Baron JA, Beach M, Mandel JS, et al. Calcium Supplements for the Prevention of Colorectal Adenomas. N Engl J Med 1999;340(2):101-7.

226. Wactawski-Wende J, Kotchen JM, Anderson GL, et al. Calcium plus Vitamin D Supplementation and the Risk of Colorectal Cancer. N Engl J Med 2006;354(7):684-96.

227. Grau MV, Baron JA, Sandler RS, et al. Vitamin D, Calcium Supplementation, and Colorectal Adenomas: Results of a Randomized Trial. J Natl Cancer Inst 2003;95(23):1765-71.

228. COMMITTEE UNS. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme. In; 2003.

229. Steele RJCP, R; Parker, J; Warner, J; Fraser, C; Mowat, N A G; Wilson, J; Alexander, F E; Paterson, J G. A demonstration pilot trial for colorectal cancer screening in the United Kingdom: a new concept in the introduction of healthcare strategies. J Med Screen 2001;8:197 - 203.

230. Grazzini GC, C; Ciabattoni, C; Franceschini, F; Giorgi, D; Gozzi, S; Mantellini, P; Lopane, P; Perco, M; Rubeca, T; Salvadori, P; Visioli, C B; Zappa, M. Colorectal cancer screening programme by faecal occult blood test in Tuscany: first round results. European Journal of Cancer Prevention 2004;13(1):19 - 26.

231. Randall WB. Colon cancer screening. Gastroenterology 2000;119(3):837-53.

232. Rhodes JM. Colorectal cancer screening in the UK: Joint Position Statement by the British Society of Gastroenterology, the Royal College of Physicians, and the Association of Coloproctology of Great Britain and Ireland. Gut 2000;46(6):746-8.

233. Campos-Outcalt D. Should you screen--or not? The latest recommendations. Journal of Family Practice 2008;57(7):469-72.

234. Steele RJC, McClements PL, Libby G, et al. Results from the first three rounds of the Scottish demonstration pilot of FOBT screening for colorectal cancer. Gut 2009;58(4):530-5.

235. Scholefield JH, Moss S, Sufi F, Mangham CM, Hardcastle JD. Effect of faecal occult blood screening on mortality from colorectal cancer: results from a randomised controlled trial. Gut 2002;50(6):840-4.

236. Benson VS, Patnick J, Davies AK, Nadel MR, Smith RA, Atkin WS. Colorectal cancer screening: A comparison of 35 initiatives in 17 countries. International Journal of Cancer 2008;122(6):1357-67.

237. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell 2000;100(1):57-70.

238. Fearon ERV, B. A Genetic Model for Colorectal Tumorigenesis. Cell 1990;61:759 - 67.

239. Vogelstein B, Fearon ER, Hamilton SR, et al. Genetic alterations during colorectal-tumor development. N Engl J Med 1988;319(9):525-32.

240. Leslie AC, F A; Pratt, N R; Steele, R J C. The colorectal adenoma-carcinoma sequence. British Journal of Surgery 2002;89(7):845-60.

241. Knudson AG. Antioncogenes and human cancer. Proceedings of the National Academy of Sciences of the United States of America 1993;90(23):10914-21.

242. Brosens LAA, Keller JJ, Offerhaus GJA, Goggins M, Giardiello FM. Prevention and management of duodenal polyps in familial adenomatous polyposis. Gut 2005;54(7):1034-43.

243. Toms JR. CancerStats Monograph. In. London: Cancer Research UK; 2004.

244. Parkin DMP, P; Ferlay, J. Estimates of the worldwideincidence of 25 major cancers in 1990. Int J Cancer 1999;80:827 - 41.

245. Boyle P, Langham JS. ABC of colorectal cancer
Epidemiology. BMJ 2000;321(7264):805 - 8.

246. Haenszen WK, M. Studies of Japanese migrants. Mortality from cancer and other diseases among Japanese in the United States. J Natl Cancer Inst 1968;40:43 - 68.

247. Willett W. The search for the causes of breast and colon cancer. Nature 1989;338:389 - 94.

248. Young GPH, Y; Le Leu, R K; Nyskohus, L. Dietary fibre and colorectal cancer: A model for environment - gene interactions. Molecular Nutrition & Food Research 2005;49(6):571-84.

249. Clevers H. Colon Cancer -- Understanding How NSAIDs Work. N Engl J Med 2006;354(7):761-3.

250. Clevers H. At the Crossroads of Inflammation and Cancer. Cell 2004;118(6):671-4.

251. Gillen CD, Andrews HA, Prior P, Allan RN. Crohn's disease and colorectal cancer. Gut 1994;35(5):651-5.

252. Nilsen TIL, Vatten LJ. Prospective study of colorectal cancer risk and physical activity, diabetes, blood glucose and BMI: exploring the hyperinsulinaemia hypothesis. Br J Cancer 2001;84(3):417-22.

253. Marchand LL, Wilkens LR, Kolonel LN, Hankin JH, Lyu L-C. Associations of Sedentary Lifestyle, Obesity, Smoking, Alcohol Use, and Diabetes with the Risk of Colorectal Cancer. Cancer Res 1997;57(21):4787-94.

254. La Vecchia C, Negri E, Decarli A, Franceschi S. Diabetes mellitus and colorectal cancer risk. Cancer Epidemiol Biomarkers Prev 1997;6(12):1007-10.

255. Larsson SC, Orsini N, Wolk A. Diabetes Mellitus and Risk of Colorectal Cancer: A Meta-Analysis. J Natl Cancer Inst 2005;97(22):1679-87.

256. Kune GA, Kune S, Watson LF. Colorectal cancer risk, chronic illnesses, operations and medications: case control results from the Melbourne Colorectal Cancer Study. Int J Epidemiol 2007;36(5):951-7.

257. Janne PA, Mayer RJ. Chemoprevention of Colorectal Cancer. N Engl J Med 2000;342(26):1960-8.

258. Taketo MM. Cyclooxygenase-2 inhibitors in tumerogenesis. J Natl Cancer Inst 1998;90:1529 - 36.

259. Eberhart CEC, R J; Radhika, A; Giardiello, F M; Ferrenbach, S; DuBois, R N. Up-regulation of cyclooxygenase 2 gene expression in human colorectal adenomas and adenocarcinomas. Gastroenterology 1994;107(4):1183 - 8.

260. Peek RMJ. Prevention of colorectal cancer through use of COX-2 selective inhibitors. Cancer Chemother Pharmacol 2004;54(Suppl 1):S50 - S6.

261. Imperiale TF. Aspirin and the Prevention of Colorectal Cancer. N Engl J Med 2003;348(10):879-80.

262. Harris RE, Beebe-Donk J, Alshafie GA. Similar reductions in the risk of human colon cancer by selective and nonselective cyclooxygenase-2 (COX-2) inhibitors. BMC Cancer 2008;8:237.

263. Gann PH, Manson JE, Glynn RJ, Buring JE, Hennekens CH. Low-Dose Aspirin and Incidence of Colorectal Tumors in a Randomized Trial. J Natl Cancer Inst 1993;85(15):1220-4.

264. Sturmer T, Glynn RJ, Lee IM, Manson JE, Buring JE, Hennekens CH. Aspirin Use and Colorectal Cancer: Post-Trial Follow-up Data from the Physicians' Health Study. Ann Intern Med 1998;128(9):713-20.

265. Thun MJ, Namboodiri MM, Heath CW. Aspirin use and reduced risk of fatal colon cancer. N Engl J Med 1991;325(23):1593-6.

266. Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, Willett WC. Aspirin Use and the Risk for Colorectal Cancer and Adenoma in Male Health Professionals. Ann Intern Med 1994;121(4):241-6.

267. Giovannucci E, Egan KM, Hunter DJ, et al. Aspirin and the Risk of Colorectal Cancer in Women. N Engl J Med 1995;333(10):609-14.

268. Giovannucci E. An Updated Review of the Epidemiological Evidence that Cigarette Smoking Increases Risk of Colorectal Cancer. Cancer Epidemiol Biomarkers Prev 2001;10(7):725-31.

269. Nair UJ, Nair J, Mathew B, Bartsch H. Glutathione S-transferase M1 and T1 null genotypes as risk factors for oral leukoplakia in ethnic Indian betel quid/tobacco chewers. Carcinogenesis 1999;20(5):743-8.

270. Vainio HB, F. Weight control and physical activity. IARC handbooks of cancer prevention 2002;6:Chapter 5.

271. Slattery ML, Edwards S, Curtin K, et al. Physical Activity and Colorectal Cancer. Am J Epidemiol 2003;158(3):214-24.

272. Campbell KL, McTiernan A. Exercise and biomarkers for cancer prevention studies. Journal of Nutrition 2007;137(1 Suppl):161S-9S.

273. Wrigley H, Roderick P, George S, Smith J, Mullee M, Goddard J. Inequalities in survival from colorectal cancer: a comparison of the impact of deprivation, treatment, and host factors on observed and cause specific survival. J Epidemiol Community Health 2003;57(4):301-9.

274. Whynes DK, Frew EJ, Manghan CM, Scholefield JH, Hardcastle JD. Colorectal cancer, screening and survival: the influence of socio-economic deprivation. Public Health 2003;117(6):389-95.

275. Fraumeni JFL, J W; Smith, E M; Wagoner, J K. Cancer mortality among nuns; role of marital status in etiology of neoplastic disease in women J Natl Cancer Inst 1969;42:455 - 68.

276. Morson CB, H J R; Day, D W; Hill, M J. Adenomas of the large bowel. Cancer Surveys 1983;2:451 - 78.

277. Hill MJ. Female sex hormones and colorectal cancer. European Journal of Cancer Prevention 1998;7:425 - 6.

278. Singh S, Sheppard MC, Langman MJ. Sex differences in the incidence of colorectal cancer: an exploration of oestrogen and progesterone receptors. Gut 1993;34(5):611-5.

279. Konstantinopoulos PA, Kominea A, Vandoros G, et al. Oestrogen receptor beta (ER[beta]) is abundantly expressed in normal colonic mucosa, but declines in colon adenocarcinoma paralleling the tumour's dedifferentiation. European Journal of Cancer 2003;39(9):1251-8.

280. Greenlee RT, Hill-Harmon MB, Murray T, Thun M. Cancer Statistics, 2001. CA Cancer J Clin 2001;51(1):15-36.

281. Potter JD, Bostick RM, Grandits GA, et al. Hormone replacement therapy is associated with lower risk of adenomatous polyps of the large bowel: the Minnesota Cancer Prevention Research Unit Case-Control Study. Cancer Epidemiol Biomarkers Prev 1996;5(10):779-84.

282. Franceschi SLV, C Colorectal cancer and hormone replacement therapy: an unexpected finding. European Journal of Cancer Prevention 1998;7:427 - 38.

283. Nelson HD, Humphrey LL, Nygren P, Teutsch SM, Allan JD. Postmenopausal Hormone Replacement Therapy: Scientific Review. JAMA 2002;288(7):872-81.

284. Dinger J, Heinemann L, Mohner S, Thai D, Assmann A. Colon cancer risk and different HRT formulations: a case-control study. BMC Cancer 2007;7(1):76.

285. Cogliano V, Grosse Y, Baan R, Straif K, Secretan B, El Ghissassi F. Carcinogenicty of combined oestrogen-progesterone contraceptives and menopausal treatment. The Lancer Oncology 2005;6(8):552 - 3.

286. Fernandez E, Vecchia CL, Balducci A, Chatenoud L, Franceschi S, Negri E. Oral contraceptives and colorectal cancer risk: a meta-analysis. Br J Cancer 2001;84(5):722-7.

287. Burkitt DP. Epidemiology of cancer of the colon and rectum. Cancer 1971;28:3 - 13.

288. Kim YI. AGA technical review: impact of dietary fiber on colon cancer occurance Gastroenterology 2000;118:1235 - 57.

