THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *28/11/2013* **par :**

NORA BOHOSSIAN

Pharmacogénomique de la sclérose en plaques: méthodes et applications

JURY

DR. EMMANUELLE GÉNIN	Rapporteur
PR. LAURENT BECQUEMONT	Rapporteur
DR. MARIA MARTINEZ	Directrice de Thèse
PR. DAVID BRASSAT	Co-Directeur de Thèse

École doctorale et spécialité :
BSB : Épidémiologie
Unité de Recherche :
Institut National de la Santé et de la Recherche Médicale (UMR 1043)
Directeur(s) de Thèse :
Maria MARTINEZ et *David BRASSAT*
Rapporteurs :
Emmanuelle GÉNIN et *Laurent BECQUEMONT*

“ 'Gainst the wall of knowledge I
All my little wisdom try. ”

“*Pangur Bán*”
Translated from Irish by Robin Flower

I dedicate this work to the MS patients and their families for the courage, understanding and patience needed to live with this condition.

ACKNOWLEDGEMENTS

I would like to thank my thesis director, Dr. Maria Martinez, and co-director, Prof. David Brassat, for this life changing experience. I am deeply indebted to you, David, for giving me this opportunity and to you, Maria, for making sure I do not waste it. David, thank you for your great sense of humor and constant positive attitude but I especially thank you for explaining to me the many complexities of MS. Maria, thank you for pushing me to excel both professionally and personally and for introducing me to the fascinating field of genetic epidemiology. Both of you encouraged me to attend courses and conferences and to develop my scientific skills. Both of you provided me with many opportunities to meet with leading scientists in the respective fields. I am very grateful to you for that. Most of all, however, I am grateful to both of you for giving me a project that I loved working on.

I sincerely thank the reviewers of my thesis, Dr. Emmanuelle Génin and Prof. Laurent Becquemont, for accepting to judge my work and for the time and effort they devoted to it. I also express my regret that Dr. Justo Lorenzo Bermejo could not serve on my jury.

I thank the members of the comité de suivi, namely, Prof. Bertrand Fontaine, Dr. Nicolas Blanchard and Dr. Frédéric Taieb, for evaluating the status of my work and advising me on its progress halfway through my studies. I especially thank Nicolas who, as my parrain de thèse, continued to advise and help me when needed.

Given the multidisciplinary nature of my project, it is no surprise that it required the collaboration of people with different expertise. While this presented challenges in itself, it made my experience richer and I thank Prof. Roland Liblau and Dr. Abdelhadi Saoudi (Team 5: Inflammatory diseases of the central nervous system: mechanisms and therapy), and Dr. Marie-Paule Roth and Dr. Hélène Coppin (Team 13: Genetics and iron metabolism regulation) for welcoming me in their respective teams.

The CRs in these teams were like mentors to me. I would like to specifically thank Lennart Mars (Team 5) for numerous fruitful discussions on my project as well as on many other topics and François Canonne-Hergaux (Team 13) for his honesty and scientific integrity. I also thank François for his valuable and refreshing company during the lunchtimes and for introducing me to Lorenne Robert and Thierry Duprat who added to the animated and varied discussions during the lunches that I much appreciated.

I thank the ladies of Prof. David Brassat's research and clinical teams, and notably the research team, Florence Bucciarelli, Béatrice Pignolet, Lise Scandella and Florence Umuhoza, for their positive attitude and encouragement especially towards the end. Florence B., I admire your efficiency and skillful management of the projects. Béatrice, your kindness on many occasions has marked me. Lise, while I missed seeing you around in the lab, the few times you brought us Coralie all made up for it. Florence U., there is such grace and style in you. Thank you all for making it a pleasure to work with you.

I further thank all my past and current officemates for creating a pleasant working environment including, recently, Myriam Hayder and Raphaelle Romieu-Mourez, who were also very supportive in these final steps.

I thank all members of Teams 5 and 13 as well as of the teams of Daniel Dunia, Nicolas Blanchard and Renaud Lesourne with whom I also interacted although on a less frequent basis throughout these

years. I thank you for heroically surviving through my statistical presentations during the Wednesday lab meetings.

I am very grateful to Andrés Legarra, from INRA Toulouse (Auzeville), for introducing me to the multi-marker models and explaining to me the software tools to conduct the multi-marker analyses which became an integral part of this thesis.

It is a pleasure to thank Institut national de la santé et de la recherche médicale, INSERM U1043, and Centre de Physiopathologie de Toulouse Purpan, CPTP, where I conducted my work for the excellent working conditions that were provided to me. I also thank the administrative staff from these institutions, especially, Dominique Rozet, Nathalie Frances and Jennifer Galtié from CPTP and Stéphanie Azalbert from INSERM U1043 for their kindness, efficiency and their indispensable help through the bureaucratic labyrinth on numerous occasions.

Similarly, I thank Christine Ferran and Sabine Aubineau from INSERM U1043 Communications with whom I enjoyed organizing the annual "rencontres chercheurs patients". I take this opportunity also to thank once again all the researchers who kindly accepted to participate in these meetings. These meetings required a certain level of scientific maturity and skill and that is why I particularly acknowledge the effort made by those in earlier stages of their careers to present their projects: Charlotte Foret, Caroline Charlier, Liliana Lucca, Myriam Hayder, Anne Dejean, Sophie Laffont-Pradines and Lennart Mars.

Throughout my PhD years, I had the absolute pleasure of meeting many MS patients professionally as well as informally. Special mention goes to Anne-Marie Blech and Sonia Mazoyer for the pleasant moments we shared.

My studies were financed by two different grants. It is a privilege for me to acknowledge them here: United Europeans for the development of PHArmacogenomics in Multiple Sclerosis (UEPHA*MS, Marie Curie Initial Training Network) and la Fondation pour l'Aide à la Recherche sur la Sclérose En Plaques.

I extend a special thank you to my friends and colleagues from Statistics Canada. Your support was much valued. I am especially grateful to my former supervisor, Pierre Daoust, and my colleague and friend, Olena Frolova, for their support during the year prior starting my PhD. Referring to that year, I also acknowledge the advice I received from my master's thesis supervisor, Prof. David Wolfson. Already seven years have passed since I last worked with you and learnt from you, David, but your influence continues to this day.

Throughout the PhD years, I faced many difficult moments but today all this is irrelevant simply because I only remember how good it felt to share these moments, as well as great ones, with the following people: Aude Saint-Pierre, Mohamad Saad, Myriam Lawand, Amit Saxena, Abdulraouf (Raouf) Ramadan, Matthias Macé, Sahar Kassem, Amine Benarbia, Elodie Persyn, Cristina Bucur, Liliana Marques and Audrey Cauchard.

I thank Aude and Mohamad for the warm welcome in their team right from day one and for constantly reminding me of the reward of perseverance. I thank Myriam and Amit for making an unforgettable ambiance in the office in my first year and for the memorable moments we have shared outside it. I have not stopped missing you. I thank Raouf for his humanity and extreme generosity. Matthias, your free spirit and intelligence are unique and precious. Sahar, I admire your maturity and mental force…and, of course, your culinary skills. Thank you for being by my side in crucial moments and for the many delicious Lebanese “goûter”-s I had the luxury to taste. Amine, you always gave me such wise advices which helped me manage stressful situations. I sincerely appreciated also your help with the French translation of this thesis abstract. Elodie, how lucky we were to have you! Cristina, not once you failed to “entertain” us with your stories. Luck is a relative concept, never think of it in absolute terms! Liliana, where do I begin…if it was not for you, I would not have had the stamina to work as hard as I did completing my thesis. Thank you for teaching me that everything should be given the time it deserves even if time is not on our side. Audrey, you know how much I appreciated your presence and the team dynamic you created. I am only sorry we crossed paths so briefly.

Being in Europe gave me the opportunity to meet again some of my Canadian friends who have returned to their roots including Danae Mahera and Daniele Toninelli. That was truly special and I thank you for keeping our friendship lit after so many years.

I enjoyed living in Toulouse and there are few places in this city the beauty and majesty of which gave me comfort at times of distress and which are engraved in my memory. I thank my flatmates in my first year, Claire-Léa Boccard and, especially, Hélène Fassolis. You were my first contact with life outside work here in Toulouse and, to this day, I find hard to believe the luck I had in finding you. In the two years that followed, I lived chez Mme. Anne-Lise Couineau and, this, quite frankly, was one of the most enriching encounters of my life. I have always said that you are an extraordinary woman, Anne-Lise. Thank you for your attentive gestures and artistic creations, for constantly reminding me of the importance of humor in our lives, for sharing your treasures - your books and written works - with me and for the long and deep discussions on so many varied topics. It is two years rich in memories, precious memories, thank you. Unforgettable is also my encounter with Jianli, my Chinese characters teacher. The world of Chinese characters is truly fascinating, Jianli, and you managed to instill in me great admiration for it but what most impressed me was your unbounded benevolence.

I have often wondered where I would be if I did not have the constant love and support from my family and, especially, my parents and my brother. You have been the backbone of every single achievement of mine and no words are enough to thank you for this. Yet, education begins at home and it is the kind of education that can be attained with no academic degree - that of love, compassion and devotion. There were times when I would cease to believe in the essence of the immense effort required behind this PhD. It is precisely the education you have given me that helped me regain my belief and keep going.

ABSTRACT

The field of genetics is rapidly expanding and evolving. As more and more is understood on the genetics of complex human traits, a natural question arises as to how these findings can be translated to the everyday medical practice. While a little more than a decade ago sequencing the entire human genome was achieved by the largest international scientific collaboration ever undertaken in biology, today it is not farfetched to expect that in the near future obtaining the genetic profile of each patient may become routine medical practice. Pharmacogenomics, a blend of pharmacology and genomics, aims to determine the most suitable treatment for each patient as a function of his or her genetic makeup. Pharmacogenomic studies have increasingly provided evidence that there are gains to be achieved by incorporating genetic information when determining the optimal treatment choice for a patient. The case of warfarin, an anticoagulant, has often been considered as one of the most motivating success stories to pursue such type of studies. The success as well as the need of such studies, however, depend on a multitude of factors and vary greatly across traits.

The objective of this thesis is to evaluate the current state of the art for Multiple Sclerosis (MS), a debilitating neurological disorder affecting primarily young adults. To date, no cure exists for MS but a number of disease-modifying therapies have been approved with varying degree of efficacy and toxicity. So far, little is known on the genetic factors that influence response to treatment in MS patients. Moreover, even if such factors are known apriori, evaluating and proving their utility at the clinical level is not as straightforward as one may be inclined to think. In this thesis, we highlight why the road to translate such findings to medical practice remains rough and challenging.

In particular, relying on the association and prediction studies that we have conducted, we expose the design and limitations of each and discuss model choice in each context. Specifically, we conducted single-marker association analysis of response to interferon-β in MS patients. We compared single-marker to multi-marker models in the context of association and also in that of prediction using both real and simulated datasets. Different approaches to multi-marker modeling exist. We focused on polygenic score analyses and Bayesian estimation methods and evaluated several of the properties of these modeling approaches.

Our findings showed that, in the context of association, the use of more complex and computationally heavy multi-marker models that has been recently advocated may lead to little, if any, benefit over the classical single-marker association analysis. On the other hand, multi-marker models that take into account the effect of many markers simultaneously clearly appear better suited to predict genetic risk. Nevertheless, focusing on polygenic score analyses, we demonstrated that many factors such as the study sample size and the heritability of the trait influence the predictive performance of a model.

Pharmacogenomic studies may revolutionize patient care. However, in all the excitement of the promise that they hold, in the concluding part of this thesis we also address the social, ethical and economic issues that they raise.

KEYWORDS: pharmacogenomics, multiple sclerosis, interferon-β, association study, prediction study, genetic markers, polygenic scores, Bayesian estimation methods, ethics

RESUME

L’expansion ainsi que l’évolution du domaine de la génétique au cours de ces dernières années a été fulgurante. Cela s’accompagne par la génération d’une masse importante d’information génétique sur les traits complexes chez l’homme. Une question naturelle est de savoir comment utiliser cette information dans la pratique médicale quotidienne. Il y a dix ans à peine le séquençage du génome humain nécessitait une collaboration scientifique d’envergure internationale entre les différents acteurs de la recherche biomédicale. Aujourd’hui, il n'est pas exclu à ce que, dans un avenir proche, on puisse obtenir le profil génétique de chaque patient dans la pratique médicale courante. La pharmacogénomique, une fusion de la pharmacologie et de la génomique, vise à déterminer le traitement le plus approprié à chaque patient en fonction de son patrimoine génétique. En effet, plusieurs études pharmacogénomiques ont pu démontrer l’intérêt d’intégrer l'information génétique du patient pour déterminer son traitement optimal. Le cas de la warfarine, un anticoagulant, a souvent été considéré comme l'un des succès les plus motivants pour poursuivre ce type d'études. Cependant, le succès ainsi que le besoin de ces études dépendent de multiples facteurs et varient considérablement selon les traits étudiés.

L'objectif de ce travail est d'évaluer l'état actuel des connaissances pour la sclérose en plaques (SEP), une maladie neurologique invalidante touchant principalement les jeunes adultes. À ce jour, il n'existe aucun remède à la SEP, mais il existe des traitements modificateurs de la maladie avec des degrés d'efficacité et de toxicité variable. Les facteurs génétiques qui influencent la réponse au traitement chez les patients atteints de SEP sont à ce jour mal connus. Même si ces facteurs peuvent être mis en évidence dans le futur, il n’en demeure pas moins que leur utilisation en routine clinique n’est pas aussi simple que supposée. Dans ce travail, nous avons essayé de mettre en évidence la complexité du passage de l’utilisation de données génétiques à grande échelle à la pratique médicale pour les traits complexes.

Nous avons mené des études d’association et de prédiction. Tout d’abord, nous exposons leurs concepts et revisitons les différences dans leurs objectifs. Plus précisément, nous avons effectué une analyse d’association simple-marqueur de la réponse à l'interféron-β chez les patients atteint de SEP. Ensuite, nous avons comparé les modèles simple-marqueur et multi-marqueur dans le contexte de la recherche d’association puis dans celui de la prédiction en utilisant des données réelles et des données simulées. Différentes approches de modélisation multi-marqueur existent. Nous nous sommes basés sur l'analyse des scores polygéniques et des méthodes d'estimation bayésienne en évaluant plusieurs des propriétés de ces approches de modélisation.

Nos résultats montrent que, dans la cadre d’une étude d’association pangénomique, les modèles multi-marqueurs, récemment préconisés, ne sont pas forcément plus puissants que les modèles classiques simple-marqueur. En revanche, les modèles multi-marqueurs qui prennent en compte l'effet de plusieurs marqueurs simultanément apparaissent clairement mieux adaptés pour prédire le risque génétique. Néanmoins, en se concentrant sur l'analyse des scores polygéniques, nous montrons que de nombreux facteurs comme la taille de l'échantillon de l'étude et l'héritabilité du trait influencent la performance prédictive d'un modèle.

Les études pharmacogénomiques peuvent révolutionner les soins aux patients. Cependant, en dehors de l'enthousiasme qu’elles peuvent susciter, nous discutons dans la dernière partie de cette thèse les questions sociales, éthiques et économiques qu'elles soulèvent.

MOTS-CLÉS: pharmacogénomique, sclérose en plaques, interféron-*β*, étude d’association, étude de prédiction, marqueurs génétiques, scores polygéniques, méthodes d'estimation bayésienne, éthique

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 The Genetic Material

"*We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.*" (*Watson and Crick, 1953*) Such is the opening of the 1953 monumental paper by James Watson and Francis Crick postulating on the structure of the molecule of life carrying our genetic information, the DNA.

The human body is composed of trillions of cells which store our genetic information. In particular, the nucleus of each cell contains two copies of 23 different chromosomes with one copy inherited from each of our parents. Of the 23 chromosomes, one is a sex chromosome (X or Y) determining the gender of an individual where females carry two copies of the X chromosome (XX) and males carry a copy of each (XY). The chromosomes comprise long strings of double-stranded DNA, made up of four nucleotide bases, namely, Cytosine (C), Adenine (A), Guanine (G), and Thymine (T). The two strands of the DNA are connected through hydrogen bonds between complementary base pairs where A always pairs with T and C always pairs with G. This is illustrated in **Figure 1.1** below.

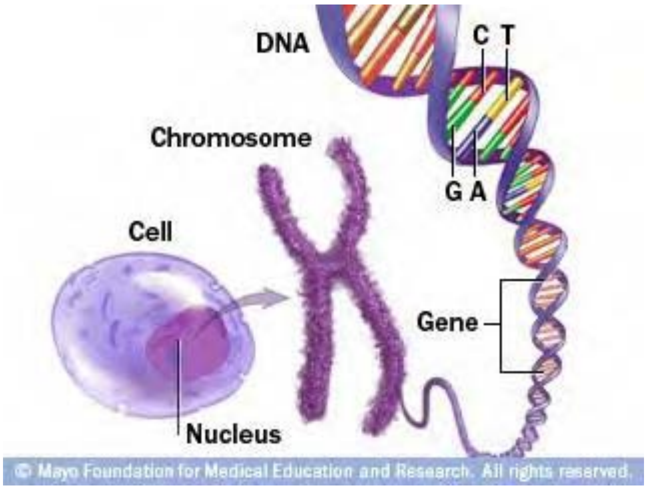

DNA
C T
Chromosome
G A
Cell
Gene
Nucleus
© Mayo Foundation for Medical Education and Research. All rights reserved.

Figure 1.1: How genetic information is stored in our bodies. Adapted from (*Mayo Clinic staff, 2011*).

The human genome consists of roughly 3 billion DNA base pairs. A specific sequence of these bases forms genes. It was long believed that genes coded for a single protein but this simplified assumption has been refuted. Specifically, the same gene can code for more than one protein or for none at all (directly transcribe to ribonucleic acid, or RNA, a single-stranded molecule similar to the DNA with the nucleotide base Thymine (T) replaced by Uracil(U)). It is complicated to come up with a precise definition of a gene and as such the estimated number of genes in the human genome vary based on the definition used (*Pennisi, 2003*). The most recent estimate lies somewhere around 20 500 genes. (*Clamp et al., 2007*) The regions of the DNA between genes are referred to as intergenic regions. The DNA comprises roughly 75% of intergenic regions. Of the remaining 25% of the DNA spanned by genes, only 1% are exons (coding for RNA or protein) while the remaining 24% are introns (non-coding sequences) (*Venter et al., 2001*).

Individuals share more than 99% of their DNA sequence. The remaining 1% or so of our genetic variation influences disease susceptibility and other complex traits and has proven important in the study of human health. Most of the genetic variation occurs in intergenic regions but some occur in genes and may thus directly impact their function. Nevertheless, even if the variations lie in intergenic regions, they may still be implicated in the susceptibility to diseases and in the phenotypic variation of other complex traits.

1.2 Biomarkers

The term biomarker is short for biological marker. In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as "*a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention*". (*Biomarkers Definitions Working Group, 2001*) There are many different applications of biomarkers leading to three major categories of biomarkers: diagnostic, prognostic and predictive. A diagnostic biomarker is a diagnostic tool for the identification of a disease. A prognostic biomarker is an indicator for disease prognosis. Lastly, a predictive biomarker predicts response to an intervention or treatment.

The United States Food and Drug Administration (US FDA) industry guidelines, proposed in 2008, narrow down the definition of a genomic biomarker as a "*measurable DNA and/or RNA characteristic that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other interventions*." (*FDA, 2008*) DNA characteristics include but are not limited to variations in a single DNA base (a Single Nucleotide Polymorphism or SNP) and other more complex forms of genetic variations discussed by (*Frazer et al., 2009*). Alternatively, RNA characteristics can refer to RNA sequences, microRNA levels and others. Thus, a genomic biomarker may simply consist of a single SNP or of a more complex combination of several DNA and/or RNA characteristics.

1.3 Pharmacogenetics and Pharmacogenomics

The terms pharmacogenetics and pharmacogenomics are often used interchangeably. The US FDA proposed industry guidelines in 2008 define pharmacogenomics as "*the study of variations of DNA* *and RNA characteristics as related to drug response*” and pharmacogenetics, as a subset of pharmacogenomics, and define it as “*the study of variations in DNA sequence as related to drug response*.” (*FDA, 2008*) Thus, pharmacogenetic studies, by definition, do not involve the study of variation in RNA characteristics. More recently, the terms are distinguished based on the scope of the study. A pharmacogenetic study focuses on variations related to drug response in a targeted gene. On the other hand, a pharmacogenomic study investigates the variations related to drug response across multiple genes or even at the genome-wide level (*Ritchie, 2012*).

Such studies may be conducted at all stages of the drug development process from drug discovery to clinical practice. Further, drug response is a broad term encompassing drug disposition (that is, absorption, distribution, metabolism, and excretion, known by the acronym ADME) and drug effects (that is, efficacy and adverse effects).

Presently, the US FDA website (*FDA, 2013*) lists roughly 120 drugs with pharmacogenomic information in their labels. For instance, the label of the drug carbamazepine (Tegretol®, Novartis), one of the most widely used and effective treatments of epilepsy, recommends against the treatment of patients carrying a specific variant in the human leukocyte antigen (HLA) region. This variant, found almost exclusively in patients of Asian ancestry, has been associated with serious side effects in these populations. (*Novartis, 2007*) Alternatively, the gene *CYP2C19* is implicated in the metabolism of many drugs. The translation of *CYP2C19* pharmacogenetics into clinical practice, however, is currently limited to a small number of functional variants although more than 2000 variants have already been discovered. (*Lee, 2012*)

1.4 Study Designs

Pharmacogenomic studies do not differ significantly from traditional epidemiological studies but important considerations specific to pharmacogenomic studies exist. Major epidemiological study designs are summarized in **Figure 1.2** below.

Two main types exist, experimental and observational. In the former, the investigators aim to control for all main forms of bias. In the latter, the information is passively observed and collected by the investigator. The most common example of an experimental study design is a randomized controlled trial (RCT) where subjects are randomly assigned to one of several treatment groups. This type of study design most closely resembles a controlled experimental setting and typically leads to the most rigorous scientific results.

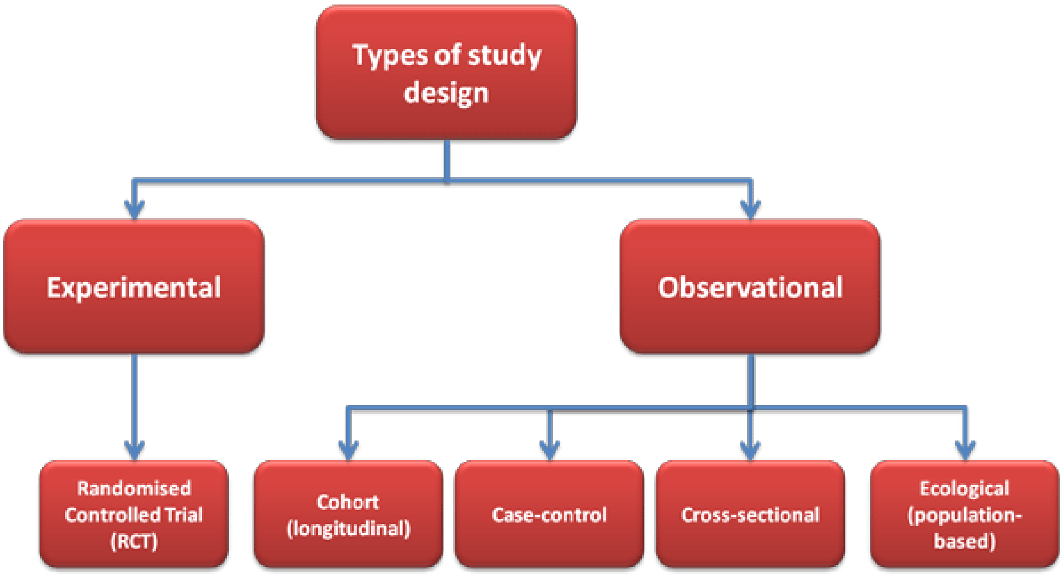


Figure 1.2: Epidemiological study designs. Adapted from (*London School of Hygiene and Tropical Medicine, 2013*).

However, oftentimes such studies are infeasible due to cost or ethical issues, so a large portion of the epidemiologic research is conducted using observational study designs the most common types being the case-control or cohort studies. In case-control studies, as their name suggests, subjects are classified into cases and controls and their risk exposure history is compared. In cohort studies, subjects are followed up examining multiple health effects of exposure. Other less common types of observational study designs include cross-sectional study designs where the relationship between exposure and disease is examined at a single point in time and ecological study designs where this relationship is compared at group rather than individual level. Cohort studies are typically prospective where the information is yet to be collected to answer a specific research question in mind, while case-control studies are typically retrospective where the information has already been collected not necessarily with the specific research question in mind.

The main study designs among those described above used in pharmacogenomic studies are RCT and case-control studies. However, given the high-dimensional data context of genetic studies atypical to traditional epidemiological studies, a third important study design that has been emerging in the field of pharmacogenomics is a type of prospective observational study design where DNA biobanks are linked to electronic health records. Such a study design, for instance, has been successfully applied for determining the appropriate dose of warfarin (an anticoagulant) to administer to patients. (*Ramirez et al., 2012*) Pharmacogenomic studies have been carried out for a wide variety of diseases. The focus of this thesis is on Multiple Sclerosis.

1.5 Multiple Sclerosis

Multiple Sclerosis (MS) was first characterized in 1868 by a French neurologist, Jean-Martin Charcot. (*Charcot, 1868*) MS is a chronic inflammatory disease of the central nervous system (CNS). Healthy nerve fibers, or axons, are surrounded with a protective covering, the myelin sheath. Myelin is a material that is primarily comprised of protein and fat and is essential for the proper functioning of the nervous system. When a loss of myelin occurs, referred to as demyelination, the functions of the implicated nerve fiber are jeopardized. Often times, the damage to myelin is reversible.

In MS, the body’s own immune system attacks the nervous system resulting in inflammation causing demyelination in many areas leaving scars (sclerosis). This may eventually result in deterioration to the nerves themselves which, however, is not reversible. Depending on the amount of damage and the nerves that are affected, the range of symptoms experienced by individuals varies from mild (sensory troubles) to severe (handicap).

MS affects two to three times more females than males and this trend is observed for other autoimmune diseases such as rheumatoid arthritis (RA). It is not clear what causes this gender difference. Most patients are diagnosed between the ages of 20 and 40 years with the peak disease onset occurring around the age of 30. MS is also diagnosed in children (younger than 18 years) and in seniors (more than 65 years).

The diagnosis of MS is very difficult and misdiagnosis can be quite common. In 2001, an international panel of neurologists derived diagnostic criteria for MS (*McDonald et al., 2001)*. These criteria were later revised in 2005 and again in 2010 (*Polman et al., 2011; Polman et al., 2005*) but, unfortunately, they remain imperfect.

The criteria are primarily based, but not limited to, the clinical presentation of at least one attack. An, attack, also referred to as flair, relapse, or exacerbation, is defined as "*patient-reported or objectively observed events typical of an acute inflammatory demyelinating event in the CNS, current or historical, with duration of at least 24 hours, in the absence of fever or infection*." (*Polman et al., 2011*) Additional data needed for MS diagnosis include radiological measures such as the presence of magnetic resonance imaging (MRI) T2 and/or gadolinium-enhancing (GD+) lesions in MS-typical regions of the CNS.

The clinical course of MS is highly heterogeneous. Many forms of the disease have been described but they can generally be grouped into four distinct types characterized by the disease progression. **Figure 1.3** below illustrates these types: relapsing-remitting MS (RRMS, panel (**A**)), progressive-relapsing MS (PRMS, panel (**B**)), secondary-progressive MS (SPMS, panel (**C**)) and primary-progressive MS (PPMS, panel (**D**)). Each figure represents disability progression on the y-axis versus time on the x-axis; each peak corresponds to an attack.

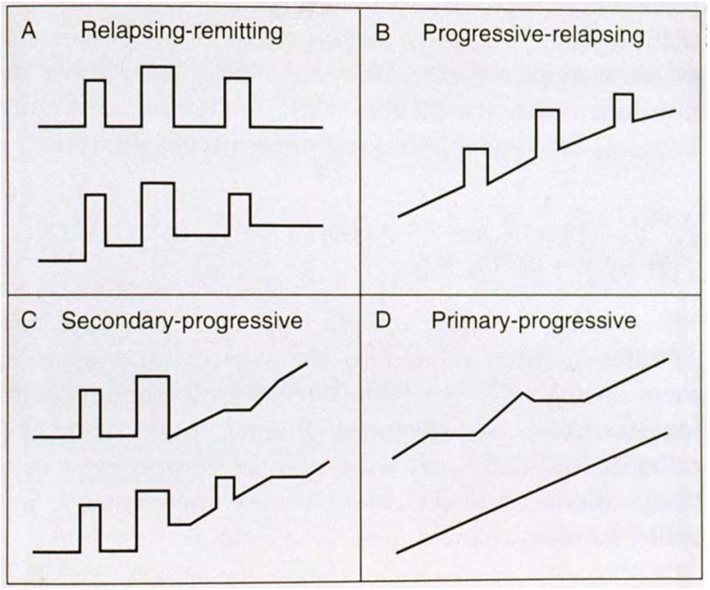


Figure 1.3: MS disease progression may be grouped into four major categories: relapsing-remitting MS (**A**), progressive-relapsing MS (**B**), secondary-progressive MS (**C**) and primary-progressive MS (**D**). Adapted from *(MS Society of Western Australia, 2013*).

The most common type of MS is RRMS with roughly 85% of MS patients suffering from it. Patients alternate between periods of attacks (relapses) followed by periods of partial or full recovery (remission). As the disease progresses, the partial recovery accumulates into disability eventually leading to the SPMS form where the worsening of the disease course continues. About 10% of MS patients have the PPMS where, contrary to the SPMS form, there is a steady progression of disability right from disease onset without periods of full recovery. Finally, in the rare type of disease progression affecting roughly 5% of the patients, PRMS, patients experience recurring relapses (attacks) and steady worsening of symptoms. (*Goldenberg, 2012*) It is also possible, however, that for some patients the disease would not progress.

To date, the cause of MS remains unknown but a number of environmental risk factors have been linked to MS. These include infectious factors such as Epstein-Barr virus infection and non-infectious factors such as vitamin D deficiency (*Ascherio and Munger, 2007a, b*). None, however, provide a definite explanation and, in fact, the list of plausible causes continues to grow such as a recent study using experimental animal models suggesting that high salt intake may, too, be linked to risk of MS. (*Kleinewietfeld et al., 2013*)

The work by *Kurtzke (2000)* illustrated that there are also geographic clues to the cause of MS. **Figure 1.4** shows the prevalence of MS around the world. The prevalence is highest in the USA, Canada and Northern Europe (more than 100 per 100 000), followed by Australia, New Zealand, Southern Europe, Russia and Latin America (between 5 and 100 per 100 000) and is rather low in Asia and Africa (less than 5 per 100 000). Variability within country exists as well. The prevalence in France ranges from 60 to 100 per 100 000 with higher prevalence in the north-eastern regions and lower in the Paris region and the south-western regions. (*Fromont et al., 2010; Vukusic et al., 2007*).

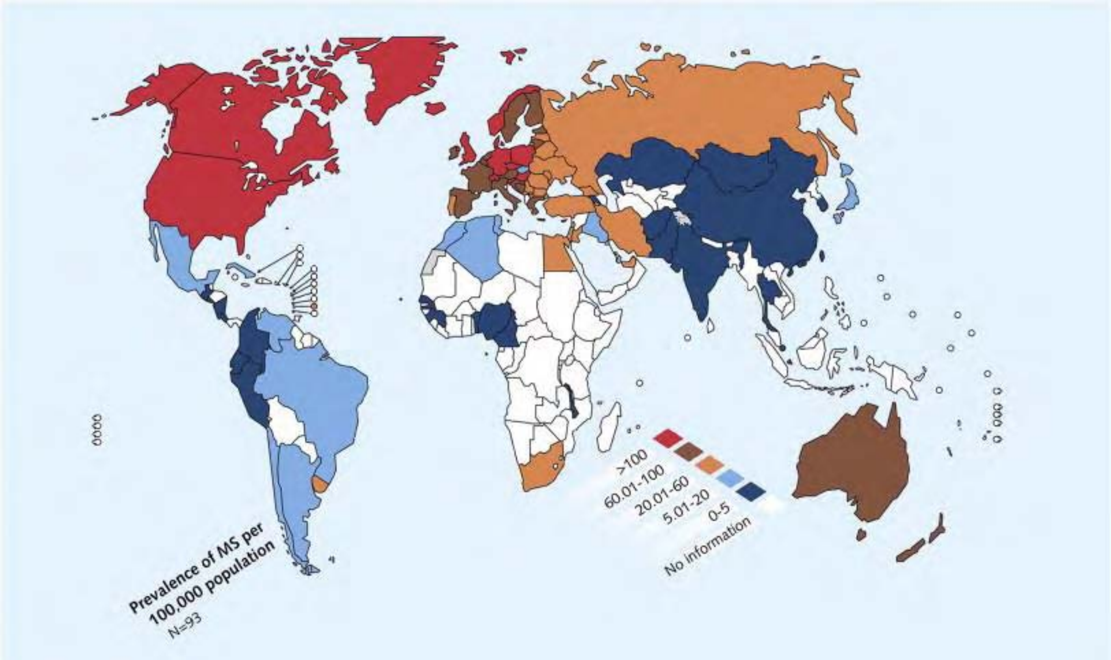


Figure 1.4: World atlas of MS prevalence. Adapted from (*World Health Organization, 2008*).

1.6 Multiple Sclerosis Genetics

The prevalence among Native Indians in Canada as well as other ethnic communities across the world is lower than the corresponding national prevalence. This suggests that genetic risk factors also appear to contribute to the risk of MS (*Rosati, 2001*). Moreover, familial studies have indicated that MS tends to aggregate in families and the risk tends to decrease with decreasing degree of relatedness. For instance, one study in a Northern European population (with MS prevalence > 0.1%, see **Figure 1.4**) estimated the age-adjusted lifetime risk at 38% for monozygotic (identical) twins and at 3-5% for dizygotic (fraternal) twins and first degree relatives such as a sibling or a child. Conversely, the risk for non-biological relatives such as adopted siblings was the same as the risk in the general population (that is, the prevalence) estimated at 0.2% in their study. (*Sadovnick et al., 1999*)

Familial recurrence risk is often measured by the sibling recurrence-risk ratio, denoted λ_S, which is the ratio of risk in siblings of affected individuals to the risk in the general population. In the study by *Sadovnick et al. (1999)* mentioned above, $\lambda_S \in (15, 25)$, that is, 3%/0.2% to 5%/0.2%. Overall, estimates of λ_S in MS vary across studies and have tended to decline over time (*Sawcer et al., 2010)* with some studies suggesting that $\lambda_S < 10$ (*Hemminki et al., 2009*).

Twin studies, aiming to evaluate the relative contribution of genetic and environmental risk factors, have also produced highly variable estimates of the genetic contribution to MS susceptibility ranging from 25% to 76%. (*Hawkes and Macgregor, 2009*) Therefore, while there is supporting evidence for a genetic component of the disease, the fine balance between the contributing genetic and environmental factors to the risk of MS is still unclear.

Prior to the era of large scale genetic association studies (genome-wide association study, GWAS), the only recognized genetic association contributing to the risk of MS was mapped to the HLA region. The first GWAS of MS in 2007 in 931 family trios (discovery dataset) and 609 family trios as well as 2322 cases/789 controls (replication dataset) confirmed this association and identified two other genes, *IL2RA* and *IL7RA* at strict genome-wide significance criteria levels (< 10^{-7}). (*Hafler et al., 2007)* In the four years that followed, six independent GWASs were conducted identifying over 20 different loci outside the HLA region (most of them at genome-wide significance). (*(Comabella et al., 2008)*; *(Baranzini et al., 2009)*; *(Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), 2009)*; *(Jakkula et al., 2010)*; *(Sanna et al., 2010)*; *(Nischwitz et al., 2010))* In

2011, the International Multiple Sclerosis Genetics Consortium in collaboration with the Wellcome Trust Case Control Consortium 2 completed the largest MS GWAS to date (roughly 9800 cases and 17400 controls) replicating almost all of the previously identified associations and further identifying 29 novel ones. (*Sawcer et al., 2011*) In 2012, another much smaller MS GWAS (296 cases and 801 controls) was conducted replicating previously reported associations (*Matesanz et al., 2012*). Many of the genetic association findings were close to immunologically relevant genes thus providing basis to the belief that MS is an immunological disorder. Despite all the discovered variants, however, a large portion of the heritability of MS risk remains unexplained. In a dataset of roughly 2 000 MS cases and 5 000 controls with close to 500 000 SNPs, *Watson et al. (2012)* found that approximately 30% of MS heritability was explained by the variants on current genome-wide SNP arrays, which includes the SNPs in the HLA region that alone account for ~8%.

1.7 Multiple Sclerosis Therapies

There is no cure for MS but currently there are eight approved disease modifying therapies on the market. In France, the escalation approach to treatment of MS, illustrated in **Figure 1.5**, is used whereby therapies with increasing effectiveness but also more severe side effects are sequentially used.

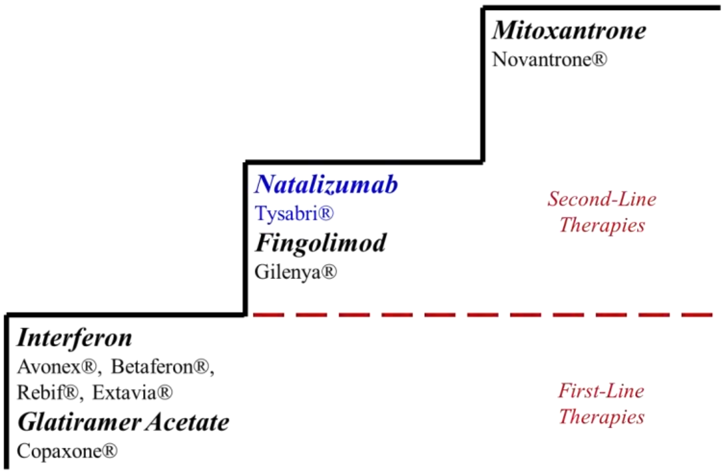


Figure 1.5: Escalation approach to MS treatment.

The first-line therapies include treatments based on interferon-β (Avonex®, Biogen Idec; Betaseron®, Bayer; Rebif®, Pfizer; Extavia®, Novartis) and glatiramer acetate (Copaxone®, Teva). Second-line therapies include natalizumab (Tysabri®, Biogen Idec), finglomod (Gilenya®; Novartis) and mitoxantrone (Novantrone®, EMD Serono). Interferon-β is the first ever therapy approved for MS dating back to the mid-90s. Fingolimod is the first oral treatment in MS and is the most recent drug approved by the European Medicines Agency.

First line therapies cause relatively mild side effects but are also less effective. Approximately half of the patients fail to respond to interferon therapy. On the other hand, second-line therapies, as more aggressive treatments, have shown to be more effective in modifying the disease course for patients but can also lead to serious and sometimes fatal side effects. Natalizumab is arguably the most effective treatment of all but unfortunately it has been linked with potentially fatal brain infection known as progressive multifocal leukoencephalopathy (PML). In fact, during less than a decade of its existence, this treatment has undergone an exceptional course, being withdrawn months after being approved only to be put back on the market under unprecedented surveillance program *(Steinman, 2005*).

The available variety of disease modifying therapies and the heterogeneous course of the disease progression pose challenges in determining the optimal treatment strategy for a MS patient. A number of treatment algorithms for MS have been proposed and have been reviewed by *Rio (2011)*. The importance of treating MS as early in the disease onset as possible has long been recognized but the poor efficacy of the available drugs seems to be related to the fact that only the early stages of the disease are targeted (*Lopez-Diego and Weiner, 2008*). Effectively, treatments of MS may be considered for treating its clinical manifestations, managing its symptoms or preventing its progression. (*Fox et al., 2006*)

There are currently at least seven medications in late phases of drug development primarily targeting the most common form of the disease, RRMS. (*Ali et al., 2013*) By 2020 the number of approved MS therapies is expected to rise significantly making the decision for the best course of treatment ever more challenging (*Huynh, 2010*). The increasing number of available therapies coupled with the potential risk of treatment failure and/or severe adverse reactions makes individualized therapy a necessity for MS. (*Río et al., 2009*)

1.8 Response Definition to Multiple Sclerosis Therapies

There is not a widely accepted definition of response to treatment in MS. The most common approach adopted in pharmacogenomic studies of MS has been to dichotomize the group of patients into Responders/Non-Responders by evaluating their response to treatment at a specific time point (for instance, one year after treatment onset) using a set of clinical and/or radiological variables.

The criteria of grouping patients has widely differed across studies but typically a patient is classified as a Responder if *all* criteria are met and those patients not classified as Responders are classified as Non-Responders (*at least one* criterion is not met). Sometimes, however, the Non-Responder group is as strictly defined (*none* of the criteria is met) as the Responder group (extreme phenotypes) leaving perhaps many patients classified as Intermediary (Suboptimal) Responders.

The number of relapses (see Section 1.5) experienced by the patient over the treatment period is commonly used to evaluate response. Another widely used clinical measure to evaluate treatment response in MS is the disability progression based on the change of the Expanded Disability Status Scale (EDSS) developed by *Kurtzke (1983)*. It is a rating scale from 0.0 (normal neurological exam) to 10.0 (death due to MS). Starting from 1.0, it goes in increments of 0.5. Despite its popularity, this measure is complicated to use and understand. While large discrepancies are unlikely, two physicians evaluating the same patient may assign different EDSS values.

For the relapsing forms of MS, response has often been assessed over the treatment period by combining these two measures (relapses and EDSS) as illustrated in **Figure 1.6** below. Sometimes one or both of these measures has been combined with radiological measures such as the presence/absence of MRI lesions to derive the response definition.

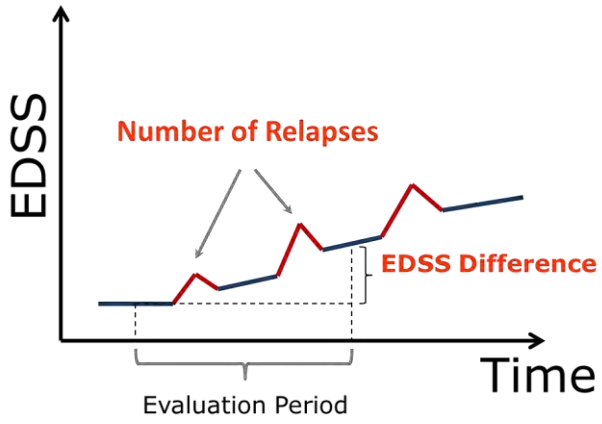


Figure 1.6: Assessing treatment response in relapsing forms of MS based on EDSS progression and number of relapses experienced over the evaluation period.

While the response classifications are usually based on the same clinical and/or radiological measures, response criteria across studies differ for the following reasons: (1) whether both clinical and radiological or only clinical measures are used to classify the patients; (2) for the same measure the threshold used to distinguish a Responder from a Non-Responder; (3) the duration of the period over which the response is evaluated (six months, one year, two years, etc.). Moreover, even if the response classification is the same across studies, the evaluation of the measures on which it is based is physician-dependent and thus subjective.

Alternatively, the response variable could be constructed as a composite score of the several variables used to classify the patients into Responders and Non-Responders. The challenges in deriving such definition of response are (1) how to build a composite score using variables measured on completely different scales (for example, EDSS versus MRI lesion load); (2) how to determine the weight that each of these variables carry in the score; and (3) the distribution of such a composite score is yet to be evaluated and its properties are yet to be established.

For the second challenge, an added complexity to defining response of MS lies in the fact that the variables used to evaluate response vary by importance with the disease duration. For instance, in the beginning of the disease, radiological measures might be the best tool to evaluate disease activity as there may be no clinical manifestations of the disease. On the other hand, as time progresses, the role of the radiological measures reduces and the disease begins to manifest clinically. *(Fox and Cohen, 2001*) Thus, time-dependent weights may need to be assigned to reflect the increasing or diminishing role of each measure in determining response to treatment in MS.

Yet another approach to defining response may rely on the principles of survival analysis. In this case, one can model the time to progression of the disease during treatment by defining progression based on one or several criteria (clinical and/or radiological). This approach is commonly used in randomized clinical trials, for example, modeling time to first relapse after treatment onset.

Thus, response to MS therapies is a highly complex outcome to evaluate. As a contrasting example, to determine the warfarin dose, physicians use the international normalized ratio (INR), which is a standardized test result evaluating the clotting tendency of blood. It is an objective measure based on blood tests and is, therefore, directly comparable nationally and internationally.

It is plausible that genetic factors play a role in determining response to MS therapies. Moreover, if such is the case, it is likely that multiple genes are involved. (*Río et al., 2009*) Therefore, pharmacogenomic studies, investigating genes implicated in drug response, have the potential to identify key genetic biomarkers and facilitate the application of individualized therapy in MS.

1.9 Pharmacogenomic Studies of Multiple Sclerosis To Date

A handful of pharmacogenomic studies of MS have been carried out to date predominantly on interferon response (summarized in **Table 1.1**). Most of these studies have investigated the role of specific genes with only two studies (on interferon response) conducting genome-wide scans. Gene names are given in greater detail in Appendix I.

MS Therapy	**Study Scope**	**Response Criteria/ Evaluation Period**	**Study Sample Size (Responders / Non-Responders)**	**Study Findings**[1]	**Reference**
Interferon-β	*HLA Class II Genes*	EDSS, Relapses / 2 years	134 Spanish patients	No association	(*Villoslada et al., 2002*)
		EDSS, Relapses / 1 year	96 Spanish patients	No association	(*Fernández et al., 2005*)
	IFNAR1, *IFNAR2*	EDSS, Relapses / 2 years	147 Spanish patients	No association	(*Sriram et al., 2003*)
		EDSS, Relapses / 1 year	147 Spanish patients[2]	No association	(*Leyva et al., 2005*)
	100 interferon-stimulated response element genes	EDSS, Relapses / 6-9 months	162 Irish patients	Associated with response: *IFNAR1*, *CTSS*, *LMP7* and *MxA*	(*Cunningham et al., 2005*)
	MxA	EDSS, Relapses, MRI / 2 years	37 US patients	No association	(*Weinstock-Guttman et al., 2007*)
	IFNG	Relapses / 2 years	110 Spanish patients	Associated with response	(*Martínez et al., 2006*)
	IL28B	EDSS, Relapses / 2 years	588 patients from US, UK, France, Spain, Italy, Germany, Serbia)	No association	(*Malhotra et al., 2011*)
	TRAIL and *TRAIL* receptor genes (*TRAIL*, *TRAILR-1*, *TRAILR-2*, *TRAILR-3* and *TRAILR-4*)	EDSS, Relapses / 2 years	509 Spanish patients (discovery); 226 Spanish patients (replication)	Associated with response: *TRAILR-1*	(*López-Gómez et al., 2013*)
	GPC5 and *HAPLN1*	EDSS, Relapses / 2 years	199 Spanish patients	Association with response: *GPC5*	(*Cénit et al., 2009*)
	Genome-wide scan (≈ 100 00 0 SNPs)	EDSS, Relapses / 2 years	206 patients from Spain and France (discovery); 81new Spanish patients for combined analyses (validation)	Associated with response*: GPC5*, *COL25A1*, *HAPLN1*, *CAST* and *NPAS3* and several SNPs in intergenic regions	(*Byun et al., 2008*)
	Genome-wide scan (≈ 430 000 SNPs)	EDSS, Relapses / 2 years	106 Spanish patients (discovery); 94 Spanish patients (validation)	Association with response: *GRIA3*, *CIT*, *ADAR*, *ZFAT*, *STARD13*, *ZFHX4*, *IFNAR2* and for 11 SNPs in intergenic regions	(*Comabella et al., 2009*)
Glatiramer Acetate	*HLA Class II Genes*	EDSS, Relapses / 2 years	44 Italian patients	Association with response	(*Fusco et al., 2001*)
	27 genes	EDSS, Relapses, MRI / 9 months; 2 years	101 patients from Europe, Canada and US (fractional cohorts from clinical trials)	Association with response: *CTSS* (corrected for multiple testing); *MBP*, *CD86*, *FAS*, *IL1R1* and *IL12RB2*	(*Grossman et al., 2007*)
Natalizumab	N/A				
Fingolimod	N/A				
Mitoxantrone	*ABC*-transporter genes	EDSS, Relapses, MRI / 9-12 months	309 Spanish and German patients	Association with response	(*Cotte et al., 2009*)

[1] Reported association findings at α=0.05 significance level. [2] Different from the study by (*Sriram et al., 2003*)

Table 1.1: List of pharmacogenomic studies carried out to date.

First-Line Therapies. Given the long established link between the HLA locus and susceptibility to MS, studies have also investigated the role of this locus in interferon response. *Villoslada et al. (2002)* and *Fernández et al.(2005)* found no association with interferon response, while *Cunningham et al.(2005)* identified an association with *LMP7*, a gene located within the HLA region.

Several studies investigated the role of interferon-α/β receptor genes, *IFNAR1* and *IFNAR2* and found no association with response. ((Sriram et al., 2003); (Leyva et al., 2005); (Cunningham et al., 2005)) A role for the *IFNAR1* gene was suggested by the study of *Sriram et al. (2003)* when response was evaluated based only on the number of relapses and by the study of *Cunningham et al. (2005)* studying a different polymorphism in that gene. The *IFNAR* genes are located in the vicinity of interferon-γ receptor genes (*IFNGR*). *Martínez et al. (2006)* found a polymorphism in the interferon-γ gene, *IFNG*, to be associated with interferon response.

All in all, *Cunningham et al. (2005)* investigated 100 interferon-stimulated response element genes. Apart from the genes already mentioned, that is, *LMP7* and *IFNAR1*, two more were found to be associated with response to interferon, *CTSS* and *MxA*. *Weinstock-Guttman et al. (2007)* evaluated the association of two SNPs from the *MxA* gene and found no relation.

Malhotra et al. (2011) investigated the role of two polymorphisms in the *IL28B* gene in interferon response but did not find an association. One of the 100 genes included in the study by *Cunningham et al. (2005)* was the *TRAIL* gene (with which no association was found). A recent study focused on *TRAIL* and *TRAIL* receptor genes (*TRAIL*, *TRAILR-1*, *TRAILR-2*, *TRAILR-3* and *TRAILR-4*) was reported suggesting a role for the *TRAILR-1* gene in interferon response. (*López-Gómez et al., 2013*)

In 2008, *Byun et al. (2008)* reported the first GWAS of response to interferon in a small cohort of 206 MS patients. Their study did not identify any of the previously reported associations (*IFNAR1, LMP7, CTSS, MxA, IFNG*). However, their study found significant (at nominal level) associations for SNPs in or close to several novel genes including *GPC5*, *COL25A1*, *HAPLN1*, *CAST* and *NPAS3* and several SNPs in intergenic regions. A study by *Cénit et al. (2009)* aimed to replicate the association findings for the two most strongly implicated genes, *GPC5* and *HAPLN1*, and was able to confirm the association only with the *GPC5* gene. A second GWAS followed in 2009 reporting significant associations for SNPs in or close to seven genes, *GRIA3*, *CIT*, *ADAR*, *ZFAT*, *STARD13*, *ZFHX4*, *IFNAR2* and for 11 SNPs in intergenic regions. (*Comabella et al., 2009*)

For glatiramer acetate, only two candidate gene studies have been conducted. *Fusco et al. (2001)* assessed the relationship between HLA and response to glatiramer acetate and, contrary to interferon therapy, results suggested that it was implicated in response to glatiramer acetate. *Grossman et al. (2007)* studied 27 candidate genes and found significant association with the *CTSS* gene after correcting for multiple testing. Nominally significant associations were reported with the following five other genes, *MBP*, *CD86*, *FAS*, *IL1R1* and *IL12RB2*.

Second-Line Therapies. To our knowledge, no genetic association study has been conducted to evaluate natalizumab or fingolimod response in MS patients. One study suggested a role for *ABC*-transporter genes in mitoxantrone response. (*Cotte et al., 2009*)

In summary, the study sample sizes were small and the response definitions differed across studies. Most of the reported association findings were weak and have not been validated in independent datasets. Therefore, these studies do not provide a clear indication of whether genetic factors influence response in MS for the investigated therapies, namely, interferon, glatiramer acetate and mitoxantrone, while this possibility is yet to be explored for natalizumab and fingolimod.

1.10 Thesis Objective

A pharmacogenomic study design needs to include two essential phases: identification and validation. In the first phase of identification, the association between genes and drug response is initially assessed. In other words, a genetic association study is conducted. If one or more genes are found to be associated with drug response, these findings then need to be validated in an independent dataset and their clinical utility in actually predicting drug response needs to be evaluated. In other words, a genetic prediction study is conducted.

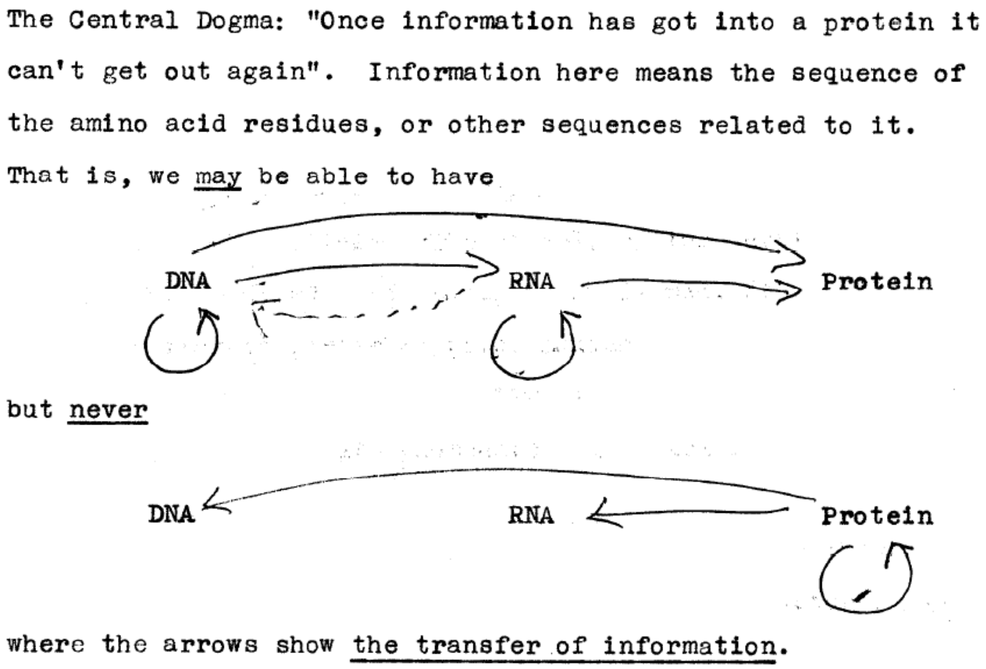
In this dissertation work, we review the basic methodological aspects behind genetic association and genetic prediction studies. To begin, we give a brief account of the multinational effort that laid the foundations to (large-scale) genetic association studies. Next, we describe these studies and present our findings from the studies we conducted. Specifically, we investigated the role of the *OAS1* gene in response to interferon-β in a multinational MS patient cohort and the role of the *ITGA4* gene in response to natalizumab in a subset of French MS patients from the BIONAT cohort. (*Outteryck et al., 2013*)

We next describe genetic prediction studies and also present findings from the prediction studies we conducted. Here, we focused on investigating the role of non-genetic factors in predicting response to natalizumab in a subset of patients from the BIONAT cohort. We note that genetic data for these patients are being generated at the time of writing this manuscript and will soon be available for analysis. Further, using a simulated dataset provided by the Genetic Analysis Workshop 18 (GAW18) on a continuous trait (diastolic blood pressure) we compared two alternative methods for detecting genetic associations and evaluated the predictive performance of each.

We conclude by contrasting genetic studies of MS susceptibility to those of response to therapy in MS with the aim of providing a perspective on the feasibility of pharmacogenomics of MS. We also briefly discuss important economic, social, legal and ethical considerations that we believe concern us all as long as genetic testing has been, is or will be used to guide clinical decision making.

2 LAYING THE FOUNDATIONS

The “central dogma of molecular biology”, an (incorrect) term coined in 1956 by Francis Crick (*Crick, 1956*), describes the flow of genetic information in a living organism (see **Figure 2.1**).



The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have

DNA
RNA
Protein

but never

DNA
RNA
Protein

where the arrows show the transfer of information.

Figure 2.1: The flow of genetic information as described by Francis Crick taken from an early draft of the original article published in 1958 ((Crick, 1956); (Crick, 1958)).

Principally, it is a unidirectional flow of information where DNA is transcribed to RNA which is then translated into a protein of a specific function. In some special cases, information can flow in the reverse direction from RNA to DNA but never from protein to RNA or from protein to DNA. Crick’s representation of the flow of genetic information has been contradicted by experimental work over the years (*Shapiro, 2009*). While Crick had later acknowledged that the use of the term “dogma” (referring to a belief that cannot be doubted) was incorrect and he had meant it more as a “hypothesis” (*Crick, 1990*), it is without question that the concept he put forward remains of fundamental importance in molecular biology.

Adopting this simplistic flow of genetic information for illustrative purposes, in **Figure 2.2** we present the type of analytical approaches performed depending on the unit of analysis at each level. In particular, genomics studies the structure and the function of the DNA. In turn, transcriptomics measures the expression level of RNA in a given cell population. Lastly, proteomics studies the functions and the structures of proteins.