289. Bingham SA. Mechanisms and experimental and epidemiological evidence relating dietary
fibre (non-starch polysaccharides) and starch to protection against large bowel. Proc Nutr Soc 1990;49:153 - 71.

290. Yuan H, Liddle FJ, Mahajan S, Frank DA. IL-6-induced survival of colorectal carcinoma cells is inhibited by butyrate through down-regulation of the IL-6 receptor. Carcinogenesis 2004;25(11):2247-55.

291. Pool-Zobel BL, Selvaraju V, Sauer J, et al. Butyrate may enhance toxicological defence in primary, adenoma and tumor human colon cells by favourably modulating expression of glutathione S-transferases genes, an approach in nutrigenomics. Carcinogenesis 2005;26(6):1064-76.

292. Riboli EK, Rudolf. The EPIC Project: Rationale and Study Design. International Journal of Epidemiology 1997;28(1 (Suppl1)):S6 - S14.

293. Peters US, R; Chatterjee, N; Subar, A F; Ziegler, R G; Kulldorff, M; Bresalier, R; Weissfeld, J L; Flood, A; Schatzkin, A; Hayes, R B. Dietary Fibre and colorectal adenoma in a colorectal cancer early detection programme. The Lancet 2003;361:1491 - 5.

294. Terry PG, E; Michels, K B; Bergkvist, L; Hansen, H; Holmberg, L; Wolk, A. Fruit, Vegetables, Dietary Fiber, and the risk of Colorectal Cancer. Journal of the National Cancer Institute 2001;93(7):525 - 33.

295. Schatzkin A, Lanza E, Corle D, et al. Lack of Effect of a Low-Fat, High-Fiber Diet on the Recurrence of Colorectal Adenomas. N Engl J Med 2000;342(16):1149-55.

296. Lanza E, Yu B, Murphy G, et al. The polyp prevention trial continued follow-up study: no effect of a low-fat, high-fiber, high-fruit, and -vegetable diet on adenoma recurrence eight years after randomization. Cancer Epidemiology, Biomarkers & Prevention 2007;16(9):1745-52.

297. Alberts DS, Martinez ME, Roe DJ, et al. Lack of Effect of a High-Fiber Cereal Supplement on the Recurrence of Colorectal Adenomas. N Engl J Med 2000;342(16):1156-62.

298. Lanza E, Schatzkin A, Daston C, et al. Implementation of a 4-y, high-fiber, high-fruit-and-vegetable, low-fat dietary intervention: results of dietary changes in the Polyp Prevention Trial. Am J Clin Nutr 2001;74(3):387-401.

299. Key TJ, Fraser GE, Thorogood M, et al. Mortality in vegetarians and non-vegetarians: a collaborative analysis of 8300 deaths among 76,000 men and women in five prospective studies. Public Health Nutrition 1998;1(01):33-41.

300. Park Y, Hunter DJ, Spiegelman D, et al. Dietary Fiber Intake and Risk of Colorectal Cancer: A Pooled Analysis of Prospective Cohort Studies. JAMA 2005;294(22):2849-57.

301. Kassie F, Uhl M, Rabot S, et al. Chemoprevention of 2-amino-3-methylimidazo[4,5-f]quinoline (IQ)-induced colonic and hepatic preneoplastic lesions in the F344 rat by cruciferous vegetables administered simultaneously with the carcinogen. Carcinogenesis 2003;24(2):255-61.

302. Nishino HT, H; Murakoshi, M; Satomi, Y; Masuda, M; Onozuka, M; Yamaguchi; Takayasu, J; Tsuruta, J; Okuda, M; Khachik, F; Narisawa, T; Takasuka, N; Yano, M. Cancer prevention by natural carotenoids. BioFactors 2000;13(1):89 - 94.

303. Ferguson LRH, P J. The dietary fibre debate: more food for thought. The Lancet 2003;361:1487 - 8.

304. Burkitt DP, Trowell H. Some implications of dietary fibre. London: London Academic; 1975.

305. Bacic A, Harris PJ, Stone BA. Structure and function of plant cell walls. New York: Academic Press; 1988.

306. Harris PJ, Ferguson LR. Dietary fibres may protect or enhance carcinogenesis. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 1999;443(1-2):95-110.

307. Mai V, Flood A, Peters U, Lacey JV, Jr., Schairer C, Schatzkin A. Dietary fibre and risk of colorectal cancer in the Breast Cancer Detection Demonstration Project (BCDDP) follow-up cohort. Int J Epidemiol 2003;32(2):234-9.

308. Schatzkin AM, T; Park, Y; Subar, A F; Kipnis, V; Hollenbeck, A; Leitzmann, M F; Thompson, F E. Dietary Fiber and whole-grain consumption in relation to colorectal cancer in the NIH-AARP Diet and Health Study. Am J Clin Nutr 2006;85:1353 - 60.

309. Byres T. Diet, colorectal adenomas, and colorectal cancer. N Engl J Med 2000;342:1206 - 7.

310. Kinzler KWV, B. Lessons from hereditary colorectal cancer. Cell 1996;87:159 - 70.

311. Kestell P, Zhu S, Ferguson LR. Mechanisms by which resistant starches and non-starch polysaccharide sources affect the metabolism and disposition of the food carcinogen, 2-amino-3-methylimidazo[4,5-f]quinoline. Journal of Chromatography B 2004;802(1):201-10.

312. Ferguson LRZ, S; Kestell, P. Contrasting effects of non starch-based diets on the disposition and excretion of the food carcinogen 2-amino-3-methylimidazo[4,5-f]quinoline (IQ), in a rat model. Food Chem Toxicol 2003;41:785 - 92.

313. Papas MAG, E; Platz, E. Fiber from fruit and colorectal neoplasia. Biomarkers Prevention 2004;13:1267 - 70.

314. Bingham SA, Norat T, Moskal A, et al. Is the association with fiber from foods in colorectal cancer confounded by folate intake? Cancer Epidemiology, Biomarkers & Prevention 2005;14(6):1552-6.

315. Bingham S. The fibre-folate debate in colo-rectal cancer. Proceedings of the Nutrition Society 2006;65(01):19-23.

316. Kim YI. Folate and carcinogenesis: Evidence, mechanisms and implications. J Nutr Biochem 1999;10:66 - 88.

317. Choi S-W, Mason JB. Folate and Carcinogenesis: An Integrated Scheme1-3. J Nutr 2000;130(2):129-32.

318. Taioli E, Garza MA, Ahn YO, et al. Meta- and Pooled Analyses of the Methylenetetrahydrofolate Reductase (MTHFR) C677T Polymorphism and Colorectal Cancer: A HuGE-GSEC Review. American Journal of Epidemiology 2009;170(10):1207-21.

319. Kim Y-I. Folate and DNA Methylation: A Mechanistic Link between Folate Deficiency and Colorectal Cancer? Cancer Epidemiol Biomarkers Prev 2004;13(4):511-9.

320. Kim YI. Folate: a magic bullet or a double edged sword for colorectal cancer prevention? Gut 2006;55(10):1387-9.

321. Miguel A. Sanjoaquin NAECAWRTJK. Folate intake and colorectal cancer risk: A meta-analytical approach. International Journal of Cancer 2005;113(5):825-8.

322. Kennedy DA, Stern SJ, Moretti M, et al. Folate intake and the risk of colorectal cancer: A systematic review and meta-analysis. Cancer Epidemiology 2011;35(1):2-10.

323. Van Guelpen B, Hultdin J, Johansson I, et al. Low folate levels may protect against colorectal cancer. Gut 2006;55(10):1461-6.

324. Cole BF, Baron JA, Sandler RS, et al. Folic Acid for the Prevention of Colorectal Adenomas: A Randomized Clinical Trial. JAMA 2007;297(21):2351-9.

325. Ulrich CM, Potter JD. Folate and Cancer--Timing Is Everything. JAMA 2007;297(21):2408-9.

326. Giovannucci E, Stampfer MJ, Colditz G, Rimm EB, Willett WC. Relationship of Diet to Risk of Colorectal Adenoma in Men. J Natl Cancer Inst 1992;84(2):91-8.

327. Norat TL, A; Ferrari, P; Riboli, E. Meat consumption and colorectal cancer risk: Dose-response meta-analysis of epidemiological studies. International Journal of Cancer 2002;98(2):241-56.

328. Alvarez-Leon EA, Roman-Vinas B, Serra-Majem L. Dairy products and health: a review of the epidemiological evidence. British Journal of Nutrition 2006;96(Supplement S1):S94-S9.

329. Cho E, Smith-Warner SA, Spiegelman D, et al. Dairy Foods, Calcium, and Colorectal Cancer: A Pooled Analysis of 10 Cohort Studies. J Natl Cancer Inst 2004;96(13):1015-22.

330. WCRF. Food, nutrition, and the prevention of cancer: a global perspective. Washington DC: American Institute for Cancer Research; 1997.

331. Cho E, Smith-Warner SA, Ritz J, et al. Alcohol Intake and Colorectal Cancer: A Pooled Analysis of 8 Cohort Studies. Ann Intern Med 2004;140(8):603-13.

332. Theodoratou E, Kyle J, Cetnarskyj R, et al. Dietary Flavonoids and the Risk of Colorectal Cancer. Cancer Epidemiol Biomarkers Prev 2007;16(4):684-93.

333. Bobe G, Sansbury LB, Albert PS, et al. Dietary flavonoids and colorectal adenoma recurrence in the Polyp Prevention Trial. Cancer Epidemiology, Biomarkers & Prevention 2008;17(6):1344-53.

334. Bjelakovic G, Nikolova D, Simonetti RG, Gluud C. Antioxidant supplements for prevention of gastrointestinal cancers: a systematic review and meta-analysis. The Lancet 2004;364(9441):1219-28.

335. Stanner SA, Hughes J, Kelly CNM, Buttriss J. A review of the epidemiological evidence for the "antioxidant hypothesis". Public Health Nutrition 2004;7(03):407-22.

336. Lynch HTdlC, A. Genetic susceptibility to non-polyposis colorectal cancer. Journal of Medical Genetics 1999;36:801 - 18.

337. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and Heritable Factors in the Causation of Cancer -- Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343(2):78-85.

338. Lengauer C, Kinzler KW, Vogelstein B. Genetic instability in colorectal cancers. Nature 1997;386(6625):623-7.

339. Grady WMM, S D. Genetic and Epigenetic Alterations in Colon Cancer. Annu Rev Genomics Hum Genet 2002;3:101 - 28.

340. Campbell WJ, Spence RA, Parks TG. Familial adenomatous polyposis. British Journal of Surgery 1994;81(12):1722 - 33.

341. Arvantis ML, Jagelman DG, Fazio VW, Lavery IC, McGannon E. Mortality in Patients with Familial Adenomatous Polyposis. Diseases of the Colon and Rectum 1990;33(8):639 - 42.

342. Knudsen ALB, Bülow, S. Attenuated familial adenomatous polyposis (AFAP): a review of the literature. Familial Cancer 2003;2(1):43 - 55.

343. Kong S, Amos CI, Luthra R, Lynch PM, Levin B, Frazier ML. Effects of Cyclin D1 Polymorphism on Age of Onset of Hereditary Nonpolyposis Colorectal Cancer. Cancer Res 2000;60(2):249-52.

344. Laurent-Puig P, Beroud C, Soussi T. APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucl Acids Res 1998;26(1):269-70.

345. Laken SJ, Petersen GM, Gruber SB, et al. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nat Genet 1997;17(1):79-83.

346. Spirio LN, Samowitz W, Robertson J, et al. Alleles of APC modulate the frequency and classes of mutations that lead to colon polyps. Nat Genet 1998;20(4):385-8.

347. Zi Qiang Yuan BGLKBNWPHGQWAPWDFMT. Polymorphisms and HNPCC: PMS2-MLH1 protein interactions diminished by single nucleotide polymorphisms. Human Mutation 2002;19(2):108-13.

348. Mitchell RJF, S M; Dunlop, M G; Campbell, H. Mismatch Repair Genes *hMLH1* and *hMSH2* and Colorectal Cancer: A HuGE Review. American Journal of Epidemiology 2002;156:885 - 902.

349. Nelson WJ, Nusse R. Convergence of Wnt, {beta}-Catenin, and Cadherin Pathways. Science 2004;303(5663):1483-7.

350. Wang SS, Esplin ED, Li JL, et al. Alterations of the PPP2R1B Gene in Human Lung and Colon Cancer. Science 1998;282(5387):284-7.

351. Park WS, Oh RR, Park JY, et al. Frequent Somatic Mutations of the {beta}-Catenin Gene in Intestinal-Type Gastric Cancer. Cancer Res 1999;59(17):4257-60.

352. Fearnhead NS, Wilding JL, Winney B, et al. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proceedings of the National Academy of Sciences of the United States of America 2004;101(45):15992-7.

353. Koppert LB, van der Velden AW, van de Wetering M, et al. Frequent loss of the AXIN1 locus but absence of AXIN1 gene mutations in adenocarcinomas of the gastro-oesophageal junction with nuclear [beta]-catenin expression. Br J Cancer 2004;90(4):892-9.

354. Liu W, Dong X, Mai M, et al. Mutations in AXIN2 cause colorectal cancer with defective mismatch repair by activating [beta]-catenin/TCF signalling. Nat Genet 2000;26(2):146-7.

355. Leung JY, Kolligs FT, Wu R, et al. Activation of AXIN2 Expression by beta -Catenin-T Cell Factor. A FEEDBACK REPRESSOR PATHWAY REGULATING Wnt SIGNALING. J Biol Chem 2002;277(24):21657-65.

356. da Costa LTH, T-C; Yu, J; Sparks, A B; Morin, P J; Polyak, K; Laken, S; Vogelstein, B; Kinzler, K W. CDX2 is mutated in a colorectal cancer with normal APC/beta-catenin signaling. Oncogene 1999;18(35):5010 - 4.

357. Soriano S, Kang DE, Fu M, et al. Presenilin 1 Negatively Regulates {beta}-Catenin/T Cell Factor/Lymphoid Enhancer Factor-1 Signaling Independently of {beta}-Amyloid Precursor Protein and Notch Processing. J Cell Biol 2001;152(4):785-94.

358. Jones S, Emmerson P, Maynard J, et al. Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C->T:A mutations. Hum Mol Genet 2002;11(23):2961-7.

359. Nichols KE, Malkin D, Garber JE, Fraumeni JF, Jr., Li FP. Germ-line p53 Mutations Predispose to a Wide Spectrum of Early-onset Cancers. Cancer Epidemiol Biomarkers Prev 2001;10(2):83-7.

360. Howe JR, Sayed MG, Ahmed AF, et al. The prevalence of MADH4 and BMPR1A mutations in juvenile polyposis and absence of BMPR2, BMPR1B, and ACVR1 mutations. J Med Genet 2004;41(7):484-91.

361. Sayed MG, Ahmed AF, Ringold JR, et al. Germline SMAD4 or BMPR1A Mutations and Phenotype of Juvenile Polyposis. Ann Surg Oncol 2002;9(9):901-6.

362. Cao X, Eu KW, Kumarasinghe MP, Li HH, Loi C, Cheah PY. Mapping of hereditary mixed polyposis syndrome (HMPS) to chromosome 10q23 by genomewide high-density single nucleotide polymorphism (SNP) scan and identification of BMPR1A loss of function. J Med Genet 2006;43(3):e13-.

363. Tomlinson IP, Houlston RS. Peutz-Jeghers syndrome. J Med Genet 1997;34(12):1007-11.

364. Jenne DE, Reomann H, Nezu J-i, et al. Peutz-Jeghers syndrome is caused by mutations in a novel serine threoninekinase. Nat Genet 1998;18(1):38-43.

365. Lim W, Hearle N, Shah B, et al. Further observations on LKB1//STK11 status and cancer risk in Peutz-Jeghers syndrome. Br J Cancer 2003;89(2):308-13.

366. Corradetti MN, Inoki K, Bardeesy N, DePinho RA, Guan KL. Regulation of the TSC pathway by LKB1: evidence of a molecular link between tuberous sclerosis complex and Peutz-Jeghers syndrome. Genes and Development 2004;18:1533 - 8.

367. Duval A, Hamelin R. Mutations at Coding Repeat Sequences in Mismatch Repair-deficient Human Cancers: Toward a New Concept of Target Genes for Instability. Cancer Res 2002;62(9):2447-54.

368. Boyer TD, Enney WC. Preparation, characterisation and properties of glutathione S-transferases New York: John Wiley and Sons; 1985.

369. Mannervik BA, Y C; Board, P G; Hayes, J D; Di Ilio, C; Ketterer, B; Listowsky, I; Morgenstern, R; Muramatsu, M; Pearson, W R; Pickett, C B; Sato, K;

Widersten, M; Wolf, C R. Nomenclature for human glutathione transferases. Biochem J 1992;282:305 - 6.

370. Ates NA, Tamer L, Ates C, et al. Glutathione S-transferase M1, T1, P1 genotypes and risk for development of colorectal cancer. Biochem Genet 2005;43(3-4):149-63.

371. Eaton DL, Bammler TK. Concise review of the glutathione S-transferases and their significance to toxicology. Toxicol Sci 1999;49(2):156-64.

372. London SJY, J M; Chung, F L; Gao, Y T; Coetzee, G A; Ross, R K. Isothiocyanates, glutathione S-transferase M1 and T1 polymorphisms, and lung cancer risk: a prospective study of men in Shanghai. Lancet 2000;356:724 - 9.

373. Zhong S, Wyllie AH, Barnes D, Wolf CR, Spurr NK. Relationship between the GSTM1 genetic polymorphism and susceptibility to bladder, breast and colon cancer. Carcinogenesis 1993;14(9):1821-4.

374. Katoh T, Nagata N, Kuroda Y, et al. Glutathione S-transferase M1 (GSTM1) and T1 (GSTT1) genetic polymorphism and susceptibility to gastric and colorectal adenocarcinoma. Carcinogenesis 1996;17(9):1855-9.

375. Deakin M, Elder J, Hendrickkse C, et al. SHORT COMMUNICATION: Glutathione S-transferase GSTT1 genotypes and susceptibility to cancer: studies of interactions with GSTM1 in lung, oral, gastric and colorectal cancers. Carcinogenesis 1996;17(4):881-4.

376. Stoehlmacher J, Park DJ, Zhang W, et al. Association Between Glutathione S-Transferase P1, T1, and M1 Genetic Polymorphism and Survival of Patients With Metastatic Colorectal Cancer. J Natl Cancer Inst 2002;94(12):936-42.

377. Gertig DM, Stampfer M, Haiman C, Hennekens CH, Kelsey K, Hunter DJ. Glutathione S-transferase GSTM1 and GSTT1 polymorphisms and colorectal cancer risk: a prospective study. Cancer Epidemiol Biomarkers Prev 1998;7(11):1001-5.

378. Lee EH, Y; Zhao, B; Seow-Choen, F; Balakrishnan, A; Chan, S H. Genetic Polymorphisms of conjugating enzymes and cancer risk: GSTM1, GSTT1, NAT1 and NAT2. The Journal of Toxicological Sciences 1998;23(Suppl II):140 - 2.

379. Cotton SC, Sharp L, Little J, Brockton N. Glutathione S-transferase Polymorphisms and Colorectal Cancer: A HuGE Review. American Journal of Epidemiology 2000;151(1):7 - 32.

380. Shi YL, J S; Galvin, K M. Everything you have ever wanted to know about Yin Yang 1. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer 1997;1332(2):49 - 66.

381. Inskip AE-C, J; Buxton, N; Dias, P S; MacIntosh, J; Campbell, D; Jones, P W; Yengi, L; Talbot, J A; Strange, R C; Fryer, A A Identification of polymorphism at the Glutathione S-transferase, GSTM3 locus: evidence for linkage with GSTM1*A. Biochem J 1995;312:713 - 6.

382. Moscow JA, Fairchild CR, Madden MJ, et al. Expression of Anionic Glutathione-S-transferase and P-Glycoprotein Genes in Human Tissues and Tumors. Cancer Res 1989;49(6):1422-8.

383. Ali-Osman F, Akande O, Antoun G, Mao J-X, Buolamwini J. Molecular Cloning, Characterization, and Expression in Escherichia coli of Full-length cDNAs of Three Human Glutathione S-Transferase Pi Gene Variants. J Biol Chem 1997;272(15):10004-12.

384. Watson MAS, R K; Smith, G B; Massey, T E; Bell, D A. Human glutathione S-transferase P1 polymorphisms: relationship to lung tissue enzyme activity and population frequency distribution. Carcinogenesis 1998;19:275 - 80.

385. Saadat M. Genetic polymorphisms of glutathione S-transferase T1 (GSTT1) and susceptibility to gastric cancer: a meta-analysis. Cancer Sci 2006;97(6):505-9.

386. Economopoulos KP, Sergentanis TN. GSTM1, GSTT1, GSTP1, GSTA1 and colorectal cancer risk: A comprehensive meta-analysis. European Journal of Cancer 2010;46(9):1617-31.

387. Raimondi S, Botteri E, Iodice S, Lowenfels AB, Maisonneuve P. Gene-smoking interaction on colorectal adenoma and cancer risk: Review and meta-analysis. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 2009;670(1-2):6-14.

388. Liabakk N-B, Talbot I, Smith RA, Wilkinson K, Balkwill F. Matrix Metalloprotease 2 (MMP-2) and Matrix Metalloprotease 9 (MMP-9) Type IV Collagenases in Colorectal Cancer. Cancer Res 1996;56(1):190-6.

389. Zucker SV, J. Role of matrix metalloproteinases (MMPs) in colorectal cancer. Cancer and Metastasis Reviews 2004;23(1-2):101 - 17.

390. Mercurio MG, Shiff SJ, Galbraith RA, Sassa S. Expression of Cytochrome P450 mRNAs in the Colon and the Rectum in Normal Human Subjects. Biochemical and Biophysical Research Communications 1995;210(2):350-5.

391. Pisani PM, N. Cooking methods, metabolic polymorphisms and colorectal cancer. . European Journal of Cancer Prevention 2002;11(1):75 - 84.

392. Little J, Sharp L, Masson LF, et al. Colorectal cancer and genetic polymorphisms of CYP1A1, GSTM1 and GSTT1: A case-control study in the Grampian region of Scotland. International Journal of Cancer 2006;119(9):2155-64.

393. Landi SG, F; Moreno, V; Gioia-Patricola, L; Chabrier, A; Guino, E; Navarro, M; de Oca, J; Capella, G; Canzian, F. A comprehensive analysis of phase I and phase II metabolism gene polymorphisms and risk of colorectal cancer. Pharmacogenetics and Genomics 2005;15(8):535 - 46.

394. Otani T, Iwasaki M, Hanaoka T, et al. Folate, vitamin B6, vitamin B12, and vitamin B2 intake, genetic polymorphisms of related enzymes, and risk of colorectal cancer in a hospital-based case-control study in Japan. Nutrition & Cancer 2005;53(1):42-50.

395. Cortessis V, Siegmund K, Chen Q, et al. A Case-Control Study of Microsomal Epoxide Hydrolase, Smoking, Meat Consumption, Glutathione S-Transferase M3, and Risk of Colorectal Adenomas. Cancer Res 2001;61(6):2381-5.

396. Wagner C. BIOCHEMICAL ROLE OF FOLATE IN CELLULAR METABOLISM1*. Clinical Research and Regulatory Affairs 2001;18(3):161-80.