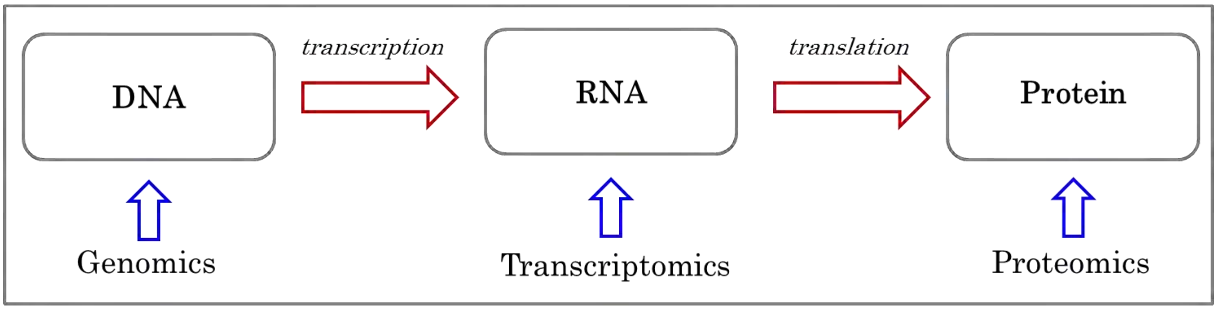


Figure 2.2: The flow of genetic information and the type of analytical approaches performed depending on the unit of analysis at each level.

Additional "-omics" terms can be added to the list depending on the unit of analysis such as metabolomics, epigenomics, glycomics, lipidomics and others. In the era of this omics data revolution, it is an increasing challenge to adopt an integrative approach to analyzing this omics information that may ultimately lead us toward personalized medicine. For instance, *Chen et al. (2012)* conducted the first-ever integrative personal omics profile analysis of a healthy individual revealing the subject's risk for various diseases such as Type 2 diabetes.

Of the omics terms, our focus in this thesis is on genomics and, specifically, on the human genome.

2.1 Understanding the Human Genome

The main goals of the Human Genome Project (HGP) were to determine the sequence of the 3 billion DNA base pairs that make up the human genome and to identify and map all of the estimated 20 500 genes. It was the largest international collaborative effort ever undertaken in biomedical research coordinated by the US Department of Energy with participants from all over the world including United Kingdom, Japan, France, Germany and China. An initial draft of the genome was reported in 2001 (*Lander et al., 2001*) and the final sequence completed in 2003 *(International Human Genome Sequencing Consortium, 2004*), 13 years after the project's initiation at a cost of more than 3 billion US dollars. Many believe that the private quest by former National Institutes of Health (NIH) scientist J. Craig Venter and his company, Celera Genomics®, to sequence the human genome challenged and accelerated the work. In 2001, only three years after commencing on his project, Venter also reported an initial draft of the human genome using samples from geographically diverse individuals as well as his, and employing a different sequencing technique at a fraction of the cost of the publicly funded project (*Venter et al., 2001*).

A large number of donors contributed blood and sperm samples to HGP only few of which were processed for DNA sequencing. Neither the researches nor the donors knew whose DNA was sequenced. In fact, the derived reference sequence did not represent a specific individual's genome but was rather based on a composite of DNA from several donors. This "representative" sequence of the human genome is freely available in public databases.

Many genetic variations were also identified during the HGP and, in October 2002, another international collaboration gave birth to the International HapMap Project (*International HapMap Consortium, 2003*). Its primary objective was to develop a haplotype map of the human genome describing the most common patterns of human DNA sequence variation with the minimum number of SNPs. By definition, a haplotype is a combination of alleles at nearby SNPs.

The project has had so far three phases. In Phase I, the haplotype map of the human genome was derived from 270 individuals from Africa, Europe, China and Japan consisting of approximately 1.3 million SNPs (*International HapMap Consortium, 2005*). In Phase II, over 3.1 million SNPs were genotyped in the same individuals (*Frazer et al., 2007*). In Phase III, the number of DNA samples was increased from 270 to 1301 obtained from more geographically diverse human populations (*Altshuler et al., 2010*).

The first complete genome sequence of a single individual (J. Craig Venter's) was published in 2007 (*Levy et al., 2007*) followed by that of James D. Watson (*Wheeler et al., 2008*) in 2008. That year, the genome sequences of two anonymous individuals from Asian and African descent were also published ((Wang et al., 2008); (Bentley et al., 2008)). Early 2008, yet another large international effort was put forth and the 1000 Genomes Project was launched the goal of which was to sequence the genomes of at least 1000 individuals from genetically diverse backgrounds. The pilot phase of the project was completed in 2010 (*Abecasis et al., 2010*) and the sequence of the genomes of 1092 anonymous individuals was published in late 2012 (*Abecasis et al., 2012*). Both the HapMap and the 1000 Genomes project databases are continuously used.

All these large-scale collaborative efforts certainly fostered biomedical research but, most importantly, built the foundation for investigating the role of genetic variants in human health and disease at the genome-wide level.

2.2 Genetic Variations

Types of genetic variants include SNPs and structural variants such as insertions-deletions, block substitutions, etc., as illustrated in **Figure 2.3** below.

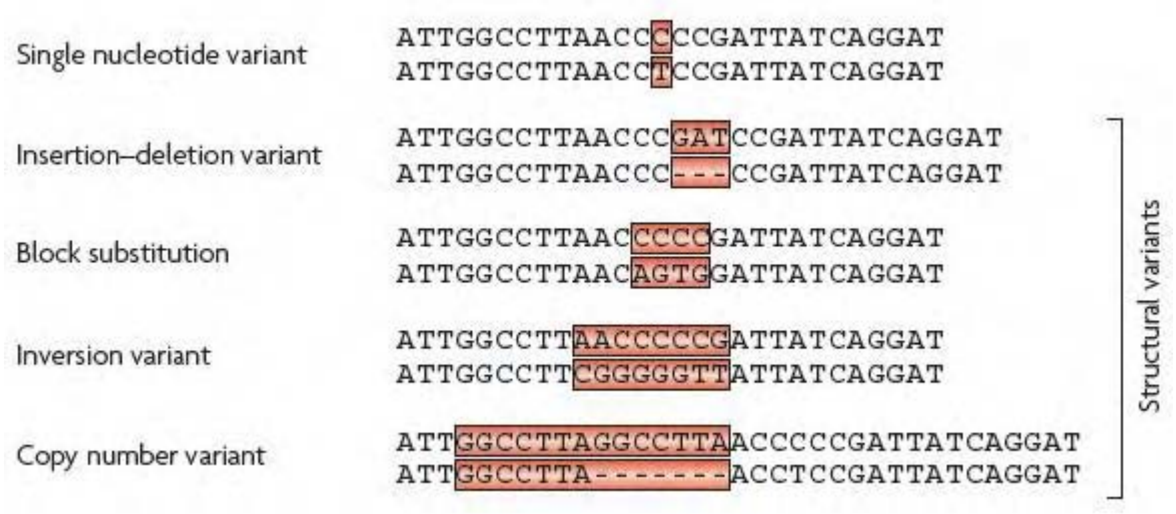

Single nucleotide variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
Insertion–deletion variant
ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT
Block substitution
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT
Inversion variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGGTTATTATCAGGAT
Copy number variant
ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA------ACCTCCGATTATCAGGAT
Structural variants

Figure 2.3: Types of human genetic variations. Adapted from (*Frazer et al., 2009*).

We focus here on the most common and simplest form of genetic variation, the SNP. The Single Nucleotide Polymorphism Database, dbSNP, is a central public repository for genetic variation (not only of SNPs unlike its name is suggesting) in humans and other species created and maintained by the US National Center for Biotechnology Information (NCBI) since 1998 (*Sherry et al., 2001*).

A SNP represents a difference in a single nucleotide base. SNPs occur on average every 300 bases giving rise to roughly 10 million SNPs in the human genome. A SNP can have up to four possible alleles (A, C, G or T) but most SNPs are bi-allelic (only two alleles are present in the population). Each individual carries two SNP alleles one for each copy of the chromosome forming the *genotype* at that single position. For instance, if the two SNP alleles are A and T, then the genotypes would be A/A, A/T and T/T. When the alleles on both chromosome copies are identical (that is, genotypes A/A or T/T), the individual is said to be *homozygous* at that locus; otherwise the individual is said to be *heterozygous* (that is, genotype A/T). The allele with the lower frequency in the population is referred to as the *minor allele*. So far, mostly *common* or *low-frequency* SNPs have been catalogued, with minor allele frequency (MAF) of at least 1%. Recent advances in sequencing technologies enable the analysis of *rare* variants with MAF < 1% (*Cirulli and Goldstein, 2010*).

2.3 Linkage Equilibrium/Disequilibrium

Obtaining fine maps of the genetic variations have not only allowed the determination of the patterns of common variation and their frequencies but have also made possible the determination of the complex correlation structures that exists between SNPs. This concept, first introduced in 1960 by Lewontin and Kojima (*Lewontin and Kojima, 1960*) and unfortunately termed Linkage Disequilibrium (LD), simply reflects the nonrandom association of alleles at two or more loci. It is a misleading term because loci that are in disequilibrium may not necessarily be linked (that is, physically close) and also loci that are linked may not necessarily be in disequilibrium. The original definition of LD referred to the non-random association between two or more loci from possibly *different* chromosomes but lately the term has been used to refer to the non-random association between two or more loci on the *same* chromosome (*Slatkin, 2008*).

Several metrics exist to measure LD. We briefly discuss the most commonly used ones here. Suppose, without loss of generality, that there are two bi-allelic SNPs with alleles A/a and B/b, respectively, resulting in the following four haplotypes: AB, Ab, aB, ab. Let *f(x)* be the frequency of *x* where *x* can refer to an allele or a haplotype. This is summarized in **Table 2.1** below.

		SNP$_2$ alleles		
		B	**b**	
SNP$_1$ alleles	**A**	*f(AB)*	*f(Ab)*	***f(A)***
	a	*f(aB)*	*f(ab)*	***f(a)***
		f(B)	***f(b)***	**1**

Table 2.1: Table of allele and haplotype frequencies for two bi-allelic SNPs, SNP$_1$ and SNP$_2$.

Then, the basic measure of LD for the pair of alleles A and B, D_{AB}, is given by

$$D_{AB} \ = \ f(AB) - f(A)f(B) \tag{2.1}$$

where D_{AB} is essentially the difference between the observed and the expected frequency (under the assumption of independence of alleles) of the haplotype AB. Since the allele frequencies at both loci have to add up to 1 and the haplotype frequencies have to add up to 1, the range of LD values is constrained. That is,

$$\begin{aligned} D_{AB} &= f(AB) - f(A)f(B) \qquad (2.2) \\ &= 1 - f(Ab) - f(aB) - f(ab) - \big(1 - f(a)\big)f(B) \\ &= 1 - f(Ab) - f(aB) - f(ab) - \big(f(B) - f(a)f(B)\big) \\ &= \ -f(aB) + f(a)f(B) + 1 - f(Ab) - f(ab) - f(B) \\ &= -f(aB) + f(a)f(B) + 1 - f(Ab) - f(ab) - f(AB) - f(aB) \\ &= \ -f(aB) + f(a)f(B) = \ -D_{aB} \end{aligned}$$

Similar derivations can be obtained for D_{Ab} and D_{ab}. In fact, the measures of LD for all pairs of alleles are related as follows

$$D_{AB} = \ - D_{Ab} = \ - D_{aB} = \ D_{ab} = D. \tag{2.3}$$

If $D = 0$, the loci are said to be in Linkage Equilibrium (LE).

While all additional metrics that have been proposed to measure LD relate to *D*, they aim to circumvent some of its limitations. For instance, when the goal is to compare the LD between different pairs of SNPs, *D* is not a good measure since its range of possible values is constrained by the specific allele frequencies. Hence, *Lewontin (1964)* defined an alternative normalized measure, D', as

$$D' = \frac{D}{D_{max}}, \tag{2.4}$$

where

$$D_{max} = \begin{cases} \min(f(A)f(B), f(a)f(b)) & if\ D < 0 \\ \min(f(A)f(b), f(a)f(B)) & if\ D > 0 \end{cases}. \tag{2.5}$$

That is, D_{max} is the maximum *D* which can be achieved given the specific allele frequencies. As such, $D' \in [-1, 1]$. When $|D'| = 1$, it is indicative of the absence of at least one of the haplotypes and the case is known as complete LD.

Another measure of LD is the correlation coefficient, r^2, between the two loci defined as follows

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}. \tag{2.6}$$

When $r^2 = 1$, exactly two of the four possible haplotypes are observed. This is known as perfect LD. In this case, knowing the genotypes at one SNP completely determines the genotypes at the other SNP. A typical LD plot with r^2 as the measure of LD is illustrated in **Figure 2.4** below. SNPs 1 and 2 are in weak LD with $r^2 = 0.02$ while SNPs 5 and 6 are in strong LD with $r^2 = 0.99$. The shading of each diamond is proportional to the r^2 value ranging from white for $r^2 = 0$ to black for $r^2 = 1.0$. (Note that a plot with D' as the measure of LD can be obtained analogously.)

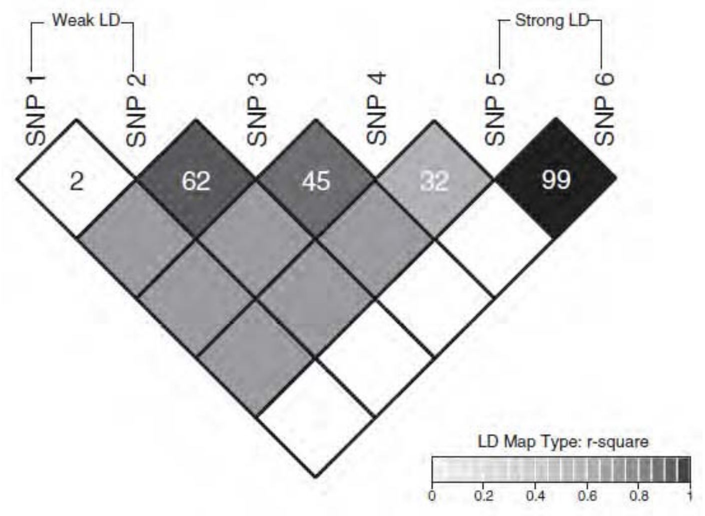


Figure 2.4: A typical LD plot with r^2 as the measure of LD. SNPs 1 and 2 are in weak LD with $r^2 = 0.02$ while SNPs 5 and 6 are in strong LD with $r^2 = 0.99$. The shading of each diamond is proportional to the r^2 value ranging from white for $r^2 = 0$ to black for $r^2 = 1.0$. Adapted from (*Lessard et al., 2012*).

LD can range over as little as few kilobases (kb, 1 kb = 1000 base pairs) to as much as 100 kb or more. (*Reich et al., 2001*) Studies on the pattern of LD have shown that it varies across the genome and across populations of different ancestry. For example, a particularly long range LD is observed at the major histocompatibility complex (MHC) on chromosome 6 over several megabases (Mb, 1Mb = 1 000 000 base pairs). Further, the LD in African populations is weaker on average than that in European or Asian populations as, for example, illustrated in **Figure 2.5** below for a region on chromosome 9.

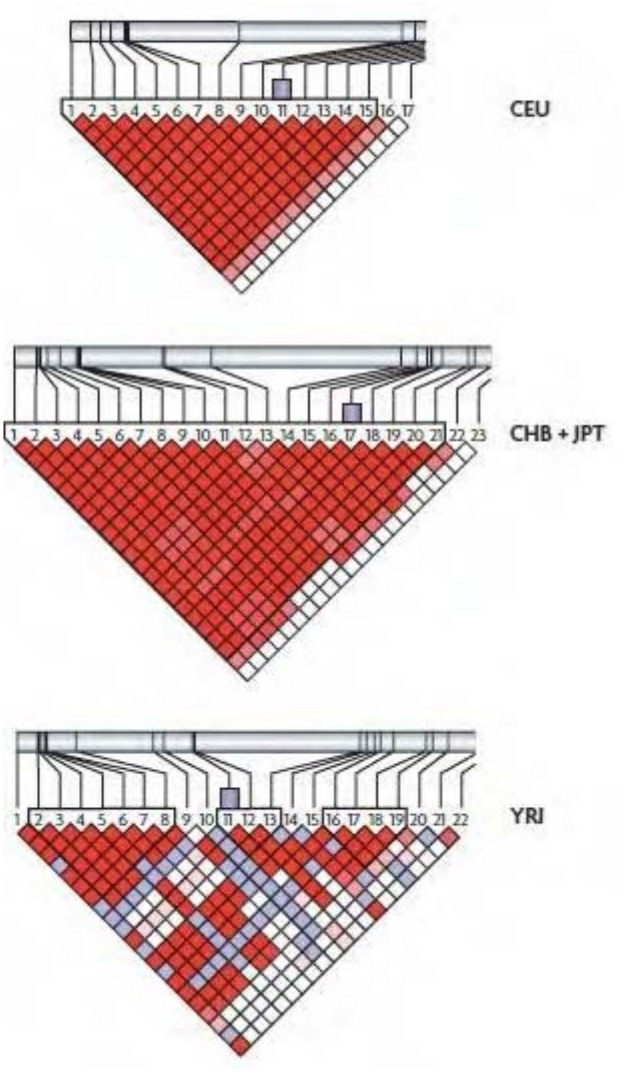


Figure 2.5: LD plots based on D' as the measure of LD for a region on chromosome 9 and for three HapMap populations: CEU (European), CHB+JPT (Asian) and YRI (African). The shading of each diamond is proportional to the D' value ranging from white for $D' = 0$ to red for $D' = 1$. Adapted from (*Frazer et al., 2009*).

The breakdown of LD is primarily driven by recombination but evolutionary forces such as mutation, genetic drift, natural selection and migration, can also influence LD.

2.3.1 Recombination

Recombination occurs during meiosis, a special type of cell division producing gametes (reproductive cells) that are genetically different from their parental types. This is illustrated in **Figure 2.6** below.

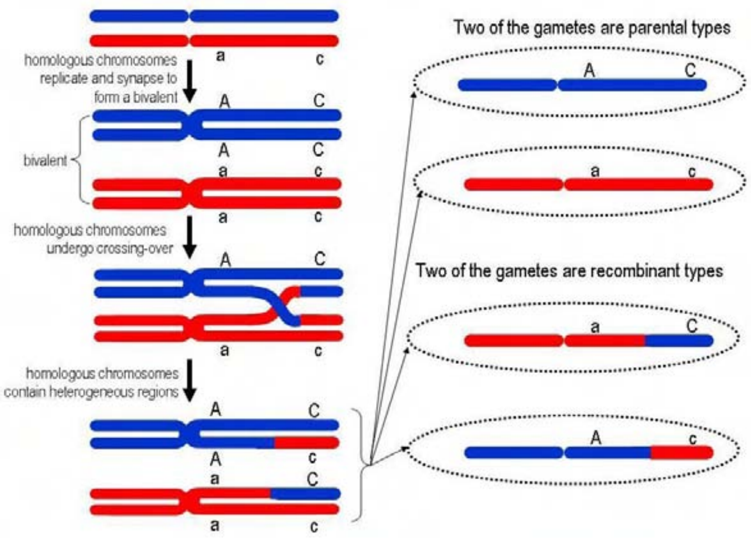

homologous chromosomes replicate and synapse to form a bivalent
bivalent
homologous chromosomes undergo crossing-over
homologous chromosomes contain heterogeneous regions
Two of the gametes are parental types
Two of the gametes are recombinant types
A
C
a
c

Figure 2.6: During meiosis, homologous chromosomes undergo crossing-over producing chromosomes containing genetically heterogeneous regions. Adapted from (*GeneticsSuite, 2013*).

The recombination frequency is given by the total number of recombinant gametes divided by the total number of all transmitted gametes. (Note that in **Figure 2.6**, the recombination frequency is 0.5.) Loci that are located close to each other tend to be inherited together (that is, they are less likely to be separated during chromosomal crossing-over). Such loci are said to be genetically linked. The unit of measure of genetic linkage between any two loci is the centimorgan (cM) and represents the *genetic* distance between the two loci. The unit, named after the Nobel laureate geneticist Thomas Hunt Morgan, refers to the distance between the two loci determined by the frequency with which recombination occurs between them. By definition, 1cM is equivalent to a recombination frequency of 1%. Note that two loci that are genetically the same distance apart may not be so physically (in terms of base pairs). In humans, 1cM corresponds, on average, to 1 million base pairs but this number varies widely across the genome.

Recombination breaks up the genomic regions over generations. Thus, the strength of LD tends to decrease with distance as well as time and this relationship is represented through the following formula

$$D_{t+1} = (1 - c)D_t, \tag{2.7}$$

where c is the recombination frequency and t is time in generations. The recombination frequency is $c = 0.5$ for unlinked loci and $c < 0.5$ for linked loci with $c \to 0$ the more linked the loci. Also, c is not uniform across the genome with some regions having a higher recombination frequency (recombination hotspots) than others. Equation (2.7) generalizes to

$$D_{t+1} = (1 - c)^t D_0 \tag{2.8}$$

where D_0 is the LD at generation 0 and D_{t+1} is the LD t generations after. From Equation (2.8) and as illustrated in **Figure 2.7**, it is easy to see that LD at closely linked loci decays at a slower rate than at loci that are far apart.

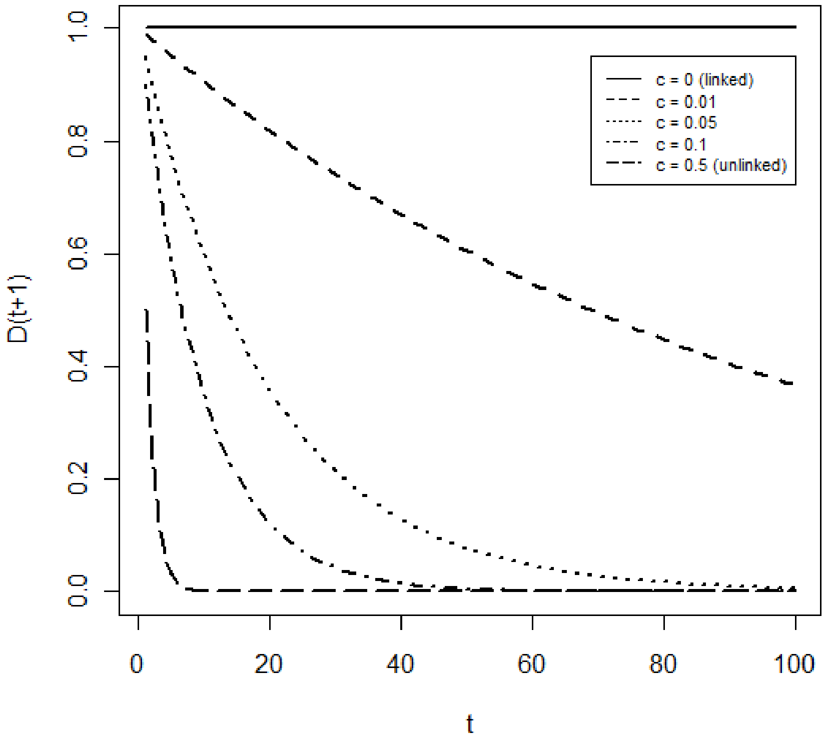


Figure 2.7: LD decay as a function of generation t and recombination frequency c.

2.3.2 Mutation, Genetic Drift, Natural Selection and Migration

The four basic mechanisms of evolutionary change are mutation, genetic drift, natural selection and migration. They lead to changes in allele and haplotype frequencies in a population and as such can influence LD.

Mutation is a change in the DNA sequence occurring in one generation and passed on to future generations. When a mutation first occurs, it creates LD with neighboring loci. Recurrent mutations, which are rare for SNPs, reduce LD.

Genetic drift is a change in allele frequencies due to random sampling of gametes in a finite population over generations. Genetic drift may lead to an eventual loss of haplotypes especially if the population is small and the haplotype is rare thus leading to increased LD.

Individuals with certain genotypes may be more likely to survive and reproduce and, thus, pass on their alleles to the next generation than individuals with other genotypes. Natural selection is the process by which allele frequencies in a population change due to these differences in survivorship and/or reproduction among individuals (genotypes). Natural selection can lead to increased or decreased LD.

Migration (or gene flow) is the transfer of alleles from one population to another. Initially, the extent of LD is proportional to the allele frequencies in each population. The larger the differences of allele frequencies between populations, the more significant the impact of migration will be on LD.

2.4 Tag SNPs

In genomic regions with high LD, the variation can be described without capturing all SNPs in the region but instead focusing on a minimal set of most informative SNPs, so called tag SNPs. This is the primary objective of the International HapMap Project that we referred to earlier in this chapter. The principle is illustrated in **Figure 2.8** below. Specifically, the panel (a) of the graph illustrates the DNA sequence of the same chromosomal region in four different individuals. These individuals differ in three nucleotide bases as shown by the colored SNPs. Panel (b) shows the haplotype for each of the four individuals composed of alleles of 20 SNPs including the three SNPs from panel (a). Panel (c) in turn shows the three SNPs that uniquely identify each of the four haplotypes.

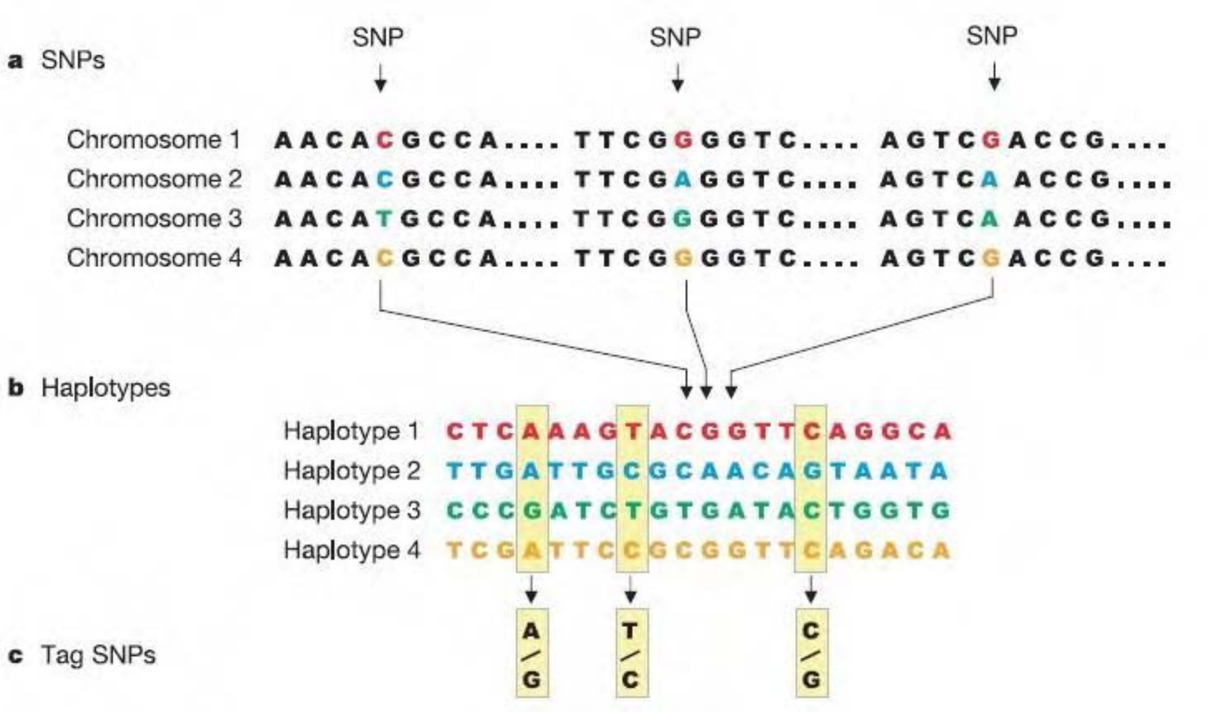

a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

Figure 2.8: Describing common patterns of human genetic variation. (**a**) DNA sequences of four individuals including three SNPs. (**b**) Haplotypes formed by nearby SNPs including the three SNPs from panel (a). (**c**) The three SNPs which uniquely identify the four haplotypes. Adapted from (*International HapMap Consortium, 2003*).

2.5 SNP Genotyping

SNP genotyping refers to the process of determining the SNP genotypes. Several genotyping platforms exist based on different technologies. The choice of platform depends on the number of SNPs to be genotyped. Microtiter (well) plates (for example, Taqman®) are typically used for small scale projects limited to few SNPs. The work by the International HapMap Project coupled with novel technological advances led to the development of microarrays or SNP chips (for example, Affimetrix® or Illumina®) enabling large-scale, whole-genome, genotyping.

2.6 SNP Chips

A common criterion to assess the quality of a SNP chip involves the evaluation of its global coverage of the genome. *Li et al. (2008)* evaluated this criterion for several commercial SNP chips available at the time for the CEU (European), CHB+JPT (Asian) and YRI (African) HapMap populations. The percentage of the genome covered by each evaluated SNP chip and for each population is given in **Table 2.2** below.

Company	SNP Chip	Number of SNPs	CEU (%)	CHB+JPT (%)	YRI (%)
Affimetrix®	SNP Array 5.0	500 568	64	66	41
	SNP Array 6.0	934 968	83	84	62
Illumina®	HumanHap300	317 511	77	66	29
	HumanHap550	555 352	87	83	50
	HumanHap650Y	660 917	87	84	60
	Human1M	1 072 820	93	92	68

Table 2.2: Global coverage of the genome for several commercial SNP chips for the CEU (European), CHB+JPT (Asian) and YRI (African) HapMap populations. Adapted from (*Li et al., 2008*).

Thus, from **Table 2.2** above, it can be seen that the global coverage of the genome depends on the number of SNPs on the chip and on the extent of LD in the population. As expected, a higher number of SNPs leads to higher coverage and fewer SNPs are needed to achieve the same coverage in populations with higher LD (for instance, CEU versus YRI, see **Figure 2.5**).

2.7 Testing for Hardy-Weinberg Equilibrium

Genotyping errors may be detected by Hardy-Weinberg Equilibrium (HWE) testing. It is an essential quality control step in genetic association studies.

The HWE states that allele and genotype frequencies in a population will remain constant from generation to generation under the assumption of random mating and in the absence of evolutionary forces (mutation, genetic drift, natural selection and migration). This principle is named after Godfrey Harold Hardy and Wilhelm Weinberg who developed it independently in 1908. (*Hardy, 1908*;*Weinberg, 1908*)

If the allele frequencies in one generation are given by $f(A) = p$ and $f(a) = q$, then the expected genotype frequencies in the next generation are $f(AA) = p^2$ for the A/A homozygote, $f(Aa) = 2pq$ for the A/a heterozygote and $f(aa) = q^2$ for the a/a homozygote as illustrated in **Table 2.3** below.

		Paternal	
		A (p)	**a (q)**
Maternal	**A (p)**	AA (p^2)	Aa (pq)
	a (q)	Aa (qp)	aa (q^2)

Table 2.3: Punnet square giving the probabilities of an offspring having a particular genotype at a bi-allelic locus in a population in Hardy-Weinberg Equilibrium.

Thus, the HWE principle relates the allelic and genotypic frequencies. The principle is presented here on a single bi-allelic locus but may be extended to loci with multiple alleles and also to multiple loci (*Hastings, 2001*).

The principle provides theoretical genotype frequencies against which the observed frequencies in a population can be compared. **Table 2.4** below summarizes the observed and the expected genotype counts under HWE in a population of size n for a single bi-allelic locus.

Genotype	AA	Aa	aa
Observed Counts	n_{AA}	n_{Aa}	n_{aa}
Expected Counts	p^2n	$2pqn$	q^2n

Table 2.4: Observed and expected (under HWE) genotype counts at a single bi-allelic locus in a population of size n.

We test for deviations from HWE in a population though the χ^2 goodness-of-fit statistic, X_{HWE}^2, given by

$$X_{HWE}^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{2.9}$$

$$= \frac{(n_{AA} - p^2 n)^2}{p^2 n} + \frac{(n_{Aa} - 2pqn)^2}{2pqn} + \frac{(n_{aa} - q^2 n)^2}{q^2 n}$$

Under the null hypothesis of HWE, $X_{HWE}^2 \sim \chi_1^2$.

Genotyping errors may impact the genotype frequencies. Thus, in large enough, randomly mating populations, where HWE is assumed to hold, significant deviations from HWE may be indicative of genotyping errors.

2.8 Where Are We Now?

Genetic variants associated with or explaining complex traits may be identified using hypothesis-driven or hypothesis-free study designs. The former incorporates prior knowledge of the potentially causal SNPs or genes focusing analyses on one candidate SNP or several of them typically lying within a specific candidate gene.

However, the obvious disadvantage that prior information may not always be available coupled with the rapid advances in technology leading to sharply falling genotyping costs, have caused a shift in the popularity of candidate SNP or gene studies to genome-wide association studies (GWASs). Nowadays, GWAS is the most widely used approach to detect genetic association with complex human traits. In a GWAS, no prior hypothesis on the potential causal SNP or gene is made (hypothesis-free) and, instead, the whole genome is scanned one SNP at a time.

The first successful GWAS was conducted in 2005 in age-related macular degeneration, a serious condition affecting old adults leading to loss of vision (*Klein et al., 2005*). Only three years after that, several hundred GWASs had been conducted identifying hundreds of association in over 80 distinct traits and diseases (*Hindorff et al., 2009*). **Figure 2.9** below illustrates identified associations as of December 2012 at stringent criteria levels for 17 trait categories (represented by different colors) locating the different findings across the genome.

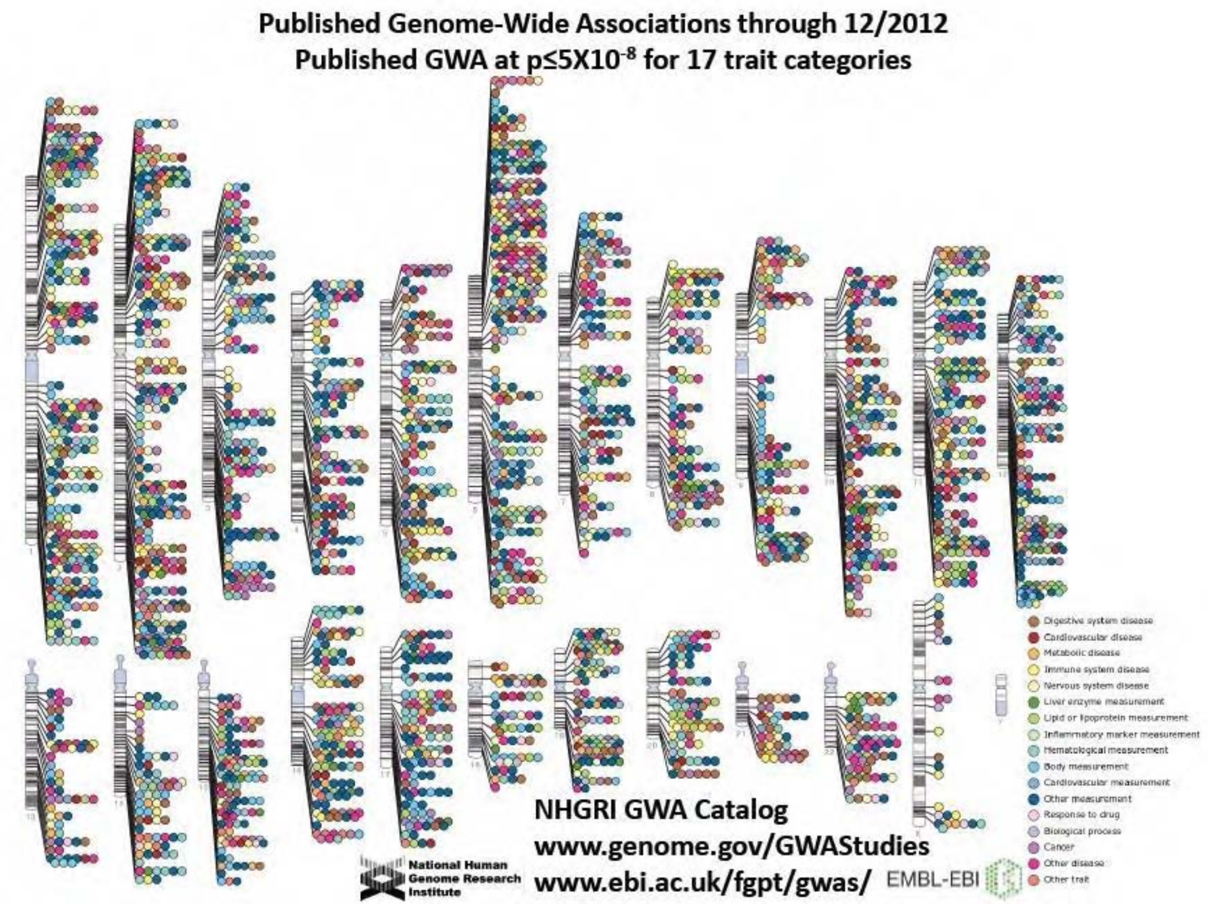

Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10-8 for 17 trait categories
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/
National Human Genome Research Institute
EMBL-EBI
Digestive system disease
Cardiovascular disease
Metabolic disease
Immune system disease
Nervous system disease
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurement
Other measurement
Response to drug
Biological process
Cancer
Other disease
Other trait

Figure 2.9: Published genome-wide associations at stringent significance criteria as of December 2012. Adapted from www.genomes.gov. Last accessed August 2013.

Basically, these findings indicate that there is a correlation between a specific genomic region and a trait. However, correlation does not imply causation which is of higher importance clinically. Thus, while our understanding of the genetic component of complex traits has deepened significantly, little benefit of that has been seen at the clinical level. Not surprisingly, skepticism on the usefulness of these studies has been on the rise. The review by *Manolio (2013)* addresses precisely this issue and suggests at least four areas where GWAS findings can be readily translated into clinical care, namely, disease prediction, disease classification, drug development and drug toxicity.

In Chapter 3, we describe the genetic association studies, while in Chapter 4, we discuss genetic prediction studies and highlight why the road to translate genetic association findings to medical practice remains rough and challenging.

3 GENETIC ASSOCIATION STUDIES

In a GWAS, the association between the SNP and the trait of interest is tested one SNP at a time. In this chapter, we describe single SNP-trait association analysis and present several limitations of GWAS.

3.1 Traits

Genetic association studies aim to relate genetic information to a clinical outcome or phenotype. The outcome can be quantitative, binary, survival, count, etc. In this thesis, we focus on quantitative and binary outcomes.

3.1.1 Quantitative Traits

A quantitative outcome is a continuous trait such as blood pressure or height. In epidemiology, assuming a linear relationship between a risk factor, x, and a continuous trait, y, measuring its effect on the trait involves the calculation of the correlation coefficient, $\rho_{x,y}$, between x and y given by

$$\rho_{x,y},= \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{E\left[(x - \mu_x)(y - \mu_y)\right]}{\sigma_x \sigma_y}, \tag{3.1}$$

where μ_x and μ_x are the mean values, and σ_x and σ_y are the standard deviations of x and y, respectively.

In the case when only a single risk factor is evaluated, the coefficient of determination, R^2, is given by the square of the correlation coefficient, that is,

$$R^2 = \rho^2_{x,y.} \tag{3.2}$$

3.1.2 Binary Traits

A binary or dichotomous outcome can take on one of two possible values typically referring to disease status, (disease versus healthy) or, in the case of response to treatment, responders versus non-responders, for example.

In epidemiology, measuring the effect of a risk factor on a dichotomous trait involves a comparison of risks or odds between the exposed (cases) and the unexposed (controls) groups. The risk is defined simply as the probability that the disease will occur, that is

$$risk = P(disease) \tag{3.3}$$

and is constrained between 0 and 1. The odds is defined as the ratio of probabilities of the disease occurring versus not occurring, that is

$$odds = \frac{P(disease)}{P(no\ disease)} \tag{3.4}$$

and can take any value between 0 and infinity. The two measures are related as follows

$$risk = \frac{odds}{1 + odds}. \tag{3.5}$$

For example, suppose that 1 in 1 000 individuals acquires a certain disease. Then, $odds = \frac{1/_{1000}}{999/_{1000}} = \frac{1}{999} = 0.001$ and, using Equation (3.5), $risk = \frac{0.001}{1 + 0.001} = 0.000999 \approx 0.001$. On the other hand, suppose for another disease, that 1 in 4 acquires it. Then, $odds = \frac{1/_4}{3/_4} = \frac{1}{3} = 0.33$ and $risk = \frac{0.33}{1+0.33} \approx 0.25$.

The above examples illustrate that, in the case of rare outcomes (diseases), the probability of disease is closely approximated by the odds of disease (that is, $risk \ \approx odds$). Alternatively, for more common outcomes, important differences between the two measures arise and this approximation loses validity.

In epidemiological studies, associations between disease and risk factors are typically expressed in terms of relative risk (RR) which is the ratio of risks in the exposed versus the unexposed group, that is

$$RR = \frac{P(disease|exposed)}{P(disease|unexposed)}. \tag{3.6}$$

In case-control studies, the odds ratio (OR) is often used as a surrogate for RR. Similarly to RR, the OR is the ratio of odds in the exposed versus the unexposed group, that is

$$OR = \frac{P(disease|exposed) \big/ P(no\ disease|exposed)}{P(disease|unexposed) \big/ P(no\ disease|unexposed)}. \tag{3.7}$$

Since the assumption that $OR \approx RR$ may not always hold, however, one must be cautious when interpreting study findings.

3.2 Heritability of Traits

3.2.1 Quantitative Traits

Variation in complex traits can be due to genetic and environmental factors and can be decomposed as follows

$$V_P = V_G + V_E + 2\,Cov(G,E), \tag{3.8}$$

where V_P is the phenotypic variance, V_G and V_E are the variance components attributable to genetic and environmental factors, respectively, and $Cov(G,E)$ is the covariance between the genetic and environmental factors. Most of the times, G and E are assumed to be independent simplifying Equation (3.8) to

$$V_P = V_G + V_E. \tag{3.9}$$

The effects of G can be further decomposed into additive, A, dominant, D, and epistatic (interaction) effects, I, leading Equation (3.9) to become

$$V_P = V_A + V_D + V_I + V_E. \tag{3.10}$$

The most widely used formulation assumes that there are no dominance and interaction effects and models only additive genetic effects. As such, Equation (3.10) reduces to

$$V_P = V_A + V_E. \tag{3.11}$$

The broad-sense heritability, H^2, is defined as the ratio of total genetic variance to total phenotypic variance and is given by

$$H^2 = \frac{V_G}{V_P}. \tag{3.12}$$

while the narrow-sense heritability, h^2, is defined as the ratio of total additive genetic variance to total phenotypic variance and is given by

$$h^2 = \frac{V_A}{V_P}. \tag{3.13}$$

3.2.2 Binary Traits

For binary traits (such as disease traits) the observed scale is 0/1 (control/case). It is assumed that such traits can be represented by an underlying normally distributed liability trait. If an individual's value exceeds a specific threshold on the liability scale, then this individual is assigned a phenotypic value of 1, otherwise he is assigned a phenotypic value of 0. The relationship between the heritability at the observed scale, h^2_{obs}, and the narrow-sense heritability on the continuous liability scale, h^2, (Equation (3.13)) is given by

$$h^2_{obs} = \frac{h^2 z_K^2}{K(1-K)}, \tag{3.14}$$

where K is the prevalence of disease in the population, $Z \sim N(0,1)$ and z_K is the standard normal quantile such that $P(Z > z_K) = K$. (*Dempster and Lerner, 1950*) The maximum value of h^2_{obs} = 0.64 when $K = 0.5$ and $h^2 = 1$. (*Visscher et al., 2008*)

As in classical epidemiology, the choice of analytical method depends on the study design which is further dictated by the study subjects (related or unrelated) and by the type of trait under investigation (dichotomous, quantitative, survival, etc.). From here onwards, we focus on the most widely used case-control study design for a dichotomous trait in unrelated individuals. Most of the discussion applies also to quantitative traits.

3.3 Penetrance (Genetic Effect)

Heritable traits are carried forth over generations through the transmission of DNA. Penetrance is defined as the probability that an individual will express the trait given the he or she carries the implicated gene (or genotype).

A Mendelian or monogenic trait is controlled by a single gene. A mutation in that gene can *cause* disease. For example, Huntington's disease and cystic fibrosis are monogenic diseases. Mutations in the *HTT* gene (Huntington's disease) or in the *CFTR* gene (cystic fibrosis) cause the respective disease. In other words, the probability that an individual will develop the disease given that he or she carries the mutations in the respective gene is 1. In this particular case, the gene is said to have complete penetrance.

Monogenic disorders are relatively rare. Most traits are complex, arising as a result of a complex interplay between genetic and environmental factors. For these traits, we study the penetrance or the effect of variants in genes that may be *associated* with the trait.

The penetrance of a gene (or genotype) may be influenced by other genes (gene-gene interactions) or by environmental factors (gene-environment interactions). For the discussion that follows, we have adopted a simplified view and assumed that no such interactions are present.

Further, penetrance may be age-related or gender-related. For instance, penetrance for cystic fibrosis is 1 at birth and for Huntington's disease is 1 by the age of 70 years. Different penetrance estimates exist for mutations for breast cancer in females and males.

Suppose, without loss of generality, that there is a single bi-allelic SNP with alleles a and A, where allele A is associated with the risk of developing a disease. An individual can carry one of three genotypes a/a, A/a and A/A with penetrance P_{aa}, P_{Aa} and P_{AA}, respectively, given by

$$\begin{aligned} P_{aa} &= \Pr(Disease \mid a/a), \\ P_{Aa} &= \Pr(Disease \mid A/a), \\ P_{AA} &= \Pr(Disease \mid A/A), \end{aligned} \tag{3.15}$$

constrained as follows

$$0 \leq P_{aa} \leq P_{Aa} \leq P_{AA} \leq 1. \tag{3.16}$$

Under the *dominant* genetic model,

$$P_{AA} = P_{Aa} > P_{aa}. \tag{3.17}$$

Under the *recessive* genetic model,

$$P_{AA} > P_{Aa} = P_{aa}. \tag{3.18}$$

Under the *additive* genetic model

$$P_{Aa} = \frac{1}{2}(P_{AA} + P_{aa}). \tag{3.19}$$

The relative risk of genotype A/a to genotype a/a is then given by

$$RR_{Aa} = \frac{P_{Aa}}{P_{aa}} \tag{3.20}$$

and, similarly, the relative risk of genotype A/A to genotype a/a is given by

$$RR_{AA} = \frac{P_{AA}}{P_{aa}}. \tag{3.21}$$

We mentioned in Section 3.1.2 that, in case-control studies, the OR is often used as a surrogate for RR. The corresponding ORs are then given by

$$OR_{Aa} = \frac{odds_{Aa}}{odds_{aa}} = \frac{{P_{Aa}}/{(1 - P_{Aa})}}{{P_{aa}}/{(1 - P_{aa})}} \tag{3.22}$$

and

$$OR_{AA} = \frac{odds_{AA}}{odds_{aa}} = \frac{P_{AA}/(1 - P_{AA})}{P_{aa}/(1 - P_{aa})}. \tag{3.23}$$

Under the dominant model, $OR_{AA} = OR_{Aa}$, under the recessive model, $OR_{AA} > OR_{Aa}$ and under the additive genetic model, $OR_{AA} = (OR_{Aa})^2$.

In the following section we describe the methods for estimating the genetic effect under the different genetic models, a summary of which is presented in **Table 3.5**.

3.4 Methods

In a case-control study, the allele or genotype frequencies are compared between disease individuals and healthy controls. Assume, for illustrative purposes, that we are analyzing the effect of the SNP in a case-control sample of n individuals. The arising contingency table of genotypic counts is illustrated in **Table 3.1** below.

	Genotype Counts			**Total**
	a/a	**A/a**	**A/A**	
Cases	n_{11}	n_{12}	n_{13}	$\boldsymbol{n_{1.}}$
Controls	n_{21}	n_{22}	n_{23}	$\boldsymbol{n_{2.}}$
Total	$\boldsymbol{n_{.1}}$	$\boldsymbol{n_{.2}}$	$\boldsymbol{n_{.3}}$	$\boldsymbol{n}$

Table 3.1: The contingency table of disease status by genotype counts for a single bi-allelic SNP with alleles a and A and for n individuals.

When the effect of a single SNP is being investigated, the analyses may also be conducted at the allelic level. The arising contingency table of allele counts is derived based on the genotypic counts in **Table 3.1** and is given in **Table 3.2** below.

	Allele Counts		
	a	**A**	**Total**
Cases	m_{11}	m_{12}	$\boldsymbol{m_{1.}}$
Controls	m_{21}	m_{22}	$\boldsymbol{m_{2.}}$
Total	$\boldsymbol{m_{.1}}$	$\boldsymbol{m_{.2}}$	$\boldsymbol{2n}$

Table 3.2: The contingency table of disease status by allele counts for a single bi-allelic SNP with alleles a and A and for n individuals with the allelic counts related to the genotype counts in **Table 3.1** as follows: $m_{i1} = n_{i1} * 2 + n_{i2}$ and $m_{i2} = n_{i3} * 2 + n_{i2}$ for $i = 1{,}2$.

3.4.1 Allelic Tests

Assuming, without loss of generality, that A is the risk allele, the estimated allelic OR_A based on the counts in **Table 3.2**, $\widehat{OR}_A$, is given by

$$\widehat{OR}_A = \frac{m_{12}m_{21}}{m_{11}m_{22}}. \tag{3.24}$$

The most popular method to estimate the precision of this estimate has been proposed by (*Woolf, 1955*). Asymptotically, the distribution of OR_A on the natural logarithm scale is approximately normally distributed. The $100(1-\alpha)\%$ confidence interval can thus be derived as

$$\exp\left(log\left(\widehat{OR}_A\right) \pm z_{\alpha/2}\, SE_{\log(\widehat{OR}_A)}\right), \tag{3.25}$$

where $z_{\alpha/2}$ denotes the $\left(1-\alpha/2\right)$ standard normal quantile and $SE_{\log(\widehat{OR}_A)}$ is the standard error of $log\left(\widehat{OR}_A\right)$ given by $SE_{\log(\widehat{OR}_A)} = \sqrt{\frac{1}{m_{11}} + \frac{1}{m_{12}} + \frac{1}{m_{21}} + \frac{1}{m_{22}}}$.

If there is no association between the SNP and the disease status, the OR_A = 1. Based on our definition (allele A is associated with risk of disease), an $OR_A \neq 1$ would indicate that the allele A is a risk allele ($OR_A > 1$, allele A increases the risk) or a protective allele ($OR_A < 1$, allele A decreases the risk). Therefore, we could test the following hypothesis

$$\begin{aligned} H_0&: OR_A = 1 \\ H_1&: OR_A \neq 1^{\cdot} \end{aligned} \tag{3.26}$$

If the $100(1-\alpha)\%$ confidence interval obtained by Equation (3.25) above contains the value of 1, then the null hypothesis cannot be rejected at the α level of significance. The hypothesis can be tested using the χ^2 test for independence of rows and columns. The test statistic, $X^2_{allelic}$, is given by

$$X^2_{allelic} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(m_{ij} - E\left[m_{ij}\right]\right)^2}{E\left[m_{ij}\right]}, \tag{3.27}$$

where $E\left[m_{ij}\right] = \frac{m_{i.}m_{.j}}{2n}, i = 1,2, j = 1,2$. Under the null hypothesis of no association, $X^2_{allelic} \sim \chi^2_1$.

In case any $m_{ij} < 5, i = 1,2, j = 1,2$, the χ^2 approximation is not valid and the Fisher's exact test is used. In this test, the exact probability of observing these cell counts is computed based on the hypergeometric distribution $f(m_{12};\ m_{1.}, m_{.2}, 2n)$ such that

$$\Pr(m_{12}) = f(m_{12};\ m_{1.}, m_{.2}, 2n)\frac{\binom{m_{1.}}{m_{12}}\binom{m_{2.}}{m_{22}}}{\binom{2n}{m_{.2}}}. \tag{3.28}$$

Next, the probability for each possible arrangement of the cell counts conditional on the marginal counts (that is, $m_{1.}, m_{.2}, 2n$) is computed. The two-sided p-value is given by

$$p - \text{value} = \textstyle\sum_y \Pr(y) \text{ such that } \Pr(y) \leq \Pr(m_{12}). \tag{3.29}$$

The effect of the SNP on the disease status may be different according to gender. Therefore, we may want to conduct stratified association analysis where we built a series of allelic contingency tables, one for each stratum (for example, females and males) as illustrated in **Table 3.3** below (a subscript k for stratum is added to the counts from **Table 3.2**).

Stratum s	**Allele Counts**		
	a	**A**	**Total**
Cases	m_{11k}	m_{12k}	$\boldsymbol{m_{1.k}}$
Controls	m_{21k}	m_{22k}	$\boldsymbol{m_{2.k}}$
Total	$\boldsymbol{m_{.1k}}$	$\boldsymbol{m_{.2k}}$	$\boldsymbol{2n_k}$

Table 3.3: Allelic contingency table per stratum $k, k = 1, \dots, K$ with n_k individuals in each stratum.

The Cochran-Mantel-Haenszel (CMH) estimate of the OR adjusted for stratum is then given by the weighted OR across all strata S.

$$\widehat{OR_{CMH}} = \frac{\sum_{k=1}^{K} \frac{m_{12k} m_{21k}}{2n_k}}{\sum_{k=1}^{K} \frac{m_{11k} m_{22k}}{2n_k}}. \tag{3.30}$$

Let OR_k denote the OR for stratum $k, k = 1, \dots, K$. The hypothesis being tested is

$$H_0 \colon OR_1 = OR_2 = \cdots = OR_K = 1 \tag{3.31}$$

$$H_1 \colon \text{there exists k such that } OR_k \neq 1$$

The CMH test statistic is given by

$$M^2 = \frac{\left[\sum_{k=1}^{K} (m_{11k} - E(m_{11k}))\right]^2}{\sum_{k=1}^{K} Var(m_{11k})}, \tag{3.32}$$

where $E(m_{11k}) = \frac{m_{1.k} m_{.1k}}{2n_k}$ and $Var(m_{11k}) = \frac{m_{1.k} m_{2.k} m_{.1k} m_{.2k}}{2n_k{}^2 (2n_k - 1)}$. Under the null hypothesis of no association, $M^2 \sim \chi_1^2$.

The CMH test assumes homogeneous association across strata (*Agresti, 2007*). That is, it assumes that

$$H_0 \colon OR_1 = OR_2 = \cdots = OR_K. \tag{3.33}$$

A χ^2 test, named the Breslow-Day (BD) test, can be used to test the homogeneity of odds ratios. The test statistic, X_{BD}^2, is given by

$$X_{BD}^2 = \sum_i \sum_j \sum_k \frac{(m_{ijk} - E(m_{ijk}))^2}{E(m_{ijk})}, \tag{3.34}$$

where $E(m_{ijk})$ is the expected cell count for cell ij in stratum k, respectively. Under the null hypothesis of homogeneous odds ratios, $X_{BD}^2 \sim \chi_{K-1}^2$.

It is straightforward to carry out stratified analyses controlling for *one* categorical variable. However, controlling for several covariates simultaneously (for example, gender, smoking status and geographic origin) leads to numerous contingency tables reducing the sample size per table (and per cell) as more covariates are added. Moreover, the covariates need to be categorical or forcibly categorized. Therefore, to adjust for several risk factors and/or for continuous factors, regression methods are used.

Lastly, the allelic test (Equation (3.27)) is valid when the true model of association is multiplicative (additive on the logarithmic scale) (*Sasieni, 1997*). The Cochran-Armitage trend test (see Equation (3.38)) is a genotype-based test that is asymptotically equivalent to the allelic test and is more robust to deviations from HWE. (*Guedj et al., 2008*)

3.4.2 Genotypic Tests

The estimated genotypic odds ratio based on the counts in **Table 3.1** for genotype A/A relative to genotype a/a, $\widehat{OR_{AA}}$, is given by

$$\widehat{OR_{AA}} = \frac{n_{13}n_{21}}{n_{11}n_{23}} \tag{3.35}$$

and for genotype A/a relative to genotype a/a, $\widehat{OR_{Aa}}$, is given by

$$\widehat{OR_{Aa}} = \frac{n_{12}n_{21}}{n_{11}n_{22}}. \tag{3.36}$$

The confidence intervals are computed in much the same way as for the allelic contingency table. For the hypothesis test, the test statistic in Equation (3.27) becomes

$$X^2_{genotypic} = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{\left(m_{ij} - E\left[m_{ij}\right]\right)^2}{E\left[m_{ij}\right]}, \tag{3.37}$$

where $E\left[m_{ij}\right] = \frac{m_{i.}m_{.j}}{2n}, i = 1,2, j = 1,2,3$. Under the null hypothesis of no association, $X^2_{genotypic} \sim \chi^2_2$.

The tests discussed so far refer to the general model without any assumption on the underlying genetic model. Based on the counts in **Table 3.1**, **Table 3.4** provides the contingency tables that arise under the dominant (**A**) and recessive (**B**) models. In those cases, the usual χ^2 test for independence of rows and columns for a $2x2$ contingency table applies as the one described for the allelic contingency table (Equation (3.27)).

(A)	Dominant	
	a/a	A/a + A/A
Cases	n_{11}	$n_{12+}n_{13}$
Controls	n_{21}	$n_{22} + n_{23}$
Total	$\boldsymbol{n}_{.1}$	$\boldsymbol{n}_{.2} + \boldsymbol{n}_{.3}$

(B)	Recessive	
	a/a + A/a	A/A
Cases	$n_{11} + n_{12}$	n_{13}
Controls	$n_{21} + n_{22}$	n_{23}
Total	$\boldsymbol{n}_{.1} + \boldsymbol{n}_{.2}$	$\boldsymbol{n}_{.3}$

Table 3.4: Contingency tables arising under the dominant (**A**) and recessive (**B**) models derived based on the genotype counts in **Table 3.1**.