397. Frosst PB, H J; Milos, R; Goyette, P; Sheppard, C A; Matthews, R G; Boers, G J; den Heijer, M; Kluijtmans, L A; van den Heuval, L P; Rozen, R. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nat Genet 1995;10:111-3.

398. Ulrich CM, Kampman E, Bigler J, et al. Lack of Association between the C677T MTHFR Polymorphism and Colorectal Hyperplastic Polyps. Cancer Epidemiol Biomarkers Prev 2000;9(4):427-33.

399. Choi SW, Mason JB. Folate status: effects on pathways of colorectal carcinogenesis. J Nutr 2002;132(8):2413S - 8S.

400. Sharp L, Little J. Polymorphisms in Genes Involved in Folate Metabolism and Colorectal Neoplasia: A HuGE Review. Am J Epidemiol 2004;159(5):423-43.

401. Ulrich CM, Kampman E, Bigler J, et al. Colorectal Adenomas and the C677T MTHFR Polymorphism: Evidence for Gene-Environment Interaction? Cancer Epidemiol Biomarkers Prev 1999;8(8):659-68.

402. van den Donk M, Buijsse B, van den Berg SW, et al. Dietary Intake of Folate and Riboflavin, MTHFR C677T Genotype, and Colorectal Adenoma Risk: A Dutch Case-Control Study. Cancer Epidemiol Biomarkers Prev 2005;14(6):1562-6.

403. Weisberg I, Tran P, Christensen B, Sibani S, Rozen R. A Second Genetic Polymorphism in Methylenetetrahydrofolate Reductase (MTHFR) Associated with Decreased Enzyme Activity. Molecular Genetics and Metabolism 1998;64(3):169-72.

404. Dimitri Trembath ALSDCVGMSKTEJLRHFSMGLJCM. Analysis of select folate pathway genes, <I>PAX3</I>, and human <I>T</I> in a midwestern neural tube defect population. Teratology 1999;59(5):331-41.

405. Trembath DS, A L; Vandyke, D C; Shaw, G M; Todoroff, K; Lammer, E J; Finnell, R H; Marker, S; Lerner, G; Murray, J C. Analysis of select folate pathway genes, *PAX3*, and human *T* in a midwestern neural tube defect population. Teratology 1999;59(5):331-41.

406. Rady PLS, S; Grady, J; Hudnall, S D; Kellner, L H; Nitowsky, H; Tyring, S K; Matalon, R K. Genetic polymorphisms of methylenetetrahydrofolate reductase (*MTHFR*) and methionine synthase reductase (*MTRR*) in ethnic populations in Texas; a report of a novel *MTHFR* polymorphic site, G1793A. American Journal of Medical Genetics 2002;107(2):162-8.

407. Li L, Plummer SJ, Thompson CL, Tucker TC, Casey G. Association between phosphatidylinositol 3-kinase regulatory subunit p85alpha Met326Ile genetic polymorphism and colon cancer risk. Clinical Cancer Research 2008;14(3):633-7.

408. Philp AJ, Campbell IG, Leet C, et al. The Phosphatidylinositol 3'-kinase p85{alpha} Gene Is an Oncogene in Human Ovarian and Colon Tumors. Cancer Res 2001;61(20):7426-9.

409. Brockton N, Little J, Sharp L, Cotton SC. N-Acetyltransferase Polymorphisms and Colorectal Cancer: A HUGE Review. Am J Epidemiol 2000;151(9):846-61.

410. de Jong MM, Nolte IM, te Meerman GJ, et al. Low-penetrance Genes and Their Involvement in Colorectal Cancer Susceptibility. Cancer Epidemiol Biomarkers Prev 2002;11(11):1332-52.

411. Hein DW, Doll MA, Fretland AJ, et al. Molecular Genetics and Epidemiology of the NAT1 and NAT2 Acetylation Polymorphisms. Cancer Epidemiol Biomarkers Prev 2000;9(1):29-42.

412. Lilla C, Verla-Tebit E, Risch A, et al. Effect of NAT1 and NAT2 Genetic Polymorphisms on Colorectal Cancer Risk Associated with Exposure to Tobacco Smoke and Meat Consumption. Cancer Epidemiol Biomarkers Prev 2006;15(1):99-107.

413. Koehne C-H, Dubois RN. COX-2 inhibition and colorectal cancer. Seminars in Oncology 2004;31(Supplement 7):12-21.

414. Marnett LJ. Aspirin and the Potential Role of Prostaglandins in Colon Cancer. Cancer Res 1992;52(20):5575-89.

415. Masferrer JL, Leahy KM, Koki AT, et al. Antiangiogenic and Antitumor Activities of Cyclooxygenase-2 Inhibitors. Cancer Res 2000;60(5):1306-11.

416. Jones MK, Sasaki E, Halter F, et al. HGF triggers activation of the COX-2 gene in rat gastric epithelial cells: action mediated through the ERK2 signaling pathway. FASEB J 1999;13(15):2186-94.

417. Zhang Z, DuBois RN. Par-4, a proapoptotic gene, is regulated by NSAIDs in human colon carcinoma cells. Gastroenterology 2000;118(6):1012-7.

418. McCarty MF, Block KI. Preadministration of High-Dose Salicylates, Suppressors of NF-{kappa}B Activation, May Increase the Chemosensitivity of Many Cancers: An Example of Proapoptotic Signal Modulation Therapy. Integr Cancer Ther 2006;5(3):252-68.

419. Bigler J, Whitton J, Lampe JW, Fosdick L, Bostick RM, Potter JD. CYP2C9 and UGT1A6 Genotypes Modulate the Protective Effect of Aspirin on Colon Adenoma Risk. Cancer Res 2001;61(9):3566-9.

420. Chan AT, Tranah GJ, Giovannucci EL, Hunter DJ, Fuchs CS. Genetic Variants in the UGT1A6 Enzyme, Aspirin Use, and the Risk of Colorectal Adenoma. J Natl Cancer Inst 2005;97(6):457-60.

421. Strassburg CP, Vogel A, Kneip S, Tukey RH, Manns MP. Polymorphisms of the human UDP-glucuronosyltransferase (UGT) 1A7 gene in colorectal cancer. Gut 2002;50(6):851-6.

422. Macarthur M, Sharp L, Hold GL, Little J, El-Omar EM. The Role of Cytokine Gene Polymorphisms in Colorectal Cancer and Their Interaction with Aspirin Use in the Northeast of Scotland. Cancer Epidemiol Biomarkers Prev 2005;14(7):1613-8.

423. Norman AW. Vitamin D Receptor: New Assignments for an Already Busy Receptor. Endocrinology 2006;147(12):5542-8.

424. Sun T, Gao Y, Tan W, et al. A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers. Nat Genet 2007;39(5):605-13.

425. Haiman CA, Le Marchand L, Yamamato J, et al. A common genetic risk factor for colorectal and prostate cancer. Nat Genet 2007;39(8):954-6.

426. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 2007;39(8):984-8.

427. Zanke BW, Greenwood CMT, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 2007;39(8):989-94.

428. Luciani MG, Campregher C, Gasche C. Aspirin blocks proliferation in colon cells by inducing a G1 arrest and apoptosis through activation of the checkpoint kinase ATM. Carcinogenesis 2007;28(10):2207-17.

429. Welfare MA, A M; Bassendine, M F; Daly, A K. Polymorphisms in GSTP1, GSTM1, and GSTT1 and susceptibility to Colorectal Cancer. Cancer Epidemiol Biomarkers Prev 1999;8:289 - 92.

430. Le Marchand L, Wilkens LR, Kolonel LN, Henderson BE. The MTHFR C677T Polymorphism and Colorectal Cancer: The Multiethnic Cohort Study. Cancer Epidemiol Biomarkers Prev 2005;14(5):1198-203.

431. Ma J, Stampfer MJ, Giovannucci E, et al. Methylenetetrahydrofolate Reductase Polymorphism, Dietary Interactions, and Risk of Colorectal Cancer. Cancer Res 1997;57(6):1098-102.

432. Ma J, Stampfer MJ, Christensen B, et al. A Polymorphism of the Methionine Synthase Gene: Association with Plasma Folate, Vitamin B12, Homocyst(e)ine, and Colorectal Cancer Risk. Cancer Epidemiol Biomarkers Prev 1999;8(9):825-9.

433. COGENT S. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40(12):1426-35.

434. Norman AW. Sunlight, season, skin pigmentation, vitamin D, and 25-hydroxyvitamin D: integral components of the vitamin D endocrine system. The American Journal of Clinical Nutrition 1998;67(6):1108 - 11.

435. Russell RM. The vitamin A spectrum: from deficiency to toxicity. The American Journal of Clinical Nutrition 2000;71(4):878 - 84.

436. Pettifor JM. Nutritional rickets: deficiency of vitamin D, calcium, or both? American Journal of Clinical Nutrition 2004;80(Suppl. 6):1725S - 9S.

437. Nordin BEM, H A. The clcium deficiency model for osteoporosis. Nutritional Review 1989;47(3):65 - 72.

438. Nishida MG, S G; Dunford, R G; Ho, A W; Trevisan, M; Genco, R J. Calcium and the Risk For Periodontal Disease. Journal of Periodontology 2000;71(7):1057 - 66.

439. Kitamura KY, T; Tanaka, H; Hashimoto, S; Yang, M; Takahashi, T. TPN-Induced Fulminant Beriberi: A Report on Our Experience and a Review of the Literature. Surgery Today 1996;26:769 - 76.

440. Butterworth RF. Pathophysiology of alcoholic brain damage: synergistic effects of ethanol, thiamine deficiency and alcoholic liver disease Metabolic Brain Disease 1995;10(1):1 - 8.

441. Powers HJ. Robiflavin (vitamin B-2) and health. The American Journal of Clinical Nutrition 2003;77(6):1352 - 61.

442. MacDonald AF, A. Nutritional deficiiencies and the skin. Clinical and Experimental Dermatology 2005;30(4):388 - 90.

443. Tahiliani AGe, C J. Pantothenic acid in health and disease. Vitam Horm 1991;46:165 - 228.

444. Rimm EBW, W C; Hu, F B. Folate and Vitamin B6 From Diet and Supplements in Relation to Risk of Coronary Heart Disease Among Women. JAMA 1998;279(5):359 - 64.

445. Campbell RK. A Critical Review of Chromium Picolinate and Biotin. US Pharmacist 2006;11:1 - 4.

446. Oh RCB, D L. Vitamin B12 Deficiency. In: American Family Physician; 2003.

447. Kamen B. Folate and antifolate pharmacology. Seminars in oncology 1997;24(5 (Suppl 18)):S18 - 30 - S18 - 39.

448. Shaw GMS, D; Velie, E M; Morland, K; Harris, J A. Periconceptional vitamin use, dietary folate, and the occurrence of neural tube defects. Epidemiology 1995;6(3):219 - 26.

449. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet 2005;37(11):1217-23.

450. Scotland H. Review of the Scottish Diet Action Plan. Edinburgh/ Glasgow: Health Scotland; 2006.

451. Fayyad UP-S, G; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. Commun ACM 1996;39(11):27-34.

452. Shaw MJ, Subramaniam C, Tan GW, Welge ME. Knowledge management and data mining for marketing. Decision Support Systems 2001;31(1):127-37.

453. Zdanowicz JS. Detecting money laundering and terrorist financing via data mining. Commun ACM 2004;47(5):53-5.

454. Bhavani T. Data mining, national security, privacy and civil liberties. SIGKDD Explor Newsl 2002;4(2):1-5.

455. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nat Genet 2002;31(3):316-9.

456. Berman JJ. Confidentiality issues for medical data miners. Artificial Intelligence in Medicine 2002;26(1-2):25-36.

457. Cios KJ, William Moore G. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002;26(1-2):1-24.

458. Zambon ML, R; Bunn, A; Powell, S. Effect of Alternative Splitting Rules on the Image Processing Using Classification Tree Analysis Photogrammetric Engineering and Remote Sensing 2006;72(1):25 - 30.

459. Quinlan JR. Induction of decision trees. Machine Learning 1986;1(1):81 - 106.

460. Therneau TMA, B; Ripley, B. rpart: Recursive Partitioning. In; 2008.

461. Liaw AW, M. Classification and Regression by randomForest. R News 2002:18 -22.

462. Everitt BSH, T. A Handbook of Statistical Analyses Using R. In: Recursive Partitioning: Large Companies and Glaucoma Diagnosis; 2008.

463. Chambers JM, Hastie TJ. Statistical Models in S. New York, London: Chapman and Hall; 1993.

464. Bahl LR, Brown PF, de Souza PV, Mercer RL. A tree-based statistical language model for natural language speech recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on 1989;37(7):1001-8.

465. Anderson NH, Titterington DM. Cross-correlation between simultaneously generated sequences of pseudo-random uniform deviates. Statistics and Computing 1993;3(2):61-5.

466. Mazumdar M, Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. Statistics in Medicine 2000;19(1):113-32.

467. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine 2006;25(1):127-41.

468. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the Practice of Dichotomization of Quantitative Variables. Psychological Methods 2002;7(1):19 - 40.

469. R: A Language and Environment for Statistical Computing. 2008. (Accessed at http://www.R-project.org.)

470. Lee E-KY, S G; Jung, Y; Namkung, J; Park, T. ORMDR. In.

471. Velez DW, BC; Motsinger, AA; Bush, WS; Ritchie MD; Williams, SM; Moore, JH. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology 2007;31:306-15.

472. Lee E-KY, S G; Jung, Y; Namkung,J; Park, T. ORMDR. In; 2008:Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions.

473. Poynter JN, Figueiredo JC, Conti DV, et al. Variants on 9p24 and 8q24 Are Associated with Risk of Colorectal Cancer: Results from the Colon Cancer Family Registry. Cancer Research 2007;67(23):11128-32.

474. Gao Y, Zhang F-M, Huang S, et al. A De Novo Mutation of STK11 Gene in a Chinese Patient with Peutz–Jeghers Syndrome. Digestive Diseases and Sciences 2010;55(4):1032-6.

475. Pomerantz MM, Ahmadiyeh N, Jia L, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 2009;41(8):882-4.

476. Tuupanen S, Turunen M, Lehtonen R, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 2009;41(8):885-90.

477. Curtin K, Lin W-Y, George R, et al. Meta Association of Colorectal Cancer Confirms Risk Alleles at 8q24 and 18q21. Cancer Epidemiology Biomarkers & Prevention 2009;18(2):616-21.

478. Abuli A, Bessa X, Gonzalez JR, et al. Susceptibility genetic variants associated with colorectal cancer risk correlate with cancer phenotype. Gastroenterology 2010;139(3):788-96.

479. Tenesa A, Farrington SM, Prendergast JGD, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 2008;40(5):631-7.

480. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 2007;39(11):1315-7.

481. Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. Nat Rev Genet 2009;10(6):353-8.

482. Pittman AM, Webb E, Carvajal-Carmona L, et al. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. Human Molecular Genetics 2008;17(23):3720-7.

483. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 2008;40(1):26-8.

484. Tomlinson IPM, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 2008;40(5):623-30.

485. Jenab M, McKay J, Bueno-de-Mesquita HB, et al. Vitamin D Receptor and Calcium Sensing Receptor Polymorphisms and the Risk of Colorectal Cancer in European Populations. Cancer Epidemiology Biomarkers & Prevention 2009;18(9):2485-91.

486. Shi Z, Johnstone D, Talseth-Palmer BA, et al. Haemochromatosis HFE gene polymorphisms as potential modifiers of hereditary nonpolyposis colorectal cancer risk and onset age. International Journal of Cancer 2009;125(1):78-83.

487. Lugli A, Zlobec I, Minoo P, et al. Prognostic significance of the wnt signalling pathway molecules APC, β-catenin and E-cadherin in colorectal cancer—a tissue microarray-based analysis. Histopathology 2007;50(4):453-64.

488. von Holst S, Picelli S, Edler D, et al. Association studies on 11 published colorectal cancer risk loci. Br J Cancer 2010;103(4):575-80.

489. Abulí A, Bessa X, González JR, et al. Susceptibility Genetic Variants Associated With Colorectal Cancer Risk Correlate With Cancer Phenotype. Gastroenterology 2010;139(3):788-96.e6.

490. Lascorz J, Forsti A, Chen B, et al. Genome-wide association study for colorectal cancer identifies risk polymorphisms in German familial cases and implicates MAPK signalling pathways in disease susceptibility. Carcinogenesis 2010;31(9):1612-9.

491. Bigler J, Ulrich CM, Kawashima T, Whitton J, Potter JD. DNA Repair Polymorphisms and Risk of Colorectal Adenomatous or Hyperplastic Polyps. Cancer Epidemiology Biomarkers & Prevention 2005;14(11):2501-8.

492. Huang W-Y, Berndt SI, Kang D, et al. Nucleotide Excision Repair Gene Polymorphisms and Risk of Advanced Colorectal Adenoma: XPC Polymorphisms Modify Smoking-Related Risk. Cancer Epidemiology Biomarkers & Prevention 2006;15(2):306-11.

493. Lee SH, Bahn JH, Whitlock NC, Baek SJ. Activating transcription factor 2 (ATF2) controls tolfenamic acid-induced ATF3 expression via MAP kinase pathways. Oncogene 2010;29(37):5182-92.

494. Dibra HK, Brown JE, Hooley P, Nicholl ID. Aspirin and alterations in DNA repair proteins in the SW480 colorectal cancer cell line. Oncology Reports 2010;24(1):37-46.

495. He T-C, Chan TA, Vogelstein B, Kinzler KW. PPAR[delta] Is an APC-Regulated Target of Nonsteroidal Anti-Inflammatory Drugs. Cell 1999;99(3):335-45.

496. Kupfer SS, Anderson JR, Hooker S, et al. Genetic Heterogeneity in Colorectal Cancer Associations Between African and European Americans. Gastroenterology 2010;139(5):1677-85.e8.

497. Zhou G-W, Hu J, Li Q. CYP2E1 PstI/RsaI polymorphism and colorectal cancer risk: a meta-analysis. World Journal of Gastroenterology 2010;16(23):2949-53.

498. Stern MC, Siegmund KD, Conti DV, Corral Rn, Haile RW. XRCC1, XRCC3, and XPD Polymorphisms as Modifiers of the Effect of Smoking and Alcohol on Colorectal Adenoma Risk. Cancer Epidemiology Biomarkers & Prevention 2006;15(12):2384-90.

499. Gao C-M, Gong J-P, Wu J-Z, et al. Relationship between growth hormone 1 genetic polymorphism and susceptibility to colorectal cancer. Journal of Human Genetics 2010;55(3):163-6.

500. Siezen CLE, Tijhuis MJ, Kram NR, et al. Protective effect of nonsteroidal anti-inflammatory drugs on colorectal adenomas is modified by a polymorphism in peroxisome proliferator-activated receptor [delta]. Pharmacogenetics and Genomics 2006;16(1):43-50.

501. Takakura Y, Hinoi T, Oue N, et al. CDX2 Regulates Multidrug Resistance 1 Gene Expression in Malignant Intestinal Epithelium. Cancer Research 2010;70(17):6767-78.

502. Guo RJS, R S; Lynch, J P. The role of Cdx proteins in intestinal development and cancer. Cancer Biology and Therapy 2004;3(7):593 - 601.

503. Hinoi TT, M; Lucas, P C; Caca, K; Dunn, R L; Macri, E, Loda, M; Appelman, H D; Cho, K R;Fearon, E R. Loss of CDX2 Expression and Microsatellite Instability Are Prominent Features of Large Cell Minimally Differentiated Carcinomas of the Colon. American Journal of Pathology, 2001;159(6):2239 - 48.

504. Barbareschi M, Murer B, Colby TV, et al. CDX-2 Homeobox Gene Expression Is a Reliable Marker of Colorectal Adenocarcinoma Metastases to the Lungs. The American Journal of Surgical Pathology 2003;27(2):141-9.

505. Wong SCC, Ng SSM, Cheung MT, et al. Clinical significance of CDX2-positive circulating tumour cells in colorectal cancer patients. Br J Cancer 2011.

506. Laurent CS, M; Flejou, J F; Chenard, M P; Duclos, B; Freund, J N; Reimund, J M. Immunohistochemical expression of CDX2, β-catenin, and TP53 in inflammatory bowel disease-associated colorectal cancer. Inflammatory Bowel Disease 2010;17(1):232 - 40.

507. McCullough ML, Bostick RM, Mayo TL. Vitamin D Gene Pathway Polymorphisms and Risk of Colorectal, Breast, and Prostate Cancer. Annual Review of Nutrition 2009;29(1):111-32.

508. Verzi MP, Hatzis P, Sulahian R, et al. TCF4 and CDX2, major transcription factors for intestinal function, converge on the same cis-regulatory regions. Proceedings of the National Academy of Sciences 2010;107(34):15157-62.

509. Fantini MC, Rizzo A, Fina D, et al. 52 Smad7 Expression in T Cells Protects from Colitis-Associated Colorectal Cancer. Gastroenterology 2009;136(5):A-8-A-9.

510. Slattery ML, Herrick J, Curtin K, et al. Increased Risk of Colon Cancer Associated with a Genetic Polymorphism of SMAD7. Cancer Research 2010;70(4):1479-85.

511. Thompson CL, Plummer SJ, Acheson LS, Tucker TC, Casey G, Li L. Association of common genetic variants in SMAD7 and risk of colon cancer. Carcinogenesis 2009;30(6):982-6.

512. ten Dijke PH, C S. New insights into TGF-β–Smad signalling Trends in Biochemical Sciences 2004;29(5):265 - 73.

513. Hong S, Lee C, Kim S-J. Smad7 Sensitizes Tumor Necrosis Factorâ€“Induced Apoptosis through the Inhibition of Antiapoptotic Gene Expression by Suppressing Activation of the Nuclear Factor-ÎºB Pathway. Cancer Research 2007;67(19):9577-83.

514. Levy L, Hill CS. Alterations in components of the TGF-[beta] superfamily signaling pathways in human cancer. Cytokine & Growth Factor Reviews 2006;17(1-2):41-58.

515. Heijink DM, Jalving M, Oosterhuis D, et al. TNF-related apoptosis-inducing ligand cooperates with NSAIDs via activated Wnt signalling in (pre)malignant colon cells. The Journal of Pathology 2010;223(3):378-89.

516. Picelli S, Zajac P, Zhou X-L, et al. Common variants in human CRC genes as low-risk alleles. European Journal of Cancer 2010;46(6):1041-8.

517. Santibanez Koref M, Wilson V, Cartwright N, et al. MLH1 Differential Allelic Expression in Mutation Carriers and Controls. Annals of Human Genetics 2010;74(6):479-88.

518. Kabat GC, Miller AB, Jain M, Rohan TE. Dietary intake of selected B vitamins in relation to risk of major cancers in women. Br J Cancer 2008;99(5):816-21.

519. Eussen SJPM, Vollset SE, Hustad S, et al. Plasma Vitamins B2, B6, and B12, and Related Genetic Variants as Predictors of Colorectal Cancer Risk. Cancer Epidemiology Biomarkers & Prevention 2010;19(10):2549-61.

520. Marchand LL, Donlon T, Hankin JH, Kolonel LN, Wilkens LR, Seifried A. B-vitamin intake, metabolic genes, and colorectal cancer risk (United States). Cancer Causes and Control 2002;13(3):239-48.