Alternatively, the Cochran-Armitage trend test (CATT - *Cochran, 1954*; *Armitage, 1955*) modifies the genotypic χ^2 test (Equation (3.37)) to incorporate a particular order of the genotypes. The test statistic is given by

$$T^2 = \frac{\left[\sum_{i=1}^{3} w_i(n_{1i}n_{2.} - n_{2i}n_{1.})\right]^2}{\frac{n_{1.}n_{2.}}{n}\left[\sum_{i=1}^{3} w_i^2 n_{.i}(n - n_{.i}) - 2\sum_{i=1}^{2}\sum_{j=i+1}^{3} w_i w_j n_{.i} n_{.j}\right]}, \tag{3.38}$$

where $w_i, i = 1,2,3$ are the weights assigned to genotypes a/a, A/a and A/A, respectively. The weights are assigned according to the assumed genetic model setting $\boldsymbol{w} = (0, 1, 1)$ for a dominant effect of allele A and $\boldsymbol{w} = (0, 0, 1)$ for a recessive effect of allele A. Most of the times, the underlying genetic model is unknown and in most genetic association studies the general additive model is used where $\boldsymbol{w} = (0, 1,2)$. CATT with $\boldsymbol{w} = (0, 1,2)$ is asymptotically equivalent to the allelic test described earlier (Equation (3.27), (*Sasieni*, *1997*)). Under the null hypothesis of no association, $T^2 \sim \chi_1^2$.

3.4.2.1 Regression Framework

All the association tests discussed so far were based on contingency table analyses. Tests based on genotype counts can also be formulated in a logistic regression framework. In logistic regression, the odds of disease is assumed to be a linear function of the intercept and p explanatory variables (for example, genotype variables) on the log scale, that is,

$$\log(odds) = \alpha + \sum_{i=1}^{p} \beta_i x_i, \tag{3.39}$$

where α is the intercept and β_i is the coefficient for the x_i variable, $i = 1, \dots, p$. The corresponding parameterization for the different genetic models is given in **Table 3.5**.

Using this formulation, it is possible to construct association tests based on the likelihood of the model. There are three main testing approaches, namely, the likelihood ratio test (LRT), the Wald test and the score test.

For the LRT, we fit two models, the null model, M_0, under which the risk of disease is not affected by the genotype, that is $\beta_i = 0, i = 1, .., p$ and the alternative model, M_1, under which the risk of disease is affected by the genotype, that is, $\exists i\ \beta_i \neq 0$ and then compare the fits. The LRT statistic, T, is given by

$$T = -2ln\left(\frac{likelihood\ M_0}{likelihood\ M_1}\right). \tag{3.40}$$

Note that M_0 is a special case of M_1 and, in fact, the LRT requires nested models. Under the null hypothesis of no association, $T \sim \chi^2_{df}$, where df is the difference between the degrees of freedom of the null and the alternative model.

The Wald test is asymptotically equivalent to the LRT and has the advantage that it requires the estimation of only one model. For a model with one explanatory variable (for instance, additive, dominant and recessive models, see **Table 3.5** below), the Wald statistic, W, is given by

$$W = \left(\frac{\hat{\beta}}{SE(\hat{\beta})}\right)^2, \tag{3.41}$$

where $\hat{\beta}$ is the maximum likelihood estimate of β and $SE(\hat{\beta})$ is the standard error of $\hat{\beta}$. Under the null hypothesis of no association, $W \sim \chi^2_1$. If many variables need to be tested simultaneously (such as for the genotypic model, see **Table 3.5** below), using matrix notation, the Wald statistic, W, is given by

$$W = \widehat{\boldsymbol{\beta}}'\Sigma^{-1}\widehat{\boldsymbol{\beta}}, \tag{3.42}$$

where $\widehat{\boldsymbol{\beta}}$ is the vector of maximum likelihood estimates of $\boldsymbol{\beta}$ and Σ is the variance matrix. Under the null hypothesis of no association, $W \sim \chi^2_{df}$, where df is the number of variables being tested simultaneously.

Lastly, the score test (often known as the Lagrange multiplier test) also requires the estimation of only one model. Contrary to the Wald test, however, the estimated model does not include the parameter(s) of interest. Let $L(\beta|x)$ be the likelihood function depending on one parameter, β, given the data x. The score, $U(\beta)$, is given by

$$U(\beta) = \frac{\partial \log(L(\beta|x))}{\partial \beta} \tag{3.43}$$

and corresponds to the slope of the log likelihood with respect to β. The variance of the score, $I(\beta)$, is known as the Fisher's information and is given by

$$I(\beta) = \mathbb{E}\left\{\left[\frac{\partial \log(L(\beta|x))}{\partial \beta}\right]^2\right\} \tag{3.44}$$

The score test statistic, $S(\hat{\beta})$, is given by

$$S(\hat{\beta}) = \frac{\left(U(\hat{\beta})\right)^2}{I(\hat{\beta})} \tag{3.45}$$

where $U(\hat{\beta})$ and $I(\hat{\beta})$ are the score and its variance evaluated at $\hat{\beta}$, the maximum likelihood estimate of β. Under the null hypothesis of no association, $S(\hat{\beta}) \sim \chi_1^2$. As for the Wald test, the multivariate version of the test statistic exists and is given by

$$S(\widehat{\boldsymbol{\beta}}) = U'(\widehat{\boldsymbol{\beta}})I^{-1}(\widehat{\boldsymbol{\beta}})U(\widehat{\boldsymbol{\beta}}), \tag{3.46}$$

where $U(\widehat{\boldsymbol{\beta}})$ is a vector of scores and $I^{-1}(\widehat{\boldsymbol{\beta}})$ is the inverse of the variance matrix. Under the null hypothesis of no association, $S(\widehat{\boldsymbol{\beta}}) \sim \chi_{df}^2$, where df is the number of variables tested by the null hypothesis. Throughout this thesis, we have used the LRT or the Wald test.

3.4.3 A Note on Quantitative Traits

Genetic associations with quantitative traits are typically conducted in a simple linear regression framework. All likelihood-based tests described so far for the binary traits are applicable to continuous traits as well. For the underlying genetic model, the same parameterization described in **Table 3.5** applies except that the linear relationship modeled is between the mean trait value and the SNP as opposed to between the logarithm of the odds of disease and the SNP.

		Genetic Model								
		Genotypic (General) $\log(odds) = \alpha + \beta_1 x_1 + \beta_2 x_2$			**Dominant** $\log(odds) = \alpha + \beta_1 x_1$		**Recessive** $\log(odds) = \alpha + \beta_1 x_1$		**Multiplicative (log-additive)** $\log(odds) = \alpha + \beta_1 x_1$	
		x_1	x_2	**OR**	x_1	**OR**	x_1	**OR**	x_1	**OR**
Genotype	**a/a**	0	0	1	0	1	0	1	0	1
	A/a	1	0	$\exp(\beta_1)$	1	$\exp(\beta)$	0	1	1	$\exp(\beta)$
	A/A	0	1	$\exp(\beta_2)$	1	$\exp(\beta)$	1	$\exp(\beta)$	2	$\exp(2\beta) = (\exp(\beta))^2$
Corresponding Contingency Table Analysis		**Table 3.1** (χ^2 test on 2df)			**Table 3.4 (A)** (χ^2 test on 1df)		**Table 3.4 (B)** (χ^2 test on 1df)		At the allelic level, **Table 3.2** (χ^2 test on 1df)	

Table 3.5: Parameterization of genetic models in a logistic regression framework and the corresponding contingency table analyses described in the text.

3.4.4 The Power of χ^2 Tests for Contingency Tables

For a contingency table with r rows and c columns, the test statistic, X^2, under the null hypothesis of independence between the rows and the columns follows a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, that is

$$X^2 \sim \chi^2_{(r-1)(c-1)}. \tag{3.47}$$

Let α be the probability of falsely rejecting the null hypothesis, that is, of committing a Type I Error, given by $\alpha = \Pr(\text{reject H}_0 \mid H_0 \text{ is true})$. The value of α is called the significance level of the test. The null hypothesis of independence is rejected if

$$X^2 > \chi^2_{(r-1)(c-1),\alpha}, \tag{3.48}$$

where $\chi^2_{(r-1)(c-1),\alpha}$ is the critical value of the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom at the α level of significance, that is, $\Pr\left(X^2 > \chi^2_{(r-1)(c-1),\alpha}\right) = \alpha$.

Next, let β be the probability of failing to reject the null hypothesis when it is false, that is of committing a Type II Error, given by $\beta = \Pr(\text{do not reject H}_0 \mid H_0 \text{ is false})$. We do not reject the null hypothesis of independence if

$$X^2 < \chi^2_{(r-1)(c-1),\alpha}. \tag{3.49}$$

The test statistic under the alternative hypothesis of association, X^2_λ, follows a non-central χ^2 distribution with non-centrality parameter λ and $(r-1)(c-1)$ degrees of freedom such that

$$X^2_\lambda \sim \chi^2_{(r-1)(c-1)}(\lambda). \tag{3.50}$$

and, hence,

$$\beta = \Pr(\text{do not reject } H_0 \mid H_0 \text{ is false}) = \Pr\left(X^2_\lambda < \chi^2_{(r-1)(c-1),\alpha}\right). \tag{3.51}$$

The power of the test at the α level of significance is then given by the probability of correctly rejecting the null hypothesis, that is,

$$\begin{aligned} Power &= \Pr(\text{reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - \Pr(\text{do not reject } H_0 \mid H_0 \text{ is false}) \\ &= 1 - \beta \\ &= 1 - \Pr\left(X_\lambda^2 < \chi^2_{(r-1)(c-1),\alpha}\right) \\ &= \Pr\left(X_\lambda^2 \geq \chi^2_{(r-1)(c-1),\alpha}\right). \end{aligned} \tag{3.52}$$

Following these principles, the power can be estimated for the allelic or the genotypic χ^2 tests that we described in Sections 3.4.1 and 3.4.2. This can be achieved by deriving the non-centrality parameter, λ. We provide here the expression for λ for the most general genotypic test (Equation (3.37)).

For simplicity, we express the genotype counts in **Table 3.1** as frequencies for n_1. cases and n_2. controls in **Table 3.6** below.

	Genotype Frequencies		
	a/a	**A/a**	**A/A**
Cases	$p_{11} = \frac{n_{11}}{n_{1.}}$	$p_{12} = \frac{n_{12}}{n_{1.}}$	$p_{13} = \frac{n_{13}}{n_{1.}}$
Controls	$p_{21} = \frac{n_{21}}{n_{2.}}$	$p_{22} = \frac{n_{22}}{n_{2.}}$	$p_{23} = \frac{n_{23}}{n_{2.}}$

Table 3.6: Genotype frequencies in cases and controls derived from the genotype counts in **Table 3.1**.

Then, *Gordon et al. (2002)* derive the non-centrality parameter, $\lambda_{genotypic}$, as

$$\lambda_{genotypic} = n_{1.}n_{2.}\left[\frac{(p_{11}-p_{21})^2}{n_{1.}p_{11} + n_{2.}p_{21}} + \frac{(p_{12}-p_{22})^2}{n_{1.}p_{12} + n_{2.}p_{22}} + \frac{(p_{13}-p_{23})^2}{n_{1.}p_{13} + n_{2.}p_{23}}\right] \tag{3.53}$$

Therefore, the power of the χ^2 tests on contingency tables at a given level of significance, α, depends on the degrees of freedom and on the non-centrality parameter λ. Through the latter, it also depends on number of cases and controls, on the genotype (allele) frequencies and on the effect size (the difference between cases and controls based on genotype).

3.5 Pitfalls and Limitations of Genome-Wide Association Studies

We discuss next several of the pitfalls and limitations of genetic association studies specifically in the context of GWASs by drawing examples from the literature as well as from our own work.

3.5.1 Multiple Testing

In a GWAS, hundreds of thousands to millions of null hypotheses of no association are being tested, one for each SNP. If we carry out one million association tests at the 0.05 significance level, we would expect approximately 50 000 (1E+6 * 0.05) false positives. Therefore, if we wish to maintain an overall Type I Error of 0.05, we need to reduce the significance level of the test.

One of the most commonly used approach to correct for multiple tests is the Bonferroni correction method whereby the pre-specified significance level, α, is adjusted to reflect the number of tests, t, that were carried out such that the new significance level, α^*, is given by $\alpha^* = \alpha/t$. For a typical GWAS, α^* is set at 1E-7 or stricter. This approach, while simple to apply, is conservative since it assumes that all t tests are independent which is not correct as SNPs are correlated due to LD.

In the previous section (Section 3.4.4), we pointed out that the power to detect association is obtained at a given significance level, α. Therefore, if the adjusted significance level is too conservative, the power is compromised.

Alternative methods exist that are less conservative such as the false discovery rate (FDR), where the proportion of rejected null hypotheses that were falsely rejected is specified instead (*Benjamini and Hochberg, 1995*).

3.5.2 Direct, Indirect and Confounded Associations

A statistically significant difference in the allele or genotype frequencies between cases and controls *may* be indicative of evidence that the SNP is associated with the disease. Moreover, even if the detected association is not spurious, it is rarely the case that the SNP's influence on the trait is direct.

For certain traits, prior investigative work has identified potentially causal SNPs. Direct association studies target these SNPs. These are powerful studies as prior knowledge is incorporated. However, most of the times, little is known on the underlying genetic mechanism of complex traits and causal SNPs are rarely directly analyzed.

In indirect association studies, the SNPs surrounding the causal SNP are analyzed. Since those SNPs are likely correlated with the causal SNP (due to LD, see Section 2.3), the association may be picked up. The concepts of direct and indirect associations are illustrated in **Figure 3.1** below.

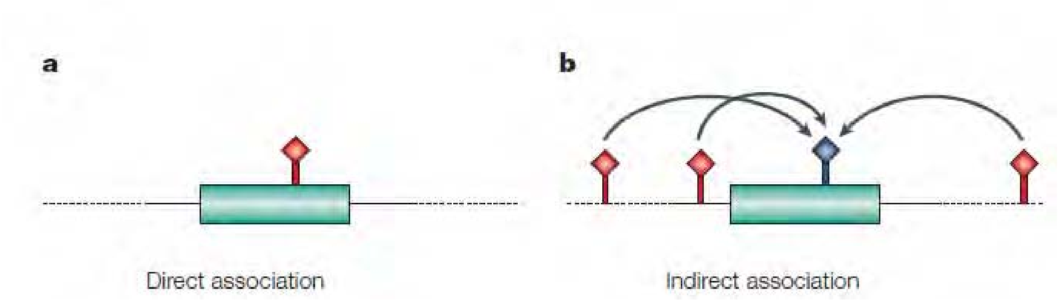


Figure 3.1: Direct (**a**) and indirect (**b**) association studies. In direct association studies, only one SNP, the causal (in red), is analyzed. In indirect association studies, several SNPs (in red) that are correlated with the causal SNP (in blue) are analyzed. Adapted from *(Hirschhorn and Daly, 2005*).

A significant SNP association from a GWAS may only be hinting to the location of the causal SNP and, therefore, should not be interpreted as the causal SNP.

Further, the power to detect association with a SNP nearby the causal SNP depends on the LD between the two SNPs. The stronger the LD between the genotyped SNP and the causal SNP is, the higher the power is to detect the association. Alternatively, *Pritchard and Przeworski (2001)* compared the sample sizes needed to achieve the same power, holding all else fixed, when the causal SNP and a SNP in LD with the causal is tested. Let n be the sample size when the causal SNP is tested and let n' be the sample size when the SNP in LD with the causal is tested. They showed that $n' = {}^{n}/_{r^2}$, where r^2 is the measure of LD as defined in Equation (2.6). Hence, to achieve roughly the same power, a much larger sample size (in the order of ${}^{1}/_{r^2}$) is necessary when the SNP in LD with the causal SNP is tested.

Lastly, if the cases and controls are not properly matched, the association may be confounded. Large study sample sizes are required for conducting GWASs (see Section 3.5.4). Major international collaborations over the years have made that possible. However, with such collaborative efforts combining samples of different ethnic origins and from different research centers may lead to biased results if not appropriately accounted for. Population stratification (also referred to as population structure) arises when the study population is non-homogeneous as, for example, when subjects from different ethnic backgrounds having different allele frequencies are included in the study and disproportionately represent the cases and controls. However, if population stratification is detected, methods exist to appropriately account for it.

3.5.3 Effect Size and Allelic Frequency

The dominating hypothesis of the GWAS era was that the genetic variability of common diseases (and traits) could be explained by few common genetic variants (with minor allele frequency, MAF ≥ 5%) exerting a small to moderate effect on the trait. This is known as the common disease-common variant hypothesis (CDCV). **Figure 3.2** below shows that the CDCV hypothesis provides a limited view of the genetic architecture of complex traits and diseases.

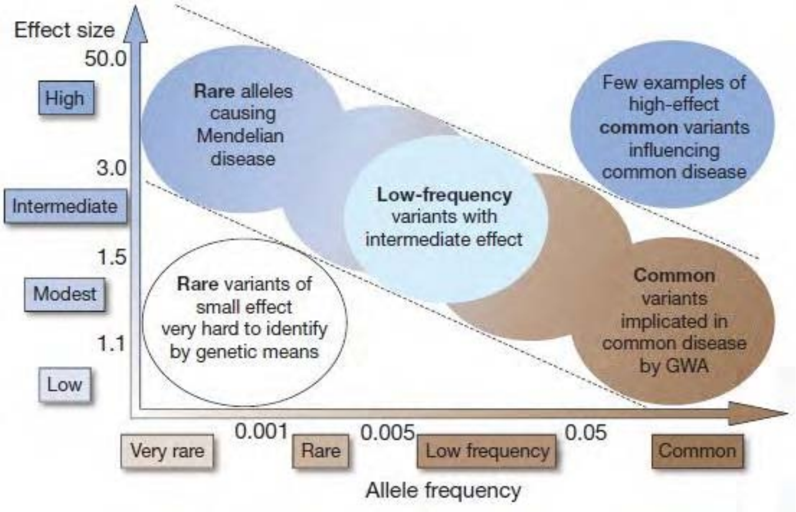

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

Figure 3.2: Effect size (OR) plotted against risk allele frequency. Adapted from (*Manolio, 2013*).

Alternatives to the CDCV hypothesis have since been proposed such as the common disease-rare variant (CDRV) hypothesis where it is assumed that genetic variation of common diseases may be explained by a large number of low frequency or rare variants (MAF < 5%) having large effect sizes (*Cirulli and Goldstein, 2010*). As we saw in Chapter 2, genotyping chips were designed to capture only common variation and, hence, rare variants are not tagged by the existing SNPs on the chips. Therefore, the GWAS was not designed to detect associations under the CDRV hypothesis. Currently, methodological focus has turned to analyzing whole-genome sequence data that would allow uncovering the role of rare variants in complex traits. As we saw in Section 3.4.4, for the association tests that we have discussed, the power depends on the effect size and on the allele frequency. Holding everything else fixed, the power increases with larger effect size and decreases with lower allele frequency.

3.5.4 Study Sample Size

Figure 3.3 below illustrates the results of a simulation study estimating the required sample size to detect associations for given effect sizes and disease-allele frequencies in order to achieve power of 80% at genome-wide significance threshold (p-value < 10^{-6}) under the multiplicative genetic model.

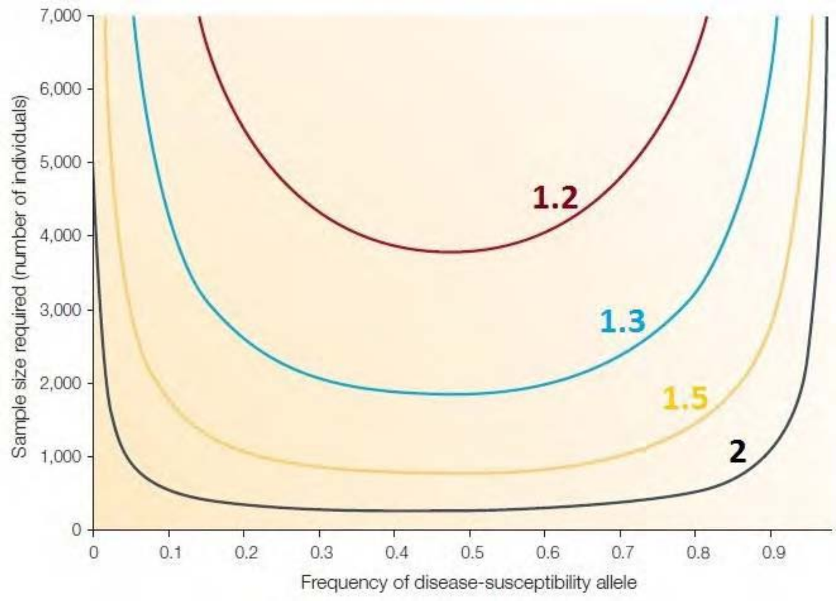


Figure 3.3: Required number of cases and controls to achieve 80% power at p-value < 10^{-6} to detect associations for given allelic frequencies and effect sizes (OR, shown near each corresponding curve). Perfect linkage disequilibrium between the test markers and the disease variants and the multiplicative genetic model is assumed. Adapted from (*Wang et al., 2005*).

Figure 3.3 above illustrates that extremely large sample sizes are required to detect variants of weaker effects and/or low allelic frequency. This provided an impetus for the formation of large international consortia to combine samples and collaborative efforts for better powered studies.

The landmark GWAS of age-related macular degeneration in 2005 was conducted on 146 subjects (96 cases and 50 controls) with roughly 120 000 genotyped SNPs *(Klein et al., 2005*). Lately, GWAS sample sizes have reached the vicinity of tens of thousands of individuals. A GWAS on blood pressure and cardiovascular disease risk reported in 2011, for example, evaluated associations between roughly 2.5 million SNPs on approximately 70 000 subjects in their discovery dataset and used a sample size of roughly 200 000 individuals for their combined discovery and replication datasets (*Ehret et al., 2011*).

3.5.5 Missing Heritability

In their study, *Ehret et al. (2011)* estimated that around 120 genetic variants contribute to blood pressure explaining 2.2% of the phenotypic variation while the 29 independent variants that they identified in their study explained only about 0.9% of the variation. Thus, despite the large sample size and large SNP map, more than half of the heritability still remained unexplained.

In fact, while GWAS have provided significant insights into the genetics of common complex traits and diseases, it is important to note that much of the genetic variance of the studied phenotypes has remained largely unexplained – the so called missing heritability problem (*Maher, 2008*). Plausible explanations for the missing heritability problem include overestimation of heritability from twin and family studies but mostly surround the limitations of GWAS.

3.5.6 Underlying Genetic Model

The power of the genetic association studies depends on the underlying genetic model. Most of the times, the underlying genetic model is unknown and typically, in GWASs, the multiplicative (additive on the log scale) genetic model is assumed. This can lead to loss of power if the underlying genetic model is not the assumed one. It is possible, of course, to use the general model but testing for it requires two degrees of freedom instead of one. We saw from Section 3.4.4 that the power depends on the number of degrees of freedom and, holding everything else constant, it decreases as the number of degrees of freedom increases.

3.5.7 Modeling Strategies

We saw in Section 3.5.4 that much larger sample sizes are needed to detect associations with rare variants and/or variants with weaker effects and, even then, the GWAS is not well powered. Alternative methods need to be considered that may improve the power in these contexts.

Although this was not the primary focus of my thesis, for the analyses of rare variants, I contributed to a comparative study evaluating the power and Type I Error of several novel statistical methods that have been proposed for the analyses of rare variants in simulated data provided by the Genetic Analysis Workshop 17 (GAW17) (*Saad et al., 2011b*).

Analogously, for the analyses of common SNPs with weak effects (where my focus lies) alternative modeling strategies to the somewhat simplistic SNP-by-SNP approach adopted in GWAS have been proposed. It has been suggested that estimating SNP effects individually is not optimal and does not lead to consistent effect estimates due to LD (see, for example, (*de los Campos et al., 2010*)). Therefore, approaches modeling the effect of all SNPs simultaneously may be better powered at identifying the set of SNPs that are associated with the trait or disease of interest. In Sections 3.6.3 and 3.6.4, we describe two different multi-marker association analyses that we conducted and how they compare to the classical single-marker approach. Briefly, though, the results of these studies did not convince us of the superiority of these more complex approaches over the classical single-marker approach to detect associations.

The remainder of this chapter describes the various genetic association studies that we have conducted.

3.6 Applications

3.6.1 Role of the 2'-5'-oligoadenylate synthetase 1 (*OAS1*) gene in interferon response in MS patients

3.6.1.1 Objective

The objective of this study was to investigate whether the SNP rs2660 in the 2'-5'-oligoadenylate synthetase 1 (*OAS1*) gene was associated with response to interferon treatment in MS patients. *OAS1*, an essential protein involved in the innate immune response to viral infection, is upregulated by interferons. Since viral infection is a potential causative factor in MS and established causative factor in relapses, *OAS1* may be playing a vital role in interferon response in MS patients. A recent study, for instance, has found a gender-related immunological action of interferon therapy in MS (*Contasta et* *al., 2012*). Therefore, we were also interested in investigating if there was a gender-related genetic action.

3.6.1.2 Dataset

The dataset consisted of 1162 unrelated MS patients from France, Germany, Italy and Spain. Patients with missing response status (1), genotypes (22), gender (22) and non-RRMS patients (3) defined as having EDSS at treatment onset > 6 were excluded from the analyses. The final study dataset consisted of 1115 patients. All patients were genotyped for the SNP rs2660 (A/G) in the *OAS1* gene.

3.6.1.3 Response Definition

Response was determined based on the EDSS progression and the number of relapses over two years of treatment (see **Figure 1.6**). Let t_0 be the time at treatment onset and let t_R be the time at which response is evaluated. Here, $t_R = 2$ years. Let $EDSS_t$ be the EDSS progression since t_0 evaluated at time t and let NR_t be the number of relapses experienced up to time t since t_0. Then, *Responders* were defined as

$$Responders = \begin{cases} EDSS_{t_R} \leq 0.5; \text{and} \\ \quad NR_{t_R} \leq 1 \end{cases} \tag{3.54}$$

All patients not classified as Responders were classified as *Non-Responders*. Clinical characteristics by response status and cohort are provided in Appendix II (**Table II.1-Table II.4**).

3.6.1.4 Methods

Prior all analyses, we tested the SNP rs2660 for deviations from HWE using the HWE exact test. We first conducted a stratified analysis by cohort using the CMH test (Equation (3.32)). We computed ORs and 95% confidence intervals (CI) with allele A as the reference allele. We tested for homogeneity of ORs using the BD test (Equation (3.34)). Pooling cohorts with homogeneous effects, we then used logistic regression to test for association including cohort as a covariate and modeling different genetic models. All genetic models we considered are described in **Table 3.5** except for the overdominant model under which both homozygous genotypes (A/A and G/G) are baseline ($OR =$

1). The LRT (Equation (3.40)) was used to evaluate the significance of these findings. Lastly, we also explored whether there was heterogeneity of genetic effects according to gender by conducting logistic regression analyses in female and male patients separately.

Analyses were carried out using PLINK (*Purcell et al., 2007*) and the R statistical software (*R Core Team, 2013*) and, specifically, the *SNPassoc* package (*González et al., 2007*).

3.6.1.5 Results

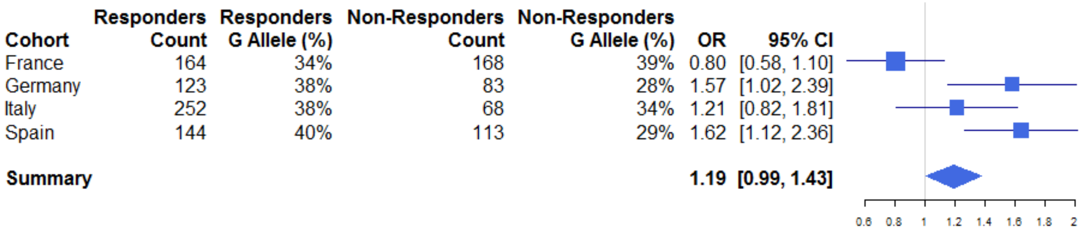
The SNP rs2660 is an A/G polymorphism with the minor allele G having a frequency of 0.36 in the European population (HapMap Release #28). The genotype distributions by response status and by cohort are given in **Table 3.7** below. No deviations from HWE were observed (HWE exact test, all p-values > 0.05).

ALL PATIENTS		Number of Patients (% Responders)	Allele		Genotype			HWE Exact Test p-value
			A	G	AA	AG	GG	
France								
	Responders	**164**	66%	34%	43%	45%	12%	1.00
	Non-Responders	**168**	61%	39%	38%	46%	16%	0.75
	Overall	***332 (49%)***	*63%*	*37%*	*40%*	*46%*	*14%*	*0.81*
Germany								
	Responders	**123**	62%	38%	39%	46%	15%	0.70
	Non-Responders	**83**	72%	28%	47%	49%	4%	0.06
	Overall	***206 (60%)***	*66%*	*34%*	*42%*	*47%*	*11%*	*0.64*
Italy								
	Responders	**252**	62%	38%	39%	45%	16%	0.43
	Non-Responders	**68**	66%	34%	47%	38%	15%	0.28
	Overall	***320 (79%)***	*63%*	*37%*	*41%*	*43%*	*16%*	*0.19*
Spain								
	Responders	**144**	60%	40%	35%	51%	14%	0.49
	Non-Responders	**113**	71%	29%	53%	36%	11%	0.25
	Overall	***257 (56%)***	*65%*	*35%*	*43%*	*45%*	*12%*	*0.89*
All Patients								
	Responders	**683**	62%	38%	39%	46%	14%	0.81
	Non-Responders	**432**	66%	34%	45%	43%	12%	0.52
	Overall	***1115 (61%)***	*64%*	*36%*	*41%*	*45%*	*13%*	*0.48*

Cell counts < 5

Table 3.7: Allele and genotype frequencies by response status and cohort for the MS patients included in the study. The p-value for the HWE exact test is given as well.

The ORs and the 95% confidence intervals per cohort and for all patients adjusted for cohort are presented in a forest plot in **Figure 3.4** below.



Cohort	Responders Count	Responders G Allele (%)	Non-Responders Count	Non-Responders G Allele (%)	OR	95% CI
France	164	34%	168	39%	0.80	[0.58, 1.10]
Germany	123	38%	83	28%	1.57	[1.02, 2.39]
Italy	252	38%	68	34%	1.21	[0.82, 1.81]
Spain	144	40%	113	29%	1.62	[1.12, 2.36]
Summary					**1.19**	**[0.99, 1.43]**

Figure 3.4: Forest plot of the stratified analyses by cohort evaluating the association between the *OAS1* SNP rs2660 and response to interferon-β.

Overall, there was no evidence that the G allele was associated with response after adjusting for cohort (CMH test, p-value = 0.0651). However, there was statistical evidence of non-homogeneous effects across cohorts (BD test, p-value = 0.016), with opposite direction of the effect in France and in the remaining cohorts.

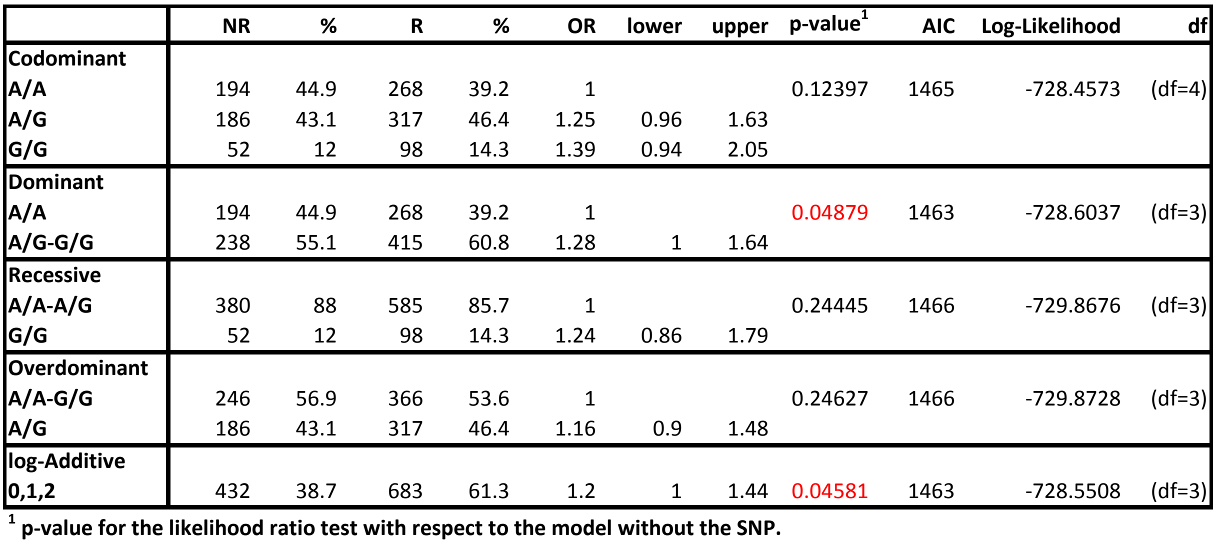
To investigate whether the heterogeneity of effects was arising due to one of the cohorts, we repeated the analyses four times, each time excluding one of the four cohorts. These analyses revealed that the genetic effects in the French cohort were significantly different from those in the remaining cohorts (BD test, p-value = 0.53810 when France was excluded from the analyses but p-value < 0.05 when France was included, see **Table 3.8**).

Excluded Cohort	Cochran-Mantel-Haenszel			Breslow-Day
	OR	95% CI	P-value	P-value
France (n=332)	1.46	(1.16, 1.83)	0.0012	0.53810
Germany (n=206)	1.11	(0.91, 1.37)	0.30	0.01581
Italy (n=320)	1.18	(0.96, 1.46)	0.11	0.00579
Spain (n=257)	1.07	(0.87, 1.33)	0.51	0.03491

Table 3.8: CMH and BD test results after excluding one cohort at a time.

We pooled the German, Italian and Spanish cohorts into one as *Non-France*. From here onwards, unless otherwise indicated, by "cohort" we refer to the France/Non-France grouping of the patients.

Table 3.9 below provides the association results from the logistic regression analyses. We observe weak evidence of association under the dominant and the log-additive models (LRT, p-values 0.049 and 0.046, respectively).



	NR	%	R	%	OR	lower	upper	p-value[1]	AIC	Log-Likelihood	df
Codominant											
A/A	194	44.9	268	39.2	1			0.12397	1465	-728.4573	(df=4)
A/G	186	43.1	317	46.4	1.25	0.96	1.63				
G/G	52	12	98	14.3	1.39	0.94	2.05				
Dominant											
A/A	194	44.9	268	39.2	1			0.04879	1463	-728.6037	(df=3)
A/G-G/G	238	55.1	415	60.8	1.28	1	1.64				
Recessive											
A/A-A/G	380	88	585	85.7	1			0.24445	1466	-729.8676	(df=3)
G/G	52	12	98	14.3	1.24	0.86	1.79				
Overdominant											
A/A-G/G	246	56.9	366	53.6	1			0.24627	1466	-729.8728	(df=3)
A/G	186	43.1	317	46.4	1.16	0.9	1.48				
log-Additive											
0,1,2	432	38.7	683	61.3	1.2	1	1.44	0.04581	1463	-728.5508	(df=3)

[1] **p-value for the likelihood ratio test with respect to the model without the SNP.**

Table 3.9: Association results for logistic regression analysis including cohort as a covariate and modeling different genetic models. Statistically significant results from the LRT (uncorrected p-value < 0.05) are highlighted in red. AIC: Akaike Information Criterion; R: Responders; NR: Non-Responders.

Gender was not associated with response (χ^2 test on 1 df, p-value=0.26), with cohort (χ^2 test on 1 df, p-value=0.13) and with the SNP (χ^2 test on 1 df, p-value=0.59). Deviations from HWE were observed in the male cohort for all patients combined (HWE exact test, p-value = 0.026).

The results from the logistic regression analyses in females (793 patients) and males (322 patients) separately are given in **Table 3.10** **(A)** and **Table 3.10** **(B)**, respectively. The association was significant under the overdominant model in male patients (LRT, p-value = 0.013) but not in female patients (LRT, p-value = 0.79). Conversely, the association was borderline significant under the recessive model in female patients (LRT, p-value = 0.048) but not in the male patients (LRT, p-value = 0.32).

(A)

	NR	%	R	%	OR	lower	upper	p-value[1]	AIC	Log-Likelihood	df
Codominant											
A/A	142	44.9	197	41.3	1			0.12726	1059	-525.4895	(df=4)
A/G	139	44	204	42.8	1.07	0.79	1.46				
G/G	35	11.1	76	15.9	1.59	1.01	2.51				
Dominant											
A/A	142	44.9	197	41.3	1			0.26836	1060	-526.9385	(df=3)
A/G-G/G	174	55.1	280	58.7	1.18	0.88	1.57				
Recessive											
A/A-A/G	281	88.9	401	84.1	1			0.04777	1057	-525.592	(df=3)
G/G	35	11.1	76	15.9	1.53	1	2.36				
Overdominant											
A/A-G/G	177	56	273	57.2	1			0.79172	1061	-527.5162	(df=3)
A/G	139	44	204	42.8	0.96	0.72	1.28				
log-Additive											
0,1,2	316	39.8	477	60.2	1.21	0.98	1.48	0.07733	1058	-525.991	(df=3)

[1] p-value for the likelihood ratio test with respect to the model without the SNP.

(B)

	NR	%	R	%	OR	lower	upper	p-value[1]	AIC	Log-Likelihood	df
Codominant											
A/A	52	44.8	71	34.5	1			0.0461	402.4	-197.1777	(df=4)
A/G	47	40.5	113	54.9	1.8	1.08	3.01				
G/G	17	14.7	22	10.7	0.96	0.45	2.04				
Dominant											
A/A	52	44.8	71	34.5	1			0.06326	403.1	-198.5297	(df=3)
A/G-G/G	64	55.2	135	65.5	1.58	0.98	2.56				
Recessive											
A/A-A/G	99	85.3	184	89.3	1			0.31633	405.5	-199.7526	(df=3)
G/G	17	14.7	22	10.7	0.7	0.35	1.41				
Overdominant											
A/A-G/G	69	59.5	93	45.1	1			0.01319	400.4	-197.1832	(df=3)
A/G	47	40.5	113	54.9	1.82	1.13	2.94				
log-Additive											
0,1,2	116	36	206	64	1.17	0.82	1.68	0.38307	405.7	-199.8742	(df=3)

[1] p-value for the likelihood ratio test with respect to the model without the SNP.

Table 3.10: Association results for logistic regression analysis in female (**A**) and male patients (**B**) including cohort as a covariate and modeling different genetic models. Statistically significant results from the LRT (uncorrected p-value < 0.05) are highlighted in red. AIC: Akaike Information Criterion; R: Responders; NR: Non-Responders.

3.6.1.6 Concluding Remarks

In the combined sample, the association of *OAS1* SNP with interferon response was borderline significant under the dominant and the log-additive models (LRT, p-values < 0.05). Using the Quanto software (*Gauderman and Morrison, 2009*), we calculated the power to detect association when the true model is log-additive or dominant in the female and male patients at the $\alpha = 0.05$ level of significance. We assumed homogeneous effects of 1.28 (dominant) and 1.20 (log-additive) for females and males. In the female patients, assuming a sample size of 793 individuals, the power was less than 40% under both the genetic models. In the male patients, assuming a sample size of 322 individuals, the power was less than 20% under both genetic models. That is, if an association exists under the dominant or the log-additive model, we have very low power to detect it. Therefore, although the separate analyses by gender appeared to suggest possible difference in the effects of *OAS1* (rs2660) in interferon response according to gender, the possibility that the underlying genetic model in both genders is the same (dominant or log-additive) cannot be excluded. At present, this study awaits finalization as an additional cohort from Italy will be added to the analyses.

A puzzling result is the evidence of heterogeneity between the French and the remaining cohorts. Indeed, *OAS1* SNP genetic effects on response to interferon were found heterogeneous with opposite direction in the France and Non-France cohorts. We performed additional analyses attempting to delineate the source of this heterogeneity.

We first investigated the epidemiological and/or clinical characteristics of the French and the non-French patients. In fact, and as shown in **Table II.5** (Appendix II), several of the tested variables significantly differ between the two cohorts. Specifically, the variables are: age at disease onset, disease severity (both EDSS at treatment onset and number of relapses 2-year prior treatment onset) and type of interferon, with French patients having an older age at disease onset, a more severe MS (higher EDSS and number of relapses) and about half of them have been treated with Avonex® (less than a third of the non-French patients received the same type of interferon). Therefore, we repeated our association analysis adjusting for clinical and epidemiological variables with and without a France/Non-France cohort variable to identify the variables that contribute the most to the evidence of heterogeneity. We excluded age at disease onset and number of relapses because of a high rate of missing data. Note that age at disease onset is, nonetheless, highly correlated with age at treatment onset in both cohorts (correlation > 0.80). Adjusting for interferon type does not improve the model containing only the SNP (LRT, p-value = 0.055) but adjusting for age at treatment onset or for EDSS at treatment onset does (LRT, p-values = 5.9E-04 and 1.7E-05, respectively). However, the models including one of these two covariates have a worse fit than the model including both of them (LRT, p-value = 2.3E-05 and = 6.9E-07 for age at treatment onset and EDSS, respectively). Yet, the France/Non-France cohort variable remains highly significant (LRT, p-value=1.1E-06) thus suggesting that factors other than age and EDSS at treatment onset underlie the genetic heterogeneity at *OAS1*.

Table 3.11 shows the estimated ORs under the different investigated models. Under all but one (including age at treatment onset) multivariate models the SNP effect was more significant.

In summary, despite adjusting for clinical/epidemiological characteristics, the genetic effects of *OAS1* on response to treatment remain heterogeneous between the France and the Non-France cohorts. From our data, we have been unable to isolate the underlying reason for this heterogeneity.

Model Including	**SNP Effect**	
	OR (95% CI)	**P-value (Wald test)**
SNP	1.20 (1.01, 1.44)	0.045
Age at treatment onset + SNP	1.19 (1.00, 1.43)	0.053
EDSS at treatment onset + SNP	1.22 (1.02, 1.47)	0.031
Age at treatment onset + EDSS at treatment onset + SNP	1.22 (1.01, 1.46)	0.036
Cohort + Age at treatment onset + EDSS at treatment onset + SNP	1.22 (1.02, 1.47)	0.033

Table 3.11: The estimated SNP effect and its significance after adjusting for various covariates.

Finally, we evaluated the sensitivity of our association results to the response definition. Various response definitions can be found in the literature. Here, we used the same definition as used by *O'Brien et al. (2010)*, *Couturier et al. (2011)* and *Malhotra et al. (2011)*. They are summarized as Alternative Definitions 1, 2 and 3, respectively, in **Table 3.12** below. These definitions apply increasingly stricter threshold criteria on $EDSS_{t_R}$ and/or NR_{t_R} as compared to our threshold criteria. Consequently, under these alternative response definitions, a larger fraction of patients are classified as “unknown” with respect to the phenotype status. For instance, over 40% of our patients are unclassified under Alternative Definition 3, thus far reducing the effective sample size of subsequent association analyses.

		Responders		**Non-Responders**		**Unclassified Patients n (%)**
		Definition	**Number of Patients**	**Definition**	**Number of Patients**	
Study Definition		$EDSS_{t_R} \leq 0.5$; and $NR_{t_R} \leq 1$	683	$EDSS_{t_R} > 0.5$; or $NR_{t_R} > 1$	432	0 (0%)
Alternative Definitions	**1**	$EDSS_{t_R} \leq 0$; and $NR_{t_R} \leq 1$	604	$NR_{t_R} > 1$	332	179 (16%)
	2	$EDSS_{t_R} \leq 0$; and $NR_{t_R} = 0$	439	$EDSS_{t_R} > 0.5$; or $NR_{t_R} > 1$	432	244 (22%)
	3	$EDSS_{t_R} \leq 0$; and $NR_{t_R} = 0$	439	$EDSS_{t_R} > 0.5$; and $NR_{t_R} \geq 1$	220	456 (41%)

Table 3.12: Alternative response definitions and how they compare to the definition we adopted for this study.

The statistical significance of *OAS1* SNP association with interferon response decreased (p-values = 0.09, 0.07 and 0.11 under definition 1, 2 and 3, respectively). This decrease is likely explained by the reduction of the effective sample size. Indeed, the SNP effect estimates remain similar (OR = 1.19, 1.20 and 1.22 under definition 1, 2 and 3, respectively) and close to the value (OR = 1.19) obtained under our response definition. These results suggest that our association finding is robust to the threshold criteria used to define response to interferon.

3.6.2 Role of the integrin alpha 4 subunit (*ITGA4)* gene in natalizumab response in MS patients

3.6.2.1 Objective

Migration of blood into tissues is facilitated by receptors such as the Very Late Antigen-4 (VLA4, α4 β1 integrin, integrin α/β complex). Natalizumab binds to VLA4 and prevents the migration of immune cells into the central nervous system. Thus, because of its mechanism, the ITGA/ITGB complex has been proposed as a reasonable candidate region to investigate the role of genetic variants in natalizumab response in MS patients. (*Pappas and Oksenberg, 2010*) The objective of this study was to investigate whether the integrin alpha 4 (*ITGA4*) gene is implicated in response to natalizumab.

3.6.2.2 Dataset

A total of 904 French RRMS patients from the BIONAT cohort *(Outteryck et al., 2013*) were genotyped for 94 SNPs in the *ITGA4* gene. During quality control (QC) analyses, noninformative SNPs, SNPs which failed the HWE or had high missing genotype rates were removed. Similarly, individuals with high missing genotype rates were removed. A total of 894 individuals genotyped at 60 SNPs were included in the study, post-QC.

3.6.2.3 Response Definition

The response definition for natalizumab-treated patients is much more complex than for interferon-treated patients. It relies on clinical as well as radiological measures. Let $t_0, t_R, EDSS_t$ and NR_t be as defined before (see Section 3.6.1.3). Further, let $MRIT2_t$ be the new T2 lesions observed at time t with respect to t_0 and let $MRIGD_t$ indicate the presence/absence of gadolinium enhancing (GD$^+$) lesions at time t. In this study, response was evaluated at one year after treatment onset ($t_R = 1$ year). *Responders, Non-Responders and Intermediary Responders* were defined as described in **Table 3.13** below.

Responders	***Non-Responders***	***Intermediary Responders***
$EDSS_{t_R} < 0.5$; and	$EDSS_{t_R} > 0.5$; or	$EDSS_{t_R} = 0.5$; or
$NR_{t_R} = 0$; and	$NR_{t_R} \geq 1$ after 3rd infusion; or	$NR_{t_R} = 1$ before 3rd infusion
$MRIT2_{t_R} = 0$	$MRIT2_{t_R} > 1$; or	-
$MRIGD_{t_R} = absence$	$MRIGD_{t_R} = presence$	-

Table 3.13: Response classification for natalizumab-treated patients.

Note that *Intermediary Responders* include patients who experienced relapses within the first three months of treatment. Natalizumab is administered through intravenous infusions on a monthly basis and a period of three months, that is, three infusions, could be allowed before evaluating treatment effectiveness.

Inclusion Criteria I. *Intermediary Responders* were excluded from this study. Only patients who received treatment for at least one year were included in the study. Thus, of the 894 patients, only 734 (431 Responders, 303 Non-Responders) met the inclusion criteria for the study.

Inclusion Criteria II. Over time patients may develop antibodies against natalizumab thus reducing treatment efficacy. Thus, in addition to Inclusion Criteria I, we imposed an additional inclusion criterion including only patients who had negative antibody status, that is, they had not develop antibodies against natalizumab. This reduced further the study sample size to 579 patients (324 Responders, 255 Non-Responders). Our conclusions were unchanged and we only report our findings under Inclusion Criteria I.

3.6.2.4 Methods

We conducted logistic regression analyses under the dominant and recessive models and also without assuming any underlying genetic model (genotypic coding, **Table 3.5**). Significance of the SNP-response association results was evaluated using the LRT (Equation (3.40)). All QC and association analyses were carried out with PLINK (*Purcell et al., 2007*). We also studied the LD structure of the investigated region based on the r^2 measure (Equation (2.6)) using Haploview (*Barrett et al., 2005*).

3.6.2.5 Results

We found that only one of all 60 tested SNPs, rs155106, was associated with natalizumab response at the nominal significance level (uncorrected p-values of 0.045 and 0.032 under the dominant and genotypic coding, respectively, and 0.34 under the recessive coding). The results under the dominant genetic model are illustrated in **Figure 3.5** below.

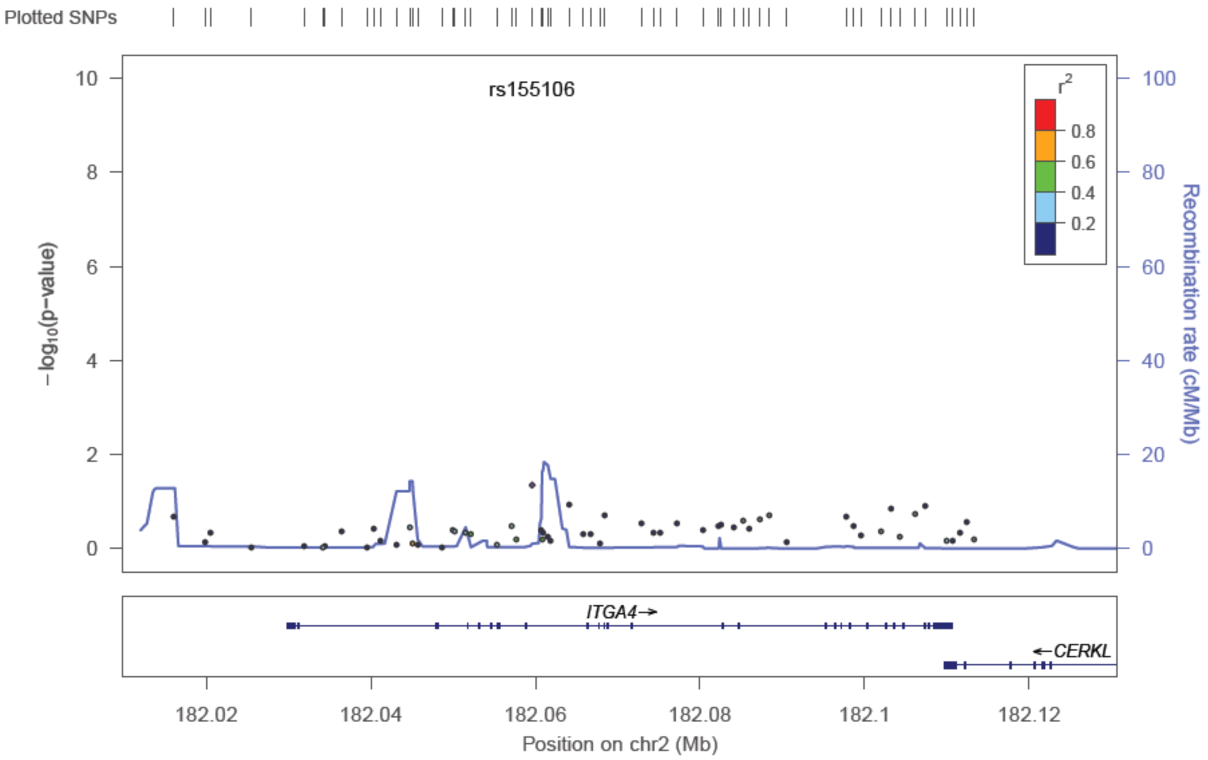


Figure 3.5: Association test results under the dominant genetic model from the *ITGA4* candidate gene study on response to natalizumab. Results are expressed as the negative logarithm of the p-value of the association test for all 60 SNPs tested. The most significant finding is indicated at the top center of the graph, the SNP rs155106. Graph generated with LocusZoom (*Pruim et al., 2010*).

The LD structure of the investigated region is presented in **Figure 3.6**. The SNP rs155106 is circled in blue on the graph. A preliminary scan on fewer patients was carried out by former colleagues (*Couturier, 2010*) and two plausible SNPs had been identified. These two SNPs are also circled, in red, in the figure below.

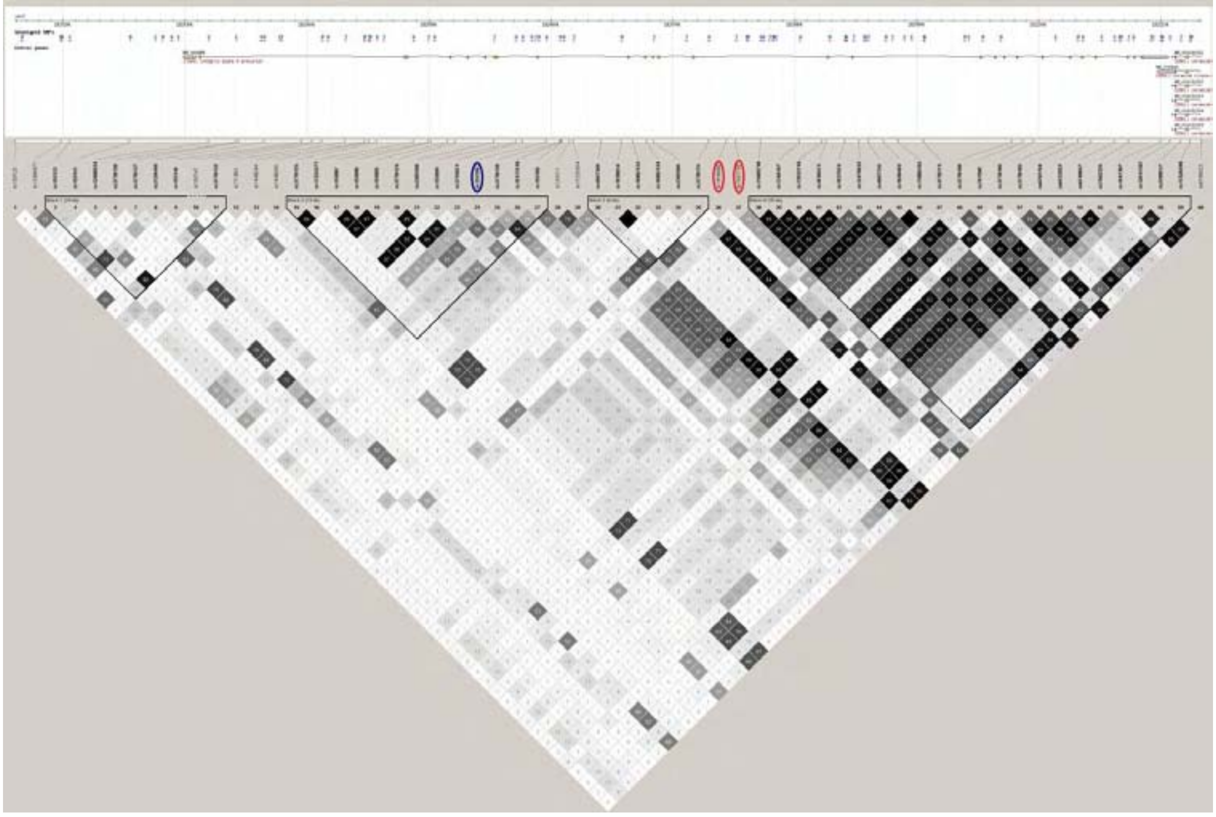


Figure 3.6: LD plot (based on r^2, Equation (2.6)) for the *ITGA4* gene. The SNP rs155106 from the candidate gene study (circled in blue) and two SNPs from the preliminary study conducted by (*Couturier, 2010*) (circled in red) are noted. Graph generated with Haploview (*Barrett et al., 2005*).

From **Figure 3.6**, we can see that the most significantly associated SNP from this study, rs155106 (circled in blue), is located in a block of weak LD. Note that this was also apparent from **Figure 3.5**. It can also be seen from **Figure 3.6** that the two previously suggested SNPs (circled in red on the graph) from the preliminary analyses appear to be independent of rs155106.

3.6.2.6 Concluding Remarks

Overall, given the large number of tests carried out (multiple SNPs and two different inclusion criteria scenarios), we concluded that this study failed to provide evidence that the *ITGA4* gene was implicated in response to natalizumab in our cohort of French MS patients.

We note that, contrary to the study on interferon response, here we evaluated only one response definition which also included radiological measures. While no consensus on the actual response definition exists, it is believed that combining radiological measures (when available) with clinical measures can lead to more precise evaluation of treatment response. (*Río et al., 2009*) This is because the disease may be active even if there are no clinical manifestations of it.

Dropping the radiological criteria from our response definition may reclassify Non-Responders into Responders but never the other way around. Using both clinical and radiological criteria, there were 431 Responders and 303 Non-Responders (under Inclusion Criteria I). Without the radiological criteria, 41 Non-Responders would be reclassified or, even worse, "misclassified" as Responders.

Genome-wide data for all patients from the BIONAT cohort (*Outteryck et al., 2013*) are being generated at the time of writing this manuscript and a GWAS of natalizumab response is planned ahead. We describe this upcoming work in greater detail in the concluding chapter of this thesis.

3.6.3 Multi-Marker Modeling of GWAS Data on Parkinson's Disease

3.6.3.1 Background

Hundreds of thousands to millions of SNPs nowadays are tested in a sample of tens of thousands of individuals (ideally). The sample sizes are unlikely to ever reach the number of SNPs tested. Thus, in order to carry out simultaneous analyses of all SNPs, we run into the high dimensionality, $p \gg n$, problem where p represents the number of SNPs and n the number of individuals. In a multivariate framework, when $p \rightarrow n$ or $p \gg n$, classical regression methods fail and alternative modeling approaches reducing the dimension of the parameter space need to be used. Two main classes of such approaches are penalized estimation methods and Bayesian estimation methods.

In penalized estimation methods, a penalty function is applied to shrink the marker effect estimates towards zero relative to their maximum likelihood estimates. Some methods apply the same large penalty to all estimates whereas others apply large penalties to some estimates and small penalties to the remaining estimates. Thus, the defining factor for these methods is the choice of the penalty function.