521. Cleary SP, Cotterchio M, Shi E, Gallinger S, Harper P. Cigarette Smoking, Genetic Variants in Carcinogen-metabolizing Enzymes, and Colorectal Cancer Risk. American Journal of Epidemiology 2010;172(9):1000-14.

522. Kaur-Knudsen D, Nordestgaard BG, Bojesen SE. CYP2C9 genotype does not affect risk of tobacco-related cancer in the general population. Cancer Epidemiology 2010;34(2):178-83.

523. Kim M-J, Nafziger A, Kashuba A, et al. Effects of fluvastatin and cigarette smoking on CYP2C9 activity measured using the probe S-warfarin. European Journal of Clinical Pharmacology 2006;62(6):431-6.

524. Tranah GJ, Chan AT, Giovannucci E, Ma J, Fuchs C, Hunter DJ. Epoxide hydrolase and CYP2C9 polymorphisms, cigarette smoking, and risk of colorectal carcinoma in the Nurses' Health Study and the Physicians' Health Study. Molecular Carcinogenesis 2005;44(1):21-30.

525. Cleary J, Shapiro G. Development of Phosphoinositide-3 Kinase Pathway Inhibitors for Advanced Cancer. Current Oncology Reports 2010;12(2):87-94.

526. Huang M-Y, Fang W-Y, Lee S-C, Cheng T-L, Wang J-Y, Lin S-R. ERCC2 2251A>C genetic polymorphism was highly correlated with early relapse in high-risk stage II and stage III colorectal cancer patients: A preliminary study. BMC Cancer 2008;8(1):50.

527. Ruzzo A, Graziano F, Loupakis F, et al. Pharmacogenetic Profiling in Patients With Advanced Colorectal Cancer Treated With First-Line FOLFOX-4 Chemotherapy. Journal of Clinical Oncology 2007;25(10):1247-54.

Appendices

Appendix 4.1

Summary of Script File for Data Generation

Removing the previous information, setting a seed and creating a variable called for the sample size

```
rm(list=ls(all=TRUE))
set.seed(1)
SampSize <- 400
```

Creating a variable to translate SEM to sd where necessary

```
rootn<-sqrt(SampSize)
```

Restrict function, to remove unrealistic values

```
restrict <- function(vec, min, max) {
        vec <- ifelse(vec <= max, vec, max)
        vec <- ifelse(vec >= min, vec, min)}
```

Male and female, basic characteristics and those that differ by sex are simulated and then combined along with the gender classification into a data frame. Simulations then follow for shared characteristics (e.g. BMI) environmental variables and SNPs. An example of a dietary variables, Vitamin D, is shown below.

```
FakeVitD<- round(rlnorm(SampSize, meanlog = 1.356,
        sdlog = 0.653),digits=1)
```

Odds Ratio Calculator to Assign Risk:

Define the proportion in the risk groups and the proportion of cases in the sample

```
p_nsaids <- 0.6634
p_gene <- 0.652
p_int <- p_nsaids*p_gene # 'y'
p_cases <- 0.5  # 'z'
```

Quadratic solver

```
A <- OR - 1
B <- (1 - OR)*SampSize*p_int + (1 - OR)*SampSize*p_cases -
        SampSize
C <- SampSize * SampSize * OR * p_int * p_cases
```

Calculate square root of the "determinant", if determinant is negative, no real solution

```
sq_det <- sqrt(B*B - 4*A*C)
```

Define the solutions as more positive (a1) and less positive (a2), then choose the correct solution (a3). The a, b, c and d refer to the cells of an OR contingency table

```
a1 <- (-B + sq_det) / (2*A)
a2 <- (-B - sq_det) / (2*A)
a3 <- ifelse(a1 > SampSize, a2, a1)
if (a3 < 0)

a <- round(a3)
b <- round(SampSize*p_int - a)
c <- round(SampSize*p_cases - a)
d <- round(SampSize - SampSize*p_int - c)
```

Assign the probability of each person being in each cell using the values for a, b, c and d. In this example the genotypes CC and CT confer an increase in risk when the person is not taking NSAIDs (n is NSAID intake).

```
CaseLabel <- c()
        for (num in 1:SampSize)
        {
                n <- Fakensaids[num]
                x <- rs2977522[num]
                if ((n == 0) && (x == "CT" || x == "TT"))
                        {
                                CaseLabel <- rbind(CaseLabel,
sample(c("2", "1"),1,replace=T, prob=c((a/SampSize), b/SampSize)))
                        }
                        else
                        {
                                CaseLabel <- rbind(CaseLabel,
sample(c("2", "1"),1,replace=T, prob=c((c/SampSize), d/SampSize)))
                        }
                }
```

Finally, all the variables and the case-control status are combined into a single data frame. A basic check is done to ensure the population have the desired OR.

Appendix 4.2

Genes associated with cancer syndromes

Gene	rs number	Polymorphism	Reference Frequency (%)	Variant Frequency (%)
APC	rs2431512	G → A	46	54
	rs2431238	A → G	30	70
	rs454886	T → C	64	36
	rs459552	A → T	20	80
PMS2	rs2286681	T → G	59	41
	rs12112229	G → T	67	33
	rs2345060	A → G	76	24
PTEN	rs1234225	G → A	63	37
	rs1234214	G → T	62	38
BMPR1A	rs10887654	T → C	70	30
	rs1124482	A → C	66	34
	rs7922846	T → A	75	25
	rs12765929	C → A	70	30
SMAD4	rs10502913	C → T	76	24
	rs8096092	G → T	62	38
STK11	rs3764640	C → A	80	20
	rs7256801	A → G	54	46
	rs1978728	G → A	63	37
	rs2075608	G → A	74	26
KIT	rs981959	C → A	43	57
	rs2017472	A → G	49	51
	rs759083	G → C	58	42
	rs6820303	C → T	54	46
	rs3020821	T → C	48	52
	rs3134889	A → G	35	65
	rs4864920	C → A	80	20
	rs1008658	A → G	34	66
PDGFRA	rs6850748	A → C	79	21
	rs7656613	A → G	75	25

Metabolic Gene Polymorphisms

Gene	rs number	Polymorphism	Reference Frequency (%)	Variant Frequency (%)
GSTM3	rs3814309	A → G	69	31
	rs1332018	C → A	43	57
GSTP1	rs1695	T → C	59	41
MTHFR	rs1476413	G → A	69	31
	rs1801131	A → C	66	34
	rs6541003	C → T	45	55
	rs1801133	C → T	69	31
	rs11121832	A → G	28	72
MTR	rs10733118	A → G	39	61
	rs883396	C → T	61	39
	rs3768149	T → C	56	44
	rs12129440	C → T	73	27
MTRR	rs3776467	G → A	22	78
	rs326121	T → C	75	25
	rs326123	G → A	37	63
	rs1532268	C → T	69	31
	rs162031	T → C	20	80
	rs161871	A → G	77	23
	rs10520873	T → C	70	30
MMP1	rs470215	A → G	65	35
	rs7125062	A → G	73	27
	rs2071232	A → G	79	21
	rs470358	A → G	39	61
MMP2	rs17301608	G → A	62	38
	rs243845	C → T	63	37
	rs2287076	A → G	54	46
	rs11639960	T → C	65	35
	rs243836	C → T	50	50
	rs7201	T → G	55	45
MMP3	rs591058	A → G	56	44
MMP7	rs2156528	A → C	80	20
	rs10750646	C → T	25	75
MMP9	rs3918253	G → A	44	56
	rs17576	T → C	64	36

MMP13	rs478927	A → G	29	71
CBS	rs706208	C → T	60	40
	rs2124458	A → G	67	33
	rs4920037	C → T	77	23
	rs234706	T → C	66	34
	rs2851391	A → G	47	53
NAT1	rs2410545	A → G	36	64
	rs13253389	T → C	26	74
	rs7003890	A → G	48	52
	rs15561	T → G	25	75
NAT2	rs1390358	A → G	63	37
	rs1112005	G → A	71	29
ALDH2	rs4767939	A → G	80	20
	rs2238151	T → C	61	39
	rs11066028	A → C	67	33
PIK3R1	rs706713	G → A	76	24
	rs7713645	T → G	52	48
	rs12652661	T → C	70	30
	rs251406	T → C	30	70
	rs173702	A → G	60	40
	rs4122269	A → G	66	34
	rs1823023	T → C	38	62
	rs173703	C → T	79	21
	rs706716	G → A	74	26
	rs13167294	T → G	76	24
	rs34309	C → T	62	38
	rs6876003	A → G	60	40
	rs3815701	T → C	80	20
	rs3756668	C → T	57	43

Genes on the Wnt Pathway

Gene	rs number	Polymorphism	Reference Frequency (%)	Variant Frequency (%)
CTNNB1	rs9813198	T → C	46	54
	rs4135385	T → C	78	22
PPP2R1B	rs736650	C → T	74	26
	rs618138	C → T	80	20
	rs647080	C → T	62	38
	rs4935790	A → G	70	30
BTRC	rs10883655	T → C	51	49
	rs4451650	C → T	34	66
	rs4436485	C → T	63	37
	rs9419923	T → C	79	21
CYCLIN D1	rs603965	C → T	49	51
	rs649392	C → T	41	59
C-MYC	rs3856557	C → T	54	46
	rs37771888	A → G	43	57
PS1	rs4213	A → C	65	35
AXIN1	rs393521	A → C	79	21
	rs214249	A → C	62	38
	rs7200589	C → T	74	26
	rs214246	T → C	54	46
	rs8063821	G → A	76	24
	rs1981492	C → T	58	42
	rs11649255	A → G	74	26
	rs3916990	T → C	67	33
	rs9921222	G → A	57	43
	rs12719801	G → A	70	30
	rs370681	G → A	44	56
	rs2885415	C → T	77	23
	rs1805105	T → C	39	61
AXIN2	rs7591	T → A	44	56
	rs11867414	A → G	35	65
	rs4074947	C → T	75	25
	rs2240308	C → T	52	48
	rs3923087	A → G	27	73
	rs3923086	T → G	48	52

	rs4541111	G → T	51	49
CDX2	rs2481952	G → A	52	48
WISP1	rs3739262	C → T	73	27
	rs16893344	G → A	65	35
	rs2977522	G → A	59	41
	rs2977533	C → T	62	38
	rs10100792	G → A	73	27
	rs2929946	T → C	39	61
	rs4330674	A → G	35	65
	rs6992383	G → A	43	57
	rs11774084	T → C	58	42
	rs6982341	T → C	42	58
	rs2929965	A → G	42	58
	rs2929967	G → A	52	48
	rs3739261	A → G	62	38
WISP2	rs6130677	C → T	78	22
	rs1230348	T → C	34	66
WISP3	rs6130677	G → A	49	51
	rs4812858	A → G	29	71
	rs6094027	T → C	42	58
	rs1061098	G → A	65	35
MYH/MUTYH	rs3219472	G → A	76	24
SFRP1	rs3242	C → T	68	32
	rs11781990	C → T	65	35
	rs7843510	T → C	58	42
	rs6651363	C → T	67	33
	rs10106678	A → G	50	50
	rs9694405	C → T	51	49
	rs968428	T → A	59	41

Other genes with inflammatory or regulatory NSAID effects

Gene	rs number	Polymorphism	Reference Frequency (%)	Variant Frequency (%)
UGT1A6	rs1604144	G → A	67	33
	rs12988520	T → G	41	59
	rs7583278	G → A	62	38
	rs4663965	A → G	55	45
	rs4148324	A → C	69	31
	rs4148328	G → A	61	39
	rs10929303	A → G	21	79
UGT1A7	rs7592624	C → T	42	58
	rs1604144	G → A	67	33
	rs6725478	G → A	63	37
	rs6431628	T → C	54	46
	rs4148324	A → C	69	31
	rs4148328	G → A	61	39
	rs10929303	A → G	21	79
CYP2C9	rs2860905	C → T	78	22
	rs7089580	T → A	76	24
	rs4918766	C → T	64	36
	rs1856908	A → C	36	64
	rs1934967	G → A	76	24
HGF	rs5745752	G → A	72	28
	rs2040968	G → A	23	77
	rs5745616	G → A	76	24
IL-10	rs3024498	A → G	71	29
	rs3024496	T → C	48	52
	rs3024490	T → G	22	78

Appendix 5.1

Computation Intensity of Analysis

There were a number of hurdles in trying to run so many datasets, each so large in nature, on a standard computer. The trouble came from saving 1000 datasets in the short term memory and then trying to run any analysis on them. A number of different approaches were tried:

1. To generate a single dataset, analyse it, save the results and clear the memory of the data itself. In practice however this led to the datasets varying depending on how often a random number generator (RNG) based application was used in the method. The mixed tree method employs a random forest step, which meant that when generating the second dataset, the random number generator was a few steps ahead in sequence than it would be when simply using logistic regression to analyse the results.