Bayesian estimation methods, on the other hand, require the specification of a prior distribution on the marker effects. There is a close relationship between these two classes of approaches. In particular, certain penalty functions lead to empirically equivalent estimates under certain priors.

We were thus interested in evaluating how the multi-marker approaches would fare compared to the single-marker approach. We note that all of these more sophisticated approaches have been primarily used for estimating breeding values in the animal and plant breeding literature and only recently their application to human genetics has been considered.

3.6.3.2 Materials and Methods

We compared several multi-marker approaches with a focus on Bayesian estimation methods using GWAS data on Parkinson's disease (PD) of 3023 French subjects (1039 cases, 1984 controls) and almost 500 000 SNPs. A recent GWAS conducted by our colleagues on this dataset had confirmed the association of known PD genes such as *SNCA* and had identified novel ones such as *BST1* in the European population (*Saad et al., 2011a*).

We treated the disease status as a continuous outcome variable. The linear model is given by

$$\boldsymbol{y} = \boldsymbol{W\alpha} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{3.55}$$

where $\boldsymbol{y}$ is $n \times 1$ vector of phenotypes, $\boldsymbol{W}$ is the $n \times q$ incidence matrix, $\boldsymbol{\alpha}$ is $q \times 1$ vector of covariates, $\boldsymbol{X}$ is a $n \times p$ matrix of observed genotypes (typically coded under the additive genetic model), $\boldsymbol{\beta}$ is the $p \times 1$ vector of SNP effects and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of random residual effects where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

In a Bayesian regression setting, a prior on $\beta_i, i = 1, \ldots, p$ that depends on the SNP-specific variance, $\sigma_{\beta_i}^2$, and on the proportion of SNPs, π, expected to be associated with the complex trait is specified. This prior function is given by

$$\beta_i|\pi,\sigma^2_{\beta_i} \sim \begin{cases} 0 \quad with\,probability\,(1-\pi) \\ N\left(0,\sigma^2_{\beta_i}\right) \quad with\,probability\,\pi \end{cases}. \tag{3.56}$$

Depending on the choice of these two parameters, namely, $\sigma^2_{\beta_i}$ and π, different Bayesian models have been proposed and have sometimes been referred to under the umbrella term "The Bayesian Alphabet" coined by (*Gianola et al., 2009*). Such models include but are not limited to Bayes A and B (*Meuwissen et al., 2001*) and later extensions Bayes C/Cπ and D/Dπ (*Habier et al., 2011*). **Table 3.14** compares the parameterization of these four models for illustrative purposes.

Model	$\boldsymbol{\sigma^2_{\beta_i}}$	$\boldsymbol{\pi}$
Bayes A	SNP-specific variance, $\sigma^2_{\beta_i}, i = 1, \dots, p$	$\pi = 1$
Bayes B	SNP-specific variance, $\sigma^2_{\beta_i}, i = 1, \dots, p$	$\pi < 1$
Bayes C π	common variance to all SNPs, $\sigma^2_{\beta_i} = \sigma^2_{\beta}, i = 1, \dots, p$	$\pi \sim Uniform(0,1)$ (i.e., treated as unknown)
Bayes D π	SNP-specific variance, $\sigma^2_{\beta_i}, i = 1, \dots, p$	$\pi \sim Uniform(0,1)$ (i.e., treated as unknown)

Table 3.14: Comparing the parameterization of Bayes A, B, C π and D π models.

The difference between the models lies in whether $\boldsymbol{\pi}$ is treated as fixed and known or as random and unknown and in whether the SNP effects follow a common or SNP-specific random distribution. Regardless, the variance in all four models is assumed to follow the scaled inverse χ^2 distribution with ν degrees of freedom.

We compared three multi-marker models. The parameter specifications of the three models are given in **Table 3.15** below. All models were fitted using the GS3 software (*Legarra et al., 2011*).

Model	$\boldsymbol{\sigma^2_{\beta_i}}$	$\boldsymbol{\pi}$	**Note**
I	common variance to all SNPs, $\sigma^2_{\beta_i} = \sigma^2_{\beta}, i = 1, \dots, p,$ fixed, known	$\pi = 1$	Reduces to the popular Best Linear Unbiased Predictor (BLUP) (*Henderson, 1975*)
II	common variance to all SNPs, $\sigma^2_{\beta_i} = \sigma^2_{\beta}, i = 1, \dots, p,$ random, unknown	$\pi \sim Uniform(0,1)$ (i.e., treated as unknown)	Bayes C π
III	common variance to all SNPs, $\sigma^2_{\beta_i} = \sigma^2_{\beta}, i = 1, \dots, p,$ random, unknown	$\pi = 0.0001$ fixed, known	Bayes C

Table 3.15: Multi-marker models we evaluated on the PD GWAS dataset.

A numerical iterative method (Gauss-Siedel) was used to derive the estimates from Model I and a Markov Chain Monte Carlo algorithm (Gibbs sampler) was used for Models II and III (as implemented in the GS3 software). We used 100 000 iterations for Model II and 500 000 iterations for Model III after ignoring the first 5 000 iterations (burn-in). Convergence criteria (difference between solutions of successive iterations less than 10^{-8}) were achieved within 100 iterations for Model I. Convergence for Models II and III was visually inspected using traceplots (plotting the sampled parameter value at each iteration against the iteration number).

Lastly, to evaluate the evidence of association of the estimated effects by the multi-marker models, we ranked the SNPs by the absolute value of their effect estimates.

3.6.3.3 Results

Focusing on *SNCA*, *BST1* as well as on two other known PD genes *(MAPT* and *LRRK2*), we compared the ranking of the SNPs in these genes obtained by the three methods and also compared them to the ranks obtained based on the p-values of the estimated effects by the GWAS. Based on the results illustrated in **Table 3.16**, the three multi-marker models appeared to detect the known PD genes.

The ranks for all SNPs were relatively highly correlated between methods. For example, the correlation between Model I and GWAS ranks was 0.43. However, the best SNP rank within a gene was not found at the same SNP. Importantly, the SNP effect estimates were on a very different scale depending on the model as illustrated, for example, in **Figure 3.7** comparing Models II and III.

Gene ± 1 Mb	**Number of SNPs**	**Model I**		**Model II**		**Model III**		**GWAS**		
		$\widehat{\beta}$	**Rank**	$\widehat{\beta}$	**Rank**	$\widehat{\beta}$	**Rank**	$\widehat{\beta}$	**P-value**	**Rank**
SNCA	323	-1.8 E-04	1	-5.0 E-04	1	-0.0537	2	0.29	1.3 E-07	1
BST1	455	1.4 E-04	4	3.6 E-04	8	0.0149	16	-0.27	2.1 E-06	6
MAPT	330	1.3 E-04	8	3.4 E-04	18	0.0098	26	-0.29	7.30 E-06	8
LRRK2	713	9.4 E-05	504	2.5 E-04	683	-0.0016	126	0.3535	0.000271	168

Table 3.16: The SNP effect estimate and the corresponding SNP rank within known PD genes (± 1Mb) for the three multi-marker models, Models I, II, III and GWAS.

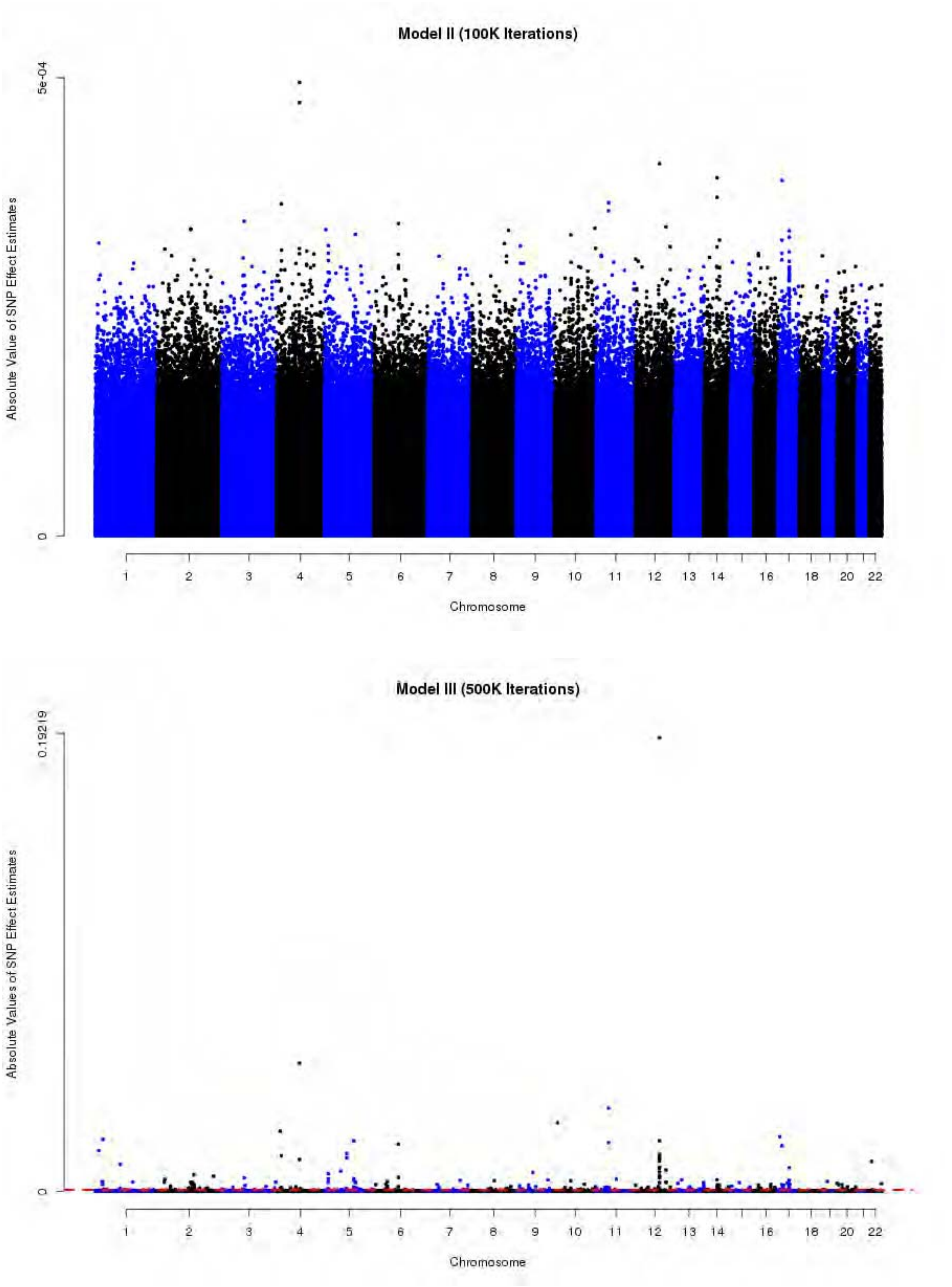


Figure 3.7: Absolute value of SNP effect estimates obtained by Model II (top) and Model III (bottom) plotted against the SNP position. The red dashed line on the graph of Model III corresponds to the maximum absolute SNP effect estimate under Model II.

It took a couple of minutes to run Model I, while it took at least 4 days per 100 000 iterations for Model II (used 100 000 iterations) and Model III (used 500 000 iterations). Running times were slower for Model II than for Model III due to the additional unknown parameter (the proportion of SNPs having an effect, π). For Model III, the SNP effect estimates remained relatively unchanged beyond 300 000 iterations. For Model II, the proportion of SNPs having an effect, π, appeared to reach a plateau at just about 100 000 iterations (results not shown).

3.6.3.4 Discussion and Conclusions

In this study, we investigated three different multi-marker models (Models I, II and III) to detect SNP-disease associations in a French Parkinson’s disease GWAS dataset. Focusing on known associated genes, we also compared the results from these models to the results from a recent GWAS conducted in this same dataset.

In Model I, the SNP effect variance was treated as a known and fixed quantity and all SNPs were assumed apriori to exert an effect on the trait. In Models II and III, the SNP effect variance was unknown and estimated from the data. In Model II, the proportion of SNPs exerting an effect was treated as unknown while, in Model III, it was assumed to be 0.0001 (only about 50 SNPs influence the trait).

Contrary to the single-marker approach, there is no formal statistical test to evaluate the significance of the SNP-specific effects in multi-marker models. To circumvent that, we ranked the SNPs by the absolute value of their effect estimate and compared the SNP ranks between the models.

Our main observations were as follows. First, focusing on four known genes (*SNCA*, *BST1*, *MAPT* and *LRRK2*), we found that the SNP ranks from Model I were closest to the SNP ranks from the GWAS (obtained by ranking the p-value of the association test). However, the best rank was not observed at the same SNP.

Second, the scale of the SNP effect estimates obtained by Model III was a couple of hundred times higher than that of the other two multi-marker models. This was not surprising since in Model I, for instance, we assumed that all SNPs (approximately half a million) exert an effect, while in Model III we limited this number to only 50 SNPs. Thus, the overall additive genetic variance was spread over fewer SNPs in Model III.

Third, the required number of iterations varied by model in the order of 100 for Model I and in the order of 100 000 for Models II and III. Consequently, the running times for Models II and III were longer, in the order of days, while Model I estimates were obtained within a couple of minutes. For Model III, 500 000 iterations appeared largely sufficient as no apparent difference in the SNP effect estimates were observed beyond 300 000 iterations. Compared to Model III, Model II had an additional unknown parameter, the proportion of SNPs exerting an effect. Although 100 000 iterations appeared to be sufficient for this model, it would have been interesting to carry out a higher number of iterations.

Overall, our study suggests that the multi-marker models we evaluated may seem a promising tool to detect association of genetic variants with complex traits. However, important considerations include the fact that there is no formal statistical test evaluating significance of the SNP effects, complex parameterization and heavy computational burden. Therefore, the advantages of multi-marker models over the classical GWAS approach remain to be fully investigated.

This work was presented as an oral communication at the European Mathematical Genetics Meeting in 2012 and as a poster at the 21st Annual International Genetic Epidemiology Society Meeting in 2012.

3.6.4 Polygenic Score Analyses of Simulated Diastolic Blood Pressure Data

3.6.4.1 Background

The one SNP at a time analytical approach adopted by GWASs does not have sufficient power to detect SNPs of weak effects at the imposed stringent genome-wide statistical significance levels. Therefore, it is possible that a substantial number of causal SNPs remain undetected by GWASs. A recent method that has drawn considerable attention so far proposes to account for SNPs having a wide spectrum of effects by aggregating them into a polygenic score (PS) and testing the PS for association with the trait. (*Purcell et al., 2009*)

Typically, the PS is constructed in two steps. First, the set of SNPs to be included in the score is selected. The criteria for SNP selection vary between studies but it is crucial that this set contains only independent SNPs to avoid the inclusion of non-independent association signals. Further, this set must exclude all established variants as they would drive the association of the PS with the trait masking the weaker effects that we are precisely looking to detect. Second, the reference alleles of these SNPs may be combined in an unweighted or weighted manner. The former approach assumes that all SNPs have the same effect size which oversimplifies the context we are trying to evaluate (that is, a mixture of different effect sizes). In the latter approach, on the other hand, each reference allele is weighted by its effect estimated in a discovery dataset.

Based on this theoretical framework, we conducted a study as part of our participation in the Genetic Analysis Workshop 18 (GAW18), where we compared PS association analyses using sets of SNPs derived from a single-marker and a multi-marker approach. We were interested in evaluating the value of PS association analysis in shedding light on the true genetic architecture of complex traits.

3.6.4.2 Materials and Methods

Study Dataset. We used the pedigree dataset provided by GAW18 on simulated blood pressure data for roughly 900 individuals (from 20 families) genotyped at more than 8.3 million SNPs. A trait, Q1, was highly heritable (heritability of 0.68) but was uninfluenced by any of the SNPs and was provided to control for Type I Error. We adjusted the simulated traits for age and gender in a linear regression framework. A total of 200 replicate datasets were made available but the genotypes were the same across the replicates.

We carried out all our analyses with knowledge of the underlying simulated model. There were 1 457 SNPs (in 288 genes) contributing to diastolic blood pressure (DBP) and/or systolic blood pressure (SBP) variability. The individual contribution of the genes ranged from as low as 0.001% for gene *ZZEF1* to as high as 6.5% for gene *MAP4* (for DBP). In addition, part of the total variability was due to 1 000 SNPs, randomly selected in each replicate.

Methods: Analyses Restricted to Associated SNPs/Genes. We first estimated the power to detect association with any of the contributing genes accounting for family relatedness using the Measured Genotype (MG) test (mixed-linear regression model) (*Boerwinkle and Sing, 1987*), as implemented in QTDT software (Abecasis et al., 2000a; Abecasis et al., 2000b). Single-marker MG test was conducted for each SNP using all 200 replicates. We found that *MAP4* was the only gene detectable (power = 96%) at the genome-wide significance level (p-value < 1E-08), which accounts for the largest percentage of the variance of DBP and contains several SNPs three of which had very strong individual effects each contributing more than 1% to the trait variability. Any of the remaining SNPs or genes were unlikely to be detected at stringent significance criteria (power < 50%).

We estimated that we would need approximately 40 days to carry out one genome-wide association study using the MG test in one replicate for one phenotype. We had 200 replicates and two phenotypes (DBP and Q1). Therefore, due to these computational constraints, we took an alternative strategy working on the traits adjusted for family relatedness. We derived these new traits using GRAMMAR (*Aulchenko et al., 2007a*) as implemented in the GenABEL add-on package *(Aulchenko et al., 2007b*) developed for the R statistical software (*R Core Team, 2013).* As such, the classical single-marker linear model could be used.

Lastly, since our goal was to evaluate whether power to detect association with SNPs with weak effects could be enhanced by pooling their effects, we further adjusted the de-correlated trait DBP for the strong effects of *MAP4* (SNPs 3_48040283 and 3_48064367).

Methods: Whole-Genome Analyses. First, in a discovery dataset (in our case it was replicate 1), we identified the set of top SNPs, S, varying the size of S, $S = \{10,\ 50, 100, 1\,000, 5\,000, 10\,000\}$. For each SNP we derived two effect estimates, one through a classical single-marker (SM) analysis whereby each effect was estimated one at a time and the other through a multi-marker (BLUP, see Model I in **Table 3.15**) analysis whereby all effects were estimated simultaneously. If the SM approach was used to obtain the SNP estimates, the SNPs were ranked by the p-value of their effect estimate. If BLUP was used, the SNP were ranked based on the absolute value of their effect estimate.

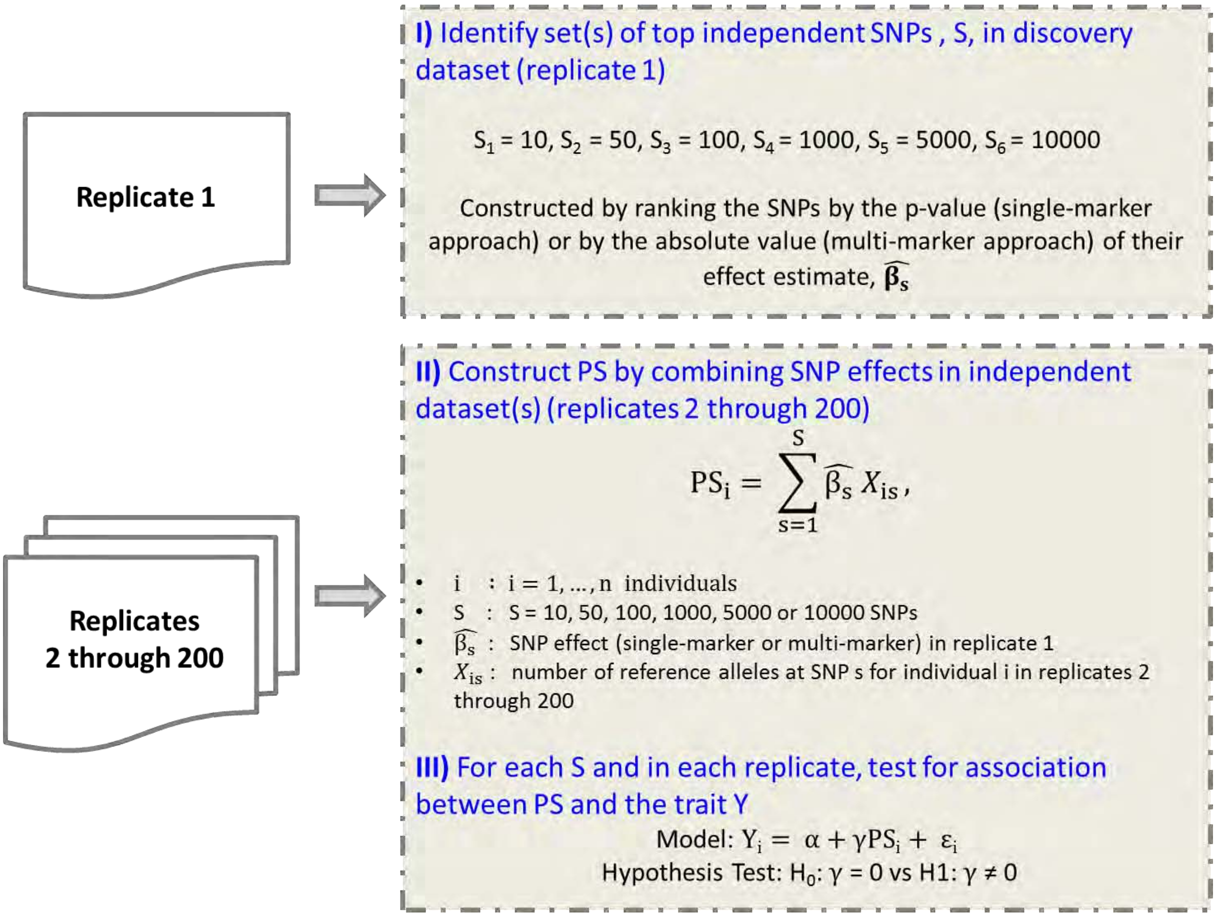
In order to ensure that S contained only independent SNPs, the best SNP over a window of 100kb was retained until the full SNP map had been covered. (We also considered larger window sizes of 1Mb and 5Mb but the results were similar and are not reported here.)

Second, we constructed the PS as follows

$$PS_i = \sum_{s=1}^{S} \hat{\beta}_s X_{is}, \tag{3.57}$$

where PS_i is the polygenic score for the i^{th} individual, S is the size of the set of SNPs to combine, $\hat{\beta}_s$ is the estimated effect of SNP s in the discovery dataset and X_{is} is the number of reference alleles of the SNP s for the i^{th} individual in an independent dataset (in our case it was replicates 2 through 200).

Third, we tested for association between PS and the analyzed trait in replicates 2 through 200 and we calculated the percentage of replicates in which the association was significant at different nominal p-values. **Figure 3.8** below provides a schematic representation of our study design.



I) Identify set(s) of top independent SNPs , S, in discovery dataset (replicate 1)

$S_1 = 10$, $S_2 = 50$, $S_3 = 100$, $S_4 = 1000$, $S_5 = 5000$, $S_6 = 10000$

Constructed by ranking the SNPs by the p-value (single-marker approach) or by the absolute value (multi-marker approach) of their effect estimate, $\widehat{\boldsymbol{\beta}_s}$

II) Construct PS by combining SNP effects in independent dataset(s) (replicates 2 through 200)

$$\mathrm{PS_i} = \sum_{s=1}^{S} \widehat{\beta_s}\, X_{is},$$

- i : i = 1, …, n individuals
- S : S = 10, 50, 100, 1000, 5000 or 10000 SNPs
- $\widehat{\beta_s}$: SNP effect (single-marker or multi-marker) in replicate 1
- X_{is} : number of reference alleles at SNP s for individual i in replicates 2 through 200

III) For each S and in each replicate, test for association between PS and the trait Y

Model: $Y_i = \alpha + \gamma PS_i + \varepsilon_i$

Hypothesis Test: H_0: $\gamma = 0$ vs H1: $\gamma \neq 0$

Figure 3.8: Schema of the study design for the GAW18 data.

Methods: Evaluating Empirical Type I Error Rates. To evaluate the significance of our findings, we carried out PS association analyses in three different scenarios. First, we evaluated the associations with the Q1 trait which, we recall, was uninfluenced by any of the genotyped SNPs. Second, we evaluated the associations after permuting the DBP trait within family once in each replicate (n=199 permutations). However, since the number of replicates/permutations was restricted to 199, we could only estimate the replication rates at nominal p-value of 0.05.Therefore, to obtain estimates of the replication rates at more stringent criteria levels (nominal p-values < 1E-04), we further permuted DBP within families 1 000 times in 100 of the replicates thus obtaining 100 000 permutations in total. Third, we also conducted PS association analyses with DBP or Q1 using a set of randomly chosen SNPs rather than SNPs selected based on their evidence of association. To compute the PS, we used the SNP effect estimates derived by each approach (SM or BLUP).

3.6.4.3 Results

Replication Rates. **Figure 3.9** below illustrates the replication rates we obtained for the DBP trait. For the SM strategy, we see that the replication rates tend to increase with the SNP set size until reaching a peak at $S = 1000$ SNPs after which they begin to decline especially at stringent significance criteria levels (p-value ≤ 1E-05). For the BLUP strategy, the peak is reached at $S = 5\ 000$ SNPs after which the replication rates tend to remain stable. Irrespective of the strategy, however, the replication rates are rather high especially at less stringent criteria (1E-03 ≤ p-value ≤ 0.05) and for larger set sizes ($S\ \geq\ 1\ 000$) where they are nearly always at or close to 100%. For smaller set sizes, ($S\ \leq\ 100$), replication rates are greater under the SM than under the BLUP strategy. The opposite trend is observed for larger set sizes ($S \geq 5\ 000$).

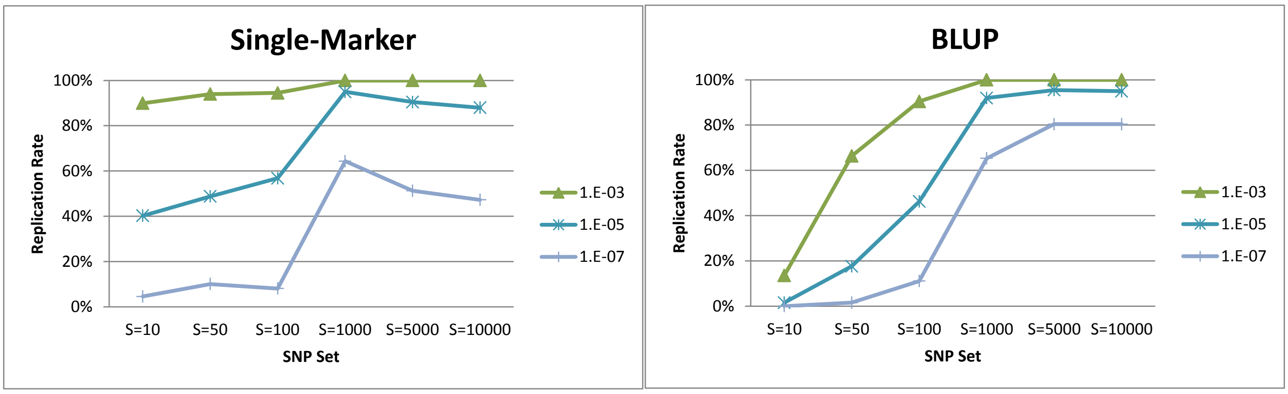


Figure 3.9: Percentage of replicates (out of replicates 2 through 200) with significant evidence of association of PS with DBP at a given nominal p-value by SNP set S derived using either Single-Marker or BLUP strategies in replicate 1.

Empirical Type I Error Rates. The results for Q1, permuted DBP and random SNP set at nominal p-value of 0.05 are presented in **Table 3.17**. The results for lower significance thresholds for the permuted DBP and the random SNP set are presented in **Table 3.18** and **Table 3.19**, respectively.

From **Table 3.17**, we see that, for Q1, the estimated replication rates did not differ significantly from their expected theoretical value of 0.05, whether the top SNPs were selected under the SM or BLUP and irrespective of the set size, S. In the case of permuted DBP, we obtained slightly inflated rates especially for larger set sizes S and under the BLUP strategy (ranging from 8% to 12% and from 6% to 17% under SM and BLUP strategies, respectively, or alternatively expressed as fold difference, ranging from 1.51 to 2.41 times higher than 0.05 under SM and from 1.21 to 3.42 times higher under BLUP). At lower significance thresholds (**Table 3.18**), except for $S = 10$ under the BLUP strategy, all replication rates were inflated and the inflation was a lot more pronounced under the SM than under the BLUP approach. For instance, at nominal p-value < 0.1%, the replication rates were between 5.55 and 12.56 times higher than expected under SM and between 0.91 and 5.48 times higher under BLUP.

Lastly, for the association analyses with DBP using a random SNP set, the replication rates increased with the SNP set size for both approaches and the rates under the SM strategy were at least as high as those under BLUP. For larger SNP sets ($S \geq 5\,000$) and irrespective of the significance threshold, the replication rates under both approaches were at or nearly 100%. However, for the smaller SNP sets ($S \leq 1\,000$), the stricter the imposed significance threshold, the larger the observed difference in the replication rates between SM and BLUP (**Table 3.19**). For instance, for $S \leq 1\,000$, the replication rate under BLUP was close to that under SM at nominal p-value of 0.05 but was almost half of that of SM at nominal p-value of 1E -04. In contrast, applying the random SNP set strategy on Q1, we obtained replication rates that were close to their theoretical values irrespective of the approach (results not shown).

n = 199			EMPIRICAL TYPE I ERROR RATES AT α = 0.05					
			S = 10	**S = 50**	**S = 100**	**S = 1000**	**S = 5000**	**S = 10000**
Q1	**Single-Marker**	**Ratio**[1]	1.11	1.41	1.21	0.90	0.90	1.31
		95% CI[2]	*(0.0235, 0.0871)*	*(0.0349, 0.1059)*	*(0.0272, 0.0934)*	*(0.0163, 0.0741)*	*(0.0163, 0.0741)*	*(0.031, 0.0996)*
	BLUP	**Ratio**	1.11	1.51	1.41	1.31	1.21	1.11
		95% CI	*(0.0235, 0.0871)*	*(0.0387, 0.1121)*	*(0.0349, 0.1059)*	*(0.031, 0.0996)*	*(0.0272, 0.0934)*	*(0.0235, 0.0871)*
Permuted DBP	**Single-Marker**	**Ratio**	2.41	2.11	2.21	2.21	1.51	1.51
		95% CI	*(0.0754, 0.1658)*	*(0.0628, 0.1482)*	*(0.067, 0.1542)*	*(0.067, 0.1542)*	*(0.0387, 0.1121)*	*(0.0387, 0.1121)*
	BLUP	**Ratio**	1.21	2.11	1.51	2.91	3.42	2.91
		95% CI	*(0.0272, 0.0934)*	*(0.0628, 0.1482)*	*(0.0387, 0.1121)*	*(0.0967, 0.1947)*	*(0.1186, 0.2232)*	*(0.0967, 0.1947)*
Random SNP Set	**Single-Marker**	**Ratio**	1.91	3.42	15.78	20.00	20.00	20.00
		95% CI	*(0.0547, 0.1363)*	*(0.1186, 0.2232)*	*(0.7322, 0.8456)*	-	-	-
	BLUP	**Ratio**	0.70	2.31	4.72	19.60	20.00	20.00
		95% CI	*(0.0096, 0.0608)*	*(0.0712, 0.16)*	*(0.1772, 0.2952)*	*(0.9604, 0.9994)*	-	-

[1] Ratio of empirical estimate to nominal value.

[2] 95% CI for empirical estimate.

Table 3.17: Empirical Type I Error rates and 95% confidence intervals (CI) under three scenarios: (1) with trait Q1 uninfluenced by any of the SNPs, (2) with permuted DBP within family thus breaking the association between the SNPs and DBP, and (3) with DBP using a randomly chosen sets of top SNPs (the same sets were used for both Single-Marker and BLUP). Rates were estimated based on 199 replicates. Due to the small number of replicates, only estimates at nominal p-value = 5% are given. Cases where the empirical estimate significantly exceeds the nominal value are underlined.

Permuted DBP, n = 100 000			EMPIRICAL TYPE I ERROR RATES AT A GIVEN THRESHOLD					
			S = 10	**S = 50**	**S = 100**	**S = 1000**	**S = 5000**	**S = 10000**
p-value < 0.05	**Single-Marker**	**Ratio**[1]	2.53	2.74	3.37	4.04	3.51	2.93
		95% CI[2]	*(0.1243, 0.1285)*	*(0.1348, 0.139)*	*(0.1664, 0.171)*	*(0.1995, 0.2045)*	*(0.1732, 0.178)*	*(0.1441, 0.1485)*
	BLUP	**Ratio**	0.99	1.44	1.53	2.35	2.46	2.47
		95% CI	*(0.0481, 0.0507)*	*(0.0704, 0.0736)*	*(0.075, 0.0782)*	*(0.1156, 0.1196)*	*(0.1211, 0.1251)*	*(0.1213, 0.1253)*
p-value < 0.01	**Single-Marker**	**Ratio**	3.67	3.91	5.14	6.80	5.65	4.72
		95% CI	*(0.0355, 0.0379)*	*(0.0379, 0.0403)*	*(0.05, 0.0528)*	*(0.0664, 0.0696)*	*(0.0551, 0.0579)*	*(0.0459, 0.0485)*
	BLUP	**Ratio**	0.96	1.75	1.88	3.35	3.53	3.59
		95% CI	*(0.009, 0.0102)*	*(0.0167, 0.0183)*	*(0.018, 0.0196)*	*(0.0324, 0.0346)*	*(0.0342, 0.0364)*	*(0.0347, 0.0371)*
p-value < 0.001	**Single-Marker**	**Ratio**	5.91	5.55	8.10	12.56	9.48	8.51
		95% CI	*(0.0054, 0.0064)*	*(0.0051, 0.0061)*	*(0.0075, 0.0087)*	*(0.0119, 0.0133)*	*(0.0089, 0.0101)*	*(0.0079, 0.0091)*
	BLUP	**Ratio**	0.91	1.94	2.39	4.99	5.38	5.48
		95% CI	*(0.0007, 0.0011)*	*(0.0016, 0.0022)*	*(0.0021, 0.0027)*	*(0.0046, 0.0054)*	*(0.0049, 0.0059)*	*(0.005, 0.006)*
p-value < 0.0001	**Single-Marker**	**Ratio**	7.80	7.70	9.80	21.00	15.00	17.00
		95% CI	*(0.0006, 0.001)*	*(0.0006, 0.001)*	*(0.0008, 0.0012)*	*(0.0018, 0.0024)*	*(0.0013, 0.0017)*	*(0.0014, 0.002)*
	BLUP	**Ratio**	0.80	2.20	2.70	7.80	8.00	7.30
		95% CI	*(0, 0.0002)*	*(0.0001, 0.0003)*	*(0.0002, 0.0004)*	*(0.0006, 0.001)*	*(0.0006, 0.001)*	*(0.0005, 0.0009)*

[1] **Ratio of empirical estimate over nominal value.**

[2] **95% CI for empirical estimate.**

Table 3.18: Empirical Type I Error rates and 95% confidence intervals (CI) for scenario (2) of **Table 3.17** with the number of permutations increased from once in each replicate (n=199) to 1 000 times in 100 replicates (n=100 000). Cases where the empirical estimate significantly exceeds the nominal value are underlined.

Random SNP Set, n = 199			**EMPIRICAL TYPE I ERROR RATES AT A GIVEN THRESHOLD**					
			S = 10	**S = 50**	**S = 100**	**S = 1000**	**S = 5000**	**S = 10000**
p-value < 0.05	**Single-Marker**	**Ratio**[1]	1.91	3.42	15.78	20.00	20.00	20.00
		95% CI[2]	*(0.0547, 0.1363)*	*(0.1186, 0.2232)*	*(0.7322, 0.8456)*	-	-	-
	BLUP	**Ratio**	0.70	2.31	4.72	19.60	20.00	20.00
		95% CI	*(0.0096, 0.0608)*	*(0.0712, 0.16)*	*(0.1772, 0.2952)*	*(0.9604, 0.9994)*	-	-
p-value < 0.01	**Single-Marker**	**Ratio**	1.01	4.02	51.76	100.00	100.00	100.00
		95% CI	*[0, 0.024)*	*(0.0129, 0.0675)*	*(0.4482, 0.587)*	-	-	-
	BLUP	**Ratio**	0.50	1.01	5.03	91.46	100.00	100.00
		95% CI	*[0, 0.0148)*	*[0, 0.024)*	*(0.0199, 0.0807)*	*(0.8758, 0.9534)*	-	-
p-value < 0.001	**Single-Marker**	**Ratio**	0.00	10.05	180.90	994.97	1000.00	1000.00
		95% CI	-	*[0, 0.024)*	*(0.1274, 0.2344)*	*(0.9852, 1.00]*	-	-
	BLUP	**Ratio**	0.00	0.00	5.03	688.44	1000.00	1000.00
		95% CI	-	-	*[0, 0.0148)*	*(0.624, 0.7528)*	-	-
p-value < 0.0001	**Single-Marker**	**Ratio**	0.00	0.00	502.51	9396.98	9949.75	10000.00
		95% CI	-	-	*(0.0199, 0.0807)*	*(0.9066, 0.9728)*	*(0.9852, 1.00]*	-
	BLUP	**Ratio**	0.00	0.00	0.00	4120.60	9798.99	9748.74
		95% CI	-	-	-	*(0.3437, 0.4805)*	*(0.9604, 0.9994)*	*(0.9532, 0.9966)*

[1] Ratio of empirical estimate over nominal value.

[2] 95% CI for empirical estimate.

Table 3.19: Empirical Type I Error rates and 95% confidence intervals (CI) for scenario (3) of **Table 3.17** for stricter significance thresholds. Cases where the empirical estimate significantly exceeds the nominal value are underlined. If the confidence limit exceeded 0 or 1, the interval was truncated (denoted by "[" or "]").

We estimated the probability of choosing a SNP lying in a true gene (as given by the simulated model) to be around 0.02 given by the proportion of SNPs lying in true genes out of all SNPs. Next, the probability that, say, in a set of ten SNPs there is at least one SNP lying in a true gene is 1-(0.98)^{10} = 0.18. Obviously, this probability increases and approaches one as the SNP set size S increases. We recall that we used windows of size 100kb to identify independent SNPs. Given the density of our SNP map, this produced roughly 13 000 windows and thus a set size of $S = 10\ 000$ (that is, top 10 000 windows) approaches full genome coverage.

Of course, not all genes exert the same influence on the trait and, among the genes implicated, not all SNPs within a gene carry the same weight (effect size). Thus, while the composition of the random SNP set might be representative of the number of SNPs that were simulated to have an effect on DBP, the association results were impacted by the strategy due to the effect estimates that these SNPs exert on the trait. The scale of the SNP weights in the PS (that is, the effect estimates) differ significantly across the two approaches. In fact, the effect estimates by SM are orders of magnitude larger than the BLUP estimates. As an example, the mean and standard deviation of the SNP effects in the set of top ten SNPs ($S = 10$) is -3.03±7.53 for SM and -7.3E-05±1.7E-04 for BLUP.

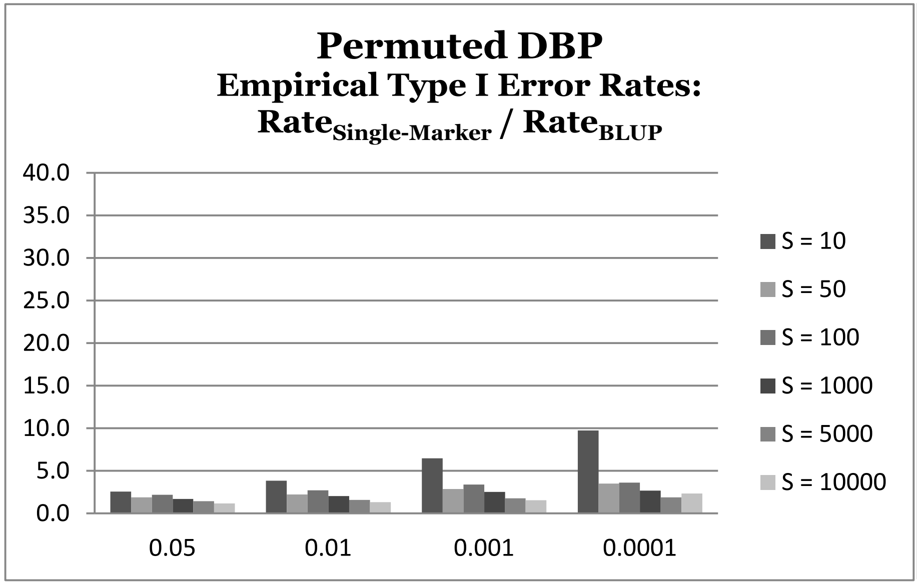
Finally, we compared the replication rates under SM and BLUP between the two Type I Error evaluating scenarios: permuted DBP and random SNP set. In absolute terms, the replication rates were very different between the scenarios (**Table 3.18** and **Table 3.19**). When using permuted DBP, the replication rates under SM increased with increasing SNP set size until $S = 1\ 000$ after which they began to decline while, when using a random SNP set, the rates continued to increase reaching 100% for large SNP sets ($S \geq 1\ 000$). For BLUP, the replication rates tended to increase with increasing SNP set size in both scenarios also reaching 100% for large SNP sets ($S \geq 1\ 000$) when using a random SNP set but not when using permuted DBP.

Irrespective of the significance threshold, for large SNP sets ($S \geq 1\ 000$), the replication rates under SM and BLUP were higher when using a random SNP set than when using permuted DBP. Conversely, for very small SNP set sizes ($S = 10$), the replication rates under SM or BLUP were lower when using a random SNP set than when using permuted DBP. Thus, for $S = 10$ or S $\geq 1\ 000$, SM and BLUP behaved similarly. However, for $S = 50$ or 100, SM and BLUP behaved differently depending on the scenario. For these two SNP sets, the replication rates under SM tended to be higher when using a random SNP set than when using permuted DBP but, under BLUP, were lower (even zero at strict significance thresholds, p-values < 0.0001).

In relative terms, we noted that the replication rates under SM were always at least as high as those under BLUP. To see this, we plotted the ratio of the replication rate in SM versus BLUP for each SNP set size and for each evaluated significance threshold in both scenarios. The results are given in **Figure 3.10** below. At the 0.05 significance threshold, a similar trend was observed whether permuted DBP or random SNP set was used, that is, the ratio decreased with increasing set size from 2.56 ($S = 10$) to 1.19 ($S = 10\,000$) when using permuted DBP and from 2.71 ($S = 10$) to 1.00 ($S = 10\,000$) when using a random SNP set. The same trend was also observed at stricter significance thresholds only when using permuted DBP. In contrast, when using a random SNP set, the ratio of replication rates in SM versus BLUP increased up until $S = 100$ and then dropped reaching one for larger SNP set sizes ($S \geq 5\,000$).

Permuted DBP ($\text{Rate}_{\text{Single-Marker}}$ / $\text{Rate}_{\text{BLUP}}$)						
p-value	S = 10	S = 50	S = 100	S = 1000	S = 5000	S = 10000
0.05	2.56	1.90	2.20	1.72	1.43	1.19
0.01	3.83	2.23	2.73	2.03	1.60	1.31
0.001	6.49	2.86	3.39	2.52	1.76	1.55
0.0001	9.75	3.50	3.63	2.69	1.88	2.33

Random SNP Set ($\text{Rate}_{\text{Single-Marker}}$ / $\text{Rate}_{\text{BLUP}}$)						
p-value	S = 10	S = 50	S = 100	S = 1000	S = 5000	S = 10000
0.05	2.71	1.48	3.34	1.02	1.00	1.00
0.01	2.00	4.00	10.30	1.09	1.00	1.00
0.001	-	-	36.00	1.45	1.00	1.00
0.0001	-	-	-	2.28	1.02	1.03



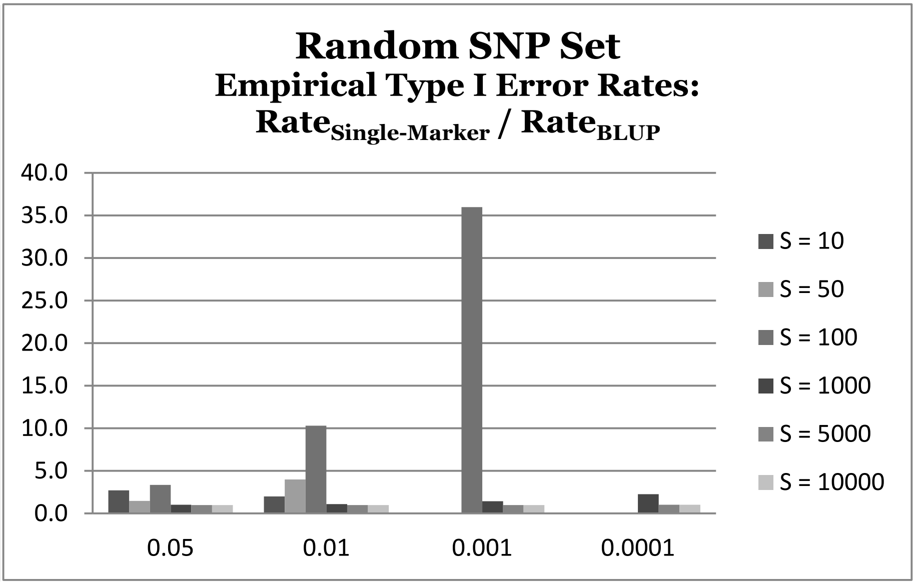


Figure 3.10: Ratio of empirical Type I Error rates of Single-Marker to BLUP for permuted DBP (**left**) and for the random SNP set (**right**). A dash "**-**" in the table above the right graph indicates that the replication rate under BLUP was zero.

3.6.4.4 Discussion and Conclusions

Overall, using a classical single-marker approach accounting for family relatedness to detect association, we found that, with the exception of the SNPs in the *MAP4* gene, there was no power to detect SNPs of weaker effect at the genome-wide significance level. Through PS association analyses, we achieved high replication rates whether SM or BLUP was used especially when large sets of SNPs (≥ 1 000) were considered. However, we noted that the replication rates under SM began to decline for larger SNP sets ($S > 1\ 000$) especially at stringent criteria levels. This may be happening due to much more noise being added with larger SNP set sizes. However, the empirical Type I Error rates were elevated and question whether the actual power of the PS approach is well estimated by these replication rates.

When no SNPs influenced the trait (Q1), the replication rates for both approaches were close to their theoretical values. When SNPs influenced the trait but these relationships were broken by permuting the trait within family, the replication rates were inflated under both SM and BLUP. This trend was exacerbated at stringent significance criteria levels (p-value < 1E-04) and more so under SM than under BLUP. Further, when SNPs were chosen randomly versus based on evidence of association, we obtained high replication rates for larger SNP sets (≥ 5 000) irrespective of the strategy to derive the SNP effect estimates. This was because with this SNP set sizes, we approached full genome coverage thus likely including all causal variants. For smaller SNP sets, however, the replication raters were higher under SM than under BLUP and this difference became more profound the stricter the significance threshold. In fact, small SNP sets were likely to contain functional variants and yet the replication rates under BLUP were zero suggesting that BLUP is possibly conservative.

We recall that the genotypes were the same across replicates making it challenging to ensure that we have arrived at evaluating correctly the Type I Error rates. While the error rates appeared well controlled for Q1 at the 0.05 significance threshold, we recall that this trait although heritable was uninfluenced by any of the SNPs. In contrast to Q1, DBP was influenced by roughly 1 500 functional SNPs and 1 000 randomly chosen SNPs in each replicate (polygenic component).

This dataset was a family dataset and we used a de-correlation technique by adjusting the trait for family relatedness to render the individuals “unrelated.” This is a required assumption for the single-marker analyses. However, in the BLUP model, the family structure can be recovered from the genome-based relationship matrix and thus adjusted for. (*Goddard, 2009*) One explanation for the particularly inflated rates we observed under SM, therefore, might be that linkage may have affected the PS association analyses. In the provided GAW18 pedigree dataset (approximately 900 related individuals from 20 pedigrees) there were roughly 150 unrelated individuals. Despite this small sample size, it could have been interesting to run our analyses on this set of individuals and evaluate whether we would observe the same trends, that is, higher replication rates under SM than under BLUP.

We conclude that the SM approach to PS association analysis should not be used with large SNP sets (≥ 5 000) – for these set sizes it performed equally well as choosing SNPs at random. For smaller sets of top SNPs, however, it remains a preferable approach to BLUP. However, if we controlled well for Type I Error, we point out that the power remained low to detect SNPs of weaker effect through PS association analyses.

Dudbridge (2013) derived a closed form expression of the power of the PS association method to detect variants of weak effect as a function of a number of parameters including the sample size for the discovery and replication datasets and the SNP weighting method used in the PS (weighted and unweighted). Using a simulation study, he found that power was high under both weight alternatives but was much higher when the SNPs were weighted by their effect estimates than when the SNPs were unweighted (that is, when each SNP carried an equal weight in the score). Also, he showed that low power could be due to small discovery sample size.

We recall that the BLUP effect estimates, contrary to the SM estimates, had very low variability and, hence, PS constructed with BLUP estimates was similar with or without weights. Hence, this could also explain why the replication rates under BLUP were lower than under SM. Moreover, our discovery sample size was small (approximately 900 individuals).

This study was further extended to the context of prediction, the results of which are described in the next chapter (Section 4.3.2).

Part of this work was presented to the genetic prediction group of GAW18 and is to appear in BMC Proceedings (Bohossian et al., 2013). I also contributed to the group summary paper submitted to Genetic Epidemiology (Ziegler et al., 2013).

4 GENETIC PREDICTION STUDIES

"*Prediction is very difficult, especially if it's about the future*" – is a famous quote attributed to the Danish physicist and Nobel laureate, Niels Bohr. We wish to make the best decision about the future with the best information available today. This is true in every aspect of our lives and especially so for health decisions - an emerging science known as evidence based medicine (EBM). EBM is "*the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients.*" (*Sackett et al., 1996*) A more precise definition has been proposed by *Donald and Greenhalgh (2000*) to emphasize the important role of mathematical approaches to achieve this objective stating that EBM is "*the enhancement of a clinician's traditional skills in diagnosis, treatment, prevention and related areas through the systematic framing of relevant and answerable questions and the use of mathematical estimates of probability and risk."*

Sackett and Rosenberg (1995) summarized the five essential steps of EBM as the following: (1) to convert our information needs into answerable questions (to formulate the question); (2) to track down, with maximum efficiency, the best evidence with which to answer these questions; (3) to appraise the evidence critically to assess its validity (closeness to the truth) and usefulness (clinical applicability); (4) to implement the results of this appraisal in our clinical practice; (5) to evaluate our performance. Most of the research has been focused on the third step. The whole process, covering all five steps, may take years, even decades.

Personalized medicine is a paradigm of EBM. Although the definition of EBM appears to have been formalized only in the mid-90s, EBM is not a new concept. Drug administration, for instance, has traditionally been tailored to patients' characteristics such as age, gender and body size. Large amount of genetic information has overwhelmed the medical community in the last decade or so changing the scope of EBM. This wealth of new information offers tremendous power for revolutionizing diagnosis and treatment and is laying the foundations of genomic medicine (*Kumar, 2007*). Understanding how our genomes affect our health may lead to more precise estimates of risk, and thus better treatment. However, predicting risk based on the genetic information that we carry has not been met with much success for complex traits. In this chapter, we review the necessary steps required to build a clinical prediction model and describe the factors influencing prediction accuracy.

4.1 Methods

Steyerberg (2009) presents seven main steps to building valid clinical prediction models. These steps apply equally in a genetic and in a non-genetic context. However, one particularity of the genetic context may be the high dimensionality of the data (if carrying out whole genome predictions). We describe these steps briefly below focusing on the genetic context and MS where applicable.

4.1.1 Careful Consideration of the Prediction Problem

Perhaps the most defining aspect of a prediction study is the outcome of interest and the available knowledge about the factors which influence it. As we discussed in the introductory chapter of this thesis (see Section 1.8), response to treatment in MS is a highly complex outcome to evaluate. Further, there is no widely accepted definition of it.

Additionally, the most commonly used treatment of MS, interferon-β, has been relatively well studied. However, to this day, limited prior knowledge exists on the factors (genetic and non-genetic) which influence interferon response in MS patients.

In contrast, we bring back the case of warfarin. The phenotype is a quantitative trait, stable warfarin dose, objectively defined. Moreover, prior building a model to predict the correct dose to administer to each patient, numerous studies had been carried out evaluating the factors, genetic and non-genetic, potentially influencing the correct dose. Thus, a lot of prior knowledge was available at the time of building the prediction model for warfarin dose. Of course, we note that warfarin has been available as a treatment since the 1950s while interferon-β, the first ever approved treatment of MS, has only been available since the late 90s.

4.1.2 Coding the Predictors in the Model

The way the predictors are coded can influence the results of the prediction model. Continuous variables could be categorized and categorical variables may be treated as continuous. For example, age of disease onset could be categorized into “early disease onset” and “late disease onset” based on a predefined threshold age value. Another example is the EDSS which we recall is an ordinal variable on a scale from 0.0 to 10.0 measuring disability in MS patients. In most studies, however, it is treated as continuous.

Further, a categorical variable may remain categorical but with reduced number of categories. Combining categories which do not have many observations in them, for example, is a common practice. Alternatively, categories could be added. For instance, missing data may be treated as a separate category although variables with a lot of missing data may need to be imputed or excluded from the study altogether. It is necessary and important to evaluate the implications of such coding strategies on the predicted outcome.

4.1.3 Specification of the Model

This is the most difficult of all steps and typically many different models are specified and evaluated. The type of outcome variable usually dictates the type of model to use such as regression for continuous variables and classification for categorical variables. The relationship between the type of outcome and the predictors may be modeled as linear or non-linear (of which the linear is a special case). Further, one may choose to work in a frequentist or in a Bayesian setting, using parametric or non-parametric approaches. Lastly, modeling techniques are often borrowed from the data mining and machine learning domains especially in the context of high dimensionality arising with genetic predictors (thousands to millions of genetic variants). Thus, there is a vast number of modeling approaches to choose from. An illustrative example of this is given by one of the prediction algorithms for warfarin dose which evaluated all of the following models: ordinary linear and polynomial regression, artificial neural networks, support vector regression with polynomial (including linear) and

Gaussian kernels, regression trees, model trees, least angle regression, Lasso and multivariate adaptive regression splines. (*Klein et al., 2009*) Of all, ordinary linear regression – the simplest - performed best (based on MAE, see Equation (4.10) below).

4.1.4 Estimation of the Model Parameters

In a frequentist setting, when the parameter space is of low dimension ($p < n$), the most commonly used estimation method is the method of maximum likelihood (ML). The maximum likelihood (ML) estimates correspond to the most likely values of the parameters given the observed data. In a high dimensional context (that is, when $p \to n$ or $p \gg n$), penalized estimation methods need to be used shrinking the estimates of the parameters towards their ML estimates.

In a Bayesian setting, the approach is quite different. Let θ be the parameter we are interested in estimating. The *prior distribution* of θ, $p(\theta)$, is defined by the investigators and reflects any prior knowledge on the value of θ, if available. Let x denote the observed data and let $p(x|\theta)$ be the likelihood. Bayesian methods estimate the *posterior distribution* of θ, $p(\theta|x)$, that is proportional to the product of the prior distribution and the likelihood such that $p(\theta|x) \propto p(x|\theta)p(\theta)$. Point estimate of θ may also be obtained using, for instance, the mode of the posterior distribution which, under certain priors, is equivalent to the ML estimate. Moreover, for most penalized estimates, there is an equivalent Bayesian prior (*de los Campos et al., 2010*).

4.1.5 Evaluation of the Model Performance

Evaluation of the model performance is determined by the type of outcome. Let y be the observed outcome and $\hat{y}$ be the outcome predicted by the model. We provide a brief review of the performance measures for quantitative and binary traits (or outcomes).

4.1.5.1 Continuous Outcomes

For continuous outcomes, the correlation between y and $\hat{y}$ is often evaluated. The estimated correlation, r, is given by

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}}, \tag{4.1}$$

where y_i and $\hat{y}_i$, $i = 1, \dots, n$ are the observed and predicted outcomes, respectively, and $\bar{y}$ is the mean of y_i, $i = 1, \dots, n$.

The most commonly used measure, however, is the square of r, the coefficient of determination, R^2, defined as

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{4.2}$$

Under certain conditions (such as when the $\hat{y}_i$'s have been derived using linear regression),

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \tag{4.3}$$

and, hence, Equation (4.2) above can be expressed as

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4.4}$$

which is the proportion of explained variance by the predictors in the model.

It has been shown that R^2 is a biased estimate of the true coefficient of determination, P^2, given by

$$P^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}, \tag{4.5}$$

where σ_e^2 and σ_y^2 are the true variances of the residuals and the dependent variable, respectively. The maximum value of the bias is given by

$$\frac{p-1}{n}, \tag{4.6}$$

where p is the number of variables in the model including the intercept. Thus, an adjusted R^2, R^2_{adj}, is often reported as well and is given by

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right). \tag{4.7}$$

(This issue is discussed at length by *Montgomery and Morrison* (1973).)

Recalling our definition of heritability from the previous chapter (Equation (3.13)), if the prediction model contains only genetic predictors with additive effects, then R^2 can be interpreted as the proportion of the phenotypic variance attributable to additive genetic factors and its upper bound is given by the heritability (narrow-sense) of the trait, h^2, such that

$$R^2 \le h^2. \tag{4.8}$$

Finally, a *loss function* is often specified measuring the "loss" in precision between the real and the predicted value. Several popular loss functions include the mean squared error (MSE), the mean absolute error (MAE) and the mean relative error (MRE) and are described below.

The MSE is defined as the mean of the squared differences between y and $\hat{y}$. It is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2. \tag{4.9}$$

Alternatively, the MAE is defined as the mean of the absolute differences between y and $\hat{y}$. It is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{4.10}$$

If $y_i \neq 0$, $i = 1, \dots, n$, the MRE can be defined as the mean of the relative differences between y and $\hat{y}$. It is given by

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|1 - \frac{\hat{y}_i}{y_i}\right|. \tag{4.11}$$

4.1.5.2 Binary Outcomes

Suppose we are interested in predicting disease status (disease/healthy). **Table 4.1** illustrates a two-by-two table with predicted versus observed outcomes.

		OBSERVED	
		Disease	**Healthy**
PREDICTED	**Disease**	TP	FP
	Healthy	FN	TN

Table 4.1: Two-by-two table of predicted versus observed disease status. TP: true positive count, FP: false positive count, FN: false negative count, TN: true negative count.

Five measures can be derived from **Table 4.1**: sensitivity, specificity, positive predictive value, negative predictive value and overall accuracy.

The *sensitivity* is defined as the proportion of correctly classified disease individuals among all observed disease individuals and is given by

$$Sensitivity = \frac{TP}{TP + FN}. \tag{4.12}$$

The *specificity* is defined as the proportion of correctly classified healthy individuals among all observed healthy individuals and is given by

$$Specificity = \frac{TN}{FP + TN}. \tag{4.13}$$

The *positive predictive value* (PPV) is defined as the proportion of correctly classified disease individuals among all predicted disease individuals and is given by

$$PPV = \frac{TP}{TP + FP}. \tag{4.14}$$

The *negative predictive value* (NPV) is defined as the proportion of correctly classified healthy individuals among all predicted healthy individuals and is given by

$$NPV = \frac{TN}{FN + TN}. \tag{4.15}$$

Lastly, the overall accuracy is defined as the proportion of correctly classified individuals (disease or healthy) among all individuals and is given by

$$Overall\ Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4.16}$$

These measures are summarized in **Table 4.2** below.

		OBSERVED		
		Disease	**Healthy**	
PREDICTED	**Disease**	TP	FP	**PPV** = TP/(TP + FP)
	Healthy	FN	TN	**NPV** = TN/(FN + TN)
		Sensitivity = TP/(TP + FN)	**Specificity** = TN/(FP + TN)	**Overall Accuracy** = (TP + TN)/(TP +FP +TN+FN)

Table 4.2: Summary of several performance measures for a binary outcome.

The accuracy of a prediction model may be assessed in several ways the most common being evaluating its discriminatory ability between the two groups (disease and healthy). The most widely used measure for that is the area under the receiver operating characteristics (ROC) curve. The ROC curve is obtained by plotting sensitivity (true positive fraction) on the y-axis versus 1-specificity (false positive fraction) on the x-axis for varying thresholds used to discriminate between disease and healthy individuals (for example, consecutive cutoffs for the probability of outcome).

Figure 4.1 below illustrates hypothetical ROC curves to demonstrate different classification accuracies as measured by the area under the curve (AUC). AUC varies between 0 and 1 but classifiers with AUC < 0.5 are typically reformulated such that AUC > 0.5 (that is if predicting disease status gives an AUC < 0.5, predicting healthy status will give an AUC > 0.5 with the same model). Thus, the focus is on the upper left diagonal such that the AUC is between 0.5 and 1 (0.5 < AUC < 1).

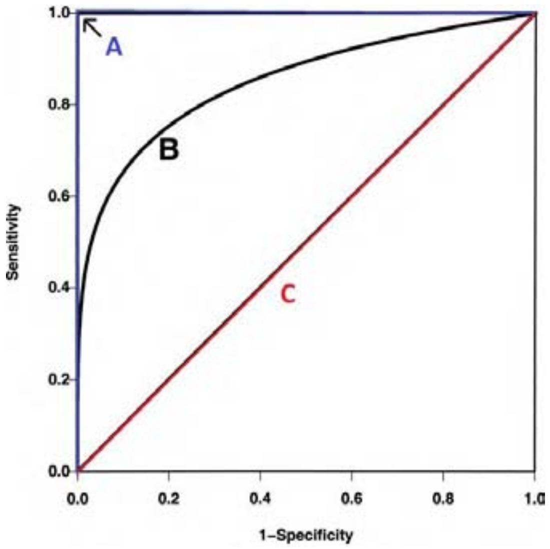


Figure 4.1: Hypothetical ROC curves. Adapted from (*Zou et al., 2007*).

For the binary classifier C, the $AUC = 0.5$ (half the area of the unit square). This classifier performs as good as a random guess, in other words, it is perfectly useless. On the other extreme, the binary classifier A has an $AUC = 1$ (full unit area) and is a perfect classifier. In other words, the model discriminates perfectly between disease and healthy individuals.

Wray et al. (2010) demonstrated that the maximum achievable AUC, AUC_{max}, when the classifier is a genetic predictor depends on the heritability at the liability scale, h^2, and on the prevalence of disease, K. Their derived expression for AUC_{max} is given by

$$AUC_{max} \approx \Phi\left(\frac{(i-v)h^2}{\sqrt{h^2\left[\left(1-\ h^2 i(i-T)\right)-\left(1-\ h^2 v(v-T)\right)\right]}}\right), \tag{4.17}$$

where

- $\Phi(x)$ is the cumulative density function of the standard normal distribution such that if $X \sim N(0,1), \Phi(x) = P(X \leq x)$;
- $i = \frac{z}{K}$ where z is the height of the standard normal curve and K is the prevalence;
- $v = \frac{iK}{(1-K)}$; and
- $T = \Phi^{-1}(1 - K)$.

Distinction should be made between discrimination and calibration, which is another component in assessing the accuracy of a prediction model. Discrimination refers to the ability of the model to distinguish between disease and healthy individuals while calibration refers to the agreement between the predicted and the observed values. For instance, for continuous outcomes, calibration can be evaluated graphically by plotting $\widehat{y_i}$ versus y_i, $i = 1, \ldots, n$. If the model is well calibrated, the scatter plot should fall closely along the diagonal. Analogous approaches have been developed for binary outcomes although they are rather imperfect.

A model may be good at discriminating but poorly calibrated. In other words, a model may be predicting a higher risk for disease individuals than for healthy individuals but the actual predicted risks by the model may be in poor agreement with the true risks. Alternatively, if there is little difference in the true risks between disease and healthy individuals, that is, the risk distribution has narrow spread, a model may not be able to discriminate well even if the predicted risks are in good agreement with the true risks. In fact, there is typically a trade-off between discrimination and calibration and the only time a model may be perfect at both is when the true and the predicted risks are 0 or 100%, that is, the risk distribution is U-shaped. (*Cook, 2007*)

Other performance metrics have been developed to circumvent the limitations of the AUC such as reclassification tables (*Cook, 2007*). The idea is to evaluate how many subjects are reclassified after including additional predictor(s) in the model. This metric can sometimes better illustrate the classification improvement which may be reflected with a small or insignificant change in the AUC.

All metrics discussed so far assume that false positives and false negatives carry equal weight and a theoretical framework to account for different weights has been developed by *Vickers and Elkin (2006)*. Net benefit, NB, is a measure that tries to accommodate different weights for each wrong decision. It is given by $NB = (TP - wFP)/n$ where w is the weight derived from the ratio of harm to benefit. Of course, in this measure, the challenging part is to come up with the value of w.

Several traditional and novel performance metrics have been reviewed by *Steyerberg et al. (2010)* who also address performance metrics for survival outcomes. We note that, for some investigators (for example, *Pepe et al., 2007*) sensitivity and specificity (and, hence, AUC) remain the classical and pertinent piece of metric to be supplied in a prediction study.

4.1.6 Generalizability of the Model

A central concern of predictive modeling is over-fitting which occurs when the model describes noise in the data rather than a true existing relationship between the predictor(s) and the outcome. In that case, the detected relationship will fail to be reproduced in an independent dataset. To guard against that, the performance metrics described in the previous step need to be evaluated in a new independent dataset. Ideally, three datasets are necessary: training sample, testing sample and a target population. The model is learned (built) on the training sample, its performance is evaluated in the testing sample and only then the model is applied in practice in a target population. In the training and testing samples, the outcome is known, while in the target population it is unknown.

Obtaining samples is a costly and lengthy process if at all possible so clever statistical approaches have been developed to circumvent this limitation as much as possible when building prediction models without limiting too much their generalizability. These approaches rely on the concept of data splitting or sample reuse techniques. Cross validation (CV), initially introduced in the mid-70s, is one of the most widely used such approach (*Stone, 1974*; *Geisser, 1975*). We discuss it at length here.

The main objectives of CV are: (1) to evaluate the error of the model giving an idea of how it can be generalized; (2) to compare the performance between models; and (3) to tune model parameters. (*Refaeilzadeh et al., 2009*) To achieve these objectives, there are many alternative ways in which the data could be split.

To illustrate the many alternative ways in which the data can be split, we generated a simple dataset with one dependent variable (y) and one independent variable (x) for 20 subjects given in **Figure 4.2** below.

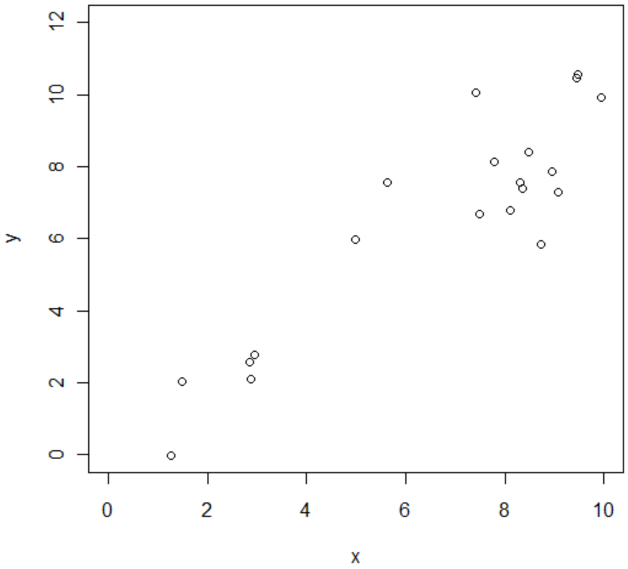


Figure 4.2: A simple dataset generated with one dependent variable (y) and one independent variable (x) for $n = 20$ subjects.

In the worst case scenario, the model is both trained and tested on the same dataset as illustrated in **Figure 4.3**. The red line in the figure illustrates the model fit and the blue lines illustrate the difference between the simulated and the predicted values (the “loss”).

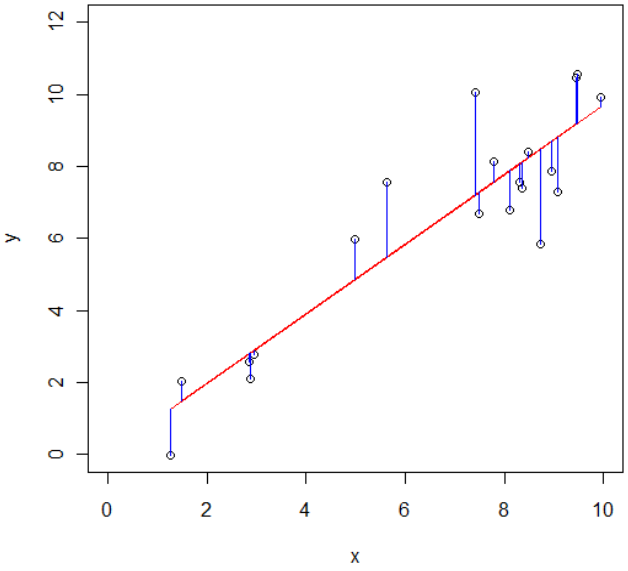


Figure 4.3: The fitted model (red line) to the dataset in **Figure 4.2**. The blue lines illustrate the difference between the simulated and the predicted values.

This procedure leads to over-fitting, where the model fits the random error rather than the relationship being modeled. To avoid the problem of over-fitting, one alternative would be to randomly split the dataset into two parts, training and testing. The model is learned on the training set and its predictive performance is evaluated in the testing set (**Figure 4.4**). The limitation of this approach is that only part of the data is used for training and as such the results may be strongly impacted by a specific split of the data. It is possible to repeat this procedure several times but this may only partially avoid the problem as the full data may remain still unused for training.

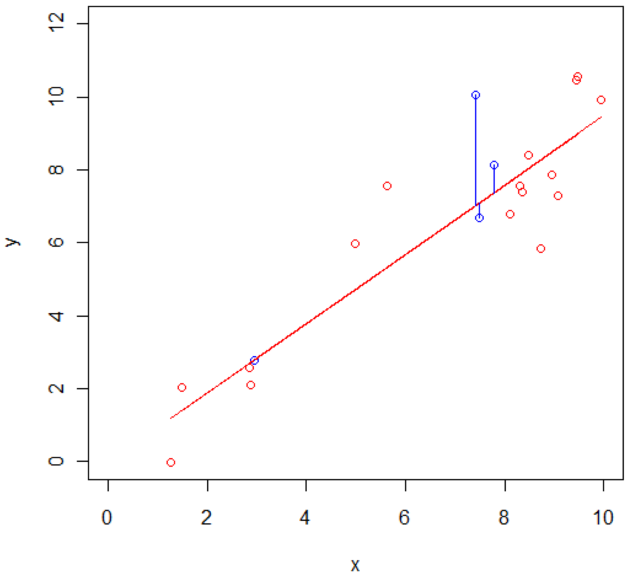


Figure 4.4: Random split of the dataset in **Figure 4.2** into training (red points) and testing (blue) points. The model is fitted (red line) on the training datasets and its loss function evaluated in the testing dataset.

An alternative to that is to use the famous k-fold cross validation. In this setting, the data are split into k equal partitions. Of these, $k-1$ partitions are used for training and the remaining partition is used for testing. This procedure is iterated k number of times (**Figure 4.5**). Then, the prediction error obtained in each iteration is averaged over all k iterations. The optimal value of k may be determined from the data or a commonly used value such as $k = 5$ or $k = 10$ may be chosen.

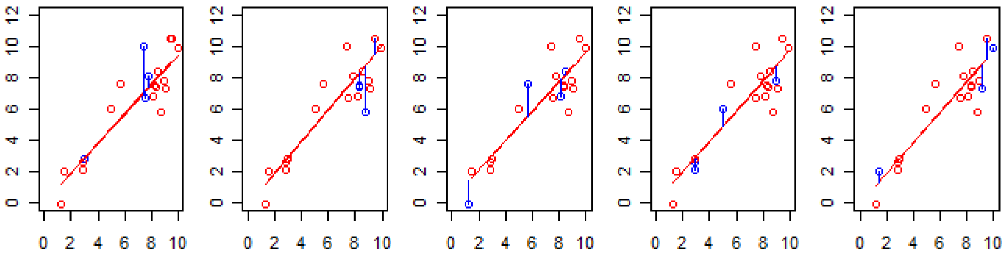


Figure 4.5: The k-fold cross validation illustrated on the dataset in **Figure 4.2**. That is, in each iteration, the model is trained on $n - \frac{n}{k}$ observations (red) and tested on $\frac{n}{k}$ observations (blue). This process is repeated until all observations in the dataset have been used for testing. Here, $k = 5$.

A special case of this is when $k = n$ (where n is the number of subjects in the full sample). This is referred to as leave-one-out CV (**Figure 4.6**). This type of approach is often used when the sample size is very small (dozen or so subjects).

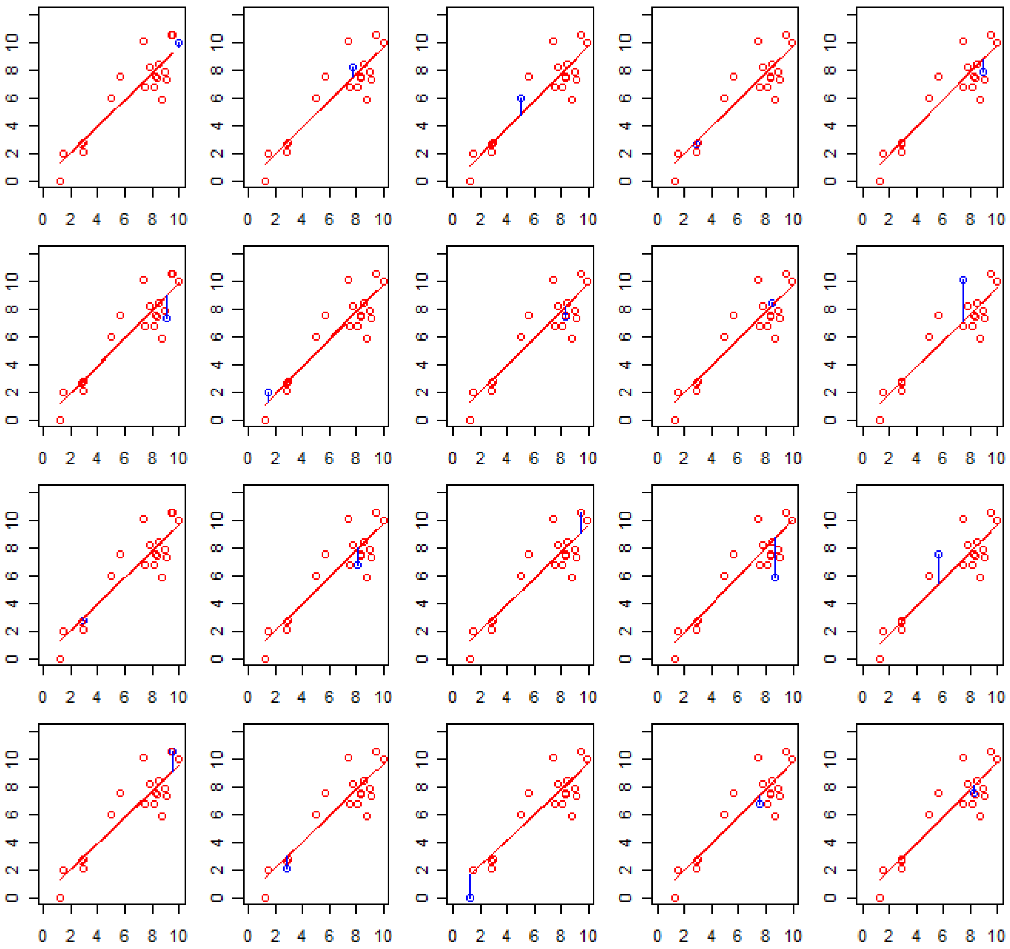


Figure 4.6: Leave-one-out cross validation illustrated on the dataset in **Figure 4.2**. In each iteration, the model is trained on $n-1$ observations (red) and tested on one observation (blue). This process is repeated until all observations in the dataset have been used for testing.

Lastly, an extension to the last method is sometimes used, where the k-fold CV is repeated Q number of times thus creating Q different sets of k partitions (**Figure 4.7**). Therefore, whereas with k-fold CV we would have k estimates of the prediction error, with Q repeated k-fold CV we have $Q * k$ estimates. In other words, this approach produces a greater number of estimates of the prediction error hopefully leading to more accurate error estimates.

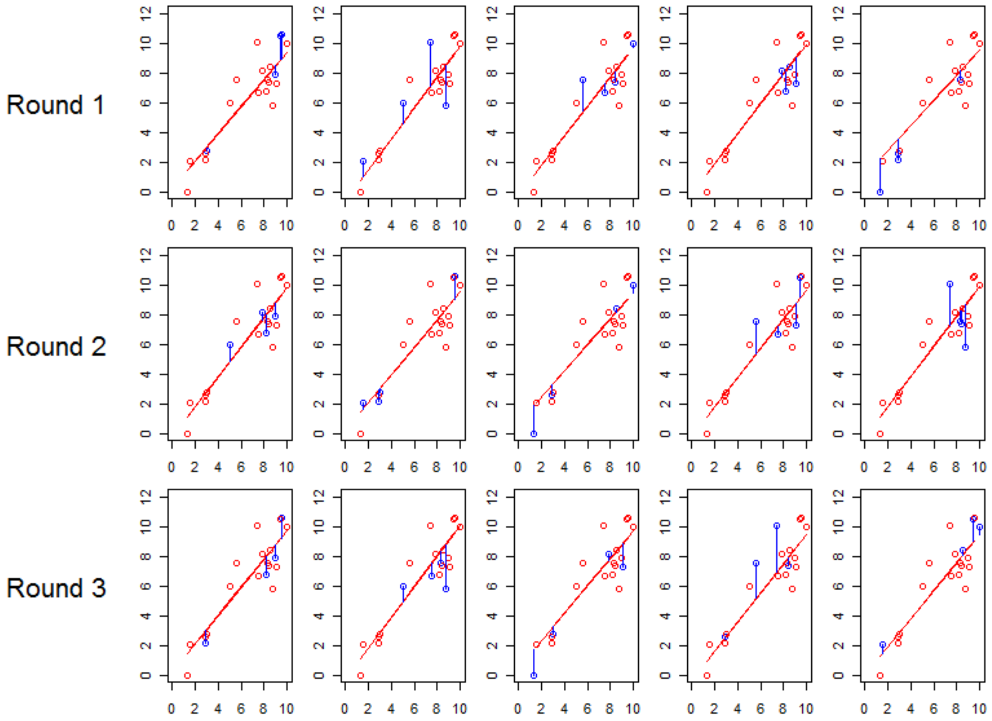
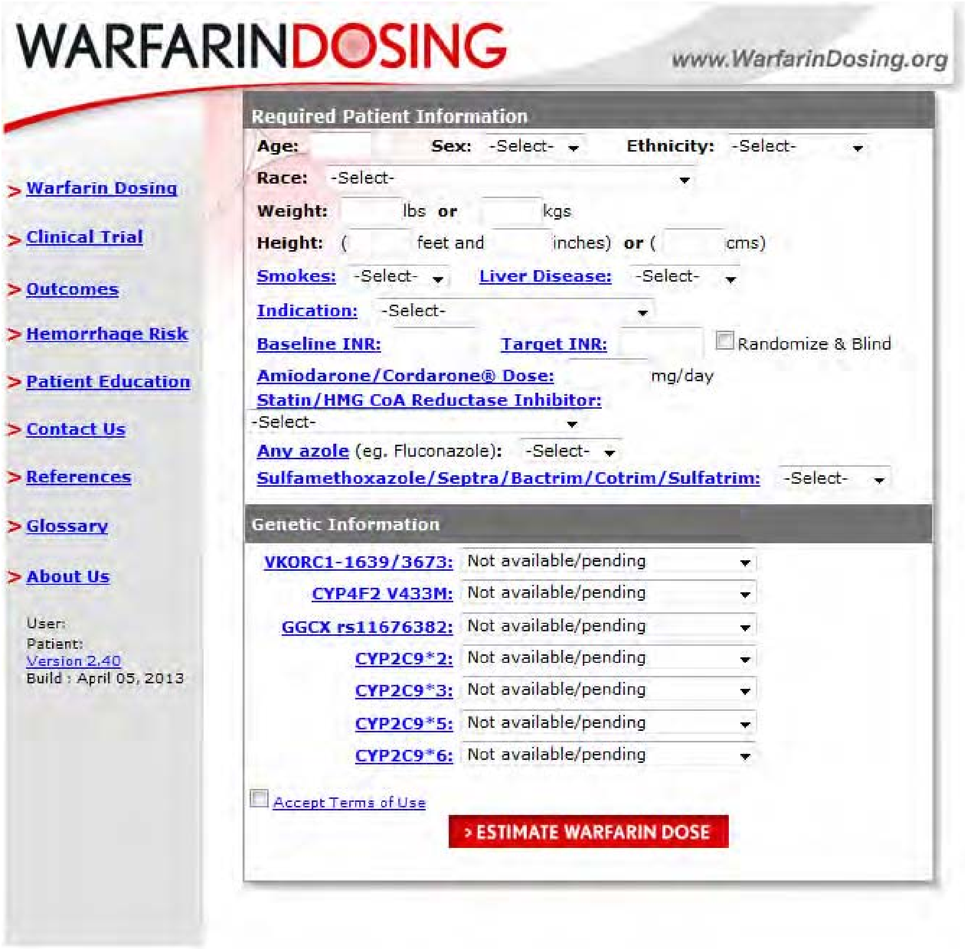


Figure 4.7: Repeated k-fold cross validation illustrated on the dataset in **Figure 4.2**. The 5-fold cross validation illustrated in **Figure 4.5** is repeated three times ($Q = 3$).

Lastly, one alternative method to CV is bootstrapping where, in its simplest form, instead of repeatedly analyzing subsets of the data, one repeatedly draws subsamples with replacement, fits the model on the subsamples and evaluates the predictive performance in the remaining observations not included in these bootstrap samples. This approach was introduced by *Efron (1983)*. There is typically a bias-variance trade-off between the two approaches with CV producing less biased but more variable estimates of the prediction error than those derived by the bootstrap approach. *Kohavi (1995)* compared CV and bootstrapping approaches and favored the stratified 10-fold CV. Since then alternatives have been developed such as the *.632+ bootstrap* method which appears to outperform CV. (*Efron and Tibshirani, 1997*) Nevertheless, in a high dimensional context, *Binder and Schumacher (2008)* recommend against approaches relying on sampling with replacement.

4.1.7 Presentation of the Model

An important final step in prediction model building is facilitating its usability. For instance, **Figure 4.8** below illustrates a snapshot of a free website created to help physicians and health professionals assess the appropriate initial warfarin dose to administer to patients based on the prediction algorithms that have been built for that purpose. Using this website, a patient's clinical and genetic information is entered and an estimate of the warfarin dose is then derived based on that information.



User:
Patient:
Version 2.40
Build : April 05, 2013

Required Patient Information

Age: Sex: -Select- Ethnicity: -Select-
Race: -Select-
Weight: lbs or kgs
Height: (feet and inches) or (cms)
Smokes: -Select- Liver Disease: -Select-
Indication: -Select-
Baseline INR: Target INR: □ Randomize & Blind
Amiodarone/Cordarone® Dose: mg/day
Statin/HMG CoA Reductase Inhibitor:
-Select-
Any azole (eg. Fluconazole): -Select-
Sulfamethoxazole/Septra/Bactrim/Cotrim/Sulfatrim: -Select-

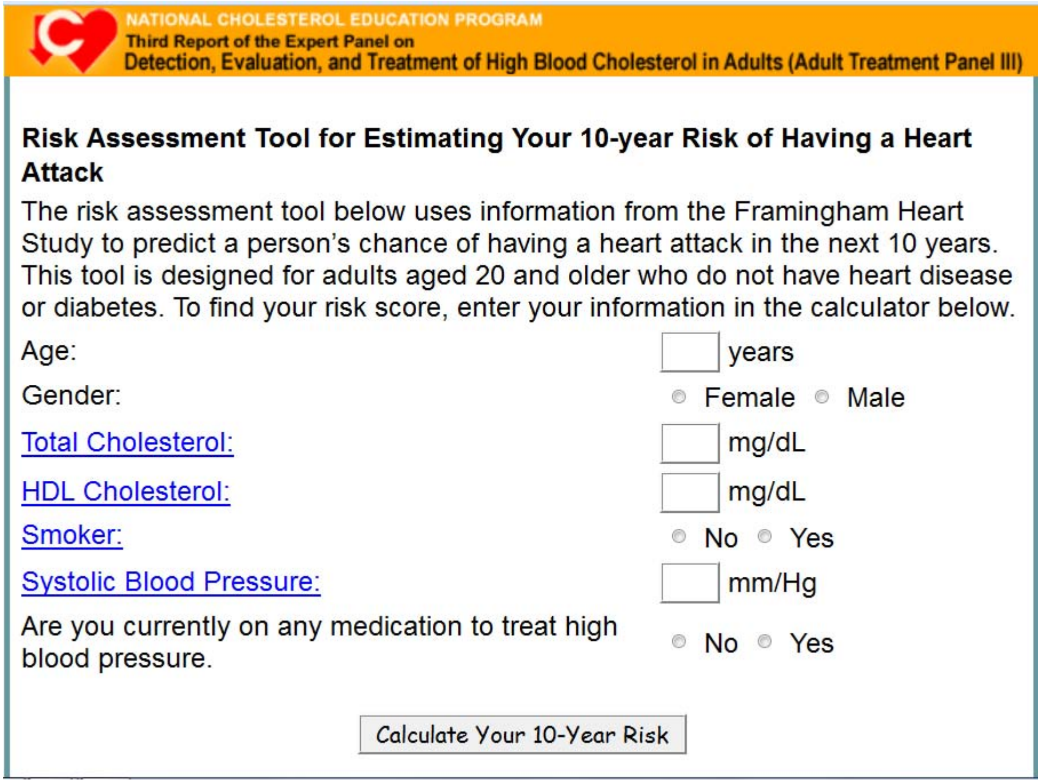
Genetic Information

VKORC1-1639/3673: Not available/pending
CYP4F2 V433M: Not available/pending
GGCX rs11676382: Not available/pending
CYP2C9*2: Not available/pending
CYP2C9*3: Not available/pending
CYP2C9*5: Not available/pending
CYP2C9*6: Not available/pending

□ Accept Terms of Use

Figure 4.8: Snapshot of the free website created to help physicians and health professionals assess the warfarin dose to administer to patients based on the predictive models that have been built for that purpose. Adapted from (*The Warfarin Dose Refinement Collaboration and International Warfarin Pharmacogenetics Consortium, 2013*).

Another example of prediction algorithms with facilitated implementation is the prediction of cardiovascular disease risk. **Figure 4.9** below illustrates the risk assessment tool based on data from the Framingham Heart Study (*Anderson et al., 1991*) available freely for anyone wishing to assess their 10-year risk of heart attack.



NATIONAL CHOLESTEROL EDUCATION PROGRAM
Third Report of the Expert Panel on
Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)

Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age:	□ years
Gender:	○ Female ○ Male
Total Cholesterol:	□ mg/dL
HDL Cholesterol:	□ mg/dL
Smoker:	○ No ○ Yes
Systolic Blood Pressure:	□ mm/Hg
Are you currently on any medication to treat high blood pressure.	○ No ○ Yes

Figure 4.9: Snapshot of the free website created for anyone wishing to assess their 10-year risk of heart attack. Adapted from (*National Heart Lung and Blood Institute, 2013*).

Both for warfarin dose and for cardiovascular risk prediction, several prediction algorithms exist and, in the case of cardiovascular risk, implementations of the different methods are also freely accessible (see for example the University of Edinburgh's cardiovascular risk calculator which allows the estimation of risk based on different risk prediction models (*The University of Edinburgh, 2010*)). Despite their complexity and variety, however, it is clear that an indispensable component of the successful application of prediction algorithms in practice is their simplicity of use.

4.2 Pitfalls and Limitations of Genetic Prediction Studies

Analogously to the previous chapter where we discussed the limitations of genetic association studies, we discuss next several of the limitations of genetic prediction studies by drawing examples from the literature as well as from our own work.

4.2.1 Direct, Indirect and Confounded Associations: Implications on Prediction Accuracy

In a training dataset, we may conduct a genetic association study to identify the genetic factors which contribute to the trait and evaluate the predictive performance of these factors in a testing dataset. For quantitative traits, Equation (4.8) illustrates that the explained variation by the model attributable to additive genetic factors is bounded by the heritability of the trait. Similarly, for binary (disease) traits, the AUC is bounded by the heritability of the trait (and the prevalence of the disease) (Equation (4.17)). This upper limit is achievable only if all causal variants contributing to the variability of the trait are known and if their effects are estimated without error (*Wray et al., 2013*).

However, as we saw in the last chapter, first of all, most of the genetic association studies are indirect association studies, where the genetic variants (SNPs) being analyzed are typically not the causal variants but rather those surrounding and correlated to the causal variants. Since not all causal variants may be tagged by neighboring SNPs on the genotyping chips, it is unlikely that all causal variants are identified in the training dataset. Second, it is possible that the effect of the actual causal SNP is larger than that estimated for its neighboring SNPs (effect gets "diluted"). Therefore, the estimated effects are also not accurate (they are underestimated).

Moreover, another factor impacting the prediction accuracy of the model is spurious associations due to population stratification and cryptic relatedness. The predictive accuracy might be inflated if the training and testing datasets arise from the same population (if, for example, using data splitting techniques to derive training and testing datasets) which is different from the target population for which the predictive model is aimed. A recent study has suggested for instance that population structure has confounded a genetic classifier for autism. (*Belgard et al., 2013*)

4.2.2 Hypothesis-Driven versus Hypothesis-Free

Prediction models in human traits incorporating genetic information have been so far restricted to the inclusion of a handful of genes the choice of which has been driven by prior knowledge. On the other hand, in animal breeding, genetic risk prediction of complex (quantitative) traits such as milk yield based on the whole-genome has been common practice. Particularly, breeding programs rely heavily on estimates of genetic values in the parent generation to predict offspring traits. The breeding value of an animal is defined as the sum of genetic effects of a breeding animal as measured by the performance of its progeny. In the absence of dominance, the genetic effects correspond to the additive genetic effects. The breeding value is estimated by summing the additive effects of the alleles and is referred to as the estimated breeding value or EBV.

The recent study by *Vazquez et al.(2012*) is one of the first to apply whole-genome prediction techniques used in animal breeding to predict risk for a complex disease phenotype in humans. The investigators compared four models to predict risk for skin cancer: (1) the baseline risk model including gender and cohort, (2) model 1 with family history added, (3) model 2 with geographic ancestry (based on genomic information) added and (4) model 1 with 41 000 SNPs across the genome. The AUCs for the four models were 0.53, 0.58, 0.62 and 0.64, respectively. That is, they found that the whole-genome prediction model outperformed all other models albeit not by much.

4.2.3 Guarding Against False Predictions

For quantitative traits, if the predictors in the model do not explain any of the phenotypic variation, then $\sigma_y^2 = \sigma_e^2$ and the population coefficient of determination (Equation (4.5)) is zero, that is, $P^2 = 0$. The estimated coefficient of determination, R^2, from the discovery (training) sample is biased with the expected value of the bias being $\frac{p-1}{n}$ (Equation (4.6)), where p is the number of variables including the intercept. That is, if randomly chosen p predictors are included in the model, the expected explained variation would be $R^2 = \frac{p-1}{n}$ with $R^2 \to 1$ as $p \to n$ even if the predictors are not associated with the phenotype of interest. Therefore, if the number of predictors, p, is large relative to the sample size, n,

R^2 may represent a significantly inflated estimate of the true explained variation to be expected in an independent (testing) dataset. This also illustrates why it is never a good idea to train a model and evaluate its predictive performance in the same dataset and that it is essential that the training and testing datasets be independent and drawn from the same target population.

4.2.4 Genetic Architecture of Complex Traits and Diseases

For monogenic (Mendelian) traits (see Section 3.3), genetic profiles provide 100% accurate predictions. Most traits, however, are not Mendelian and variants associated with these traits cover a wide spectrum of penetrance values. Common variants, such as those identified by GWASs, have typically low penetrance, while rare variants have high penetrance.

Predicting non-Mendelian traits is therefore based on probabilistic modeling and certainly involves errors in prediction accuracy. The more frequent the genetic variant, the lower its penetrance. The lower the penetrance, the weaker its role and the stronger the role of environmental factors in predicting the trait.

For example, the *BRCA1* and *BRCA2* mutations occur in less than 1% of the US population. Their penetrance for breast cancer at age 70 years was estimated at 0.57 and 0.49, respectively. (*Chen and Parmigiani, 2007)*. Nevertheless, family history remains one of the strongest predictors of the disease.

4.2.5 Study Sample Size and Marker Panel Coverage

Since the SNPs included in genome-wide SNP chips are typically not the causal SNPs the variation they capture does not fully explain the phenotypic variation that is due to genetic factors. Let h_M^2 denote the genetic variation captured by the genotyped SNPs and let h^2 be the heritability in the narrow-sense as defined before (Equation (3.13)). Then, typically, $h_M^2 \leq h^2$. This is especially true if rare variants contribute to the phenotypic variation as current marker panels capture only common genetic variation.

4.2.5.1 Quantitative Traits

For quantitative traits, we recall that the estimated proportion of phenotypic variation due to genetic factors is bounded by the heritability, that is, $R^2 \leq h^2$ (Equation (4.8)). It depends on the number of independently measured genetic variants, M, the proportion of total variance that they explain, h_M^2, and the sample size in the training (discovery) dataset, N_d. It is given by (*Wray et al., 2013*; *Daetwyler et al., 2008*)

$$R^2 = \frac{h_M^2}{1 + \frac{M}{N_d h_M^2}(1 - R^2)}. \tag{4.18}$$

Figure 4.10 below illustrates the variation of R^2 as a function of N_d, the discovery sample size, for different proportions of total variance explained by the genotyped SNPs, h_M^2. We see that high marker panel coverage and very large sample sizes in the discovery set are needed to achieve high R^2.

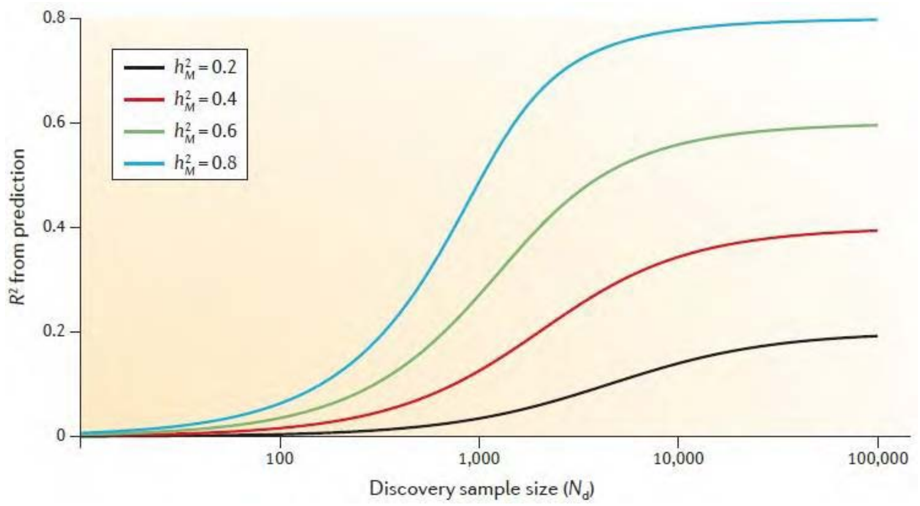


Figure 4.10: The explained variation, R^2, as a function of discovery sample size, N_d, for different proportions of total variance explained by the genotyped SNPs, h_M^2. Adapted from (*Wray et al., 2013*).

In other words, **Figure 4.10** above illustrates that unless the discovery sample sizes are very large ($N_d > 10\ 000$), $R^2 < h_M^2 \leq h^2$, that is, R^2 underestimates the heritability.

4.2.5.2 Binary Traits

For binary (disease) traits, we give some examples in **Table 4.3** and **Table 4.4** to illustrate the impact heritability, disease prevalence and discovery sample size have on the maximum achievable AUC. All results are taken from the analyses conducted on polygenic score studies by *Dudbridge (2013*).

Table 4.3 illustrates the maximum achievable AUC for two similarly heritable diseases, coronary artery disease and Crohn's disease but the former is roughly 50 times more prevalent than the latter.

		Coronary Artery Disease	**Crohn's Disease**
Prevalence, K		**0.056**	**0.001**
Heritability, h^2		0.72	0.76
Discovery Sample Size, N_d		~2000 cases / 1480 controls	
Variance explainable by markers, h_M^2	$h_M^2 = \frac{1}{2}h^2$	*0.547 (0.843)*	*0.620 (0.948)*
	$h_M^2 = h^2$	*0.592 (0.948)*	*0.727 (0.995)*

Table 4.3: Maximum achievable AUC (in italic) for two diseases, coronary artery disease (CAD) and Crohn's disease, with similar heritability but with very different prevalence. The maximum achievable AUC is given under the current discovery sample sizes and for marker panels explaining half ($h_M^2 = \frac{1}{2}h^2$) or full ($h_M^2 = h^2$) heritability. In parenthesis, the AUC achievable with infinite discovery sample sizes, that is, when $N_d \to \infty$, is given. Values adapted from Table 2 in (*Dudbridge, 2013*).

Alternatively, **Table 4.4** illustrates the maximum achievable AUC for two diseases, schizophrenia and prostate cancer, with similar prevalence but the former being almost twice more heritable than the latter.

		Schizophrenia	Prostate Cancer
Prevalence, K		0.01	0.024
Heritability, h^2		**0.80**	**0.44**
Discovery Sample Size, N_d		3322 cases / 3587 controls	1164 cases / 1113 controls
Variance explainable by markers, h_M^2	$h_M^2 = \frac{1}{2}h^2$	*0.62 (0.91)*	*0.52 (0.80)*
	$h_M^2 = h^2$	*0.72 (0.99)*	*0.54 (0.90)*

Table 4.4: Maximum achievable AUC (in italic) for two diseases, schizophrenia and prostate cancer, with similar prevalence but with very different heritability. The maximum achievable AUC is given for the current discovery sample sizes and for marker panels explaining half ($h_M^2 = \frac{1}{2}h^2$) or full ($h_M^2 = h^2$) heritability. In parenthesis, the AUC achievable with infinite discovery sample sizes, that is, when $N_d \to \infty$, is given. Values adapted from Table 3 in (*Dudbridge, 2013*).

We see from **Table 4.3** and **Table 4.4** that the maximum achievable AUC increases with higher marker coverage and that this increase is more significant for more heritable and/or diseases having a low prevalence. However, the sample size of the discovery dataset has by far the largest impact on the maximum achievable AUC where AUC levels deemed clinically useful (AUC > 0.75) are achieved only with infinite (hypothetical) sample sizes.

In the extreme, if we compare a rare, highly heritable disorder such as Crohn's disease ($K = 0.001, h^2 = 0.76$) with a common, modestly heritable disease such as breast cancer ($K = 0.036, h^2 = 0.44$), we see from **Figure 4.11** that extremely large sample sizes are needed to achieve clinically useful AUC levels in the order of tens of thousands for the former disease and hundreds of thousands to a million for the latter.

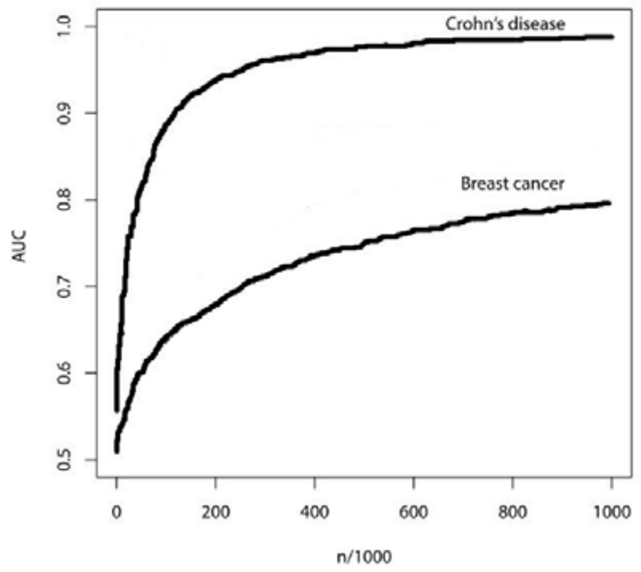


Figure 4.11: Maximum achievable AUC as a function of sample size (number of cases and controls) for Crohn's disease (rare, highly heritable) and for breast cancer (common, modestly heritable). Adapted from Figure 4 in (*Dudbridge, 2013*).

4.2.6 Phenotype Issues

It is clear that more precise measures of the phenotype can lead to improved prediction accuracy (due to reduced noise in the model). Nevertheless, we saw that the prediction performance of a model is positively correlated with the sample size of the discovery set and heritability and, in the case of binary traits, it is negatively correlated with the prevalence. While the sample size could be a component of the study design that can be controlled for, the properties of the studied phenotype such as its heritability and prevalence cannot be modified. It is, therefore, important to recognize the limitations of the study imposed by the phenotype being investigated and adjust performance expectations accordingly.

4.2.7 Clinical Use of the Findings

A final remark is needed on the correspondence between the OR of a binary predictor and its predictive accuracy. **Figure 4.12** below illustrates hypothetical accuracy curves corresponding to different ORs for a binary marker.

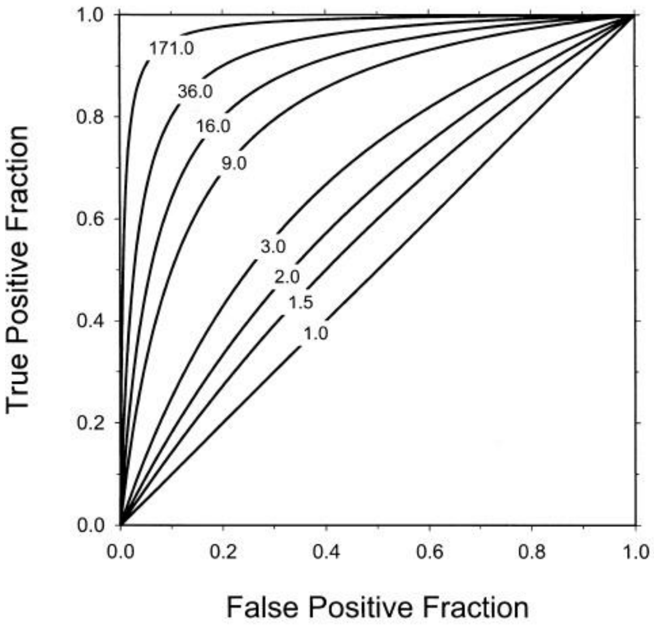


Figure 4.12: Correspondence between the accuracy curves and the ORs for binary markers with ORs as indicated on the various curves. Adapted from (*Pepe et al., 2004*).

Figure 4.12 illustrates that binary predictors with OR as high as 3 which is considered strong association in traditional epidemiological studies will be very poor classifiers. From the graph, it is apparent that very strong ORs ($OR \geq 9$) are needed to achieve clinically meaningful classification performance. Most of the genetic variants that have been identified through GWASs exert a small to moderate effects ($1.0 < OR \leq 1.5$) on the traits. Moreover, as we mentioned in the previous chapter, the OR is an estimate of the true risk ratio, RR, and $OR \approx RR$ only if the disease under study is rare, otherwise it overestimates it.

Based on family history alone, siblings would have the same risk estimate. Incorporating genetic information, therefore, may lead to differential risk estimates. Family history and genetic profiling have typically been regarded as competing sources of information (one or the other) in estimating risk but recent studies have shown that there are benefits to be gained by incorporating genetic information into risk prediction models that have already accounted for family history. (*Do et al., 2012*) In many cases, however, these improvements would be marginal and perhaps of limited clinical utility (*Ware, 2006*).

4.2.8 Modeling Strategies

We mentioned earlier in Section 4.1.3, that the most difficult step in building a clinical prediction model lies in the specification of the model. Oftentimes, many different models need to be specified and evaluated. Typically, there is a trade-off to be made between model complexity and prediction accuracy.

We conducted two studies to evaluate the prediction performance of different modeling approaches for a binary and a quantitative trait. We compared classical logistic regression with a data mining approach (binary tree with recursive partitioning) to predict natalizumab response one year after treatment onset in MS patients using non-genetic factors. Further, we extended our work on the simulated blood pressure data from GAW18 to the prediction context. We describe these two studies in Section 4.3.1 and Section 4.3.2, respectively.

4.3 Applications

4.3.1 Logistic Regression versus Binary Tree with Recursive Partitioning

4.3.1.1 Objective

A recent study on 48 MS patients investigated several potential clinical predictors of natalizumab response and found that the number of relapses one year prior treatment onset (alternatively expressed as the number of relapses per year and referred to as the annualized relapse rate) was the only good predictor of response. (*Sargento-Freitas et al., 2013*) So far, however, little is known on the non-genetic factors that influence natalizumab response. The objective of this study was to compare the performance of two modeling approaches, logistic regression and a classification tree algorithm, to predict response to natalizumab one year after treatment onset based on clinical, biological and radiological measures in French MS patients.

4.3.1.2 Dataset

A total of 531 patients from the BIONAT cohort (*Outteryck et al., 2013*) were included in our study. Detailed description of the clinical, biological and radiological characteristics by response status for these patients is provided in Appendix III (**Table III.1-Table III.3**). (The data shown in these tables are taken from the most recent version of the BIONAT database, frozen November 2013, and may not necessarily correspond to the data we had at hand at the time this study was started.)

All clinical variables, namely, gender, EDSS at treatment onset, number of relapses one year prior treatment onset, disease duration at treatment onset, previous immuno-modulatory treatment use and previous immuno-suppressory treatment use, were included in the study. Additionally, select biological and radiological variables at treatment onset were also included, namely, specific types of white blood cell (CD4, CD8 and CD19) counts and gadolinium (MRI GD+) enhancing lesions. For the majority of these patients, genotype data (SNPs from the *ITGA4* gene) were available. However, given the unconvincing role of the *ITGA4* in natalizumab response that we observed in the genetic association study that we conducted (Section 3.6.2), we did not include any genetic variants from this gene in the prediction model.

4.3.1.3 Response Definition

In **Table 3.13**, we defined *Responders, Non-Responders* and *Intermediary Responders* for natalizumab-treated MS patients one year after treatment onset. In the *ITGA4* candidate gene study, we excluded *Intermediary Responders*. In this study, *Intermediary Responders* were included along with the *Responders* group (that is, the *Responder*-classification criteria were relaxed). Any patient who had developed antibodies against natalizumab was excluded from the study. Moreover, only patients whose clinical, biological and radiological data had been validated by two independent neurologists were included in the study.

4.3.1.4 Methods

We compared the performance of two modeling techniques, a logistic regression (LR) model with stepwise backward selection and a binary tree with recursive partitioning (BT) algorithm to predict natalizumab response based on the non-genetic factors that we included in the study. All analyses were carried out with the R statistical software package (*R Core Team, 2013*).

We used the *rpart* R package (*Therneau and Atkinson, 2012*) for the implementation of the BT model. Specifically, the procedure involves splitting the dataset recursively until an optimal classification of the Responders/Non-Responders is achieved based on pre-defined parameters controlling the fit of the tree.

A tree consists of linked nodes. Any node in a binary tree can have at most two child nodes and at most one parent node. The node without a parent node is called the root node (at the top of the tree) and the node without any child nodes is called a terminal node. The depth of a node is the number of nodes on the path to its root node. The root node has a depth of zero. The resulting model fit is represented as a tree where the terminal nodes indicate the predicted response status.

At each node, except for terminal nodes, the algorithm identifies the best variable on which to split, that is, the variable that best classifies the patients compared to all other remaining variables. For each possible split variable, the algorithm computes an impurity measure. This measure is at a maximum when the split classifies an equal number of patients as Responders and Non-Responders and is at a minimum when the split results in only one class (either Responders or Non-Responders). The chosen split variable then is the one which has the minimum impurity. Various impurity measures are implemented in the *rpart* package. We used a measure based on the Gini index (GI) which is closely related to the AUC (discussed in Section 4.1.5.2) through the formula $GI = 2 * AUC - 1$. (*Gail and Pfeiffer, 2005*)

The largest advantage of BT over LR is its treatment of missing data. As long as the response status and at least one predictor variable are not missing, the patient can be included in the study. This is achieved as follows. First, if a potential split variable has missing values, the impurity measure is computed only over the observations which are not missing. Next, if a chosen split variable has missing values, *rpart* "imputes" the missing values by applying the partitioning algorithm to predict the split outcome based on the remaining independent variables. This approach could be particularly advantageous in studies with clinical data where missing data are the norm rather than the exception and our study is no different. **Table 4.5** below lists all predictor variables in our study dataset as well as the number and proportion (out of 531 patients) of missing values for each variable.

Predictors	**Missing n (%)**
CD19 count	124 (23)
CD8 count	53 (10)
CD4 count	51 (10)
Disease Duration	0 (0)
MRI GD$^+$ lesions	0 (0)
Relapses prior treatment	0 (0)
EDSS	0 (0)
Gender	0 (0)
Previous immuno-modulatory treatment use	0 (0)
Previous immuno-suppressory treatment use	0 (0)

Table 4.5: The number and proportion (out of 531 patients) of missing values per predictor variable in our study dataset.

Thus, in the BT analyses all 531 patients were included, while in the LR analyses only patients with non-missing data in all ten predictors were included.

Several parameters can be specified which influence the resulting model fit. These include

- the minimum number of observations in a node required to partition the data (defaults to 20);
- a parameter to control the selection of surrogate variables. For instance, the default option is to select surrogate variables leading to the highest number of correct classifications. Alternatively, surrogates may be selected based on the highest *proportion* of correct classifications calculated after excluding the number of missing values of the specific surrogate variable;
- the maximum depth of any node (defaults to 30).

For this study, we restricted the node depth to 3, a somewhat arbitrary choice. We used the default settings for the remaining parameters.

We used repeated two-fold cross validation with 100, 500 or 1 000 repetitions. In each repetition, we trained each of the two modeling techniques on one partition of the data and we evaluated its predictive performance in the other partition. We then evaluated the sensitivity, specificity, PPV, NPV and overall accuracy (Equations (4.12) through (4.16)) for both approaches averaging them over each validation run and then over the repetitions.

We also constructed ROC curves (using the *ROCR* R package, (*Sing et al., 2009*)) in one repetition chosen at random. The curves were constructed by varying the threshold of the predicted probabilities of being a Responder by each approach and evaluating the sensitivity and the specificity at each threshold. Lastly, we evaluated the number of times each predictor was chosen by the model (that is, selected by the backward selection algorithm for the LR approach or selected as a split variable for the BT approach) over the repetitions. We then ranked the predictors from most selected to least selected and compared the ranks across the approaches.

4.3.1.5 Results

The performance metrics obtained for each of the two modeling approaches are illustrated in **Table 4.6** below. The number of repetitions (100, 500 or 1 000) had little impact on the results and here we report only the results for 1000 repetitions.

On 1000 Repetitions	**LR**	**BT**
Sensitivity	0.92	0.83
Specificity	0.07	0.17
PPV	0.61	0.62
NPV	0.23	0.36
Accuracy	0.58	0.58

Table 4.6: Sensitivity, specificity, PPV, NPV and overall accuracy for the logistic regression (LR) and the binary trees with recursive partitioning (BR) models averaged over 1000 repetitions of two-fold cross validation runs.

From **Table 4.6** we see that LR classifies better Responders (sensitivity is 0.92 for LR versus 0.83 for BT), while BT classifies better Non-Responders (specificity is 0.07 for LR versus 0.17 for BT). Both approaches lead to similar PPVs while the BT predicts slightly better the Non-Responder status than LR (NPV is 0.36 for BT versus 0.23 for LR). Both approaches, however, have the same and quite poor overall accuracy of 0.58 (a random guess has an expected overall accuracy of 0.50).

The ROC curves generated from the predicted probabilities in a randomly chosen repetition are illustrated in **Figure 4.13** below. The ROC curve was smoother under the LR approach than under the BT approach due to more variable predicted probabilities although the range for both approaches was similar (from 0.36 to 0.79 for LR and from 0.44 to 0.82 for BT – right vertical axis in **Figure 4.13**). For both approaches, the curves were quite close to the diagonal line which, as we saw in **Figure 4.1** with classifier C, is equivalent to a random guess ($AUC \approx 0.5$).

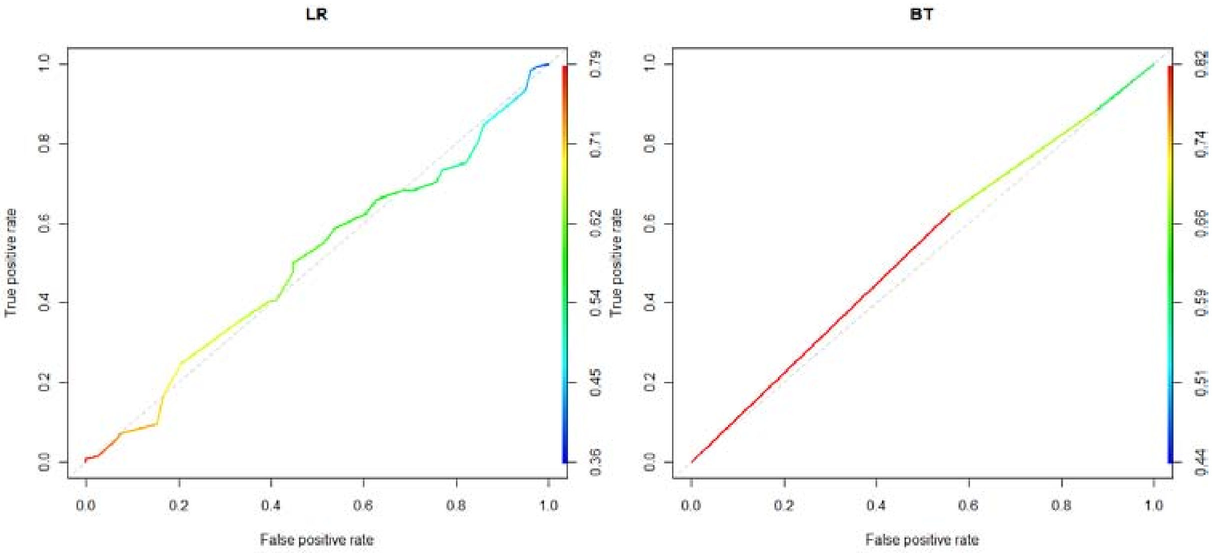


Figure 4.13: ROC curves for the LR and the BT approaches for a randomly chosen repetition.

Lastly, **Table 4.7** below illustrates the ranking of the predictors selected by each approach from most often to least often over 1000 repetitions.