2. Adapt the first suggestion so that after a dataset is generated it is first analysed using the mixed tree method, then logistic regression, then MDR, saving all the results appropriately. However the difference in data format required for MDR and the removal of the MDR function from the more recent version of R made this impossible.

3. Using a UNIX machine, with compatible code for the simulations steps. This however, turned out to take longer than the windows version to run comparable runs.

The following table shows the time taken to run the analysis by different number of datasets and environmental variables. Figure 5.1x shows this relationship in graphical form.

Table 5.1x Timescale of Multiple Simulations

Number of variables	Number of datasets	Time taken (mins)	Notes
1	1	1	Results when 1000 datasets had been simulated and a selection run. Simulating less improved speed only marginally.
1	50	26	
1	100	71	
1	200	137	
1	500	184	
1	1000	357	
2	1	1	Similarly 1000 simulated. The faster run of 100 than 50 is simply down to random variation.
2	50	76	
2	100	73	
2	200	164	
2	500	504	
2	1000	1324	
5	1	4	Noticeably slower to observe than datasets containing only 1 or 2 variables..
5	50	162	
5	100	412	
5	200	1030	
5	500	2650	
5	1000	5112	
21	1	18	The first few variables, particularly binary ones, were processed quicker than later variables – showing strain on the short term computer memory.
21	50	421	
21	100	1267	
21	200	4720	
21	500	9231	
21	1000	18117	

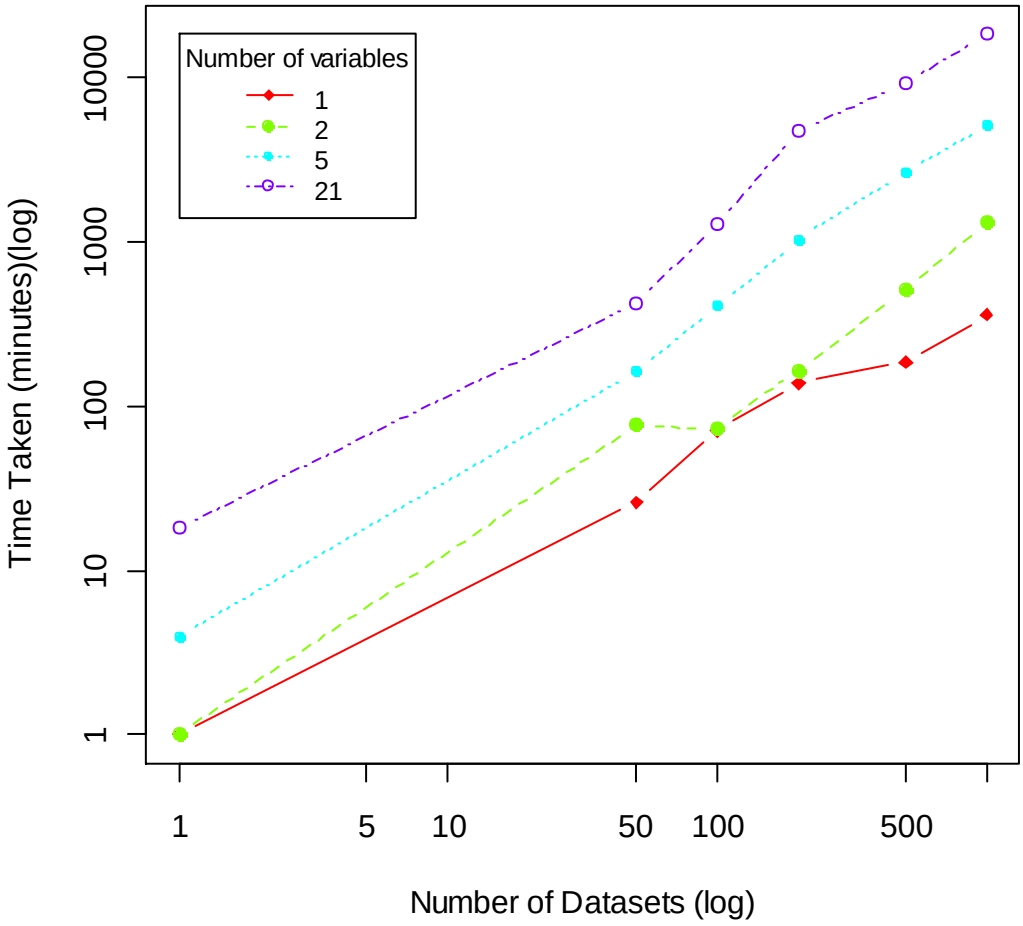


Figure 5.1x Computational Intensity, time taken (mins)

There are three factors that influence how long a run takes: number of variables; number of repetitions; and the number of datasets generated and being saved in the short term memory of the computer. Once the false error rate has been established there is no analytical advantage to running large number of variables alongside the target variable for each run. As long as there is one variable with an effect and one that doesn't have a set effect with which to compare, then it is possible to gain a full picture of what the results would be in such a situation without adding too much time to the simulation.

Appendix 6.1

When using multiple set seeds, it is important to ensure that the numbers being generated do not share any sort of common patterns [465]. With the size of a number of simulations exceeding the maximum possible with the available system memory, it was necessary to divide the simulation into sections and run them separately. In order to be able to repeat these simulations with other methods, set seeds were used and recorded for each section. When the simulation was done in four parts, the set seeds used were: 29, 17, 88 and 63; when two, only 29 and 17. For all other simulations, 29 was used as the set seed throughout.

To check for any potential patterns, each set seed was tasked with generating 100 random numbers between 1 and 10,000 to two decimal places. These numbers were then plotted on a number of scatter plots, each comparing one set seed with another, so six in all shown below as figures 6.1x1 – 6.1x6

As can be seen from the scatter plot, there were no repeating patterns across the set seeds, so using them together will not detract from the random nature of the data generation.

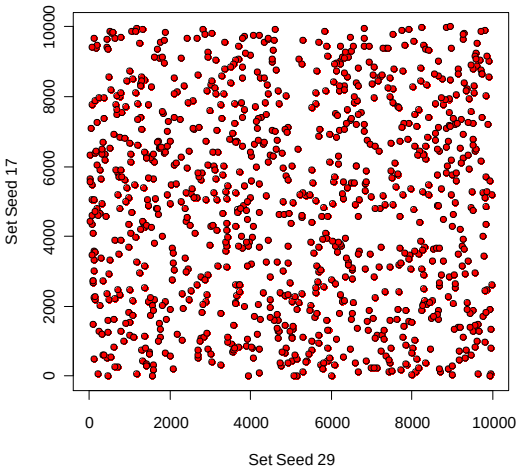


Figure 6.1x1 Set seeds 29 and 17

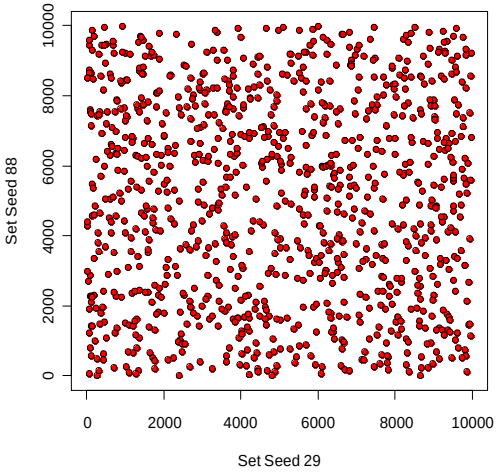


Figure 6.1x2 Set seeds 29 and 88

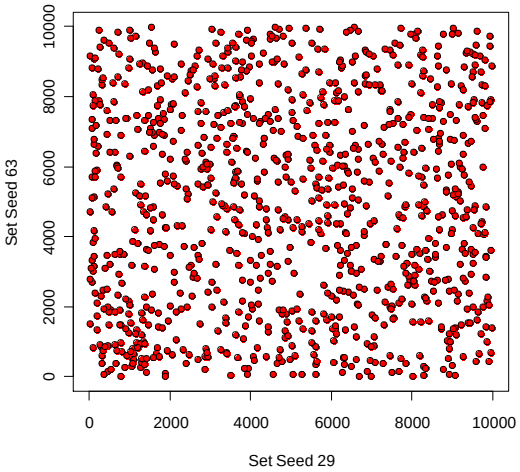


Figure 6.1x3 Set seeds 29 and 63

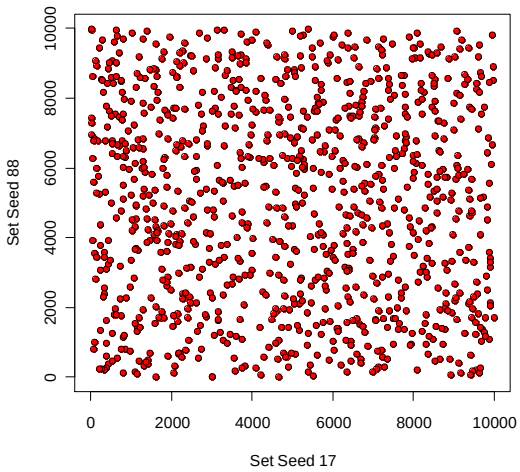


Figure 6.1x4 Set seeds 17 and 88

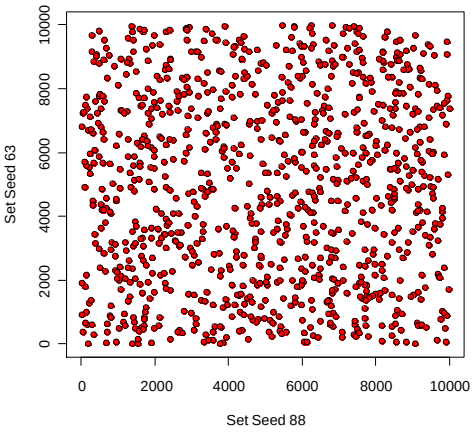


Figure 6.1x5 Set seeds 17 and 63

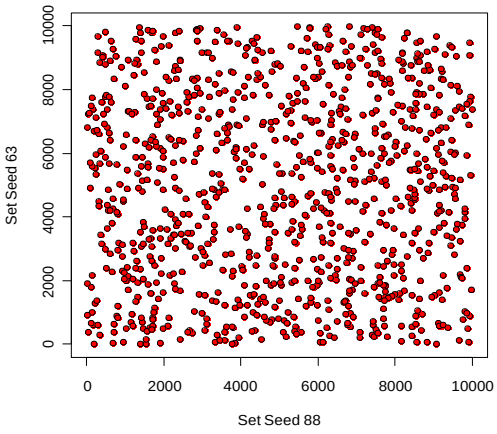


Figure 6.1x6 Set seeds 88 and 63

Appendix 6.2: Sample Size by Effect Size

Table 6.2x1 Percentage of Datasets Identifying Interaction at Different Effect Sizes, Sample Size 500