LR		BT	
Top Predictor	**Selection Rate Over 1000 Repetitions (%)**	**Top Predictor**	**Selection Rate Over 1000 Repetitions (%)**
MRI GD$^+$ lesions	32	CD8 count	67
Relapses prior treatment	28	CD19 count	59
Disease duration	20	CD4 count	39
CD4 count	18	Disease duration	22
CD19 count	17	MRI GD$^+$ lesions	21
EDSS	12	Relapses prior treatment	18
Previous immuno-modulatory treatments	12	EDSS	4
CD8 count	8	Gender	4
Gender	7	Previous immuno-suppressory treatments	3
Previous immuno-suppressory treatments	6	Previous immuno-modulatory treatments	< 1

Table 4.7: Predictors ranked by the percentage of times over the 1000 repetitions that they were selected when using the logistic regression with backward selection (LR) or the binary tree with recursive partitioning model (BT).

As **Table 4.7** illustrates, the order as well as the selection rate of top predictors varies by approach. Interestingly, the most selected predictor by the BT approach, the CD8 count selected 67% of the times, was among the least selected by the LR approach (8%). Disease duration, on the other hand, was the only predictor with comparable selection rates between the two approaches (selection rate of 20% by LR and of 22% by BT). Overall, CD4 count, CD19 count and disease duration were identified among the top five predictors by both approaches. We note that the correlation (estimated in the complete dataset) between the CD4 and CD19 counts was 0.35, between CD4 count and disease duration was 0.004 and between CD19 count and disease duration was -0.05. Additionally, CD4 and CD8 counts were correlated (correlation = 0.51) as well as disease duration and EDSS (correlation = 0.33).

4.3.1.6 Concluding Remarks

In summary, from our study we were unable to identify any potential predictors of response of clinical usefulness irrespective of whether LR or BT was used. Nevertheless, from the variables included, both approaches consistently selected the immune cell counts and disease duration suggesting perhaps that these variables may be playing an important role in determining response. Referring back to the study by *Sargento-Freitas et al. (2013)*, we note that the number of relapses was more consistently selected by the LR approach than by the BT approach. Interestingly, *Sargento-Freitas et al. (2013)* also used logistic regression (with stepwise forward selection) to build their prediction model.

Of the two methodologies we evaluated, we recommend always considering the BT approach due to its significant advantage of missing data treatment even if in this instance it did not bring much benefit over the classical logistic regression. We note, however, that BT consistently selected the variables with the highest missing rates as predictors. In fact, one of the flaws of BT in its way of treating missing data is that a potential split variable with only two observations (extreme scenario of missing data) would be assigned an impurity measure of zero guarantying its selection as a split variable. The authors of the package acknowledge this bias of the method towards selection of variables with missing data and mention that it is unclear how it carries through to less extreme cases of missing data. (Section 5.1 in *Therneau and Atkinson (2012)*)

BT is more heavily parameterized than LR. Although we did not do it in this study, it is desirable to evaluate the impact on the results using several different parameterizations of BT. For instance, we could have evaluated whether increasing the depth of the tree from our arbitrarily set choice of 3 would have improved or worsened classification performance. Deeper trees may be expected to lead to improved performance but, of course, the deeper the tree, the larger the risk for over-fitting. Finally, we note that the BT approach can be easily applied to continuous and survival outcomes.

The clinical, biological and radiological information for this cohort is being continuously revised and updated and missing information, where possible, filled in. A much more ambitious study incorporating also genomic data for each patient is planned in the near future (see Section 5.2).

*Part of this work was presented as a poster at the UEPHA*MS Final Network Conference "Multiple Sclerosis and the Omics Spring" in April 2012.*

4.3.2 Single-Marker and Multi-Marker Models for Polygenic Score Analyses

4.3.2.1 Background

We are interested in building a genetic (based on SNPs) predictor of a trait. To achieve this we follow a typical two-step procedure. In the first step, the SNP effects are estimated and the SNPs that are associated with the trait are identified in a training set. In the second step, a model based on the SNPs identified in the first step is built and the model's predictive performance is evaluated in an independent testing set. In both steps, it is necessary to make important decisions on the choice of modeling approach that might have important implications on the performance of the prediction model.

Specifically, we discussed in the last chapter a recent trend to move away from the simplistic GWAS approach to estimating SNP effects individually to more complex whole-genome regression methods that allow the estimation of SNP effects simultaneously. Earlier in this chapter, we also discussed the fact that the accuracy of the SNP effect estimates has important implications on the prediction accuracy of the model (Section 4.2.1). Thus, any methodology that reduces the SNP effect estimation error in the first step of the procedure is likely to lead to improved predictive performance.

In the second step, the issue is how to combine the different SNP effects. It has been argued that whole-genome prediction methods are expected to achieve better prediction accuracy than methods which impose significance thresholds (such as, polygenic scores) and as such fail to capture all the genetic variability. (*Daetwyler et al., 2008*)

4.3.2.2 Material and Methods

As an extension to our study described in the previous chapter (see Section 3.6.4), we evaluated the R^2 (Equation (4.2)) and the MSE (Equation (4.9)) in the simulated diastolic blood pressure (DBP) and Q1 traits from the GAW18 dataset averaged over replicates 2 through 200.We recall that we constructed PS based on sets of varying number of top SNPs (10, 50, 100, 1 000, 5 000 and 10 000) with effect estimates derived by a single-marker (SM) approach and a multi-marker (BLUP) approach and evaluated the association of the PS with the respective trait of interest (DBP or Q1).

We also computed the average R^2 and MSE values using top SNPs from the *MAP4* gene alone. We recall that this gene contributed the most to DBP variation (6.5%). For SM, we used the two SNPs for which we adjusted the DBP trait for the effects of *MAP4* (SNPs 3_48040283 and 3_48064367 – SNP name refers to <chromosome>_<position>). For BLUP, we refitted the model but this time on chromosome 3 only (where the *MAP4* gene is located) and identified the top two SNPs which happened to be from that gene (SNPs 3_48024629 and 3_48096735).

Finally, for the BLUP approach only, we computed the average R^2 and MSE values using the expected breeding value (EBV) (see Section 4.2.2) which is essentially a PS computed with all 8.3 million SNPs (no selection of SNPs).

4.3.2.3 Results

Trait: DBP. **Figure 4.14 (A)** and **(B)** show, respectively, the estimated R^2 and MSE for the DBP trait averaged over replicates 2 through 200 for each of the six different SNP set sizes that we evaluated.

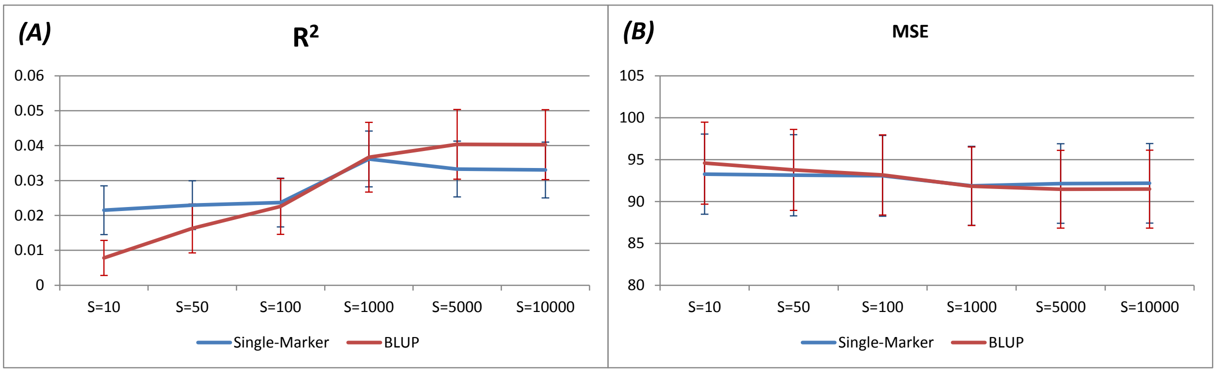


Figure 4.14: The mean explained variation, R^2 (**A**), and MSE (**B**) averaged over replicates 2 through 200 of the PS constructed using sets of top SNPs derived by the single-marker and the BLUP approaches for the DBP trait. Error bars indicate the standard deviations of the R^2 and the MSE values in each set.

We observed higher R^2 under SM than BLUP for smaller SNP set sizes ($S \leq 100$) while the reverse was true for larger SNP sets ($S > 1\,000$). The two approaches did not seem to differ on the MSE measure. Both approaches gave almost identical results for $S = 1\,000$, where $R^2 \cong 0.037$ and $MSE \cong 92$.

For the analyses including only the top two SNPs from the *MAP4* gene, we obtained mean R^2 of 0.0586 (± 0.013 standard deviations) and of 0.0178 (± 0.007 standard deviations) for the SM and BLUP approaches, respectively. Similarly, the MSE was 99 (± 4.8 standard deviations) and 103 (± 4.9 standard deviations) for the SM and BLUP approaches, respectively. These results are summarized in **Table 4.8** below.

DBP	**R^2(± standard deviations)**		**MSE (± standard deviations)**	
	Single-Marker	**BLUP**	**Single-Marker**	**BLUP**
Top two SNPs from *MAP4* gene	0.0586 (± 0.013)	0.0178 (± 0.007)	99 (± 4.8)	103 (± 4.9)
EBV (full genome - > 8.3 million SNPs)	N/A	0.0405 (± 0.009)	N/A	92 (± 4.7)

Table 4.8: Summary of mean R^2 and MSE values for DBP under the single-marker and BLUP approaches when taking only top two independent SNPs from the MAP4 gene, and under BLUP when taking all SNPs. N/A: Not Applicable.

On the other extreme, **Table 4.8** also gives the average R^2 and MSE values under BLUP when the EBV (> 8.3 million SNPs) was used as a predictor. For DBP, the mean R^2 was 0.0405 (± 0.009 standard deviations) and the MSE was 92 (± 4.7 standard deviations).

Trait: Q1. For the trait Q1, within each approach there was little difference across the various SNP set sizes in the mean R^2 and the MSE. On average, the mean R^2 was 0.001238 (± 0.002 standard deviations) and 0.001364 (± 0.002 standard deviations) for the SM and BLUP approaches, respectively. We evaluated the MSE at 114 (± 5.6 standard deviations) under both approaches.

We recall that Q1 was highly heritable but was uninfluenced by any of the genotyped SNPs. Therefore, we obtained R^2 values consistent with the expected bias in R^2 (see Section 4.2.3). Specifically, the PS association analyses were conducted on roughly 900 individuals (857 precisely) for whom both phenotypic and genotypic data were available. The linear regression model included the intercept and the PS obtained either by SM or BLUP and thus there were $p = 2$ parameters.

Following Equation (4.6), the expected bias, therefore, was ${(2-1)}/{857} = 0.001167$. Also, not surprisingly, the MSE values for Q1 were higher than those for DBP at 114 (± 5.6 standard deviations) under both approaches.

4.3.2.4 Discussion and Conclusions

For smaller SNP sets, ($S \leq 100$), higher R^2 values were obtained under SM than under BLUP. The reverse was true for larger SNP sets ($S \geq 5\ 000$). Based on the MSE metric, there was no difference between SM and BLUP. When no SNPs were associated with the trait (Q1), the R^2 and the MSE values were similar between the two approaches and across the different SNP set sizes. Further, the observed R^2 was close to its expected value and the MSE was higher than the MSE obtained for the DBP trait.

Based on the R^2 metric, for larger SNP set sizes, the BLUP approach would therefore be preferable. However, under this approach, it does not seem necessary to carry out SNP selection apriori. When the full genome was included in the PS (that is, the EBV was computed), the obtained R^2 (and MSE) was essentially identical to that for $S = 5\ 000$ and $S = 10\ 000$. In other words, under BLUP, including all SNPs (and thus correlated) versus only independent SNPs did not impact the results. This is not surprising due to the way BLUP estimates the SNP effects. In fact, in BLUP, the effects are spread over all SNPs which results in very small effects for a single SNP. (*Meuwissen, 2009*)

On the other hand, when only the two independent SNPs with strong effects from the *MAP4* gene were included, the R^2 was much higher than that of any of the SNP set sizes under SM and was also higher for the smaller SNP sets under BLUP ($S = 10$ and 50). In other words, under SM, including few SNPs with strong effects appears to predict at least as well or better than including thousands of small to moderate effect sizes. This comes at a price, however, because the MSE was much higher but more so under BLUP than under SM.

In conclusion, we come back once again to the underlying genetic architecture of the trait. Our results seem to suggest that if few SNPs exert a strong effect and a large number exert small or moderate effects on the trait, there is little benefit in terms of predictive performance in including the SNPs with small or moderate effect sizes in the PS using SM-derived estimates as weights. On the other hand, if the SNP sets sizes are allowed to include many SNPs, the BLUP-derived estimates are better suited as weights, and for that matter, also without the necessity to apply any apriori SNP selection criterion.

5 CONCLUSION

While the hype of genetic association studies is slowly winding down, that of prediction studies is just starting to pick up. What have we learned about MS in this era of genetic discoveries and of what utility have these findings been in the understanding and treatment of this debilitating disease?

5.1 What have we learned?

When trying to answer this question, it is necessary to distinguish between the susceptibility of the disease and response to its treatment as our knowledge and understanding of these two areas have not evolved in a similar fashion. It is of interest to note that genetic association studies for alternative phenotypes related to MS have also been conducted such as a GWAS of disease severity (*International Multiple Sclerosis Genetics Consortium, 2011*) and a GWAS of brain lesion distribution (*Gourraud et al., 2013*).

5.1.1 Susceptibility to Multiple Sclerosis

While genetic variation appears to be an important determinant of susceptibility to MS, as we discussed in the introduction, estimates of its heritability vary widely between 25% and 76%. (*Hawkes and Macgregor, 2009)* Despite that many genetic variants have been associated with susceptibility to MS through GWASs, as with other complex diseases, a large portion of its heritability still remains unexplained. Alternative methods to GWAS have been explored with the aim of explaining this missing heritability.

The successful application of polygenic score analyses in explaining the missing heritability in susceptibility to schizophrenia (*Purcell et al., 2009*) motivated a similar study in susceptibility to MS (*Bush et al., 2010)*. The investigators found that, through polygenic score analyses, which relax significantly the SNP inclusion significance threshold from the typically imposed genome-wide significance thresholds, they were able to explain approximately 3% of the variance in MS risk in another independent MS GWAS dataset. Another study using a multi-step logistic regression approach provided consistent evidence supporting a polygenic model of inheritance for MS risk (*Wang et al., 2011*).

To date, several studies to predict susceptibility to MS by incorporating a handful of known genetic risk variants have been carried out. For instance, *De Jager et al.* (2009) constructed a polygenic score based on 16 known MS susceptibility genes and evaluated the AUC for various models including the score with and without taking into consideration additional environmental factors. The obtained AUC for the purely genetic model was 0.70 and improved to 0.74 when gender was added (1340 cases / 1109 controls). In a second validation cohort (143 cases / 281 controls), the investigators obtained an AUC of 0.64 which improved to 0.68 after adding information on smoking and exposure to Epstein-Barr virus. The higher achieved AUC in the first validation cohort could be due to the much larger sample size. Nevertheless, in both validation cohorts, the models incorporating environmental factors improved the prediction accuracy.

Jafari et al. (2011) also used a polygenic score approach including a varying number of identified SNPs by GWASs without considering any environmental factors. The AUC increased as more SNPs were included ranging from 0.64 (six SNPs) to 0.69 (53 SNPs). The investigators also conducted a simulation study illustrating that many more common variants (at least 50) with weak effect sizes would be needed to achieve AUC of 0.85. This led the authors to conclude that even if new MS susceptibility variants are continuously being identified they will have limited utility in the clinical setting.

5.1.2 Response to Multiple Sclerosis Therapies

In the introductory chapter of this thesis (see Section 1.9), we reviewed the current state of the art for genetic association studies of response in MS. First and foremost, in contrast to MS susceptibility studies, mostly candidate gene studies have been conducted for treatment response in MS. Second, sample sizes were orders of magnitude smaller. For instance, the two GWASs on interferon response were carried out in 200-300 patients (discovery and replication combined). Third, the treatment response definition is extremely complex and is subjectively chosen by the investigator as no standardized response definition exists.

Fourth, a total of eight drugs have been approved for MS treatment based on five different acting mechanisms the most widely studied being interferon-β. Despite the many studies conducted, however, the genetic basis for interferon response is yet to be fully determined. Our study on interferon response (Section 3.6.1) and a polymorphism in the *OAS1* gene awaits finalization but, based on the results we obtained so far, if an association is detected, it is unlikely that it will be a strong one and will need replication in an independent dataset. To our knowledge, no genetic association studies have been carried out for natalizumab or fingolimod. The candidate gene association study we conducted (Section 3.6.2) did not show any evidence that *ITGA4* polymorphisms were associated with natalizumab response.

Fifth, the environmental factors influencing MS susceptibility have been widely studied. In the case of response to therapy in MS, not only the genetic associations are weak but also no non-genetic factors are known to date to be influencing response. Some studies have suggested that response in interferon is gender-specific but this remains to be fully investigated. Apart from a recent study that found the number of relapses to be a good predictor of natalizumab response (*Sargento-Freitas et al., 2013*), little is known on the influence of non-genetic variables on response. The results from our prediction study of natalizumab response based on non-genetic factors also did not lead to any definite conclusions (Section 4.3.1).

Finally, all studies on treatment response discussed so far were association studies. To our knowledge, no genetic prediction study has been reported.

In summary, the need for personalized treatment in MS is widely acknowledged. (*Sormani and De Stefano, 2013*) Nevertheless, many challenges need to be overcome before we consider this approach plausible in the context of MS and further studies in this area are clearly needed to achieve that. Such is the goal of a study designed as part of this dissertation work and which we discuss next.

5.2 Future Directions

The BIONAT cohort consists of roughly 1200 French MS patients with clinical, biological and radiological information. (*Outteryck et al., 2013*) The data management for this cohort requires a substantial effort to which I have also contributed.

A GWAS of natalizumab response is planned on this cohort. The GWAS data are being generated at the time of writing this manuscript. (The GWAS was scheduled for earlier throughout my thesis but there were unexpected delays in generating the genotype data).

Using this vast dataset, the main objective is to investigate the relationship between the different variables and response to natalizumab and to build a model predicting response by combining genetic with clinical, biological and radiological variables.

The model may include the top genetic variants identified through the GWAS (restricted to low dimensional setting) or include the whole genome (high-dimensional setting) similarly to the study by *Vazquez et al. (2012)* discussed in the previous chapter (see Section 4.2.2). Contrary to the model in the low-dimensional setting, with this approach it is not necessary to identify top SNPs apriori. On the other hand, however, the study by *Vazquez et al. (2012)* was conducted on more than 5100 subjects genotyped at 41 000 SNPs across the genome. From the BIONAT cohort, there may be at best 1200 subjects (not all patients may have sufficiently complete information to be included in the model) with at least half a million SNPs each. Thus, with these resources, the high-dimensional approach also has serious limitations.

It is unlikely that the derived prediction model would achieve the desirable properties to be directly useful in a clinical setting. In the best case scenario, any leading findings would have to await replication in a completely independent dataset and perhaps in subjects of different origin (non-

French). Such datasets are rare, expensive and take years to compile. Nevertheless, as *Kitsios and Kent (2012)* argue, "*we still lack even a basic framework that permits the multiple patient attributes that influence the effect of treatment (be they clinical, genetic, biological, or environmental) to be meaningfully integrated to support personalized decision making*". Therefore, this work could make an important contribution in constructing such a framework to develop integrated tools for personalized medicine in MS.

Finally, it is important to recognize that it is not only study limitations that influence the successful transition of scientific findings into medical practice. Many other factors are at play and, although not exhaustively, we try to address them as well.

5.3 Other Considerations of Pharmacogenomic Studies

So far we have been focused on methodological aspects of pharmacogenomic studies. We devote the remaining pages to important ethical, social, economic and legal considerations relevant to these studies.

Personal genetic testing is on the rise. There are already several companies to choose from which offer direct-to-consumer (DTC) genetic testing such as 23andme®, Navigenics®, Knome® and deCodeme® to name a few. Anyone who wishes to know their genetic risk of MS (and of numerous other traits or diseases) can simply mail in a sample of their saliva and, shortly after, will receive a report of their estimated risks. This report is based on the sort of risk-SNP association findings that we discussed in the third chapter of this thesis and that have been extensively compiled from the scientific literature by these companies. While recognizing that genetic testing can be a "*valuable tool to aid in diagnostic and therapeutic decisions*", the American Medical Association (AMA) warned against DTC genetic testing arguing that "*patients may spend money on direct-to-consumer genetic tests needlessly or misinterpret the results of the tests, causing them to make unnecessary or unhealthy lifestyle changes.*" (*Todd and Craine, 2011*)

The truth is that, as *Elton (2009*) put it, "*personal genomics is a wildly unregulated and woefully immature field.*" It took US legislature 13 years to approve the Genetic Information Nondiscrimination Act (GINA) in 2008. While previous protections existed before GINA, this was the first piece of law at the nationwide level to protect against genetic discrimination in health insurance and employment. GINA is far from perfect, however, as it does not address for instance other insurance policies such as life or disability insurance. It is clear that the substantial investment in genetic science that we have seen to date needs to be matched by innovation in regulation. (*Hudson et al., 2008*) The regulatory framework in Europe is as much, if not more, complex and challenging. (*Borry et al., 2012*)

In fact, such regulations would only foster collaboration from the public. Individuals might be more willing to participate in medical research that involves genetic testing and participate in projects such as the Personal Genome Project (*PersonalGenomes.org, 2013*) as they would no longer fear genetic discrimination.

Of course, protecting the information once it has been made available is one issue. The debate continues on whether genetic testing should be carried out only under the supervision of a medical professional trained to interpret genetic test results. France, for instance, does not allow the provision of DTC genetic tests, while Belgium does. (*Borry et al., 2012*) Moreover, how much of the genetic test results should be disclosed to the individual? If the test was carried out for one purpose in mind but substantial risk was noted for another condition, a so called "incidental finding", should the individual be made aware of it especially if no measures can be taken to prevent it?

The overall attitude towards genetic testing to optimize therapy decisions is somehow different from that to evaluate disease risk. An increasing number of FDA-approved drugs recommend genetic testing in their labels while, to date, genetic tests are required for three drugs (*Cohen et al., 2013*). The observed benefits so far of pharmacogenomics studies in patient stratification (for example, breast cancer and trastuzumab), in predicting adverse events (for example, HIV/AIDS and abacavir) and in determining optimal dose (for example, blood clotting and warfarin) have perhaps contributed to this more positive outlook. However, we must be cautiously optimistic.

Translating scientific findings into the clinical practice is a slow process. *Kitsios and Kent (2012)* described the process as comprising of the following three clinical translation phases: (1) basic biomedical research; (2) clinical research; and (3) clinical application. While warfarin, the most popular success case of pharmacogenomic studies, has passed phases (1) and (2), it has had very limited impact in phase (3). Despite recommended genetic testing by the FDA, Centers for Medicare and Medicaid Services (CMS) has refused to routinely pay for these tests arguing that "*the available evidence does not demonstrate that pharmacogenomic testing of CYP2C9 or VKORC1 alleles to predict warfarin responsiveness improves health outcomes in Medicare beneficiaries*." (*Centers for Medicare and Medicaid Services, 2013*) That is, despite mounting scientific evidence of the benefit of incorporating genetic information into warfarin dose prediction, for the purposes of health insurance policies this evidence may still fail to be sufficiently convincing.

Drug therapy remains suboptimal for a significant proportion of individuals (*Wolpe, 2009*). One of the promises of pharamacogenomics is that it will reduce health care costs by optimizing treatment options. Many are skeptical. It is true that the cost of genetic testing is quickly reaching affordable levels. However, the drug development process was complex, long and expensive even prior to the genomics era. Thus, it is not clear that incorporating genetic information into this process would make it more efficient. In fact, drugs with pharmacogenomics tests may be even more expensive targeting smaller populations and their cost-effectiveness would be far from certain.

In fact, personalized medicine reflects a fundamental shift in the conceptual basis of drug development moving away from blockbuster drugs to targeted therapies. (*Olivier et al., 2008*) One serious repercussion of this is that drug companies would be driven by "profitable genotypes" focusing research and development of drugs for the most prevalent genotypes. (*Wolpe, 2009*) Entire sections of a population and/or ethnic and racial groups may be left out in this process.

In conclusion, as we continue to improve our understanding of the genetic information we carry and develop the necessary technologies and methodologies to analyze it appropriately, we believe that it is only a matter of time before the clinical utility of genetic testing in improving the effectiveness and efficiency of preventive interventions for many conditions is attained. Nevertheless, as has been already suggested, measuring only the clinical utility of genetic testing is somewhat a restrictive concept of the overall benefit that it may lead to. (*Grosse and Khoury, 2006*) Economical, ethical, social and legal issues, of which we have only scratched the surface here, need to be considered in evaluating the net balance between benefit and harm of genetic testing, whether treatment optimization or risk prediction is the goal.

REFERENCES

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean, and G. P. Consortium, 2010, A map of human genome variation from population-scale sequencing: Nature, v. 467, p. 1061-73.

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. McVean, and G. P. Consortium, 2012, An integrated map of genetic variation from 1,092 human genomes: Nature, v. 491, p. 56-65.

Abecasis, G. R., L. R. Cardon, and W. O. Cookson, 2000a, A general test of association for quantitative traits in nuclear families: Am J Hum Genet, v. 66, p. 279-92.

Abecasis, G. R., W. O. Cookson, and L. R. Cardon, 2000b, Pedigree tests of transmission disequilibrium: Eur J Hum Genet, v. 8, p. 545-51.

Agresti, A., 2007, An introduction to categorical data analysis, 400 p.

Ali, R., R. S. Nicholas, and P. A. Muraro, 2013, Drugs in development for relapsing multiple sclerosis: Drugs, v. 73, p. 625-50.

Altshuler, D. M., R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, P. E. Bonnen, P. I. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, C. Gonzaga-Jauregui, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, Q. Zhang, M. J. Ghori, R. McGinnis, W. McLaren, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O. Wang, D. Reddy, C. N. Rotimi, C. D. Royal, R. R. Sharp, C. Zeng, L. D. Brooks, J. E. McEwen, and I. H. Consortium, 2010, Integrating common and rare genetic variation in diverse human populations: Nature, v. 467, p. 52-8.

Anderson, K. M., P. M. Odell, P. W. Wilson, and W. B. Kannel, 1991, Cardiovascular disease risk profiles: Am Heart J, v. 121, p. 293-8.

Armitage, P., 1955, Tests for linear trends in proportions and frequencies: Biometrics, v. 11, p. 375-386.

Ascherio, A., and K. L. Munger, 2007a, Environmental risk factors for multiple sclerosis. Part I: the role of infection: Ann Neurol, v. 61, p. 288-99.

Ascherio, A., and K. L. Munger, 2007b, Environmental risk factors for multiple sclerosis. Part II: Noninfectious factors: Ann Neurol, v. 61, p. 504-13.

Aulchenko, Y. S., D. J. de Koning, and C. Haley, 2007a, Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis: Genetics, v. 177, p. 577-85.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. van Duijn, 2007b, GenABEL: an R library for genome-wide association analysis: Bioinformatics, v. 23, p. 1294-6.

Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), 2009, Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20: Nat Genet, v. 41, p. 824-8.

Baranzini, S. E., J. Wang, R. A. Gibson, N. Galwey, Y. Naegelin, F. Barkhof, E. W. Radue, R. L. Lindberg, B. M. Uitdehaag, M. R. Johnson, A. Angelakopoulou, L. Hall, J. C. Richardson, R. K. Prinjha, A. Gass, J. J. Geurts, J. Kragt, M. Sombekke, H. Vrenken, P. Qualley, R. R. Lincoln, R. Gomez, S. J. Caillier, M. F. George, H. Mousavi, R. Guerrero, D. T. Okuda, B. A. Cree, A. J. Green, E. Waubant, D. S. Goodin, D. Pelletier, P. M. Matthews, S. L. Hauser, L. Kappos, C. H. Polman, and J. R. Oksenberg, 2009, Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis: Hum Mol Genet, v. 18, p. 767-78.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly, 2005, Haploview: analysis and visualization of LD and haplotype maps: Bioinformatics, v. 21, p. 263-5.

Belgard, T. G., I. Jankovic, J. K. Lowe, and D. H. Geschwind, 2013, Population structure confounds autism genetic classifier: Mol Psychiatry.

Benjamini, Y., and Y. Hochberg, 1995, Controlling the false discovery rate: a practical and powerful approach to multiple testing: Journal of the Royal Statistical Society. Series B (Methodological), v. 57, p. 289-300.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, et al., 2008, Accurate whole human genome sequencing using reversible terminator chemistry: Nature, v. 456, p. 53-9.

Binder, H., and M. Schumacher, 2008, Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples: Stat Appl Genet Mol Biol, v. 7, p. Article12.

Biomarkers Definitions Working Group, 2001, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework: Clin Pharmacol Ther, v. 69, p. 89-95.

Boerwinkle, E., and C. F. Sing, 1987, The use of measured genotype information in the analysis of quantitative phenotypes in man. III. Simultaneous estimation of the frequencies and effects of the apolipoprotein E polymorphism and residual polygenetic effects on cholesterol, betalipoprotein and triglyceride levels: Ann Hum Genet, v. 51, p. 211-26.

Bohossian, N., M. Saad, A. Legarra, and M. Martinez, 2013, Single-marker and multi-marker mixed models for polygenic score analysis in family-based data, BMC Proceedings. To appear.

Borry, P., R. E. van Hellemondt, D. Sprumont, C. F. Jales, E. Rial-Sebbag, T. M. Spranger, L. Curren, J. Kaye, H. Nys, and H. Howard, 2012, Legislation on direct-to-consumer genetic testing in seven European countries: Eur J Hum Genet, v. 20, p. 715-21.

Bush, W. S., S. J. Sawcer, P. L. de Jager, J. R. Oksenberg, J. L. McCauley, M. A. Pericak-Vance, J. L. Haines, and I. M. S. G. C. (IMSGC), 2010, Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come: Am J Hum Genet, v. 86, p. 621-5.

Byun, E., S. J. Caillier, X. Montalban, P. Villoslada, O. Fernández, D. Brassat, M. Comabella, J. Wang, L. F. Barcellos, S. E. Baranzini, and J. R. Oksenberg, 2008, Genome-wide pharmacogenomic analysis of the response to interferon beta therapy in multiple sclerosis: Arch Neurol, v. 65, p. 337-44.

Centers for Medicare and Medicaid Services, 2013, Pharmacogenomic testing to predict warfarin responsiveness.

Charcot, J.-M., 1868, Histologie de la sclérose en plaques., Gazette Hôpitaux, p. 554, 557–558, 566.

Chen, R., G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. Lam, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, Y. Cheng, M. J. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O'Huallachain, J. T. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A. P. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M. A. Blasco, P. L. Greenberg, P. Snyder, T. E. Klein, R. B. Altman, A. J. Butte, E. A. Ashley, M. Gerstein, K. C. Nadeau, H. Tang, and M. Snyder, 2012, Personal omics profiling reveals dynamic molecular and medical phenotypes: Cell, v. 148, p. 1293-307.

Chen, S., and G. Parmigiani, 2007, Meta-analysis of BRCA1 and BRCA2 penetrance: J Clin Oncol, v. 25, p. 1329-33.

Cirulli, E. T., and D. B. Goldstein, 2010, Uncovering the roles of rare variants in common disease through whole-genome sequencing: Nat Rev Genet, v. 11, p. 415-25.

Clamp, M., B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, M. Kellis, K. Lindblad-Toh, and E. S. Lander, 2007, Distinguishing protein-coding and noncoding genes in the human genome: Proc Natl Acad Sci U S A, v. 104, p. 19428-33.

Cochran, W. G., 1954, Some methods for strengthening the common chi-squared tests: Biometrics, v. 10, p. 417-451.

Cohen, J., A. Wilson, and K. Manzolillo, 2013, Clinical and economic challenges facing pharmacogenomics: Pharmacogenomics J, v. 13, p. 378-88.

Comabella, M., D. W. Craig, M. Camiña-Tato, C. Morcillo, C. Lopez, A. Navarro, J. Rio, X. Montalban, R. Martin, and B. S. Group, 2008, Identification of a novel risk locus for multiple sclerosis at 13q31.3 by a pooled genome-wide scan of 500,000 single nucleotide polymorphisms: PLoS One, v. 3, p. e3490.

Comabella, M., D. W. Craig, C. Morcillo-Suárez, J. Río, A. Navarro, M. Fernández, R. Martin, and X. Montalban, 2009, Genome-wide scan of 500,000 single-nucleotide polymorphisms among responders and nonresponders to interferon beta therapy in multiple sclerosis: Arch Neurol, v. 66, p. 972-8.

Contasta, I., R. Totaro, P. Pellegrini, T. Del Beato, A. Carolei, and A. M. Berghella, 2012, A gender-related action of IFNbeta-therapy was found in multiple sclerosis: J Transl Med, v. 10, p. 223.

Cook, N. R., 2007, Use and misuse of the receiver operating characteristic curve in risk prediction: Circulation, v. 115, p. 928-35.

Cotte, S., N. von Ahsen, N. Kruse, B. Huber, A. Winkelmann, U. K. Zettl, M. Starck, N. König, N. Tellez, J. Dörr, F. Paul, F. Zipp, F. Lühder, H. Koepsell, H. Pannek, X. Montalban, R. Gold, and A. Chan, 2009, ABC-transporter gene-polymorphisms are potential pharmacogenetic markers for mitoxantrone response in multiple sclerosis: Brain, v. 132, p. 2517-30.

Couturier, N., 2010, Pharmacogenetics of MS: Response to Tysabri Treatment.

Couturier, N., F. Bucciarelli, R. N. Nurtdinov, M. Debouverie, C. Lebrun-Frenay, G. Defer, T. Moreau, C. Confavreux, S. Vukusic, I. Cournu-Rebeix, R. H. Goertsches, U. K. Zettl, M. Comabella, X. Montalban, P. Rieckmann, F. Weber, B. Müller-Myhsok, G. Edan, B. Fontaine, L. T. Mars, A. Saoudi, J. R. Oksenberg, M. Clanet, R. S. Liblau, and D. Brassat, 2011, Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility: Brain, v. 134, p. 693-703.

Crick, F. H., 1956, Ideas on Protein Synthesis.

Crick, F. H., 1958, On protein synthesis: Symp Soc Exp Biol, v. 12, p. 138-63.

Crick, F. H., 1990, What Mad Pursuit: A Personal View of Scientific Discovery, Basic Books.

Cunningham, S., C. Graham, M. Hutchinson, A. Droogan, K. O'Rourke, C. Patterson, G. McDonnell, S. Hawkins, and K. Vandenbroeck, 2005, Pharmacogenomics of responsiveness to interferon IFN-beta treatment in multiple sclerosis: a genetic screen of 100 type I interferon-inducible genes: Clin Pharmacol Ther, v. 78, p. 635-46.

Cénit, M. D., F. Blanco-Kelly, V. de las Heras, M. Bartolomé, E. G. de la Concha, E. Urcelay, R. Arroyo, and A. Martínez, 2009, Glypican 5 is an interferon-beta response gene: a replication study: Mult Scler, v. 15, p. 913-7.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008, Accuracy of predicting the genetic risk of disease using a genome-wide approach: PLoS One, v. 3, p. e3395.

De Jager, P. L., L. B. Chibnik, J. Cui, J. Reischl, S. Lehr, K. C. Simon, C. Aubin, D. Bauer, J. F. Heubach, R. Sandbrink, M. Tyblova, P. Lelkova, E. Havrdova, C. Pohl, D. Horakova, A. Ascherio, D. A. Hafler, E. W. Karlson, S. c. o. t. B. study, S. c. o. t. B. study, S. c. o. t. L. study, and S. c. o. t. C. study, 2009, Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score: Lancet Neurol, v. 8, p. 1111-9.

de los Campos, G., D. Gianola, and D. B. Allison, 2010, Predicting genetic predisposition in humans: the promise of whole-genome markers: Nat Rev Genet, v. 11, p. 880-6.

Dempster, E. R., and I. M. Lerner, 1950, Heritability of Threshold Characters: Genetics, v. 35, p. 212-36.

Do, C. B., D. A. Hinds, U. Francke, and N. Eriksson, 2012, Comparison of family history and SNPs for predicting risk of complex disease: PLoS Genet, v. 8, p. e1002973.

Donald, A., and T. Greenhalgh, 2000, A hands-on guide to evidence based healthcare: practice and implementation: Oxford, Blackwell Science.

Dudbridge, F., 2013, Power and predictive accuracy of polygenic risk scores: PLoS Genet, v. 9, p. e1003348.

Efron, B., 1983, Estimating the error rate of a prediction rule: improvement on cross-validation: Journal of the American Statistical Association, v. 78, p. 316-331.

Efron, B., and R. Tibshirani, 1997, Improvements on cross-validation: the .632+ bootstrap method: Journal of the American Statistical Association, v. 92, p. 548-560.

Ehret, G. B., P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, A. V. Smith, M. D. Tobin, G. C. Verwoert, S. J. Hwang, V. Pihur, P. Vollenweider, P. F. O'Reilly, N. Amin, J. L. Bragg-Gresham, A. Teumer, N. L. Glazer, L. Launer, J. H. Zhao, Y. Aulchenko, S. Heath, S. Sõber, A. Parsa, J. Luan, P. Arora, A. Dehghan, F. Zhang, G. Lucas, A. A. Hicks, A. U. Jackson, J. F. Peden, T. Tanaka, S. H. Wild, I. Rudan, W. Igl, Y. Milaneschi, A. N. Parker, C. Fava, J. C. Chambers, E. R. Fox, M. Kumari, M. J. Go, P. van der Harst, W. H. Kao, M. Sjögren, D. G. Vinay, M. Alexander, Y. Tabara, S. Shaw-Hawkins, P. H. Whincup, Y. Liu, G. Shi, J. Kuusisto, B. Tayo, M. Seielstad, X. Sim, K. D. Nguyen, T. Lehtimäki, G. Matullo, Y. Wu, T. R. Gaunt, N. C. Onland-Moret, M. N. Cooper, C. G. Platou, E. Org, R. Hardy, S. Dahgam, J. Palmen, V. Vitart, P. S. Braund, T. Kuznetsova, C. S. Uiterwaal, A. Adeyemo, W. Palmas, H. Campbell, B. Ludwig, M. Tomaszewski, I. Tzoulaki, N. D. Palmer, T. Aspelund, M. Garcia, Y. P. Chang, J. R. O'Connell, N. I. Steinle, D. E. Grobbee, D. E. Arking, S. L. Kardia, A. C. Morrison, D. Hernandez, S. Najjar, W. L. McArdle, D. Hadley, M. J. Brown, J. M. Connell, A. D. Hingorani, I. N. Day, D. A. Lawlor, J. P. Beilby, R. W. Lawrence, R. Clarke, et al., 2011, Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk: Nature, v. 478, p. 103-9.

Elton, C., 2009, The burden of knowing, Boston Magazine.

FDA, 2008, E15 Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories: Guidance for Industry.

FDA, 2013, Table of Pharmacogenomic Biomarkers in Drug Labels.

Fernández, O., V. Fernández, C. Mayorga, M. Guerrero, A. León, J. A. Tamayo, A. Alonso, F. Romero, L. Leyva, G. Luque, and E. de Ramón, 2005, HLA class II and response to interferon-beta in multiple sclerosis: Acta Neurol Scand, v. 112, p. 391-4.

Fox, R. J., F. Bethoux, M. D. Goldman, and J. A. Cohen, 2006, Multiple sclerosis: advances in understanding, diagnosing, and treating the underlying disease: Cleve Clin J Med, v. 73, p. 91-102.

Fox, R. J., and J. A. Cohen, 2001, Multiple sclerosis: the importance of early recognition and treatment: Cleve Clin J Med, v. 68, p. 157-71.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, W. Sun, H. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, et al., 2007, A second generation human haplotype map of over 3.1 million SNPs: Nature, v. 449, p. 851-61.

Frazer, K. A., S. S. Murray, N. J. Schork, and E. J. Topol, 2009, Human genetic variation and its contribution to complex traits: Nat Rev Genet, v. 10, p. 241-51.

Fromont, A., C. Binquet, E. A. Sauleau, I. Fournel, A. Bellisario, J. Adnet, A. Weill, S. Vukusic, C. Confavreux, M. Debouverie, L. Clerc, C. Bonithon-Kopp, and T. Moreau, 2010, Geographic variations of multiple sclerosis in France: Brain, v. 133, p. 1889-99.

Fusco, C., V. Andreone, G. Coppola, V. Luongo, F. Guerini, E. Pace, C. Florio, G. Pirozzi, R. Lanzillo, P. Ferrante, P. Vivo, M. Mini, M. Macrì, G. Orefice, and M. L. Lombardi, 2001, HLA-DRB1*1501 and response to copolymer-1 therapy in relapsing-remitting multiple sclerosis: Neurology, v. 57, p. 1976-9.

Gail, M. H., and R. M. Pfeiffer, 2005, On criteria for evaluating models of absolute risk: Biostatistics, v. 6, p. 227-39.

Gauderman, J., and J. Morrison, 2009, Quanto: A computer program for power and sample size calculations for genetic-epidemiology studies.

Geisser, S., 1975, The predictive sample reuse method with applications: Journal of the American Statistical Association, v. 70, p. 320-328.

GeneticsSuite, 2013, Recombination: Creating variation in gametes.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009, Additive genetic variability and the Bayesian alphabet: Genetics, v. 183, p. 347-63.

Goddard, M., 2009, Genomic selection: prediction of accuracy and maximisation of long term response: Genetica, v. 136, p. 245-57.

Goldenberg, M. M., 2012, Multiple sclerosis review: P T, v. 37, p. 175-84.

González, J. R., L. Armengol, X. Solé, E. Guinó, J. M. Mercader, X. Estivill, and V. Moreno, 2007, SNPassoc: an R package to perform whole genome association studies: Bioinformatics, v. 23, p. 644-5.

Gordon, D., S. J. Finch, M. Nothnagel, and J. Ott, 2002, Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms: Hum Hered, v. 54, p. 22-33.

Gourraud, P. A., M. Sdika, P. Khankhanian, R. G. Henry, A. Beheshtian, P. M. Matthews, S. L. Hauser, J. R. Oksenberg, D. Pelletier, and S. E. Baranzini, 2013, A genome-wide association study of brain lesion distribution in multiple sclerosis: Brain, v. 136, p. 1012-24.

Grosse, S. D., and M. J. Khoury, 2006, What is the clinical utility of genetic testing?: Genet Med, v. 8, p. 448-50.

Grossman, I., N. Avidan, C. Singer, D. Goldstaub, L. Hayardeny, E. Eyal, E. Ben-Asher, T. Paperna, I. Pe'er, D. Lancet, J. S. Beckmann, and A. Miller, 2007, Pharmacogenetics of glatiramer acetate therapy for multiple sclerosis reveals drug-response markers: Pharmacogenet Genomics, v. 17, p. 657-66.

Guedj, M., G. Nuel, and B. Prum, 2008, A note on allelic tests in case-control association studies: Ann Hum Genet, v. 72, p. 407-9.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011, Extension of the bayesian alphabet for genomic selection: BMC Bioinformatics, v. 12, p. 186.

Hafler, D. A., A. Compston, S. Sawcer, E. S. Lander, M. J. Daly, P. L. De Jager, P. I. de Bakker, S. B. Gabriel, D. B. Mirel, A. J. Ivinson, M. A. Pericak-Vance, S. G. Gregory, J. D. Rioux, J. L. McCauley, J. L. Haines, L. F. Barcellos, B. Cree, J. R. Oksenberg, S. L. Hauser, and I. M. S. G. Consortium, 2007, Risk alleles for multiple sclerosis identified by a genomewide study: N Engl J Med, v. 357, p. 851-62.

Hardy, G. H., 1908, MENDELIAN PROPORTIONS IN A MIXED POPULATION: Science, v. 28, p. 49-50.

Hastings, A., 2001, Hardy–Weinberg Theorem, Encyclopedia of Life Sciences, Macmillan Publishers Ltd, Nature Publishing Group.

Hawkes, C. H., and A. J. Macgregor, 2009, Twin studies and the heritability of MS: a conclusion: Mult Scler, v. 15, p. 661-7.

Hemminki, K., X. Li, J. Sundquist, J. Hillert, and K. Sundquist, 2009, Risk for multiple sclerosis in relatives and spouses of patients diagnosed with autoimmune and related conditions: Neurogenetics, v. 10, p. 5-11.

Henderson, C. R., 1975, Best linear unbiased estimation and prediction under a selection model: Biometrics, v. 31, p. 423-447.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, 2009, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits: Proc Natl Acad Sci U S A, v. 106, p. 9362-7.

Hirschhorn, J. N., and M. J. Daly, 2005, Genome-wide association studies for common diseases and complex traits: Nat Rev Genet, v. 6, p. 95-108.

Hudson, K. L., M. K. Holohan, and F. S. Collins, 2008, Keeping pace with the times--the Genetic Information Nondiscrimination Act of 2008: N Engl J Med, v. 358, p. 2661-3.

Huynh, T., 2010, The multiple sclerosis market: Nat Rev Drug Discov, v. 9, p. 759-60.

International HapMap Consortium, 2003, The International HapMap Project: Nature, v. 426, p. 789-96.

International HapMap Consortium, 2005, A haplotype map of the human genome: Nature, v. 437, p. 1299-320.

International Human Genome Sequencing Consortium, 2004, Finishing the euchromatic sequence of the human genome: Nature, v. 431, p. 931-45.

International Multiple Sclerosis Genetics Consortium, 2011, Genome-wide association study of severity in multiple sclerosis: Genes Immun, v. 12, p. 615-25.

Jafari, N., L. Broer, C. M. van Duijn, A. C. Janssens, and R. Q. Hintzen, 2011, Perspectives on the use of multiple sclerosis risk genes for prediction: PLoS One, v. 6, p. e26493.

Jakkula, E., V. Leppä, A. M. Sulonen, T. Varilo, S. Kallio, A. Kemppinen, S. Purcell, K. Koivisto, P. Tienari, M. L. Sumelahti, I. Elovaara, T. Pirttilä, M. Reunanen, A. Aromaa, A. B. Oturai, H. B. Søndergaard, H. F. Harbo, I. L. Mero, S. B. Gabriel, D. B. Mirel, S. L. Hauser, L. Kappos, C. Polman, P. L. De Jager, D. A. Hafler, M. J. Daly, A. Palotie, J. Saarela, and L. Peltonen, 2010, Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene: Am J Hum Genet, v. 86, p. 285-91.

Kitsios, G. D., and D. M. Kent, 2012, Personalised medicine: not just in our genes: BMJ, v. 344, p. e2161.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh, 2005, Complement factor H polymorphism in age-related macular degeneration: Science, v. 308, p. 385-9.

Klein, T. E., R. B. Altman, N. Eriksson, B. F. Gage, S. E. Kimmel, M. T. Lee, N. A. Limdi, D. Page, D. M. Roden, M. J. Wagner, M. D. Caldwell, J. A. Johnson, and I. W. P. Consortium, 2009, Estimation of the warfarin dose with clinical and pharmacogenetic data: N Engl J Med, v. 360, p. 753-64.

Kleinewietfeld, M., A. Manzel, J. Titze, H. Kvakan, N. Yosef, R. A. Linker, D. N. Muller, and D. A. Hafler, 2013, Sodium chloride drives autoimmune disease by the induction of pathogenic TH17 cells: Nature, v. 496, p. 518-22.

Kohavi, R., 1995, A study of cross-validation and bootstrap for accuracy estimation and model selection: 14th International Joint Conference on Artificial Intelligence, p. 1137-1143.

Kumar, D., 2007, From evidence-based medicine to genomic medicine: Genomic Med, v. 1, p. 95-104.

Kurtzke, J. F., 1983, Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS): Neurology, v. 33, p. 1444-52.

Kurtzke, J. F., 2000, Multiple sclerosis in time and space--geographic clues to cause: J Neurovirol, v. 6 Suppl 2, p. S134-40.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T.

Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al., 2001, Initial sequencing and analysis of the human genome: Nature, v. 409, p. 860-921.

Lee, S. J., 2012, Clinical Application of CYP2C19 Pharmacogenetics Toward More Personalized Medicine: Front Genet, v. 3, p. 318.

Legarra, A., A. Ricard, and O. Filangi, 2011, GS3: Genomic Selection - Gibbs Sampling - Gauss Seidel (and Bayes Cpi).

Lessard, C. J., J. A. Ice, I. Adrianto, G. B. Wiley, J. A. Kelly, P. M. Gaffney, C. G. Montgomery, and K. L. Moser, 2012, The genomics of autoimmune disease in the era of genome-wide association studies and beyond: Autoimmun Rev, v. 11, p. 267-75.

Levy, S., G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y. H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and J. C. Venter, 2007, The diploid genome sequence of an individual human: PLoS Biol, v. 5, p. e254.

Lewontin, R. C., 1964, The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models: Genetics, v. 49, p. 49-67.

Lewontin, R. C., and K.-i. Kojima, 1960, The Evolutionary Dynamics of Complex Polymorphisms: Evolution, v. 14, p. 458-472.

Leyva, L., O. Fernández, M. Fedetz, E. Blanco, V. E. Fernández, B. Oliver, A. León, M. J. Pinto-Medel, C. Mayorga, M. Guerrero, G. Luque, A. Alcina, and F. Matesanz, 2005, IFNAR1 and IFNAR2 polymorphisms confer susceptibility to multiple sclerosis but not to interferon-beta treatment response: J Neuroimmunol, v. 163, p. 165-71.

Li, M., C. Li, and W. Guan, 2008, Evaluation of coverage variation of SNP chips for genome-wide association studies: Eur J Hum Genet, v. 16, p. 635-43.

London School of Hygiene and Tropical Medicine, 2013, Basic Epidemiology module: Study Design Types.

Lopez-Diego, R. S., and H. L. Weiner, 2008, Novel therapeutic strategies for multiple sclerosis--a multifaceted adversary: Nat Rev Drug Discov, v. 7, p. 909-25.

López-Gómez, C., A. Pino-Ángeles, T. Órpez-Zafra, M. J. Pinto-Medel, B. Oliver-Martos, J. Ortega-Pinazo, C. Arnáiz, C. Guijarro-Castro, J. Varadé, R. Álvarez-Lafuente, E. Urcelay, F. Sánchez-Jiménez, Ó. Fernández, and L. Leyva, 2013, Candidate gene study of TRAIL and TRAIL receptors: association with response to interferon beta therapy in multiple sclerosis patients: PLoS One, v. 8, p. e62540.

Maher, B., 2008, Personal genomes: The case of the missing heritability: Nature, v. 456, p. 18-21.

Malhotra, S., C. Morcillo-Suárez, D. Brassat, R. Goertsches, J. Lechner-Scott, E. Urcelay, O. Fernández, J. Drulovic, A. García-Merino, F. Martinelli Boneschi, A. Chan, K. Vandenbroeck, A. Navarro, M. F. Bustamante, J. Río, D. A. Akkad, G. Giacalone, A. J. Sánchez, L. Leyva, R. Alvarez-Lafuente, U. K. Zettl, J. Oksenberg, X. Montalban, and M. Comabella, 2011, IL28B polymorphisms are not associated with the response to interferon-β in multiple sclerosis: J Neuroimmunol, v. 239, p. 101-4.

Manolio, T. A., 2013, Bringing genome-wide association findings into clinical use: Nat Rev Genet, v. 14, p. 549-58.

Martínez, A., V. de las Heras, A. Mas Fontao, M. Bartolomé, E. G. de la Concha, E. Urcelay, and R. Arroyo, 2006, An IFNG polymorphism is associated with interferon-beta response in Spanish MS patients: J Neuroimmunol, v. 173, p. 196-9.

Matesanz, F., A. González-Pérez, M. Lucas, S. Sanna, J. Gayán, E. Urcelay, I. Zara, M. Pitzalis, M. L. Cavanillas, R. Arroyo, M. Zoledziewska, M. Marrosu, O. Fernández, L. Leyva, A. Alcina, M. Fedetz, C. Moreno-Rey, J. Velasco, L. M. Real, J. L. Ruiz-Peña, F. Cucca, A. Ruiz, and G. Izquierdo, 2012, Genome-wide association study of multiple sclerosis confirms a novel locus at 5p13.1: PLoS One, v. 7, p. e36140.

Mayo Clinic staff, 2011, How genetic disorders are inherited.

McDonald, W. I., A. Compston, G. Edan, D. Goodkin, H. P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. van den Noort, B. Y. Weinshenker, and J. S. Wolinsky, 2001, Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis: Ann Neurol, v. 50, p. 121-7.

Meuwissen, T. H., 2009, Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping: Genet Sel Evol, v. 41, p. 35.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001, Prediction of total genetic value using genome-wide dense marker maps: Genetics, v. 157, p. 1819-29.

Montgomery, D. B., and D. G. Morrison, 1973, A note on adjusting R-squared: The Journal of Finance, v. 28, p. 1009-1013.

MS Society of Western Australia, 2013, Types of MS.

National Heart Lung and Blood Institute, 2013, Risk assessment tool for estimating your 10-year risk of having a heart attack.

Nischwitz, S., S. Cepok, A. Kroner, C. Wolf, M. Knop, F. Müller-Sarnowski, H. Pfister, D. Roeske, P. Rieckmann, B. Hemmer, M. Ising, M. Uhr, T. Bettecken, F. Holsboer, B. Müller-Myhsok, and F. Weber, 2010, Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis: J Neuroimmunol, v. 227, p. 162-6.

Novartis, 2007, Tegretol: prescribing information.

O'Brien, M., R. Lonergan, L. Costelloe, K. O'Rourke, J. M. Fletcher, K. Kinsella, C. Sweeney, G. Antonelli, K. H. Mills, C. O'Farrelly, M. Hutchinson, and N. Tubridy, 2010, OAS1: a multiple sclerosis susceptibility gene that influences disease severity: Neurology, v. 75, p. 411-8.

Olivier, C., B. Williams-Jones, B. Godard, B. Mikalson, and V. Ozdemir, 2008, Personalized medicine, bioethics and social responsibilities: re-thinking the pharmaceutical industry to remedy inequities in patient care and international health: Current Pharmacogenomics and Personalized Medicine, v. 6, p. 108-120.

Outteryck, O., J. C. Ongagna, B. Brochet, L. Rumbach, C. Lebrun-Frenay, M. Debouverie, H. Zéphir, J. C. Ouallet, E. Berger, M. Cohen, S. Pittion, D. Laplaud, S. Wiertlewski, P. Cabre, J. Pelletier, A. Rico, G. Defer, N. Derache, W. Camu, E. Thouvenot, T. Moreau, A. Fromont, A. Tourbah, P. Labauge, G. Castelnovo, P. Clavelou, O. Casez, P. Hautecoeur, C. Papeix, C. Lubetzki, B. Fontaine, N. Couturier, N. Bohossian, M. Clanet, P. Vermersch, J. de Sèze, D. Brassat, and a. C. BIONAT network, 2013, A prospective observational post-marketing study of natalizumab-treated multiple sclerosis patients: clinical, radiological and biological features and adverse events. The BIONAT cohort: Eur J Neurol.

Pappas, D. J., and J. R. Oksenberg, 2010, Multiple sclerosis pharmacogenomics: maximizing efficacy of therapy: Neurology, v. 74 Suppl 1, p. S62-9.

Pennisi, E., 2003, Bioinformatics. Gene counters struggle to get the right answer: Science, v. 301, p. 1040-1.

Pepe, M. S., H. Janes, and J. W. Gu, 2007, Letter by Pepe et al regarding article, "Use and misuse of the receiver operating characteristic curve in risk prediction": Circulation, v. 116, p. e132; author reply e134.

Pepe, M. S., H. Janes, G. Longton, W. Leisenring, and P. Newcomb, 2004, Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker: Am J Epidemiol, v. 159, p. 882-90.

PersonalGenomes.org, 2013, The Personal Genome Project.

Polman, C. H., S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, F. D. Lublin, X. Montalban, P. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, E. Waubant, B. Weinshenker, and J. S. Wolinsky, 2011, Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria: Ann Neurol, v. 69, p. 292-302.

Polman, C. H., S. C. Reingold, G. Edan, M. Filippi, H. P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, B. G. Weinshenker, and J. S. Wolinsky, 2005, Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria": Ann Neurol, v. 58, p. 840-6.

Pritchard, J. K., and M. Przeworski, 2001, Linkage disequilibrium in humans: models and data: Am J Hum Genet, v. 69, p. 1-14.

Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer, 2010, LocusZoom: regional visualization of genome-wide association scan results: Bioinformatics, v. 26, p. 2336-7.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, 2007, PLINK: a tool set for whole-genome association and population-based linkage analyses: Am J Hum Genet, v. 81, p. 559-75.

Purcell, S. M., N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, P. Sklar, and I. S. Consortium, 2009, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder: Nature, v. 460, p. 748-52.

R Core Team, 2013, R: A language and environment for statistical computing.

Ramirez, A. H., Y. Shi, J. S. Schildcrout, J. T. Delaney, H. Xu, M. T. Oetjens, R. L. Zuvich, M. A. Basford, E. Bowton, M. Jiang, P. Speltz, R. Zink, J. Cowan, J. M. Pulley, M. D. Ritchie, D. R. Masys, D. M. Roden, D. C. Crawford, and J. C. Denny, 2012, Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record: Pharmacogenomics, v. 13, p. 407-18.

Refaeilzadeh, P., L. Tang, and H. Liu, 2009, Cross-validation, *in* L. Liu, and M. Tamer Özsu, eds., Encyclopedia of Database Systems, p. 532-538.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander, 2001, Linkage disequilibrium in the human genome: Nature, v. 411, p. 199-204.

Ritchie, M. D., 2012, The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era: Hum Genet, v. 131, p. 1615-26.

Rosati, G., 2001, The prevalence of multiple sclerosis in the world: an update: Neurol Sci, v. 22, p. 117-39.

Río, J., M. Comabella, and X. Montalban, 2009, Predicting responders to therapies for multiple sclerosis: Nat Rev Neurol, v. 5, p. 553-60.

Río, J., M. Comabella, and X. Montalban, 2011, Multiple sclerosis: current treatment algorithms: Curr Opin Neurol, v. 24, p. 230-7.

Saad, M., S. Lesage, A. Saint-Pierre, J. C. Corvol, D. Zelenika, J. C. Lambert, M. Vidailhet, G. D. Mellick, E. Lohmann, F. Durif, P. Pollak, P. Damier, F. Tison, P. A. Silburn, C. Tzourio, S. Forlani, M. A. Loriot, M. Giroud, C. Helmer, F. Portet, P. Amouyel, M. Lathrop, A. Elbaz, A. Durr, M. Martinez, A. Brice, and F. P. s. D. G. S. Group, 2011a, Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population: Hum Mol Genet, v. 20, p. 615-27.

Saad, M., A. S. Pierre, N. Bohossian, M. Macé, and M. Martinez, 2011b, Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data: BMC Proc, v. 5 Suppl 9, p. S33.

Sackett, D. L., and W. M. Rosenberg, 1995, The need for evidence-based medicine: J R Soc Med, v. 88, p. 620-4.

Sackett, D. L., W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson, 1996, Evidence based medicine: what it is and what it isn't: BMJ, v. 312, p. 71-2.

Sadovnick, A. D., A. Dircks, and G. C. Ebers, 1999, Genetic counselling in multiple sclerosis: risks to sibs and children of affected individuals: Clin Genet, v. 56, p. 118-22.

Sanna, S., M. Pitzalis, M. Zoledziewska, I. Zara, C. Sidore, R. Murru, M. B. Whalen, F. Busonero, A. Maschio, G. Costa, M. C. Melis, F. Deidda, F. Poddie, L. Morelli, G. Farina, Y. Li, M. Dei, S. Lai, A. Mulas, G. Cuccuru, E. Porcu, L. Liang, P. Zavattari, L. Moi, E. Deriu, M. F. Urru, M. Bajorek, M. A. Satta, E. Cocco, P. Ferrigno, S. Sotgiu, M. Pugliatti, S. Traccis, A. Angius, M. Melis, G. Rosati, G. R. Abecasis, M. Uda, M. G. Marrosu, D. Schlessinger, and F. Cucca, 2010, Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis: Nat Genet, v. 42, p. 495-7.

Sargento-Freitas, J., S. Batista, C. Macario, F. Matias, and L. Sousa, 2013, Clinical predictors of an optimal response to natalizumab in multiple sclerosis: J Clin Neurosci, v. 20, p. 659-62.

Sasieni, P. D., 1997, From genotypes to genes: doubling the sample size: Biometrics, v. 53, p. 1253-61.

Sawcer, S., M. Ban, J. Wason, and F. Dudbridge, 2010, What role for genetics in the prediction of multiple sclerosis?: Ann Neurol, v. 67, p. 3-10.

Sawcer, S., G. Hellenthal, M. Pirinen, C. C. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, S. Edkins, E. Gray, D. R. Booth, S. C. Potter, A. Goris, G. Band, A. B. Oturai, A. Strange, J. Saarela, C. Bellenguez, B. Fontaine, M. Gillman, B. Hemmer, R. Gwilliam, F. Zipp, A. Jayakumar, R. Martin, S. Leslie, S. Hawkins, E. Giannoulatou, S. D'alfonso, H. Blackburn, F. Martinelli Boneschi, J. Liddle, H. F. Harbo, M. L. Perez, A. Spurkland, M. J. Waller, M. P. Mycko, M. Ricketts, M. Comabella, N. Hammond, I. Kockum, O. T. McCann, M. Ban, P. Whittaker, A. Kemppinen, P. Weston, C. Hawkins, S. Widaa, J. Zajicek, S. Dronov, N. Robertson, S. J. Bumpstead, L. F. Barcellos, R. Ravindrarajah, R. Abraham, L. Alfredsson, K. Ardlie, C. Aubin, A. Baker, K. Baker, S. E. Baranzini, L. Bergamaschi, R. Bergamaschi, A. Bernstein, A. Berthele, M. Boggild, J. P. Bradfield, D. Brassat, S. A. Broadley, D. Buck, H. Butzkueven, R. Capra, W. M. Carroll, P. Cavalla, E. G. Celius, S. Cepok, R. Chiavacci, F. Clerget-Darpoux, K. Clysters, G. Comi, M. Cossburn, I. Cournu-Rebeix, M. B. Cox, W. Cozen, B. A. Cree, A. H. Cross, D. Cusi, M. J. Daly, E. Davis, P. I. de Bakker, M. Debouverie, M. B. D'hooghe, K. Dixon, R. Dobosi, B. Dubois, D. Ellinghaus, I. Elovaara, F. Esposito, et al., 2011, Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis: Nature, v. 476, p. 214-9.

Shapiro, J. A., 2009, Revisiting the central dogma in the 21st century: Ann N Y Acad Sci, v. 1178, p. 6-28.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, 2001, dbSNP: the NCBI database of genetic variation: Nucleic Acids Res, v. 29, p. 308-11.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer, 2009, ROCR: Visualizing the performance of scoring classifiers.

Slatkin, M., 2008, Linkage disequilibrium--understanding the evolutionary past and mapping the medical future: Nat Rev Genet, v. 9, p. 477-85.

Sormani, M. P., and N. De Stefano, 2013, Defining and scoring response to IFN-β in multiple sclerosis: Nat Rev Neurol, v. 9, p. 504-12.

Sriram, U., L. F. Barcellos, P. Villoslada, J. Rio, S. E. Baranzini, S. Caillier, A. Stillman, S. L. Hauser, X. Montalban, and J. R. Oksenberg, 2003, Pharmacogenomic analysis of interferon receptor polymorphisms in multiple sclerosis: Genes Immun, v. 4, p. 147-52.

Steinman, L., 2005, Blocking adhesion molecules as therapy for multiple sclerosis: natalizumab: Nat Rev Drug Discov, v. 4, p. 510-8.

Steyerberg, E. W., 2009, Clinical prediction models: a practical approach to development, validation, and updating: Statistics for Biology and Health, Springer, 500 p.

Steyerberg, E. W., A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, 2010, Assessing the performance of prediction models: a framework for traditional and novel measures: Epidemiology, v. 21, p. 128-38.

Stone, M., 1974, Cross-validatory choice and assessment of statistical predictions: Journal of the Royal Statistical Society. Series B (Methodological), v. 36, p. 111-147.

The University of Edinburgh, 2010, Cardiovascular Risk Calculator.

The Warfarin Dose Refinement Collaboration, and International Warfarin Pharmacogenetics Consortium, 2013, Warfarin Dosing.

Therneau, T. M., and B. Atkinson, 2012, rpart: Recursive Partitioning.

Todd, H. L., and B. Craine, 2011, AMA to FDA: genetic testing should be conducted by qualified health professionals.

Vazquez, A. I., G. de los Campos, Y. C. Klimentidis, G. J. Rosa, D. Gianola, N. Yi, and D. B. Allison, 2012, A comprehensive genetic approach for improving prediction of skin cancer risk in humans: Genetics, v. 192, p. 1493-502.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V.

A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, et al., 2001, The sequence of the human genome: Science, v. 291, p. 1304-51.

Vickers, A. J., and E. B. Elkin, 2006, Decision curve analysis: a novel method for evaluating prediction models: Med Decis Making, v. 26, p. 565-74.

Villoslada, P., L. F. Barcellos, J. Rio, A. B. Begovich, M. Tintore, J. Sastre-Garriga, S. E. Baranzini, P. Casquero, S. L. Hauser, X. Montalban, and J. R. Oksenberg, 2002, The HLA locus and multiple sclerosis in Spain. Role in disease susceptibility, clinical course and response to interferon-beta: J Neuroimmunol, v. 130, p. 194-201.

Visscher, P. M., W. G. Hill, and N. R. Wray, 2008, Heritability in the genomics era--concepts and misconceptions: Nat Rev Genet, v. 9, p. 255-66.

Vukusic, S., V. Van Bockstael, S. Gosselin, and C. Confavreux, 2007, Regional variations in the prevalence of multiple sclerosis in French farmers: J Neurol Neurosurg Psychiatry, v. 78, p. 707-9.

Wang, J., W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, J. Zhang, J. Li, Y. Guo, B. Feng, H. Li, Y. Lu, X. Fang, H. Liang, Z. Du, D. Li, Y. Zhao, Y. Hu, Z. Yang, H. Zheng, I. Hellmann, M. Inouye, J. Pool, X. Yi, J. Zhao, J. Duan, Y. Zhou, J. Qin, L. Ma, G. Li, G. Zhang, B. Yang, C. Yu, F. Liang, W. Li, S. Li, P. Ni, J. Ruan, Q. Li, H. Zhu, D. Liu, Z. Lu, N. Li, G. Guo, J. Ye, L. Fang, Q. Hao, Q. Chen, Y. Liang, Y. Su, A. San, C. Ping, S. Yang, F. Chen, L. Li, K. Zhou, Y. Ren, L. Yang, Y. Gao, G. Yang, Z. Li, X. Feng, K. Kristiansen, G. K. Wong, R. Nielsen, R. Durbin, L. Bolund, X. Zhang, and H. Yang, 2008, The diploid genome sequence of an Asian individual: Nature, v. 456, p. 60-5.

Wang, J. H., D. Pappas, P. L. De Jager, D. Pelletier, P. I. de Bakker, L. Kappos, C. H. Polman, L. B. Chibnik, D. A. Hafler, P. M. Matthews, S. L. Hauser, S. E. Baranzini, J. R. Oksenberg, and A. a. N. Z. M. S. G. C. (ANZgene), 2011, Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data: Genome Med, v. 3, p. 3.

Wang, W. Y., B. J. Barratt, D. G. Clayton, and J. A. Todd, 2005, Genome-wide association studies: theoretical and practical concerns: Nat Rev Genet, v. 6, p. 109-18.

Ware, J. H., 2006, The limitations of risk factors as prognostic tools: N Engl J Med, v. 355, p. 2615-7.

Watson, C. T., G. Disanto, F. Breden, G. Giovannoni, and S. V. Ramagopalan, 2012, Estimating the proportion of variation in susceptibility to multiple sclerosis captured by common SNPs: Sci Rep, v. 2, p. 770.

Watson, J. D., and F. H. Crick, 1953, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid: Nature, v. 171, p. 737-8.

Weinberg, W., 1908, Über den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl. Naturkd.: Württemb., v. 64, p. 369–382.

Weinstock-Guttman, B., M. Tamaño-Blanco, K. Bhasi, R. Zivadinov, and M. Ramanathan, 2007, Pharmacogenetics of MXA SNPs in interferon-beta treated multiple sclerosis patients: J Neuroimmunol, v. 182, p. 236-9.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg, 2008, The complete genome of an individual by massively parallel DNA sequencing: Nature, v. 452, p. 872-6.

Wolpe, P. R., 2009, Personalized Medicine and its Ethical Challenges: World Medical & Health Policy, v. 1, p. 47-55.

Woolf, B., 1955, On estimating the relation between blood group and disease: Ann Hum Genet, v. 19, p. 251-3.

World Health Organization, 2008, Atlas: Multiple Sclerosis Resources in the World.

Wray, N. R., J. Yang, M. E. Goddard, and P. M. Visscher, 2010, The genetic interpretation of area under the ROC curve in genomic profiling: PLoS Genet, v. 6, p. e1000864.

Wray, N. R., J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher, 2013, Pitfalls of predicting complex traits from SNPs: Nat Rev Genet, v. 14, p. 507-15.

Ziegler, A., N. Bohossian, V. P. Diego, and C. Yao, 2013, Genetic prediction in the Genetic Analysis Workshop 18 sequencing data: Genetic Epidemiology. Submitted.

Zou, K. H., A. J. O'Malley, and L. Mauri, 2007, Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models: Circulation, v. 115, p. 654-7.

I. APPENDIX I: Gene Names for Genes Cited Throughout the Thesis

GENE SYMBOL	GENE NAME
ABC	ATP-binding cassette
ADAR	adenosine deaminase, RNA-specific
BST1	bone marrow stromal cell antigen 1
CAST	calpastatin
CD86	T-lymphocyte activation antigen CD86
CFTR	cystic fibrosis transmembrane conductance regulator
CIT	citron (rho-interacting, serine/threonine kinase 21)
COL25A1	collagen, type XXV, alpha 1
CTSS	cathepsin S
CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9
FAS	Fas (TNF receptor superfamily member 6)
GPC5	glypican 5
GRIA3	glutamate receptor, ionotropic, AMPA 3
HAPLN1	hyaluronan and proteoglycan link protein 1
HLA Class II	major histocompatibility complex, class II
HTT	huntingtin
IFNAR1	interferon (alpha and beta) receptor 1
IFNAR1	interferon (alpha and beta) receptor 1
IFNAR2	interferon (alpha, beta and omega) receptor 2
IFNG	interferon gamma
IL12RB2	interleukin 12 receptor, beta 2
IL1R1	interleukin 1 receptor, type I
IL28B	interleukin 28B (interferon, lambda 3)
IL2RA	interleukin 2 receptor, alpha
IL7RA	interleukin 7 receptor
ITGA4	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)
LMP7	large multifunctional protease 7
LRRK2	leucine-rich repeat kinase 2
MAP4	microtubule-associated protein 4
MAPT	microtubule-associated protein tau
MBP	myelin basic protein
MxA	myxovirus (influenza) resistance A
NPAS3	neuronal PAS domain protein 3
OAS1	2'-5'-oligoadenylate synthetase 1
SNCA	synuclein, alpha (non A4 component of amyloid precursor)
STARD13	StAR-related lipid transfer (START) domain containing 13
TRAIL	tumor necrosis factor (TNF) related apoptosis inducing ligand
TRAILR-1	TRAIL receptor 1
TRAILR-2	TRAIL receptor 2
TRAILR-3	TRAIL receptor 3
TRAILR-4	TRAIL receptor 4
VKORC1	vitamin K epoxide reductase complex, subunit 1
ZFAT	zinc finger and AT hook domain containing
ZFHX4	zinc finger homeobox 4
ZZEF1	zinc finger, ZZ-type with EF-hand domain 1

Table I.1: Gene symbols and corresponding names. (*Source:* http://www.ncbi.nlm.nih.gov/gene/)

II. APPENDIX II: Clinical Characteristics of the Cohorts Included in the *OAS1* Study

France

			France		P-value[1]
			Responders	**Non-Responders**	
			164 (49%)	168 (51%)	
Gender					***0.26***
	NA		-	-	
	Female	**n (%)**	127 (51%)	120 (49%)	
	Male	**n (%)**	37 (44%)	48 (56%)	
Age (years)					
	At disease onset				***0.009***
		NA	1	1	
		mean (sd)	30.24 (8.65)	27.87 (8.59)	
		median (range)	30.00 (13.00, 52.00)	27.00 (12.00, 51.00)	
	At treatment onset				***0.01046***
		NA	8	8	
		mean (sd)	35.51 (9.31)	32.72 (9.16)	
		median (range)	36.00 (15.00, 58.00)	33.00 (15.00, 55.00)	
Disease Severity					
	EDSS at treatment onset				***0.27***
		NA	-	-	
		mean (sd)	2.07 (1.24)	2.17 (1.33)	
		median (range)	2.00 (0.00, 6.00)	2.00 (0.00, 5.00)	
	Relapses 2-year prior treatment onset				***0.05885***
		NA	1	2	
		mean (sd)	2.29 (1.01)	2.52 (1.20)	
		median (range)	2.00 (0.00, 5.00)	2.00 (7.00)	
Type of Interferon- β					***0.71***
	NA		1	-	
	Avonex	**n (%)**	84 (50%)	85 (50%)	
	Betaferon	**n (%)**	45 (52%)	42 (48%)	
	Rebif	**n (%)**	34 (45%)	41 (55%)	

[1] Chi-square test used for categorical variables (Gender, Type of Interferon-β) and Mann-Whitney test used for the remaining variables.

Table II.1: Clinical characteristics of the French cohort in the *OAS1* study.

Germany

			Germany		P-value[1]
			Responders	Non-Responders	
			123 (60%)	83 (40%)	
Gender					***0.0614***
	NA		-	-	
	Female	**n (%)**	87 (56%)	69 (44%)	
	Male	**n (%)**	36 (72%)	14 (28%)	
Age (years)					
	At disease onset				***0.19***
		NA	93	73	
		mean (sd)	28.6 (8.74)	24.80 (6.46)	
		median (range)	28.00 (13.00, 51.00)	23.50 (16.00, 37.00)	
	At treatment onset				***0.57***
		NA	-	-	
		mean (sd)	36.15 (9.09)	35.27 (10.58)	
		median (range)	36.00 (19.00, 57.00)	36.00 (16.00, 63.00)	
Disease Severity					
	EDSS at treatment onset				***0.001***
		NA	-	-	
		mean (sd)	1.71 (1.21)	2.16 (1.22)	
		median (range)	1.50 (0.00, 6.00)	2.00 (0.00, 6.00)	
	Relapses 2-year prior treatment onset				***0.62***
		NA	94	74	
		mean (sd)	1.64 (0.94)	1.78 (0.83)	
		median (range)	1.50 (0.00, 5.00)	2.00 (1.00, 3.00)	
Type of Interferon- β					***0.74***
	NA		1	-	
	Avonex	**n (%)**	20 (57%)	15 (43%)	
	Betaferon	**n (%)**	73 (58%)	52 (42%)	
	Rebif	**n (%)**	29 (64%)	16 (36%)	

[1] **Chi-square test used for categorical variables (Gender, Type of Interferon-β) and Mann-Whitney test used for the remaining variables.**

Table II.2: Clinical characteristics of the German cohort in the *OAS1* study.

Italy

			Italy		P-value[1]
			Responders	**Non-Responders**	
			252 (79%)	68 (21%)	
Gender					***0.06077***
	NA				
	Female	**n (%)**	168 (86%)	54 (14%)	
	Male	**n (%)**	84 (76%)	14 (24%)	
Age (years)					
	At disease onset				***8.53E-04***
		NA	-	-	
		mean (sd)	29.78 (9.09)	25.63 (7.59)	
		median (range)	28.70 (13.80, 60.00)	23.80 (10.90, 45.30)	
	At treatment onset				***0.001972***
		NA	-	-	
		mean (sd)	34.53 (9.31)	30.58 (8.51)	
		median (range)	34.05 (15.60, 63.80)	29.25 (17.10, 59.30)	
Disease Severity					
	EDSS at treatment onset				***0.96***
		NA	-	-	
		mean (sd)	1.66 (0.66)	1.70 (0.78)	
		median (range)	1.50 (0.00, 4.50)	1.50 (0.00, 4.00)	
	Relapses 2-year prior treatment onset				***0.01166***
		NA	-	-	
		mean (sd)	1.80 (0.95)	2.24 (1.24)	
		median (range)	2.00 (0.00, 5.00)	2.00 (0.00, 7.00)	
Type of Interferon- β					***0.67***
	NA		-	-	
	Avonex	**n (%)**	81 (82%)	18 (18%)	
	Betaferon	**n (%)**	13 (76%)	4 (24%)	
	Rebif	**n (%)**	158 (77%)	46 (23%)	

[1] Chi-square test used for categorical variables (Gender, Type of Interferon-β) and Mann-Whitney test used for the remaining variables.

Table II.3: Clinical characteristics of the Italian cohort in the *OAS1* study.

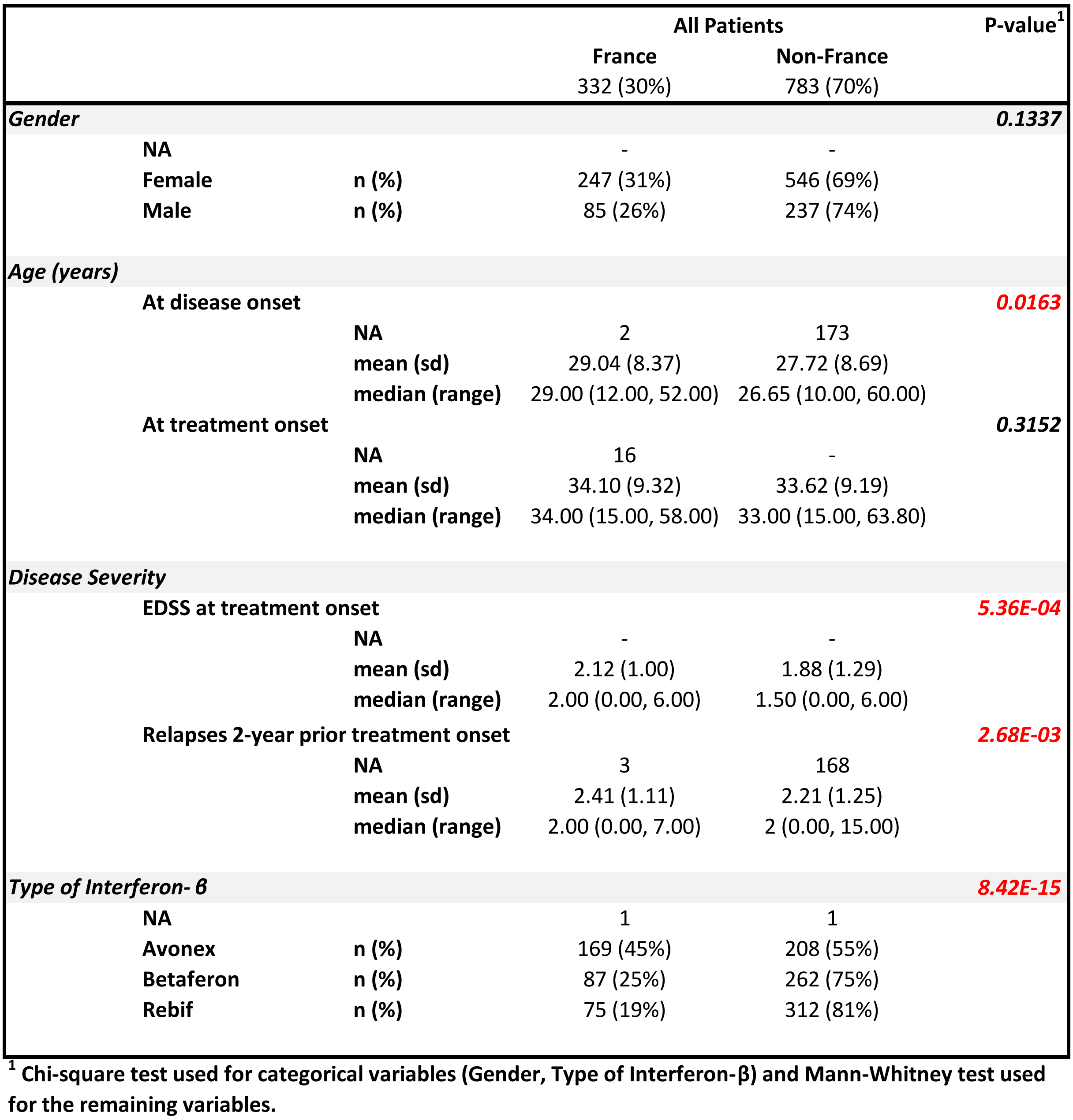
Spain

			Spain		P-value[1]
			Responders	**Non-Responders**	
			144 (56%)	113 (44%)	
Gender					***0.92***
	NA				
	Female	**n (%)**	95 (56%)	73 (44%)	
	Male	**n (%)**	49 (55%)	40 (45%)	
Age (years)					
	At disease onset				***0.82***
		NA	7	-	
		mean (sd)	26.26 (7.51)	26.16 (7.14)	
		median (range)	26.00 (10.00, 47.00)	24.00 (13.00, 45.00)	
	At treatment onset				***0.62***
		NA	-	-	
		mean (sd)	32.01 (8.10)	31.50 (8.47)	
		median (range)	31.00 (15.00, 55.00)	31.00 (16.00, 51.00)	
Disease Severity					
	EDSS at treatment onset				***6.36E-04***
		NA	-	-	
		mean (sd)	1.97 (1.07)	2.34 (1.02)	
		median (range)	2.00 (0.00, 5.50)	2.00 (0.00, 5.50)	
	Relapses 2-year prior treatment onset				***0.17***
		NA	-	-	
		mean (sd)	2.54 (1.04)	2.88 (1.71)	
		median (range)	2.00 (0.00, 5.00)	3.00 (0.00, 15.00)	
Type of Interferon- β					***0.74***
	NA		-	-	
	Avonex	**n (%)**	39 (53%)	35 (47%)	
	Betaferon	**n (%)**	70 (58%)	50 (42%)	
	Rebif	**n (%)**	35 (56%)	28 (44%)	

[1] Chi-square test used for categorical variables (Gender, Type of Interferon-β) and Mann-Whitney test used for the remaining variables.

Table II.4: Clinical characteristics of the Spanish cohort in the *OAS1* study.

All Patients (France/Non-France)



			All Patients		P-value[1]
			France	Non-France	
			332 (30%)	783 (70%)	
Gender					***0.1337***
	NA		-	-	
	Female	n (%)	247 (31%)	546 (69%)	
	Male	n (%)	85 (26%)	237 (74%)	
Age (years)					
	At disease onset				***0.0163***
		NA	2	173	
		mean (sd)	29.04 (8.37)	27.72 (8.69)	
		median (range)	29.00 (12.00, 52.00)	26.65 (10.00, 60.00)	
	At treatment onset				***0.3152***
		NA	16	-	
		mean (sd)	34.10 (9.32)	33.62 (9.19)	
		median (range)	34.00 (15.00, 58.00)	33.00 (15.00, 63.80)	
Disease Severity					
	EDSS at treatment onset				***5.36E-04***
		NA	-	-	
		mean (sd)	2.12 (1.00)	1.88 (1.29)	
		median (range)	2.00 (0.00, 6.00)	1.50 (0.00, 6.00)	
	Relapses 2-year prior treatment onset				***2.68E-03***
		NA	3	168	
		mean (sd)	2.41 (1.11)	2.21 (1.25)	
		median (range)	2.00 (0.00, 7.00)	2 (0.00, 15.00)	
Type of Interferon- β					***8.42E-15***
	NA		1	1	
	Avonex	n (%)	169 (45%)	208 (55%)	
	Betaferon	n (%)	87 (25%)	262 (75%)	
	Rebif	n (%)	75 (19%)	312 (81%)	

[1] **Chi-square test used for categorical variables (Gender, Type of Interferon-β) and Mann-Whitney test used for the remaining variables.**

Table II.5: Clinical characteristics of the France/Non-France cohorts in the *OAS1* study.

III. APPENDIX III: Baseline Characteristics of the Patients Included in the Prediction Study on Natalizumab Response (Section 4.3.1)

Clinical Characteristics

CLINICAL CHARACTERISTICS			All Patients	Responders	Non-Responders	P-value[1]
			531	331 (62%)	200 (38%)	
Gender						*0.73*
	NA		**-**	-	-	
	female	**n (%)**	**396**	249 (63%)	147 (37%)	
	male	**n (%)**	**135**	82 (61%)	53 (39%)	
EDSS at treatment onset						*0.78*
	NA		**-**	-	-	
	mean (sd)		**3.36 (1.66)**	3.37 (1.64)	3.33 (1.68)	
	median (range)		**3.50 (0.00, 7.50)**	3.00 (0.00, 7.50)	3.50 (0.00, 7.00)	
Relapses 1-year prior treatment onset						*0.15*
	NA		**-**	-	-	
	mean (sd)		**2.10 (1.11)**	2.04 (1.06)	2.19 (1.17)	
	median (range)		**2.00 (0.00, 8.00)**	2.00 (0.00, 6.00)	2.00 (0.00, 8.00)	
Disease duration (years)						*0.27*
	NA		**-**	-	-	
	mean (sd)		**8.94 (7.00)**	9.08 (6.84)	8.71 (7.26)	
	median (range)		**8.00 (0.00, 42.00)**	8.00 (0.00, 41.00)	7.00 (0.00, 42.00)	
Previous immuno-suppressant use						*0.95*
	NA		**-**	-	-	
	yes	**n (%)**	**111**	70 (63%)	41 (37%)	
	no	**n (%)**	**420**	261 (62%)	159 (38%)	
Previous immuno-modulatory use						*0.86*
	NA		**-**	-	-	
	yes	**n (%)**	**475**	295 (62%)	180 (38%)	
	no	**n (%)**	**56**	36 (64%)	20 (36%)	

[1] Chi-square test used for categorical variables, Mann-Whitney test used for numeric variables. NA: Not available.

Table III.1: Clinical characteristics of a subset of patients from the BIONAT cohort (database version November 2013) included in the prediction study on natalizumab response (Section 4.3.1.2).

Biological Characteristics

BIOLOGICAL CHARACTERISTICS	All Patients	Responders	Non-Responders	P-value[1]
	531	331 (62%)	200 (38%)	
IgG at treatment onset				*0.039*
NA	**64**	39	25	
mean (sd)	**9.94 (2.54)**	10.07 (2.51)	9.72 (2.58)	
median (range)	**9.80 (0.90, 19.70)**	10.00 (0.90, 17.20)	9.35 (3.00, 19.70)	
IgA at treatment onset				*0.10*
NA	**64**	39	25	
mean (sd)	**2.00 (0.80)**	2.04 (0.81)	1.94 (0.77)	
median (range)	**1.92 (0.20, 5.52)**	1.98 (0.20, 5.52)	1.84 (0.31, 5.04)	
IgM at treatment onset				*0.53*
NA	**66**	40	26	
mean (sd)	**1.38 (1.10)**	1.39 (1.16)	1.37 (0.98)	
median (range)	**1.19 (0.21, 18.00)**	1.21 (0.21, 18.00)	1.17 (0.24, 10.94)	
Lymphocytes at treatment onset				*0.28*
NA	**59**	41	18	
mean (sd)	**2.38 (3.15)**	2.42 (3.50)	2.30 (2.52)	
median (range)	**1.80 (0.19, 33.80)**	1.80 (0.59, 33.80)	1.85 (0.19, 27.20)	
CD3 count at treatment onset				*0.12*
NA	**255**	149	106	
mean (sd)	**1402 (530.91)**	1364 (529.78)	1475 (528.29)	
median (range)	**1326 (372, 4970)**	1298 (372, 4970)	1373 (594, 2919)	
CD4 count at treatment onset				*0.09*
NA	**51**	35	16	
mean (sd)	**965 (418.06)**	947.40 (436.08)	993.20 (386.75)	
median (range)	**897 (237, 4596)**	862 (237, 4596)	938 (252, 2250)	
CD8 count at treatment onset				*0.04291*
NA	**53**	36	17	
mean (sd)	**441.30 (211.97)**	432.60 (223.56)	455.40 (191.55)	
median (range)	**395 (43, 1850)**	379 (106, 1850)	408 (43, 1162)	
CD19 count at treatment onset				*0.73*
NA	**124**	83	41	
mean (sd)	**298.20 (264.00)**	284.60 (148.13)	319.60 (379.93)	
median (range)	**249.50 (33.00, 4442.00)**	252 (44, 990)	240 (33, 4442)	
JC virus at treatment onset				*0.27*
NA	**70**	35	35	
detected n (%)	**268**	166 (62%)	102 (38%)	
not detected n (%)	**193**	130 (67%)	63 (33%)	

[1] Chi-square test used for categorical variables, Mann-Whitney test used for numeric variables. NA: Not available.

Table III.2: Biological characteristics of a subset of patients from the BIONAT cohort (database version November 2013) included in the prediction study on natalizumab response (Section 4.3.1.2).

Radiological Characteristics

RADIOLOGICAL CHARACTERISTICS			All Patients	Responders	Non-Responders	P-value[1]
			531	331 (62%)	200 (38%)	
GD+ enhancing lesions at treatment onset						*0.02323*
	NA		**-**	-	-	
	yes	**n (%)**	**308**	205 (67%)	103 (33%)	
	no	**n (%)**	**223**	126 (56%)	97 (44%)	
T2 superior to 9 at treatment onset						*0.54*
	NA		**3**	3	-	
	yes	**n (%)**	**482**	297 (62%)	185 (38%)	
	no	**n (%)**	**46**	31 (67%)	15 (33%)	
T2 confluent lesions at treatment onset						*0.29*
	NA		**-**	-	-	
	yes	**n (%)**	**36**	19 (53%)	17 (47%)	
	no	**n (%)**	**495**	312 (63%)	183 (37%)	
Black holes at treatment onset						*0.91*
	NA		**181**	109	72	
	yes	**n (%)**	**112**	71 (63%)	41 (37%)	
	no	**n (%)**	**238**	151 (63%)	87 (37%)	

[1] Chi-square test used for categorical variables. NA: Not available.

Table III.3: Radiological characteristics of a subset of patients from the BIONAT cohort (database version November 2013) included in the prediction study on natalizumab response (Section 4.3.1.2).

IV. APPENDIX IV: Scientific Production

Selected Oral Presentations

1. Bohossian N, Brassat D. "**A multi-center study evaluating long-term natalizumab treatment: the France study.**" Biosignature Workshop, Muenster, Germany, October 2011.
2. Bohossian N, Saad M, Legarra A, Martinez M. "**Exploring models for simultaneous analysis of all SNPs in genome-wide association data of human complex phenotypes.**" European Mathematical Genetics Meeting, Gottingen, Germany, April 2012.

Selected Poster Presentations

1. Bohossian N, Martinez M, Brassat D. "**Predicting response to natalizumab in multiple sclerosis patients one year after treatment onset**." Final UEPHA*MS Network Conference, Bilbao, Spain, April 2012.
2. Bohossian N, Saad M, Legarra A, Martinez M. "**A multi-marker genome-wide association study: the story of Bayes Cπ**." The 21st Annual Meeting of the International Genetic Epidemiology Society, Stevenson, Washington, USA, October 2012.

Publications

1. Bohossian N*, Saad M*, Legarra A, Martinez M. **Single-marker and multi-marker mixed models for polygenic score analysis in family-based data.** *BMC Proc.* To appear.
2. Ziegler A, Bohossian N, Diego VP, Yao C. **Genetic prediction in the genetic analysis workshop 18 sequencing data.** *Genet Epidemiol.* To appear.
3. Outteryck O, Ongagna JC, Brochet B, Rumbach L, Lebrun-Frenay C, Debouverie M, Zéphir H, Ouallet JC, Berger E, Cohen M, Pittion S, Laplaud D, Wiertlewski S, Cabre P, Pelletier J, Rico A, Defer G, Derache N, Camu W, Thouvenot E, Moreau T, Fromont A, Tourbah A, Labauge P, Castelnovo G, Clavelou P, Casez O, Hautecoeur P, Papeix C, Lubetzki C, Fontaine B, Couturier N, Bohossian N, Clanet M, Vermersch P, de Sèze J, Brassat D, and BIONAT network, and CFSEP. **A prospective observational post-marketing study of natalizumab-treated multiple sclerosis patients: clinical, radiological and biological features and adverse events. The BIONAT cohort.** *Eur J Neurol.* 2013 doi: 10.1111/ene.12204.
4. Saad M, Pierre AS, Bohossian N, Macé M, Martinez M. **Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data.** *BMC Proc.* 2011 Nov 29;5 Suppl 9:S33.

[*] These authors contributed equally to this work.

Single-marker and multi-marker mixed models for polygenic score analysis in family-based data

Nora Bohossian[1,2*§], Mohamad Saad[1,2*], Andrés Legarra[3], Maria Martinez[1]

[1]Inserm UMR1043-CPTP, CHU Purpan, Toulouse, France
[2]University of Toulouse III – Paul Sabatier, Toulouse, France
[3]Inra UR631-SAGA, Castanet-Tolosan, France

*These authors contributed equally to this work
[§]Corresponding author

Email addresses:

NB: nora.bohossian@inserm.fr
MS: mohamad.saad@inserm.fr
AL: andres.legarra@toulouse.inra.fr
MM: maria.martinez@inserm.fr

Abstract

Genome-wide association studies have proven successful but they remain underpowered to detect variants of weaker effect. Alternative methods propose instead to test for association an aggregate score combining the effects of the most associated variants. The set of variants that are to be aggregated may come from two modeling approaches: single-marker or multi-marker. The goal of this paper is to evaluate this alternative strategy by using sets of SNPs identified by the two modeling approaches in the simulated pedigree dataset provided for the Genetic Analysis Workshop 18. Here, we focused on quantitative traits association analysis of diastolic blood pressure and of Q1 that served to control the statistical significance of our results. We carried out all analyses with knowledge of the underlying simulation model. We found that the probability to replicate association with the aggregate score depends on the SNP set size and, for smaller sets (≤ 100), also on the modeling approach. Nonetheless, assessing the statistical significance of these results in this dataset was challenging likely due to linkage since we are analyzing pedigree data and also because the genotypes were the same across the replicates. Further methods need to be developed to facilitate the application of this alternative strategy in pedigree data.

Background

Genome-Wide Association Studies (GWAS) have proven successful in identifying common SNPs associated with complex traits but the underlying genetic architecture of these traits has remained largely unknown. This classical approach is restricted to analyzing one SNP at a time and only those reaching genome-wide significance ($\alpha \leq$ 1E-08) are retained for further analyses. As such, for many complex traits, only few SNPs have been identified so far leaving a large part of the trait heritability unexplained. This is partially because a wide spectrum of effects may be implicated most of which are not detected at the stringent significance criteria levels set in GWAS. One way to circumvent this limitation has been through using larger sample sizes thus increasing the power to detect SNPs of weaker effect. Recently, following their successful application in few studies [1, 2], attention has been turned instead to the use of alternative methods that propose to aggregate the effects of several SNPs into a polygenic score (PS) and test PS for association with the trait. Typically, the PS is constructed in two steps. First, the set of SNPs to be included in the score is selected. The criteria for SNP selection vary between studies but it is crucial that this set contains only independent variants to avoid over-representing the same signal. Further, since we are working in the context of detection rather than prediction, this set must exclude all established variants as they would drive the association of the PS with the trait masking the weaker effects that we are precisely looking to detect. Second, the reference alleles of these variants are combined in an unweighted or weighted manner. The former approach assumes that all SNPs have the same effect size which oversimplifies the context we are trying to evaluate (a mixture of different effect sizes). In the latter approach, on the other hand, each reference allele is weighted by its effect estimated in an independent dataset. The effect estimates could be obtained through a classical single-marker analysis whereby each effect is estimated one at a time or, alternatively, through a multi-marker analysis whereby all effects are estimated simultaneously. Studies have suggested that the latter approach may outperform single-marker analysis in detection [3]. The goal of this study is, therefore, to compare the PS analysis using sets of SNPs derived from single-marker and multi-marker analyses and evaluate the value of this novel analytical approach to shed light on the true genetic architecture of a complex quantitative trait in family-based data.

Methods

Data and phenotypes

We used the pedigree dataset provided for the Genetic Analysis Workshop 18 (GAW18) with knowledge of the simulated model. We focused on the simulated quantitative trait diastolic blood pressure measured at exam 1 (DBP_1). We used the trait Q1 to control for Type I error. In the simulated model, there were 1457 SNPs (in 288 genes) contributing to DBP and/or systolic blood pressure (SBP) variability. Their individual contribution ranged from as low as 0.001% for gene *ZZEF1* to as high as 6.5% for gene *MAP4* (for DBP). Part of the total heritability was due to polygenic alleles from 1000 SNPs, randomly selected in each replicate. The trait Q1 was uninfluenced by any of the provided genotyped SNPs. There were 200 replicates but the genotypes were the same across the replicates.

We adjusted the traits for age and sex in a linear regression framework. Let Y_i denote the trait adjusted for age at exam 1 and sex for individual i ($i = 1,\ldots,N$ individuals). We used the full SNP map ($j = 1,\ldots,J; J = 8{,}348{,}674$ SNPs). We denote the observed $N \times J$ genotype matrix by X. All genotypes were coded under the additive genetic model. Since we worked with models that treated the SNP effects as fixed or random, to distinguish between the two, we use β to denote the $J \times 1$ vector of fixed SNP effects and α to denote the $J \times 1$ vector of random SNP effects. Finally, let u be the $N \times 1$ vector of random polygenic effects with $u \sim N(0, \sigma_u^2 K)$ where K is twice the $N \times N$ kinship coefficient relationship matrix based on pedigree information and let ε be the $N \times 1$ vector of random residual effects with $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$ where I is the $N \times N$ identity matrix.

Single-marker mixed linear model (analysis limited to the SNPs /genes associated with DBP)

We first estimated the power to detect association of Y with any of these associated variants using the Measured Genotype test [4] (mixed-linear regression model), as implemented in QTDT software (http://www.sph.umich.edu/csg/abecasis/QTDT/). Single-marker MG test was conducted for each SNP using all 200 replicates. We found that *MAP4* was the only gene detectable (Power = 96%) at the genome-wide significance level (α < 1E-08), which accounts for the largest percentage of the variance of DBP and contains SNPs with very strong individual effects. Any of the remaining SNPs or genes were unlikely to be detected at stringent significance criteria (Power < 50%). We estimated that it will take ~40 days to analyze the whole-genome data (> 8 million SNPs) using the MG test by replicate and by phenotype. Due to these computational constraints, we derived a new trait adjusted for family relatedness using GRAMMAR [5] as implemented in the GenABEL add-on package developed for the R statistical software [6]. As such, the single-marker linear model could be used. Lastly, since our goal was to evaluate whether power to detect association with SNPs with weak effects could be enhanced by pooling their effects, we further adjusted the de-correlated trait DBP_1 for the strong effects of *MAP4* (SNPs 3_48040283 and 3_48064367).

Single-Marker (SM) linear model

For the de-correlated trait we tested for association using simple linear model (without the random polygenic component) with PLINK version 1.07 [7].

Polygenic Score (PS)

Polygenic scores (PS) were built as follows:

$$PS_i = \sum_{s=1}^{S} \hat{\gamma}_s X_{is} \tag{1}$$

where PS_i is the polygenic score for the *i*th individual, S is the size of the set of SNPs to combine, $\hat{\gamma}_s$ is the estimated effect of SNP s in a discovery dataset and X_{is} is the number of minor alleles of the SNP s for the *i*th individual in an independent dataset (replication). The PS is computed after excluding genome-wide significant SNPs and including only independent SNPs (not in LD). The PS values were computed with PLINK using the "--score" option. By default, if missing, the number of reference alleles was imputed from the sample allele frequency.

Multi-marker Mixed Linear (MML) model

Here, all SNP effects are estimated simultaneously. For the original traits not adjusted for family relatedness, Y, the model is formulated as follows (using matrix notation):

$$Y = \mu 1 + X\beta + u + \varepsilon \tag{2}$$

where μ is the fixed mean effect and 1 is a vector of 1's. Since the above model analyzes all J markers jointly, it can, therefore, account for the covariance structures between individuals through the realized genome-based relationship matrix [8] and can be formulated equivalently as a random regression approach as follows:

$$Y = \mu 1 + X\alpha + \varepsilon \tag{3}$$

where α is now a $J \times 1$ vector of *random* SNP effects with $\alpha \sim N(0, \sigma_\alpha^2 I)$. This is the widely used Best Linear Unbiased Predictor (BLUP) model (with only one fixed effect) [9]. We worked in a penalized regression framework (l_2 penalty) setting the penalty parameter, λ, as $\lambda = \sigma_\varepsilon^2 / \sigma_\alpha^2$. We derived σ_α^2 from the relationship $V_A = \sum_{j=1}^{J} 2p_j q_j \sigma_\alpha^2$, where V_A is the total additive genetic variance, p_j is the minor allele frequency and $q_j = 1 - p_j$.[10] These analyses were carried out using the GS3 software (http://snp.toulouse.inra.fr/~alegarra/).

Polygenic score analyses

Here, we were interested in comparing PS analysis with sets of SNPs derived from SM and BLUP models. We used replicate 1 as the discovery dataset and each of the remaining 199 replicates to replicate the association of PS with the analyzed trait. Under SM, all SNPs were ranked by their P-values. Under BLUP, all SNPs were ranked by the magnitude of their effect estimate. The best ranked S SNPs were used for computing the PS values. Here, we let S vary as 10, 50, 100, 1000, 5000 and 10000 SNPs. To ensure that S contained only independent SNPs, the best SNP over a window of 100kb was retained until the full SNP map has been covered. (We also considered larger window sizes of 1Mb and 5Mb but the results were similar and are not reported here.) These independent SNPs were then ranked. We conducted the same analyses on Q1. Further, we also evaluated the association of PS but this time by permuting DBP_1 within families in each replicate.

Results and Discussion

Figure 1 illustrates the PS association results using the two strategies, SM and BLUP, for selecting the top S SNPs in replicate 1. The results are expressed as the percentage of replicates (out of replicates 2 through 200, from here onwards referred to as the replication rate) with significant evidence of association of PS with DBP_1 at different nominal P-values and by SNP set size S. For the SM strategy, we see that the replication rates tend to increase with the SNP set size until reaching a peak at $S = 1000$ SNPs after which they begin to decline especially at stringent significance criteria levels (P-value ≤ 1E-05). (This may be happening due to much more noise being added with larger S (>1000).) For the BLUP strategy, the peak is reached at $S = 5000$ SNPs after which the replication rates tend to remain stable. Irrespective of the strategy, however, the replication rates are rather high especially at nominal P-values ≥ 1E-03 and for larger set sizes (S ≥ 1000) where they are nearly always at 100%. For smaller set sizes, (S ≤ 100), replication rates are greater under SM than under BLUP strategy. The opposite trend is observed for larger set sizes (S ≥ 5000). To evaluate the significance of these findings, we carried out the PS association analyses on Q1. Note that because of the small number of available replicates (199), replication rates could not be estimated at stringent criteria levels (i.e., nominal P-values < 1%). Estimates of the replication rates were close to the theoretical values, whether the top SNPs were selected under SM or BLUP and irrespective of set size S (ranging from 5% to 8% at P-value=5%). PS association analyses conducted on the permuted DBP_1 trait yielded to slightly inflated rates (results not shown), especially for larger set sizes S and under the BLUP strategy (ranging from 8% to 12% and from 6% to 17% under SM and BLUP strategies, respectively). From these results it is not clear if the distribution of the PS association test appropriately follows the theoretical distribution in the pedigree dataset even though the traits were de-correlated.

Conclusions

Using a classical approach to detect association, we found that, with the exception of the SNPs in the *MAP4* gene, it had no power to detect SNPs of weaker effect at the genome-wide significance level. In our study, we aimed to evaluate PS association analysis as a method to detect SNPs of weaker effect that fail to reach genome-wide significance in classical genome-wide association studies. We used a single-marker approach and a multi-marker approach to derive the top SNP sets. In summary, both strategies lead to relatively high replication rates especially when large sets of SNPs (≥ 1000) were considered. Our study presents some weaknesses and limitations. PS analysis was conducted using linear regression on the de-correlated traits. Thus, the way we constructed the sets of independent top SNPs or the fact that the genotypes were the same in all replicates may have led to biased and inflated estimates of power rates. Type I error rates were found close to the theoretical values when analyzing Q1 but not when analyzing the permutated DBP trait. It appears, therefore, that linkage may have affected our PS analyses even if we worked with de-correlated traits. These results suggest that the presently available methods need to be extended to address the challenges of PS association analyses in pedigree data.

List of abbreviations

BLUP: Best Linear Unbiased Predictor; DBP: diastolic blood pressure; DBP_1: diastolic blood pressure at exam 1; MML: Multi-marker Mixed Linear Model; PS: Polygenic Score; SBP: systolic blood pressure; SM: Single-Marker.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

MM designed the overall study, AL contributed to the design and implemented the software for the BLUP model, NB, MS and MM conducted statistical analyses and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

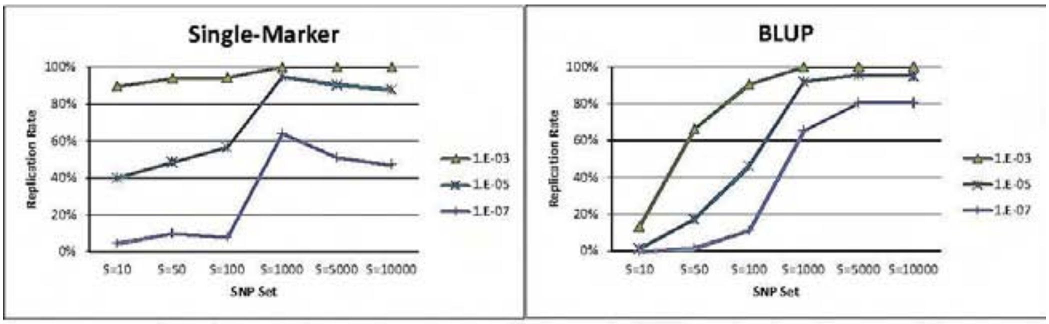
This project was partially supported by the platform bioinformatics Toulouse Midi-Pyrenees. NB has received funding from the European Community's Seventh Framework Programme ([FP7/2007-2013] under grant agreement no 212877 (UEPHA*MS)) and from La Fondation pour l'Aide à la Recherche sur la Sclérose En Plaques. MS has received funding from France PARKINSON Association. NIH grant for the data of GAW18 (to be determined).

References

1. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P: **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.** *Nature* 2009, 460(7256):748-52.
2. Evans DM, Visscher PM, Wray NR: **Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk.** *Hum Mol Genet* 2009, 18(18):3525-31.
3. Ayers KL, Cordell HJ: **SNP selection in genome-wide and candidate gene studies via penalized logistic regression.** *Genet Epidemiol* 2010, 34(8):879-91.
4. Boerwinkle E, Chakraborty R, Sing CF: **The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods.** *Ann Hum Genet* 1986, 50(Pt 2):181-94.
5. Aulchenko YS, de Koning DJ, Haley C: **Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis.** *Genetics* 2007, 177(1):577–585.
6. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics* 2007, 23(10):1294-6.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, 81(3):559-75.
8. Goddard M: **Genomic selection: prediction of accuracy and maximisation of long term response.** *Genetica* 2009, 136(2):245-57.
9. Henderson CR: **Sire evaluation and genetic trends.** *J Anim Sci* 1973, 10-41.
10. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R: **Additive genetic variability and the Bayesian alphabet.** *Genetics* 2009, 183(1):347-63.

Figures

Figure 1 - Polygenic score association results on DBP_1.



Percentage of replicates (out of replicates 2 through 200) with significant evidence of association of PS with DBP_1 at a given nominal P-value by SNP set S derived using either Single-Marker or BLUP strategies in replicate 1.

Genetic Prediction in the Genetic Analysis Workshop 18 Sequencing Data

Andreas Ziegler[1,2], Nora Bohossian[3,4], Vincent P. Diego[5], Chen Yao[6]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

[2]Center for Clinical Trials, University of Lübeck, Germany

[3]Inserm UMR1043-CPTP, CHU Purpan, Toulouse, France

[4]University of Toulouse III – Paul Sabatier, Toulouse, France

[5]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, U.S.A.

[6]Department of Dairy Science, University of Wisconsin, Madison, WI, U.S.A.

Running Title: Genetic Prediction

Corresponding Author

Univ.-Prof. Dr. Andreas Ziegler
Institut für Medizinische Biometrie und Statistik
Universität zu Lübeck
Universitätsklinikum Schleswig-Holstein, Campus Lübeck
Ratzeburger Allee 160, Haus 24
23538 Lübeck
Germany
Phone: +49 (451) 500-2780
Fax: +49 (451) 500-2999
Email: ziegler@imbs.uni-luebeck.de

Abstract

High-throughput sequencing data can be used to predict phenotypes from genotypes, and this corresponds to establishing a prognostic model. In extended pedigrees, the relatedness of subjects provides additional information so that genetic values and heritability can be estimated. Under specific assumptions of the underlying statistical model, fixed or random genetic components can also be estimated. At the Genetic Analysis Workshop 18, the genetic prediction group dealt with both establishing a prognostic model, and in one contribution standard logistic regression was compared with robust logistic regression in a sample of unrelated affected or unaffected individuals. Results of both logistic regression approaches were similar. All other contributions to this group were using extended family data, in general using the quantitative trait blood pressure as example. The individual contributions varied in several important aspects, such as the estimation of the kinship matrix or the estimation method. Contributors chose various approaches for model validation, including different versions of cross-validation or within-family validation. Within-family validation included model building in the first generations and validation in later generations. The choice of the statistical model and the computational algorithm had substantial effects on computation time. If decorrelation approaches are applied, the computational burden is substantially reduced. Some software packages estimated negative eigenvalues although eigenvalues of correlation matrices should be non-negative. It has become clear that the standard animal breeding models are problematic in human genetics because of the different family structure. For all these models improved implementations are required.

Key words: Decorrelation, Genetic prediction, Linear mixed model, next-generation sequencing, personalized medicine, prediction, robust regression

Introduction

After more than a decade of frustration and almost stagnation with the use of non-replicable linkage studies [Risch and Merikangas 1996], great hope for rapid progress has been raised with the publication of the landmark genome-wide association study (GWAS) paper by the Wellcome Trust Case Control Consortium [2007]. Indeed, GWAS have led to a great boost in knowledge about genetic factors contributing to disease. However, the biology of most of the findings is not well understood, and it might take a long time to unravel the biological function of many of the GWAS hits. For example, the best replicated locus over all GWAS is probably the chromosome 9p21.3 region, a "genetic desert", which has been associated with coronary artery disease in several GWAS [Helgadottir et al., 2007; McPherson et al., 2007; Samani et al., 2007]. It took, however, four years to understand the function at this locus [Harismendy et al., 2011].

Even if the biology is not well understood for all loci contributing to the genetic disease, "the genetic information" … might already tell "people about their risk for various diseases and which medications they should use or avoid" [Ioannidis 2009]. Although it is clear that genetic variants contributing to the pathogenesis of complex genetic diseases need to be distinguished from pharmacogenetic markers which guide treatment selection [Ziegler et al., 2012], terms are sometimes used interchangeably and might lead to confusion. For example, **a** single nucleotide polymorphism (SNP) is called to be predictive (**predictive SNP**) if it forecasts the likely response to treatment. In contrast, **predictive genetic tests** concern healthy people and aim at determining individuals at risk for a disease, and the process is called genetic prediction. The term genetic prediction is used in the same way in animal breeding. If multiple markers are used for disease prediction, the statistical model is termed **predictive model**. It is important to note that the term genetic prediction is also used, when quantitative traits are considered, such as blood pressure or body weight. Here, the aim is to predict the value of the quantitative trait instead of a dichotomous disease phenotype.

All papers in the Genetic Prediction group of the Genetic Analysis Workshop 18 dealt with genetic prediction in the sense just described. Interestingly, only one individual contribution to this group investigated prediction models for unrelated subjects, and they considered the binary endpoint hypertension [Kesselmeier et al., 2013]. Their approach will be discussed first.

All other contributions in this group considered family data of large pedigrees [Bohossian et al., 2013; Quillen et al., 2013; Yang et al., 2013; Yao et al., 2013], and most papers used a quantitative trait only.

All articles working with the family data used a specific linear mixed model (LMM) for analysis. In the following section, we introduce the basic LMM and derive the variants of the LMM as used in the individual contributions. We discuss the estimation aims, several approaches to determine the correlation between family members, and some maximization approaches. When compared with estimating equations for independent individuals, estimating equations for correlated subjects are more difficult to solve, need more computational time, and are less numerically stable. One approach to overcome these obstacles is to decorrelate observations, which means that phenotypes –sometimes also the genotypes – are transformed to make observations uncorrelated. If phenotypes are normally distributed, family members are independent after decorrelation. Finally, we consider the properties of some of the estimation methods. By investigating the eigenvalues of some of the solutions we see that not all variances need to be positive. Furthermore, it might be that some variances cannot be reliably estimated and turn out to be negative. We draw the conclusion that improved implementations for LMM are required for application in human genetics.

However, before we start to describe the family studies, we consider the single paper dealing with unrelated individuals.

Influential points and logistic regression versus robust logistic regression

Logistic regression is the standard statistical approach for estimating disease probability from independent subjects. Several crucial assumptions, such as the correct specification of the logistic regression model and the absence of outliers, limit this approach. Specifically, few observations may have substantial influence on the parameter estimates. In their contribution, using Cook's distance [Hosmer et al., 1991], Kesselmeier et al. [2013] first showed that several observations had to be termed influential – the authors termed them outliers –, and these observations substantially affected the parameter estimates.

Cantoni and Ronchetti [2001] proposed a new class of robust regression models using M-estimators as an alternative to standard generalized linear models. The authors showed that their robust logistic regression is "robust" in the sense that a small amount of contamination at a point x has bounded influence on the asymptotic type I error level and the power of the test. However, the authors did not derive any robustness results in terms of changes in parameter estimates. This might be the explanation for the fact that Kesselmeier et al. [2013] observed similar results for the logistic regression and the robust logistic regression.

The basic linear mixed model (LMM)

All remaining papers in the genetic prediction group investigated extended pedigrees using the LMM. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})'$ denote the vector of dependent variables of family i, $i = 1, \ldots, n$, having T_i family members. Let $\boldsymbol{X}_i = (x'_{i1}, \ldots, x'_{iT_i})'$ be the $T_i \times p$ matrix of covariates, and family member $t = 1, \ldots, T_i$. The basic LMM is given by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{G}_i\boldsymbol{\alpha} + \boldsymbol{u}_i + \boldsymbol{\varepsilon}_i\,, \qquad (1)$$

where

- $\boldsymbol{X}_i\boldsymbol{\beta}$ denotes the fixed component of covariates, such as environmental variables, where $\boldsymbol{\beta}$ is the p-dimensional parameter vector of interest.
- $\boldsymbol{G}_i\boldsymbol{\alpha}$ is a fixed genetic marker component, where $\boldsymbol{\alpha}$ is the q-dimensional parameter vector of interest, and $\boldsymbol{G}_i$ is of dimension $T_i \times q$.
- $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \sigma_u^2 \boldsymbol{K}_i)$ is the polygenic component.
- $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I})$ is the error term.
- $\boldsymbol{K}_i$ reflects the correlation within pedigree i on the polygenic component, and it is twice the expected kinship matrix, i.e., $\boldsymbol{K}_i = 2\,\Phi_i$. It has been termed genetic relationship matrix or genetic relationship kernel (GRK) by Blangero et al. [2013].
- σ_u^2 is the variance of the polygenic component, which is assumed to be homoscedastic.
- σ_ε^2 is the variance of the random error, and the random error is assumed to be independent and homoscedastic.