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Effect Size	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**
Target Int	1.6	7.5	8.4	37.5	50.2	0.3	1.7	2.2	13.3	22.7	0.3	1.2	1.3	9.2	17.6	0.1	0.2	0.5	4.8	8.3
Target SNP	7.0	33.5	28.1	86.9	95.1	2.7	19.4	21.1	72.1	85.9	1.9	14.2	18.2	66.7	82.8	0.7	6.6	11.6	51.7	71.9
Neither Int or SNP	92.8	66.2	71.9	12.8	4.7	97.1	80.0	78.8	27.8	13.9	97.9	85.3	81.7	33.1	17.0	99.2	93.3	88.3	48.1	28.0
Other Int	30.8	25.2	26.9	22.0	21.5	6.3	4.6	4.7	4.7	4.6	3.3	3.0	3.1	2.8	2.6	0.8	0.9	1.1	0.5	0.7
Other SNP	32.6	23.9	8.2	5.5	3.8	9.6	6.3	1.3	1.7	1.2	4.9	3.9	0.9	1.0	0.5	0.9	0.9	0	0.2	0.1
Splitting on environmental variable with no effect (smoking)																				
Target SNP	2.8	11.9	10.0	47.7	62.0	1.1	5.7	4.4	29.4	41.9	0.7	3.5	3.4	24.0	35.5	0.2	1.1	2.2	12.7	21.7
Target SNP and smoking	0.7	1.1	0.5	1.4	1.2	0.1	0.1	0	0.3	0.3	0.1	0.1	0	0.1	0.1	0.1	0	0	0	0
Other Int	27.1	25.9	28.5	25.0	22.2	6.3	5.9	3.0	3.7	3.7	3.8	3.8	2.2	3.3	2.7	0.7	1.0	0.5	1.1	0.8
Other SNP	33.9	29.4	13.9	20.0	17.6	10.2	9.4	2.1	5.5	5.5	5.8	5.4	1.7	4.0	3.6	1.3	1.8	0.3	1.1	0.9

Table 6.2x2 Percentage of Datasets Identifying Interaction at Different Effect Sizes, Sample Size 2000

	Significance Threshold																			
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Effect Size	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**
Target Int	16.4	64.0	90.1	98.8	99.9	4.9	36.6	72.9	92.1	98.5	3.7	30.4	65.2	88.1	97.8	1.4	19.8	49.4	78.4	93.4
Target SNP	53.8	99.1	100	100	100	35.6	95.6	100	100	100	29.3	94.0	99.9	100	100	18.0	88.9	99.8	100	100
Neither Int or SNP	45.5	0.9	0	0	0	63.8	4.3	0	0	0	70.1	5.8	0.1	0	0	81.6	10.9	0.1	0	0
Other Int	23.1	22.1	20.7	21.0	21.5	6.7	5.8	5.2	6.0	5.3	4.5	3.5	3.4	3.8	3.3	1.2	1.2	1.1	1.2	0.8
Other SNP	11.9	2.1	1.9	2.2	1.7	4.3	0.2	0.2	0.5	0.2	2.5	0.1	0	0.3	0.1	1.0	0	0	0	0.1
Splitting on environmental variable with no effect (smoking)																				
Target SNP	23.8	81.0	97.4	99.8	100	11.7	63.1	91.7	98.3	99.9	8.5	57.1	88.2	97.8	99.9	3.4	41.2	80.1	95.2	99.8
Target SNP and smoking	1.4	1.6	1.7	1.5	1.6	0.1	0.1	0.1	0.3	0.5	0	0	0	0.2	0.2	0	0	0	0	0
Other Int	24.6	24.5	25.4	26.1	27.1	5.6	5.0	4.8	6.4	5.6	3.4	3.6	3.2	3.3	3.1	1.2	1.3	0.9	0.9	1.1
Other SNP	24.0	13.5	17.7	21.5	24.9	6.1	3.3	3.7	5.4	6.2	3.6	1.7	2.0	2.8	3.6	1.4	0.5	0.8	0.4	0.9

Appendix 6.3: Interaction Type

Table 6.3x1 Percentage of Datasets Identifying Interaction for Interaction Model A

Significance Threshold																				
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	0	0.2	0.3	0.2	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0
Target SNP	0	0.1	0.4	0.3	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0
Neither Int or SNP	100	100	99.5	99.7	100	100	100	99.9	100	100	100	100	100	100	100	100	100	100	100	100
Other Int	26.0	29.8	30.7	30.9	30.3	5.2	5.8	5.4	7.2	7.3	3.1	5.2	3.2	3.9	4.7	1.2	1.3	1.2	0.8	1.9
Other SNP	34.9	38.1	38.3	37.8	36.6	8.7	9.0	9.5	9.3	10.5	4.5	5.7	4.1	4.3	5.7	0.9	1.3	1.1	0.7	1.5
Splitting on environmental variable with no effect (smoking)																				
Target SNP	0.7	3.5	4.5	0.3	0	0.7	2.3	1.7	0.1	0	0.7	1.9	1.4	0.1	0	0.7	1.0	0.5	0	0
Target SNP and smoking	0.2	0.5	1.0	0	0.1	0.1	0.1	0.2	0	0	0.1	0.1	0.1	0	0	0.1	0	0	0	0
Other Int	30.0	30.0	28.3	28.2	30.8	6.3	7.5	5.1	5.8	7.1	3.7	4.5	3.0	3.0	4.0	1.0	0.8	0.4	0.7	0.9
Other SNP	33.5	34.3	31.1	34.5	31.6	9.2	10.1	10.8	8.9	8.2	6.1	6.3	6.9	4.8	4.7	1.4	1.9	1.6	0.9	0.8

Table 6.3x2 Percentage of Datasets Identifying Interaction for Interaction Model B

Significance Threshold																				
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	11.2	80.8	61.8	8.6	0.2	6.8	59.3	36.3	1.6	0.1	5.6	52.1	30.0	1.0	0	4.0	33.7	18.2	0.4	0
Target SNP	13.2	94.8	96.6	29.0	0.7	13.1	93.9	89.8	14.3	0.2	12.9	93.3	86.4	10.6	0.2	12.0	90.0	76.1	4.7	0
Neither Int or SNP	86.8	5.2	3.4	70.1	99.1	86.9	6.0	9.9	85.3	99.7	87.1	6.6	13.3	88.9	99.8	88.0	9.9	23.7	95.1	100
Other Int	26.1	17.0	18.9	24.8	27.1	6.0	3.7	3.6	5.1	5.3	3.6	1.8	2.4	3.2	2.7	0.9	0.5	0.5	0.6	0.5
Other SNP	31.7	3.5	3.4	21.2	37.9	10.4	0.8	0.9	6.4	10.6	6.5	0.6	0.6	3.6	5.3	1.6	0.2	0.2	1.3	1.7
Splitting on environmental variable with no effect (smoking)																				
Target SNP	0.8	48.9	63.0	8.4	0.2	0.7	42.2	42.9	3.2	0.2	0.7	37.5	35.9	2.2	0.2	0.7	26.7	22.8	1.1	0.1
Target SNP and smoking	0.1	0.9	1.6	1.0	0.3	0	0.1	0.3	0.2	0.2	0	0.1	0.2	0.1	0.1	0	0.1	0	0.1	0
Other Int	30.4	25.7	28.3	28.0	30.1	7.6	5.6	5.9	6.8	5.2	4.5	2.8	3.2	3.6	2.7	0.6	0.5	0.8	0.8	0.8
Other SNP	34.1	21.9	18.3	32.3	35.1	9.3	6.7	5.6	8.4	9.3	5.8	3.6	3.9	5.0	5.5	0.9	0.6	1.1	1.1	0.9

Table 6.3x3 Percentage of Datasets Identifying Interaction for Interaction Model C

Significance Threshold																				
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	11.6	83.3	65.7	11.2	0.1	8.1	61.8	37.4	3.1	0	6.1	53.1	30.2	2.2	0	3.2	37.3	17.1	0.4	0
Target SNP	13.9	98.6	98.2	35.7	1.0	13.8	98.5	93.8	15.4	0.6	13.7	98.0	91.7	12.5	0.4	13.6	95.6	80.5	6.7	0.4
Neither Int or SNP	86.1	1.4	1.6	62.6	98.9	86.1	1.5	6.0	83.8	99.4	86.2	2.0	8.2	87.0	99.6	86.3	4.3	19.5	93.3	99.6
Other Int	17.7	11.9	12.7	10.6	17.0	3.3	2.3	2.1	1.6	3.1	2.4	1.2	1.2	0.7	1.9	0.8	0.4	0.2	0.1	0.4
Other SNP	18.0	1.4	2.7	10.4	20.5	3.4	0	0.9	2.1	3.4	1.6	0	0.6	1.2	2.0	0.4	0	0	0.3	0.5
Splitting on environmental variable with no effect (smoking)																				
Target SNP	1.2	64.8	73.2	11.6	0.1	1.0	56.8	49.3	4.6	0.1	0.9	53.2	41.2	3.5	0.1	0.7	43.0	27.2	1.5	0
Target SNP and smoking	0.1	0.5	1.1	0.9	0.4	0	0	0	0	0.2	0	0	0	0	0.1	0	0	0	0	0.1
Other Int	19.3	18.0	20.0	18.3	19.8	3.5	4.0	2.6	3.4	2.9	2.3	1.7	1.1	1.5	2.0	0.4	0.4	0.2	0.1	0.1
Other SNP	21.2	12.1	10.9	15.3	18.6	3.2	3.2	2.7	3.4	3.9	1.8	1.8	1.0	1.6	1.8	0.4	0.4	0.2	0.5	0.4

Table 6.3x4 Percentage of Datasets Identifying Interaction for Interaction Model E

Significance Threshold																				
	0.01					0.001					0.0005					0.0001				
Splitting on Target Environmental Variable (NSAIDs)																				
Allele frequency	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**
Target Int	12.1	59.2	42.9	5.7	0	7.2	33.6	19.0	0.6	0	5.5	27.2	14.6	0.4	0	2.8	15.5	7.1	0.1	0
Target SNP	15.8	89.9	90.6	23.5	0	15.8	87.9	77.3	10.7	0	15.7	86.0	72.2	7.7	0	14.9	78.8	57.9	2.9	0
Neither Int or SNP	84.2	10.1	9.3	75.8	100	84.2	12.1	22.4	89.2	100	84.3	14.0	27.4	92.3	100	85.1	21.1	42.0	97.1	100
Other Int	28.5	20.1	19.6	23.6	28.5	5.7	4.8	4.4	5.5	3.3	3.3	2.6	3.0	2.4	2.8	0.6	0.7	0.5	0.8	0.5
Other SNP	29.8	4.3	4.8	21.8	38.0	8.1	1.2	1.2	7.0	10.0	4.8	1.0	0.8	4.4	6.4	1.2	0.3	0	1.2	1.6
Splitting on environmental variable with no effect (smoking)																				
Target SNP	2.0	47.2	35.1	7.1	0.2	1.9	39.2	33.7	2.5	0.2	1.8	35.3	27.4	1.6	0.2	1.5	24.3	14.4	0.7	0.2
Target SNP and smoking	0.2	1.6	1.3	0.6	0.2	0	0.3	0.3	0.2	0	0	0.1	0.2	0.2	0	0	0	0.1	0.1	0
Other Int	30.0	25.2	25.3	29.0	28.6	6.5	5.3	6.0	5.9	6.1	3.8	2.4	3.3	3.1	4.2	0.8	0.5	1.0	0.9	1.1
Other SNP	31.9	22.1	17.2	31.5	37.4	8.0	6.6	5.1	3.7	8.1	4.9	3.2	2.9	2.1	5.0	0.8	0.5	1.0	1.0	1.0