With these assumptions $\boldsymbol{y}_i$ is normally distributed with mean $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{G}_i\boldsymbol{\alpha}$ and covariance matrix $\Sigma_i = \sigma_u^2\boldsymbol{K}_i + \sigma_\varepsilon^2\boldsymbol{I}$.

A few remarks on this LMM are required. First, the independence of the errors within a family statse that $\boldsymbol{X}_i\boldsymbol{\beta}$ models all environmental factors, and that $\boldsymbol{G}_i\boldsymbol{\alpha} + \boldsymbol{u}_i$ captures all genetic information. Second, the homoscedasticity of the polygenic component means that the variance of the polygenes is constant over generations and even within nuclear families of the same large family. Third, the polygenic component captures the additive genetic variance only; dominance and epistasis variance caused by polygenic components are neglected in this model. Fourth, the fixed genetic component $\boldsymbol{G}_i\boldsymbol{\alpha}$ captures all effects from genetic markers using standard coding. For example, for a SNP j with alleles A and a and additive coding, g_{itj} equals 0, 1, and 2 if subject t of family i has 0, 1, and 2 a alleles. Other codings can be used, such as recessive, dominant or heterozygous advantage [Ziegler and König 2010], and coding schemes for rare variants [Dering et al., 2011] may also be employed. Fifth, the model is only identifiable if polygenic effects $\boldsymbol{u}_i$ and the error term $\boldsymbol{\varepsilon}_i$ are uncorrelated. The final aspect is that the LMM of Eq. (1) does not include a gene by environment interaction component although this can be easily modeled through $\boldsymbol{X}_i$ and $\boldsymbol{G}_i$.

Different modifications of this basic LMM of Eq. (1) have been used in the genetic prediction group. For example, Bohossian et al. [2013] used a random genetic marker $\boldsymbol{G}_i\boldsymbol{\alpha}_i$ with $\boldsymbol{\alpha}_i \sim N(0, \sigma_\alpha^2 \boldsymbol{I})$. Yao et al. [2013] did not only look at cross-section data but included panel aspect of the data. To this end, they considered the model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{T}_i\boldsymbol{c}_i + \boldsymbol{\varepsilon}_i\,, \tag{2}$$

where $\boldsymbol{c}_i \sim N(0, \sigma_c^2 \boldsymbol{I})$ is the random effect for repeated measurements. In Eq. (2), the polygenic component $\boldsymbol{u}_i$ is combined with a design matrix, sometimes termed incidence matrix $\boldsymbol{Z}_i$ to capture structure from the repeatedness of the measurements.

Finally, the simplified model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{u}_i + \boldsymbol{\varepsilon}_i \tag{3}$$

without a fixed genetic effect has been used in the genetic prediction group as well.

Likelihood, decorrelation, choice of the genetic relationship kernel

Estimation of the model from Eq. (1) is burdensome as it invariably requires repeated inversion of a potentially large covariance matrix, and it is reasonable to make use of the eigenvalue decomposition (EVD) of the covariance matrix $\Sigma_i = \boldsymbol{U}_i\Lambda_i\boldsymbol{U}_i^{-1}$ into a diagonal matrix $\Lambda_i = \mathrm{diag}(\lambda_{it})$ of eigenvalues and an orthogonal matrix $\boldsymbol{U}_i$ of eigenvectors. The EVD is an orthogonal transformation of the data vector with the aim to linearly transform the vector of non-independent observations $\boldsymbol{y}_i$ to a vector of transformed but independent observations $\tilde{\boldsymbol{y}}_i$. The EVD simplifies the likelihood because transformed observations are independent, and it thus allows simplified calculations. For example, Yang et al. [2013] rewrote the log-likelihood of Eq. (1) as

$$\begin{aligned} L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_u^2, \sigma_\varepsilon^2) = -\frac{1}{2}\Bigg(& n \ln(2\pi\sigma_u^2) + \ln\det\Big(\Lambda_i + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\boldsymbol{I}\Big) \\ & + \frac{1}{\sigma_u^2}\Big(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{G}_i\boldsymbol{\alpha}_i\Big)'\boldsymbol{U}_i\Big(\Lambda_i + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\boldsymbol{I}\Big)\boldsymbol{U}_i\Big(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{G}_i\boldsymbol{\alpha}_i\Big)\Bigg)^{,} \end{aligned}$$

to which they added some penalty so that the penalized log-likelihood was given by $PL_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_u^2, \sigma_\varepsilon^2, \kappa) = L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_u^2, \sigma_\varepsilon^2) - \kappa\, p(\boldsymbol{\alpha})$ for some penalty function $p(\boldsymbol{\alpha})$ and the Lagrange multiplier κ.

Decorrelation, also termed whitening, leads to substantial simplifications in the estimating equations. The estimation process is simpler, faster, and numerically more stable. It is, however, questionable how the GRK $\boldsymbol{K}_i$ should be determined. In the group, three different approaches have been used (Table 1). All groups used the kinship coefficient without genetic marker information, which is based on the reported pedigree relationships. Improvements can be obtained by using estimated relationships instead of reported ones. To this end, Diego et al. [unpublished] used pathway-specific genetic marker-based identity by descent estimates, and both Quillen et al. [2013] and Yao et al. [2013] used genome-wide marker-based estimates. Interestingly, Yao et al. based their GRK on the observed non-centered moment matrix – which corresponds to identity by state values –, not on estimated identity by descent values. Other methods are available as well, see, e.g., VanRaden [2008].

A different decorrelation approach has been used by Bohossian et al. [2013], following the work of Aulchenko et al. [2007]. In their approach, termed GRAMMAR, the linear model $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{u}_i + \boldsymbol{\varepsilon}_i$ is fit in the first step, and the estimated residual $\hat{\boldsymbol{\varepsilon}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{u}}_i$ is used in the second step for genetic marker analyses.

Eigenvalues and heritability

Another simplification using decorrelation can be obtained from Eq. (3) under the assumption that phenotypes are standardized, i.e., $\mathbb{V}ar(y_{it}) = \sigma_y^2 = \sigma_u^2 + \sigma_\varepsilon^2 = 1$ [Blangero et al., 2013].

In this case, the covariance matrix $\boldsymbol{\Sigma}_i = \mathbb{V}ar(\boldsymbol{y}_i)$ is identical to $h^2\,\boldsymbol{K}_i + (1 - h^2)\,\boldsymbol{I}$, where $h^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$ denotes the narrow sense heritability of the polygenic component, and the log likelihood function reduces to

$$L_i = -\frac{1}{2}\sum_{t=1}^{T_i}\Big(1 + h^2(\lambda_{it} - 1)\Big) - \frac{1}{2}\sum_{t=1}^{T_i}\frac{\tilde{y}_{it}^2}{1 + h^2(\lambda_{it} - 1)},$$

where $\tilde{\boldsymbol{y}}_i = \boldsymbol{U}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)$ is the vector of transformed residuals having components $\tilde{y}_{it}$. λ_{it} are obtained by applying the EVD to the GRK $\boldsymbol{K}_i = \tilde{\boldsymbol{U}}_i\tilde{\Lambda}_i\tilde{\boldsymbol{U}}_i^{-1}$additive genetic eigenvalues, and $\tilde{\Lambda}_i = \mathrm{diag}(\lambda_{it})$.

Blangero et al. [2013] and Diego et al. [unpublished] have shown for this model that the expected likelihood ratio statistic (LRT), when the estimate $\hat{h}^2$ exceeds 0, is given by

$$\mathbb{E}(\mathrm{LRT})\big|_{\hat{h}^2>0} = -\sum_{t=1}^{T_i}\ln\Big(1 + h^2(\lambda_{it} - 1)\Big). \qquad (4$$

This means that the expected value of the test statistic depends on the eigenvalues λ_{it} of the GRK, i.e., the kinship matrix. Eigenvalues of selected relative pairs are displayed in Table 2. It is important to note that eigenvalues of monozygotic (MZ) twin pairs are either 0 or 2. Consequently, when the heritability is exactly equal to 1, the ELRT –in consequence the likelihood function – becomes degenerate. However, this can be dealt with by bounding the heritability slightly less than 1. The value of the ELRT is that can be used to compute power in likelihood analysis [Brown et al., 1999]. Diego et al. [unpublished] exploited this utility of the ELRT to show that different GRK algorithms can be compared in terms of their respective power functions and asymptotic relative efficiencies [Blangero et al., 2013].

Estimation and negative eigenvalues

The fixed effects can be estimated by best linear unbiased estimation (BLUE), while random effects are estimated by best linear unbiased prediction (BLUP). BLUE and BLUP may be computed simultaneously by solving the so-called mixed model equations (MME), given e.g., in Searle et al. [1992], and several alternative approaches are available [VanRaden 2008]. Several estimation approaches are available, such as maximum likelihood (ML), restricted ML (REML), penalized ML or gradient methods, and they are implemented in various packages used by group members (Table 1).

The estimator $\hat{\boldsymbol{u}}_i$ from Eq. (1) is called estimated genetic value (EGV) in human genetics, and it is the equivalent to the concept of estimated breeding values (EBV) in animal and plant genetics. If a model with a fixed incidence matrix $\boldsymbol{Z}_i$ is used, estimates of $\hat{\boldsymbol{u}}_i$ can be obtained in this case [VanRaden 2008].

The GRK reflects a correlation matrix, and its eigenvalues should be non-negative. In fact, it has been shown above that one eigenvalue of an MZ twin pair is expected to be 0. However, negative eigenvalues can be observed when ML is used for estimation (PLINK, GCTA, KINGr), and Quillen et al. [2013] therefore bounded the estimates to the admissible parameter space. If some eigenvalues are estimated to be substantially smaller than 0, others will be estimated to be substantially larger than 1 – for balancing, and this results in extreme heritability estimates (Eq. (4)) [Blangero et al., 2013]. Investigators should therefore carefully choose a proper algorithm for estimating the GRK, and some animal and plant breeders have clearly recommended the use of REML instead of ML; see, e.g., Piepho et al. [2008]. Furthermore, the assumption of a homogeneous variance σ_u^2 for the genetic effect is not justifiable because $\mathbb{V}ar\left(\hat{\boldsymbol{u}}_i\right)$ is a function of kin, and the more relatives in the pedigree are, the lower is the estimated standard error of $\hat{\boldsymbol{u}}_i$.

Model validation

Various approaches have been used for model validation (Table 1). Kesselmeier et al. [2013], who dealt with the case-control data in unrelated individuals, used a leave-one-out cross-validation (CV) approach. Yang et al. [2013] ran a 10-fold CV 20 times.

Bohossian et al. [2013] used the first replicate for training and the other 199 replicates for testing. This approach has been criticized in the group discussions because sequences, thus genotypes, were fixed for all replicates and not generated anew for each replicate.

An interesting validation approach was taken by Yao et al. [2013]. In one validation approach, they used the first three generations of each family for model development and the fourth and fifth generation from the same families for model validation. However, this approach can suffer from a high rate of missing data in the younger generations, and there could be secular trends between generations. Finally for late-onset diseases testing the model in the younger generation may lack power. In another validation approach, they used a 5-fold CV. However, CV samples were not generated randomly but families were ordered by family size instead, and every 5 families were randomly assigned to 5 different folds from the CV.

Discussion

The group discussions of the genetic prediction group were intense and different from the previous experience of group members. The LMM used by many authors of this group were technically challenging, and most of the discussion time was dedicated to details of the models and technical details, such as implementations. The results from the actual analysis of the GAW data were secondary. One reason was the late data delivery combined with the high computational burden of some of the models employed.

For example, when the decorrelation approach was used, the analysis of 1 chromosome with 60,000 SNPs using REML with iterations took about half an hour on a desktop PC with a single core. In contrast, the BayesCπ and GBLUP, standing for genomic BLUP, required 2 days for 12,000 MCMC iterations on a cluster when all 960,000 SNPs from all chromosomes were analyzed. Specifically, the BayesCπ model treats the SNP effects as well as their variance as random variables. Furthermore, not all SNP effects are assumed to follow the same probability distribution, but only a fraction, π, of the SNPs.

This difference in computational speed illustrates that decorrelation approaches greatly reduce the computational burden. It is, however, still unclear how to best choose the GRK because the genome-wide estimation of identity by descent values is challenging. However, it would be even better if local identity by descent information, i.e., from smaller chromosomal regions could be used for estimation.

Another important aspect why most time of the discussions was about the models can be explained by the overall negative findings in the group. For example, Bohossian et al. [2013] concluded that their approach had poor ability to identify associated loci in the pedigree data provided. Similar conclusions were drawn by Yang et al. [2013] and Yao et al. [2013]. This consistent finding was reported by one of the reviewers during the review process of the individual articles.

Nevertheless, Quillen et al. [2013] found in their study that the removal of environmental factors, i.e., the extra variability caused by environmental variables, shows promise for increasing the power of associations. This finding seems to be somewhat contradictory to the results from genome-wide association studies, where it has been found that the power of identifying an association may be substantially reduced if adjustments are made for covariates [Pirinen et al., 2012].

An important aspect in addition to computational burden is the validity of results. It has been shown analytically that some eigenvalues are 0 if MZ twin pairs are included in the model. As a result, the normal distribution is degenerate even if the reported pedigree structure is used for determining the GRK. However, when the GRK was estimated, negative eigenvalues were obtained by some packages using ML estimation. Improved implementations of LMM for human pedigrees are urgently required.

Acknowledgments

The authors have learnt much from all participants of the group discussions. The colleagues participating in the discussions were John Blangero, Nora Bohossian, Vincent Diego, Wei Guo, Guan-Hua Huang, Miriam Kesselmeier, Kristine Lee, Maria Martinez, Ellen Quillen, Saroja Voruganti, Can Yang, and Yin Yao Shugart. The Genetic Analysis Workshops are funded by National Institutes of Health grant R01 GM031575. The work of AZ was funded by the German Ministry of Education and Research (grant no.: 0315536F).

References

Aulchenko YS, de Koning DJ, Haley C. 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177(1):577-85.

Blangero J, Diego VP, Dyer TD, Almeida M, Peralta J, Kent JW, Jr., Williams JT, Almasy L, Goring HH. 2013. A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees. Adv Genet 81:1-31.

Bohossian N, Saad M, Legarra A, Martinez M. 2013. Single-marker and multi-marker mixed models for polygenic score analysis in family-based data. BMC Proc in press.

Brown BW, Lovato J, Russell K. 1999. Asymptotic power calculations: description, examples, computer code. Stat Med 18(22):3137-51.

Cantoni E, Ronchetti E. 2001. Robust inference for generalized linear models. J Am Stat Assoc 96(455):1022-1030.

Dering C, Hemmelmann C, Pugh E, Ziegler A. 2011. Statistical analysis of rare sequence variants: An overview of collapsing methods. Genet Epidemiol 35(S1):S12-S17.

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG and others. 2011. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature 470(7333):264-8.

Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A and others. 2007. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316:1491-1493.

Hosmer DW, Taber S, Lemeshow S. 1991. The importance of assessing the fit of logistic regression models: a case study. Am J Public Health 81(12):1630-5.

Ioannidis JP. 2009. Personalized genetic prediction: too limited, too expensive, or too soon? Ann Intern Med 150(2):139-41.

Kesselmeier M, Legrand C, Peil B, Kabisch M, Fischer C, Hamann U, Lorenzo Bermejo J. 2013. Practical investigation of the performance of robust logistic regression to predict the genetic risk of hypertension BMC Proc in press.

McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR and others. 2007. A common allele on chromosome 9 associated with coronary heart disease. Science 316:1488-91.

Piepho HP, Möhring J, Melchinger AE, Büchse A. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161(1-2):209-228.

Pirinen M, Donnelly P, Spencer CC. 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. Nat Genet 44(8):848-51.

Quillen EE, Voruganti VS, Chittoor G, Rubicz R, Peralta JM, Almeida MAA, Kent Jr JW, Dyer TD, Comuzzie AG, Göring HHH and others. 2013. Evaluation of estimated genetic values and their application to genome-wide investigation of blood pressure measurements. BMC Proc in press.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273(5281):1516-7.

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann H-E and others. 2007. Genomewide association analysis of coronary artery disease. N Engl J Med 357(5):443-53.

Searle SR, Casella G, McCulloch CE. 1992. Variance componentsedition. New York: John Wiley & Sons.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414-23.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661-78.

Yang C, Li C, Chen M, Chen X, Hou L, Zhao H. 2013. A penalized linear mixed model for genomic prediction using pedigree structures. BMC Proc in press.

Yao C, Leng N, Weigel KA, Lee KE, Engelman CD, Meyers KJ. 2013. Prediction of genetic contributions to complex traits using whole genome sequencing data. BMC Proc in press.

Ziegler A, Koch A, Krockenberger K, Großhennig A. 2012. Personalized medicine using DNA biomarkers: a review. Hum Genet 131(10):1627-1638.

Ziegler A, König IR. 2010. A Statistical Approach to Genetic Epidemiology: Concepts and Applications. Second edition. Weinheim: Wiley-VCH.

Table 1. Characteristics of the contributed articles using extended pedigrees in the genetic prediction group.

First author	Linear mixed model	Focus of estimation	Decorrelation	GRK	Estimation method	Implementations used	Model validation
Bohossian	Eq. (1) with random genetic marker effect	Random genetic component $\hat{\alpha}$	GRAMMAR	Kinship	BLUP	GS3	Replicate-based (training rep 1; test reps 2-200)
Diego*	Eq. (1)	Heritability based on σ_u^2	Eigen decomp	Kinship; pathway-specific marker-based	ML	GCTA, KING (KINGh, KINGr), PLINK	None
Quillen	Eq. (1)	EGV	Eigen decomp	Kinship genome-wide marker-based	ML	SOLAR GAUSS scripts	None
Yang	Eq. (1)	Fixed genetic component $\hat{\alpha}$	Eigen decomp	Kinship; genome-wide marker-based	Penalized ML	Customized implementation	20 times 10-fold CV
Yao	Eq. (2)	EGV	None	Kinship	Stage I: REML Stage II: GBLUP, BayesCπ	Stage I: Pedigreemm (R package) Stage II: MCMC	5-fold CV with ordering by family size Within-family validation (train generations 1-3; test generations 4-5)

* article not submitted for GAW18 Proceedings; BLUP: best linear unbiased prediction; CV: cross-validation; EGV: estimated genetic value (identical to estimated breeding value in animal breeding); Eigen decomp: eigenvalue decomposition; GBLUP: genomic BLUP; GRAMMAR: see Aulchenko et al. [2007]; GRK: genetic relationship kernel; MCMC: Markov chain Monte Carlo; ML: maximum likelihood; REML: restricted maximum likelihood.

Table 2. Eigenvalues and expected likelihood ratio statistics ($\mathbb{E}(LRT)$) for selected relative pairs.

Relationship	**Eigenvalue 1**	**Eigenvalue 2**	$\mathbb{E}(LRT)$
Monozygotic twins	2	0	0.13880
Sib pair	3/2	1/2	0.03294
Sibship of size s	$(n_s+1)/2$	$(n_s-1)/2$	0.25148

doi:10.1111/ene.12204

EUROPEAN JOURNAL OF NEUROLOGY
THE OFFICIAL JOURNAL OF THE EUROPEAN FEDERATION OF NEUROLOGICAL SOCIETIES

A prospective observational post-marketing study of natalizumab-treated multiple sclerosis patients: clinical, radiological and biological features and adverse events. The BIONAT cohort

O. Outteryck[a], J.C. Ongagna[b], B. Brochet[c], L. Rumbach[d,†], C. Lebrun-Frenay[e], M. Debouverie[f], H. Zéphir[a], J.C. Ouallet[c], E. Berger[d], M. Cohen[e], S. Pittion[f], D. Laplaud[g], S. Wiertlewski[g], P. Cabre[h], J. Pelletier[i], A. Rico[i], G. Defer[j], N. Derache[j], W. Camu[k], E. Thouvenot[l], T. Moreau[m], A. Fromont[m], A. Tourbah[n], P. Labauge[k], G. Castelnovo[l], P. Clavelou[o], O. Casez[p], P. Hautecoeur[q], C. Papeix[r], C. Lubetzki[r], B. Fontaine[r], N. Couturier[s], N. Bohossian[s], M. Clanet[s], P. Vermersch[a], J. de Sèze[b] and D. Brassat[s] on behalf of the BIONAT network, and CFSEP

[a]*Neurologie, Université de Lille Nord de France (EA2686), Hôpital Roger Salengro CHRU Lille, Lille;* [b]*Neurologie, Hôpital Civil, Strasbourg;* [c]*Neurologie, CHU Pellegrin, Bordeaux;* [d]*Neurologie, CHU Besancon, Besancon;* [e]*Neurologie, Hôpital Pasteur, Nice;* [f]*Neurologie, CHU Nancy, Nice;* [g]*Neurologie, CHU Nantes, Nantes;* [h]*Neurologie, CHU Fort de France, Fort de France;* [i]*Neurologie, Hôpital de la Timone, Marseille;* [j]*Neurologie, CHU Caen, Caen;* [k]*Neurologie, CHU Montpellier, Montpellier;* [l]*Neurologie, CHU Nimes, Nimes;* [m]*Neurologie, CHU Dijon, Dijon;* [n]*Neurologie, CHU Reims, Reims;* [o]*Neurologie, CHRU Clermont Ferrand, Clermont Ferrand;* [p]*Neurologie, CHU Grenoble, Grenoble;* [q]*Neurologie, CH St Vincent, GHICL, Lille;* [r]*Neurologie, Hôpital de la Salpêtrière, Paris; and* [s]*Pole des neurosciences CHU Purpan, INSERM U1043, Toulouse, France*

Keywords:
biomarker, BIONAT, JCV, multiple sclerosis, natalizumab, progressive multifocal leukoencephalopathy

Received 27 January 2013
Accepted 30 April 2013

Background and purpose: BIONAT is a French multicentric phase IV study of natalizumab (NTZ)-treated relapsing–remitting multiple sclerosis (MS) patients. The purpose of this study was to collect clinical, radiological and biological data on 1204 patients starting NTZ, and to evaluate the clinical/radiological response to NTZ after 2 years of treatment.

Methods: Patients starting NTZ at 18 French MS centres since June 2007 were included. Good response to NTZ was defined by the absence of clinical and radiological activity. Data analysed in this first report on the BIONAT study focus on patients who started NTZ at least 2 years ago ($n = 793$; BIONAT$_{2Y}$).

Results: NTZ was discontinued in 17.78% of BIONAT$_{2Y}$. The proportion of patients without combined disease activity was 45.59% during the first two successive years of treatment. Systematic dosage of anti-NTZantibodies (Abs) detected only two supplementary patients with anti-NTZ Abs compared with strict application of recommendations. A significant decrease of IgG,M concentrations at 2 years of treatment was found.

Conclusions: The efficacy of NTZ therapy on relapsing–remitting MS in a real life setting is confirmed in the BIONAT cohort. The next step will be the identification of biomarkers predicting response to NTZ therapy and adverse events.

Introduction

BIONAT is a French multicentric phase IV study of natalizumab (NTZ)-treated relapsing–remitting multiple sclerosis (MS) patients. The main objective of the BIONAT study is to prospectively collect clinical, radiological and biological data on more than 1000 patients starting NTZ therapy, in the post-marketing setting, and to identify biological markers associated with the response to NTZ therapy and adverse events connected with it. NTZ is the first of a new class of selective adhesion molecule inhibitors used in relapsing–remitting MS. It was approved in France in April 2007. The results of the AFFIRM study demonstrated that this α4-integrin antagonist reduced the annualized relapse rate (ARR) by 68% over 2 years and the risk

Correspondence: O. Outteryck, Department of Neurology, University Hospital of Lille, Université Lille Nord de France, Hôpital Roger Salengro, 1 rue Emile Laine, 59037 Lille Cedex, France (tel.: −33-3 20 44 68 46; fax: −33-3 20 44 44 84; e-mail: olivier.outteryck@chru-lille.fr).

[†]Deceased

of sustained disability progression in relapsing–remitting MS by 42% 54% [1]. Up to now, the concept of freedom from disease activity in MS (no radiological and no clinical disease activity) has received little attention from the neurology community [2] but a *post hoc* analysis of the AFFIRM study showed that this goal could be attainable for 37% of patients in the NTZ arm versus 7% in the placebo group [3]. As new and increasingly effective treatments for MS are developed, it can be anticipated that 'disease activity free' status in MS will be an attainable goal for greater proportions of patients. Here, the results of 1204 NTZ-treated patients are presented with a focus on the proportion of patients free of disease activity at 2 years of NTZ treatment and biological parameters.

Methods

Patients with relapsing–remitting MS starting NTZ therapy at 18 MS centres in France since June 2007 were included and were followed prospectively. These 18 MS centres were located in various areas of France. Thus the BIONAT cohort could be considered as representative of MS French patients treated by NTZ. All patients gave their written informed consent to participate in the BIONAT study. The decision to treat with NTZ was based on the French guidelines for all the patients [at least one relapse during well conducted interferon-β therapy plus nine T2 lesions and/or gadolinium enhanced lesion(s) on brain MRI or two relapses plus T2 lesion load increasing and/or gadolinium enhanced lesion(s)]. NTZ was infused intravenously once every 4 weeks. All patients were followed up clinically [complete clinical and neurological evaluation for relapses; disability scored on the Expanded Disability Status Scale (EDSS) every 6 months]. All patients receiving NTZ underwent brain magnetic resonance imaging (MRI) just before initiation of the treatment and after completion of 1 and 2 years of treatment. Biological samples were collected at baseline before the first NTZ infusion, at 1, 2 years and planned at 5 years.

Good response to NTZ therapy was defined by the absence of disease activity. The same criteria as in the *post hoc* analysis of the AFFIRM study were used to determine 'clinical disease activity': occurrence of new relapses and/or sustained disability progression. Progression was defined as an increase in the EDSS of at least 1.5 from a baseline score of 0, at least 1.0 point from a baseline score between 1.0 and 5.0 and at least 0.5 from a baseline score higher than or equal to 5.5 [3,4]. Regarding radiological disease activity, all brain MRI performed included axial T2 fluid attenuated inversion recovery and axial T1-weighted gadolinium-enhanced sequences. These examinations were not all performed in the same MRI centre and were not analysed with T2 lesion load measurement software. All brain MRI were analysed by two experienced neurologists in each MS centre. 'Radiological disease activity' was defined as the appearance of any new T2 hyperintense lesions and/or the presence of any gadolinium-enhancing lesions. It was not possible to measure enlarging T2 hyperintense lesions. Patients without clinical activity and without radiological activity were considered as free of combined activity and designated as disease activity free patients.

A 'patient with highly active disease' was defined as a patient presenting at least two relapses in the year before NTZ initiation and at least one gadolinium-enhanced lesion on brain MRI within the 3 months preceding NTZ initiation. All other patients were considered as 'non-highly active' [5].

Disease activity during the first year, during the second year and after two successive years of treatment were analysed.

NTZ was interrupted for various reasons. Pregnancy or planned pregnancy and a patient's decision (not related to MS or because of fear of NTZ potential side effects) were not considered as treatment failure. For any other causes of treatment discontinuation in what was defined as 'the worst-case scenario' patients were considered as treatment failure with clinical and radiological disease activity in the 2 years of treatment if discontinuation occurred in the first year, and only in the second year if discontinuation occurred in the second year of treatment. In 'the best-case scenario', patients discontinuing NTZ were considered without clinical and radiological disease activity in the 2 years of treatment if discontinuation occurred in the first year, and only in the second year if discontinuation occurred in the second year of treatment. In 'the intermediate-case scenario', patients discontinuing NTZ were not considered for the calculation of clinical and radiological disease activity in the 2 years of treatment if discontinuation occurred in the first year, and only in the second year if discontinuation occurred in the second year of treatment.

The data analysed in this first report of the BIONAT study were extracted from our database in May 2012. The proportion of missing data is reported.

Anti-John Cunningham virus (JCV) antibody (Ab) prevalence was measured in the first group of patients included in BIONAT ($n = 811$ patients) with the first generation two-step validated ELISA test at baseline.

Tests for anti-NTZ Abs were performed systematically after 1 year of treatment and also after 6 months of treatment as recommended when a patient presented a clinical evolution [relapse(s) and/or disability progression] or repeated adverse events.

Lymphocyte count and immunoglobulin concentrations were recorded at baseline for the entire BIONAT cohort. In one MS centre of the BIONAT network ($n = 133$), prospective follow-up of lymphocyte count (CD3, CD4, CD8, CD19) and immunoglobulin (Ig; types G, A and M) concentrations was performed during 2 years of continuous treatment with NTZ. These analyses were performed on site with the same method throughout the follow-up.

Statistical analyses were performed with SPSS 15.0 software (Chicago, IL, USA). A Student's t test was used to compare mean values and the χ^2 test to compare data distributions. When multiple comparisons were performed, a Bonferroni correction was applied. The level of statistical significance (P) was set at 0.05.

Results

Description of the BIONAT cohort

The BIONAT study included 1204 patients. At the time of analysis, 65.9% of patients ($n = 793$) had started NTZ therapy at least 2 years previously. Amongst these patients, 41.4% ($n = 328$) were classified as having highly active disease.

Demographic characteristics of the BIONAT cohort, the cohort already at 2 years of treatment (BIONAT_{2Y}) and the subgroup with highly active disease, itself already at 2 years of treatment (BIONAT_{HA2Y}), are shown in Table 1.

BIONAT and BIONAT_{2Y} compared with the AFFIRM pivotal trial

BIONAT and BIONAT_{2Y} cohorts were very similar to the NTZ arm of the AFFIRM pivotal trial and had similar mean age and gender distributions. However, they had significantly higher pre-NTZ ARR and EDSS ($P < 0.01$) with less radiological activity, measured by the number of gadolinium-enhanced lesion(s) ($P < 0.01$).

BIONAT_{HA2Y} versus BIONAT/BIONAT_{2Y}

The BIONAT_{HA2Y} cohort was younger ($P < 0.01$) with higher ARR and higher radiological activity ($P < 0.01$) than the BIONAT/BIONAT_{2Y} cohorts.

Table 1 Demographic, clinical and radiological characteristics of the BIONAT cohort

	AFFIRM	BIONAT	BIONAT_{2Y} patients at least 2 years from NTZ initiation	BIONAT_{HA2Y} patients with highly active disease and at least 2 years from NTZ initiation
Number	627	1204	793	328
Age (NTZ initiation; years)				
Mean ± SD	35.6 ± 8.5	36.67 ± 9.79	36.61 ± 9.71	34.45 ± 9.29
Median [min;max]	[18;50]	36 [15;72]	36 [17;72]	33 [17;60]
Gender (sex ratio F/M)	2.52	3.21	3.03	3.37
Disease duration (years)				
Mean ± SD		11.57 ± 6.92	12.15 ± 6.87	11.06 ± 6.53
Median [min;max]	5 [0;34]	10 [1;45]	11 [2;45]	9 [2;33]
ARR				
Mean ± SD	1.53 ± 0.91	2.05 1.1	2.1 ± 1.09	2.7 ± 0.95
Median [min;max]	[0;12]	2 [1;8]	2 [1;8]	2 [2;8]
EDSS				
Mean ± SD	2.3 ± 1.2	3.35 ± 1.66	3.36 ± 1.67	3.26 ± 1.68
Median [min;max]	[0;6]	3.25 [0;8]	3.5 [0;8]	3 [0;8]
Brain MRI				
Number of Gd+ lesions	2.2 ± 4.7	1.18 ± 3.03	1.07 ± 2.52	3.07 ± 3.78
Proportion of patients with Gd+ lesions	51%	63.8%	64.4%	100%
Proportion of patients with IMD > 6 months	0%	88.9%	90.5%	84.2%
Number of IMD				
Mean ± SD	NA	1.3 ± 0.73	1.33 ± 0.72	1.23 ± 0.75
Median [min;max]	NA	1 [0;4]	1 [0;4]	1 [0;3]
Proportion of patients with ISD	NA	20.09%	21.4%	22.7%
Number of ISD				
Mean ± SD	NA	0.26 ± 0.57	0.28 ± 0.6	0.3 ± 0.61
Median [min;max]	NA	0 [0;3]	0 [0;3]	0 [0;3]

ARR, annualized relapse rate; EDSS, Expanded Disability Status Scale; MRI, magnetic resonance imaging; Gd+, gadolinium enhancement; IMD, immunomodulatory drugs; ISD, immunosuppressive drugs; NA, not applicable.

NTZ discontinuations

Table 2 summarizes the numerous causes and various times of discontinuation in BIONAT$_{2Y}$. Discontinuation occurred in 17.78% ($n = 141$) of BIONAT$_{2Y}$, mostly in the first year (60.28%) and with mean treatment duration at NTZ discontinuation of 10.89 ± 6.56 months. The most frequent causes, together representing more than half the discontinuations, were pregnancy planning (24.82%), cutaneous allergy (17.02%), always occurring in the first year, conversion to a secondary progressive form of MS (12.06%) and serious adverse event (8.51%).

NTZ tolerance and antibodies against NTZ

Serious adverse events in BIONAT$_{2Y}$ were one progressive multifocal leukoencephalopathy (PML) (19th infusion), five neoplasms (two breast, one rectal, one uterine, one basocellular carcinoma), one lethal Hurst acute encephalopathy (reported on autopsy analysis), one autoimmune haemolytic anaemia, one pulmonary embolism, one myelodysplasic syndrome, one case of repeated bronchopneumonia and one case of elevated hepatic cytolysis. Five cases of PML were also observed in the BIONAT cohort but after the first 2 years of treatment (39th, 44th, 44th, 45th and 50th infusion; see Table 3).

Table 2 Causes, time and distribution of NTZ discontinuation in BIONAT$_{2Y}$ and BIONAT$_{HA2Y}$

Causes of NTZ discontinuation and number of patient(s)	Year 0–1	Year 1–2	Year 0–2
Pregnancy planning	16 (9)	19 (9)	35 (18)
Pregnancy	1 (0)	4 (4)	5 (4)
Moving away	1 (0)	1 (0)	2 (0)
Patient wish without validated medical reason	5 (1)	3 (2)	8 (3)
Cutaneous allergy	24 (10)	0 (0)	24 (10)
SPMS	9 (2)	8 (4)	17 (6)
Serious adverse event	7 (6)	5 (1)	12 (7)
Lack of efficacy without anti-NTZ Ab	5 (4)	1 (0)	6 (4)
Unknown	3 (1)	6 (5)	9 (6)
Psychiatric disorder	3 (0)	2 (2)	5 (2)
Fear of PML	1 (0)	3 (2)	4 (2)
Relapse(s) + anti-NTZ Ab	5 (3)	2 (1)	7 (4)
Adverse event + anti-NTZ Ab	2 (1)	0 (0)	2 (1)
Intolerance without anti-NTZ Ab	2 (1)	0 (0)	2 (1)
Absence of compliance	1 (0)	1 (1)	2 (1)
Total	85 (38)	56 (31)	141 (69)

SPMS, secondary progressive multiple sclerosis; PML, progressive multifocal leukoencephalopathy; Ab, antibody; values for BIONAT$_{HA2Y}$ are indicated in parentheses.

Table 3 Demographic data of PML patients

Sex	F	F	F	F	M	F
Number of NTZ infusions at PML diagnosis	19	39	44	44	45	50
Pre-NTZ treatment history	Interferon β1a IM	Interferon β1a IM Copolymer acetate Mitoxantrone Cyclophosphamide	Interferon β1a IM	Interferon β1a IM	Interferon β1a SC Interferon β1a IM	Interferon β1a IM Interferon β1a SC
JCV status	+	+	+	+	+	+

PML, progressive multifocal leukoencephalopathy; JCV, John Cunningham virus; IM, intramuscular; SC, subcutaneous.

Tests for anti-NTZ Abs were performed in clinical practice after 6 months of treatment as recommended when a patient presented clinical evolution (mainly relapses) or repeated adverse events. Detection of neutralizing Abs against NTZ involved 1.1% (9/793) of BIONAT$_{2Y}$ and represented 6.4% (9/141) of discontinuations in BIONAT$_{2Y}$. The Abs were mainly detected in the first year (77.8% of BIONAT$_{2Y}$) and most often detected after a relapse rather than for repeated adverse event occurrence.

A systematic assay of anti-NTZ Abs at 1 year of treatment was also performed, which was positive in 2.6% (21/793) of cases in BIONAT$_{2Y}$ versus 1.1% by following recommendations. The other 1.5% were represented by nine patients with cutaneous allergy during the first injections which stopped NTZ therapy, two patients without any clinical or radiological activity during the first and also the second year of treatment and one patient with pregnancy planned during the first year of treatment. Finally, systematic testing for anti-NTZ Abs detected only two supplementary patients with anti-NTZ Abs compared with strict application of the recommendations. In the BIONAT$_{HA2Y}$ cohort the causes, time and distribution of NTZ discontinuation were fully comparable with BIONAT$_{2Y}$.

Analysis of disease activity under NTZ

The proportions of patients free of disease activity in BIONAT$_{2Y}$ and BIONAT$_{HA2Y}$ are reported and compared with the *post hoc* AFFIRM study in Fig. 1, which considered only the worst-case scenario. At the time of analysis, to evaluate clinical, radiological and combined disease activity, missing data were noted. Missing data were observed in 8%, 22% and 22.5% for clinical disease activity at year 0 1, year 1 2 and year 0 2 respectively; in 13%, 25% and 26% for radiological disease activity at year 0 1, year 1 2 and year 0 2 respectively; in 15%, 27% and 28% for combined disease activity at year 0 1, year 1 2 and year 0 2 respectively. Proportions of patients without clinical, radiological and combined disease activity at year 0 2 in BIONAT$_{2Y}$ according to the worst-, intermediate- and best-case scenarios are reported in Table 4. Only six patients presented radiological activity without clinical activity during their first 2 years of treatment. These patients represented only 1.4% of patients from the BIONAT$_{2Y}$ cohort with two complete years of treatment and full data available (6/432).

None of the demographical, clinical or radiological baseline data of the BIONAT$_{2Y}$ cohort were predictive of a better response to NTZ, whatever the scenario considered (Table 5).

Annualized relapse rates of BIONAT$_{2Y}$ at year 0 1 and year 1 2 were 0.45 ± 0.78 and 0.32 ± 0.61 respectively. Thus reduction of ARR in BIONAT$_{2Y}$ was evaluated as 78.6% in the first year, with an additional reduction of 28.9% in the second year. ARRs of BIONAT$_{HA2Y}$ at year 0 1 and year 1 2 were 0.45 ± 0.81 and 0.29 ± 0.58 respectively. Thus reduction of ARR in BIONAT$_{HA2Y}$ was higher than in BIONAT$_{2Y}$ and evaluated as 83.3% in the first year, with an additional reduction of 35.6% in the second year.

Table 4 Proportion of patients without clinical, radiological and combined disease activity in the worst versus the best scenario

	Worst scenario	Intermediate scenario	Best scenario
Clinical (%)			
Year 0–1	62.41	67.32	69.71
Year 1–2	57.64	68.48	73.46
Year 0–2	47.43	56.42	63.36
Radiological (%)			
Year 0–1	88.88	96.31	96.59
Year 1–2	82.25	98.67	98.89
Year 0–2	80.85	96.88	97.40
Combined (%)			
Year 0–1	61.28	66.56	69.20
Year 1–2	56.41	67.95	73.40
Year 0–2	45.59	55.09	62.84

Biological analysis

The anti-JCV Ab prevalence of our first 811 patients included in the BIONAT study (BIONAT$_{JCV}$) was evaluated: 57.7% of these patients had anti-JCV antibodies. The mean ages in years (± SD) of anti-JCV

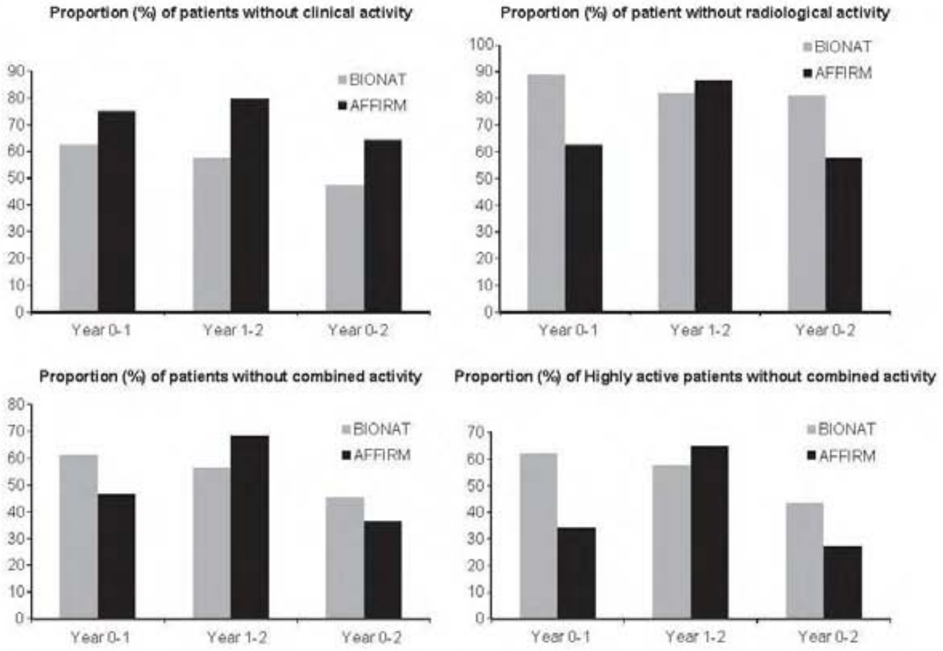


Figure 1 Proportion of patients without radiological, clinical and combined disease activity in BIONAT$_{2Y}$ and BIONAT$_{HA2Y}$ according to the worst-case scenario.

Table 5 Demographic, clinical and radiological baseline characteristics in subgroups of BIONAT$_{2Y}$ according to the best-/intermediate-/worst-case scenario and their disease activity status

	Best-case scenario		Intermediate-case scenario		Worst-case scenario	
	Free	Not free	Free	Not free	Free	Not free
Number	328	194	238	194	238	284
Age (years)						
Mean ± SD	37.22 ± 9.99	37.39 ± 9.52	37.13 ± 10.31	37.39 ± 9.52	37.13 ± 10.31	37.40 ± 9.39
Median [min;max]	37 [18;65]	37 [17;64]	37 [18;65]	37 [17;64]	37 [18;65]	37 [17;64]
Gender (sex ratio)	2.95	2.88	2.55	2.88	2.55	3.30
Disease duration (years)						
Mean ± SD	12.53 ± 7.26	12.79 ± 6.78	12.54 ± 7.32	12.79 ± 6.78	12.54 ± 7.32	12.70 ± 6.89
Median [min;max]	11 [2;45]	11 [3;37]	11 [3;45]	11 [3;37]	11 [3;45]	11 [2;40]
ARR						
Mean ± SD	2.05 ± 1.04	2.21 ± 1.20	1.99 ± 0.99	2.21 ± 1.20	1.99 ± 0.99	2.22 ± 1.18
Median [min;max]	2 [0;6]	2 [0;8]	2 [0;5]	2 [0;8]	2 [0;5]	2 [0;8]
EDSS						
Mean ± SD	3.51 ± 1.78	3.42 ± 1.64	3.40 ± 1.76	3.42 ± 1.64	3.40 ± 1.76	3.53 ± 1.71
Median [min;max]	3.5 [0;8]	3.5 [0;8]	3.5 [0;8]	3.5 [0;8]	3.5 [0;8]	3.5 [0;8]
Brain MRI						
Number of Gd − lesion(s)	1.09 ± 2.42	0.85 ± 2.88	1.06 ± 2.46	0.85 ± 2.88	1.06 ± 2.46	0.96 ± 2.72
Proportion of patients with Gd − lesion(s) (%)	55.90	46.9	53.39	46.9	53.39	51.79

ARR, annualized relapse rate; EDSS, Expanded Disability Status Scale; MRI, magnetic resonance imaging; Gd −, Gadolinium enhancement. Patients discontinuing NTZ for pregnancy or planned pregnancy and personal decision (not related to MS or because of fear of NTZ potential side effects) and patients without complete available data at 2 years of treatment are not included in this analysis. No significant difference was noted.

seronegative and seropositive patients were 34.56 ± 9.67 and 37.53 ± 9.72 respectively and were shown to be statistically different by the Student *t* test ($P < 0.01$). Statistical significance was also observed when the distribution of JCV serological status on either side of the mean age of our cohort (≤ 36 years and > 36 years) was compared by the χ^2 test ($P < 0.01$). Positive anti-JCV serostatus was not associated with the use of immunosuppressive or immunomodulatory pretreatment(s). No difference in terms of anti-JCV serological status between women (57.02% of anti-JCV seropositive) and men (57.67% anti-JCV seropositive) was observed. The mean ages (± SD) of women (36.77 ± 9.83 years) and men (35.61 ± 10.33 years) in BIONAT$_{JCV}$ were not statistically different ($P = 0.15$). Because serological analysis was performed on patients' serum at baseline, it was not possible to evaluate whether NTZ exposure duration was a risk factor for being anti-JCV seropositive. Nevertheless, an evaluation of the annual rate of JCV seroconversion in the BIONAT cohort is planned.

At baseline, mean IgA,G,M values could be considered as normal according to laboratory normative values (IgA, 0.88 4.10 g/l; IgG, 6.90 14.00 g/l; IgM, 0.40 2.40 g/l). After 1 year of continuous NTZ therapy, significant increases in all lymphocyte subtypes, a significant decrease in IgM concentrations and a trend towards decreasing IgG concentrations were observed. These biological modifications persisted and remained or became significant at 2 years of treatment (Table 6).

Discussion

Analysis of disease activity is of prime importance in the BIONAT study because the identification of biomarkers of good response to NTZ therapy will depend on it. In this first report of the BIONAT study, missing data were reported at reasonable levels and did not interfere with data interpretation. The BIONAT cohort was comparable with the AFFIRM study in terms of age and sex ratio [1] but, as in many postmarketing studies already published, patients of the BIONAT cohort also presented more severe disease with higher pre-NTZ EDSS score and ARR [6 10]. Real life NTZ practice in France and other European countries shows that NTZ is, in fact, mainly proposed as a second-line therapy.

NTZ discontinuation causes are most often associated with a bad response to NTZ. All three scenarios excluded from the disease activity analysis were patients discontinuing NTZ for pregnancy or pregnancy planning and also patients' decision not related to MS or to fear of NTZ potential side effects. Thus it could be considered that the worst-case scenario is closer to the real life NTZ efficiency than the other scenarios. The proportion of patients without clinical or radiological disease activity in BIONAT$_{2Y}$ seemed

Table 6 Biological follow-up of lymphocytic immunophenotyping and Ig G, A, M concentrations in BIONAT ($n = 1204$) and in a subgroup of BIONAT_{2Y} ($n = 133$) at baseline, first and second years of NTZ therapy

	T0 ($n = 1204$)	T0* ($n = 133$)	T1*	T2*	T0* vs. T1	T1* vs. T2	T0* vs. T2*
IgG (g/l)	9.89 ± 2.55 9.77 [0.47–19.7]	10.07 ± 2.53 10.19 [5.6–16.5]	9.58 ± 2.40 9.6 [4.6–16]	9.10 ± 2.24 8.9 [3.6–15.6]	NS	NS	$P = $ **0.024**
IgA (g/l)	2.01 ± 0.86 1.9 [0.09–9.38]	2.05 ± 0.76 1.975 [0.78–4.15]	1.89 ± 0.74 1.73 [0.58–4.55]	1.92 ± 0.81 1.67 [0.6–4.76]	NS	NS	NS
IgM (g/l)	1.34 ± 0.91 1.2 [0.05–18]	1.27 ± 0.70 1.17 [0.24–5.03]	0.79 ± 0.51 0.69 [0.15–3.45]	0.75 ± 0.48 0.61 [0.14–3.00]	$P <$ **0.001**	NS	$P <$ **0.001**
Total lymphocytes (cells/mm^3)	2230 ± 2430 1.81 [0.19–33.8]	1913 ± 722 1798 [792–4200]	3365 ± 1033 3281 [1279–7992]	3489 ± 991 3406 [1699–7440]	$P <$ **0.001**	NS	$P <$ **0.001**
CD3 (/mm^3)	NA	1389 ± 535 1283 [385–2919]	2182 ± 738 2042 [767–4955]	2263 ± 727 2103 [927–4587]	$P <$ **0.001**	NS	$P <$ **0.001**
CD4 (/mm^3)	NA	921 ± 373 856.5 [276–1933]	1419 ± 486 1376 [558–3157]	1468 ± 471 1365 [638–2842]	$P <$ **0.001**	NS	$P <$ **0.001**
CD8 (/mm^3)	NA	457 ± 213 408 [129–1124]	852 ± 335 810 [270–1972]	886 ± 334 818 [293–1928]	$P <$ **0.001**	NS	$P <$ **0.001**
CD19 (/mm^3)	NA	271 ± 131 261 [51–813]	779 ± 348 741 [96–2107]	776 ± 327 740 [88–1953]	$P <$ **0.001**	NS	$P <$ **0.001**

No statistical differences were observed between T0 and T0*. Significant differences are indicated in bold. NS, non-significant.

to be similar to that reported in the *post hoc* analysis of the AFFIRM study. However, substantial differences could be noted. Reduction of clinical disease activity was less marked in BIONAT_{2Y} and reduction of radiological activity was greater in BIONAT_{2Y} than in the *post hoc* study of AFFIRM. It was also noted that NTZ-treated patients with radiological activity but without clinical activity are fewer than in AFFIRM/SENTINEL pivotal studies [11] and another post-marketing study (17%) [12]. It could be speculated that this difference is due to the high clinical disease activity of our cohort and to a potential underestimation of radiological activity because it was not possible to determine enlarging T2 lesions. The greatest reduction of ARR was observed in BIONAT_{HA2Y}, which concerns younger patients with higher ARR pre-NTZ and more gadolinium-enhanced lesions on brain MRI. These previous demographic and clinical characteristics are known to be in favour of a better response to NTZ [5]. However, younger age and higher ARR pre-NTZ were not found to be predictive of a better response to NTZ whichever scenario was considered.

As in all previous studies [13 17], seroprevalence for anti-JCV antibodies increased with age. In contrast to some previous studies [14,16,17] but in agreement with others [13, 15, 18], males were not found to be over-represented in the anti-JCV seropositive population. In the STRATIFY-1 cohort male gender was a risk factor for being JCV seropositive but there was only a trend towards significance in the TYGRIS-US cohort [14] as in a large German NTZ-treated MS cohort [16]. The mean ages of women and men cohorts are not known and a confounding effect of age could not be excluded. In the BIONAT cohort, past history of immunosuppressive or immunomodulatory drugs was not associated with risk of being anti-JCV seropositive like most post-marketing cohorts [14 16,18] but unlike one study [19].

The frequency of anti-NTZ Abs in BIONAT_{2Y} was lower than reported for the NTZ pivotal studies [20]. Our data confirm the interest of anti-NTZ Ab screening in the case of relapse or repeated adverse events but do not argue for a systematic screening of anti-NTZ Abs.

It is well known that duration of NTZ exposure is a PML risk factor for MS patients [21]. A decrease of serum Ig concentrations in MS patients under NTZ therapy is reported for the first time. The decrease of IgM and IgG concentrations at 2 years observed in a subsequent analysis of the BIONAT cohort could be considered to be in agreement with maximum PML risk after 2 years of treatment. Nevertheless, these data were not compared with those of healthy controls and a slight hypergammaglobulinemia in our MS patients before NTZ initiation that could be corrected with efficient therapy cannot be excluded. Nevertheless, MS is not a known cause of hypergammaglobulinemia and baseline Ig values of our cohort are in normal range. In MS patients, NTZ significantly increases the numbers of B cells and particularly immature pre-B cells [22] that highly express VLA4 antigen but that are also decreased in the blood and increased in cerebrospinal fluid in the earliest clinical stage of the disease [23]. The modification of B cell distribution under NTZ is concomitant with others' findings: disappearance of oligoclonal bands in MS patients [24], decreasing IgG [25,26] and IgM intrathecal secretion [25]. To the best of our knowledge, there is no previous report about a significant decrease of IgG and IgM rates under NTZ. Although NTZ is associated with the occurrence of PML [21], it has never been clearly demonstrated that it could induce other opportunistic infections or facilitate other infections in general. Humoral immune response in NTZ-treated patients seems comparable with that achieved in healthy individuals [27]. At this time, humoral changes under NTZ are not well understood.

This is the first report of the BIONAT study. It is planned to complete all missing values at 2 years but also project clinical, radiological and biological follow-up at 5 years. The next step of the BIONAT study (planned for 2014) is a genome and transcriptome screening of the 1204 patients and, of course, potential identification of biomarkers predicting response to NTZ therapy as planned in the Best-MS network (Best therapeutic choice for MS, a 2012 FP7 Health Innovation 1 supported network). Pre-PML samples from six patients up to 50 months before PML diagnosis were also collected. The aim is to determine whether any biomarkers are associated with PML prediction.

Acknowledgements

The authors would like to thank CFSEP (Club Francophone de la SEP, Dr Heinzleff), the REFGENSEP (Réseau francais de Recherche sur la Génétique de la SEP, Professor Fontaine), TYSEDMUS and the Eugène Devic Foundation (Professors Confavreux and Vukusic) for their help and constant support. BIONAT is funded by the French Ministry of Health (Projet Hospitalier de Recherche Clinique, PHRC 2008) the French MS Society (ARSEP 2007, 2008, 2010) and an FP7 Health Innovation 1 grant (2012). JC virus serology and 1-year neutralizing antibody determination were funded by Biogen Idec at FOCUS Diagnostics (CA, USA) and Pasteur France.

Disclosure of conflict of interest

Please refer to Data S1 for the full list of disclosures.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Author disclosures.

References

1. Polman CH, O'Connor PW, Havrdova E, *et al.* A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med* 2006; **354:** 899–910.
2. Havrdova E, Galetta S, Stefoski D, *et al.* Freedom from disease activity in multiple sclerosis. *Neurology* 2010; **74** (suppl 3)**:** S3–S7.
3. Havrdova E, Galetta S, Hutchinson M, *et al.* Effect of natalizumab on clinical and radiological disease activity in multiple sclerosis: a retrospective analysis of the natalizumab safety and efficacy in relapsing–remitting multiple sclerosis (AFFIRM) study. *Lancet Neurol* 2009; **8:** 254–260.
4. Rio J, Nos C, Tintoré M, *et al.* Defining the response to interferon-β in relapsing–remitting multiple sclerosis patients. *Ann Neurol* 2006; **59:** 344–352.
5. Hutchinson M, Kappos L, Calabresi PA, *et al.* The efficacy of natalizumab in patients with relapsing multiple sclerosis: subgroup analyses of AFFIRM and SENTINEL. *J Neurol* 2009; **256:** 405–415.
6. Outteryck O, Ongagna JC, Zéphir H, *et al.* Demographic and clinic characteristics of French patients treated with natalizumab in clinical practice. *J Neurol* 2010; **257:** 207–211.
7. Prosperini L, Borriello G, Fubelli F, *et al.* Natalizumab treatment in multiple sclerosis: the experience of S. Andrea MS Centre in Rome. *J Neurol Sci* 2010; **3:** 303–307.
8. Fernandez O, Alvarenga MP, Guerrero M, *et al.* The efficacy of natalizumab in patients with multiple sclerosis according to level of disability: results of an observational study. *Mult Scler* 2011; **17:** 192–197.
9. Mancardi GL, Tedeschi G, Amato MP, *et al.* Three years of experience: the Italian registry and safety data update. *Neurol Sci* 2011; **31**(suppl 3)**:** S295–S297.
10. Piehl F, Holmen C, Hillert J, *et al.* Swedish natalizumab (Tysabri) multiple sclerosis surveillance study. *Neurol Sci* 2011; **31**(Suppl 3)**:** S289–S293.
11. Bates D, Bartholomé E. Treatment effect of natalizumab on relapse outcomes in multiple sclerosis patients despite ongoing MRI activity. *J Neurol Neurosurg Psychiatry* 2012; **83:** 55–60.
12. Prosperini L, Gianni C, Barletta V, *et al.* Predictors of freedom from disease activity in natalizumab treated-patients with multiple sclerosis. *J Neurol Sci* 2012; **323:** 104–112.
13. Egli A, Infanti L, Dumoulin A, *et al.* Prevalence of polyomavirus BK and JC infection and replication in 400 healthy blood donors. *J Infect Dis* 2009; **199:** 836–846.
14. Bozic C, Richman S, Plavina T, *et al.* Anti-John Cunningham virus antibody prevalence in multiple sclerosis patients: baseline results of STRATIFY-1. *Ann Neurol* 2011; **70:** 742–750.
15. Outteryck O, Ongagna JC, Duhamel A, *et al.* Anti-JCV antibody prevalence in a French cohort of MS patients under natalizumab therapy. *J Neurol* 2012; **259:** 2293–2298.
16. Trampe AK, Hemmelmann C, Stroet A, *et al.* Anti-JC virus antibodies in a large German natalizumab-treated multiple sclerosis cohort. *Neurology* 2012; **78:** 1736–1742.

17. Olsson T, Achiron A, Alfredsson L, *et al.* Anti-JC virus antibody prevalence in a multinational multiple sclerosis cohort. *Mult Scler* 2013. [Epub ahead of print]
18. Warnke C, Dehmel T, Posevitz-Fejfàr A, *et al.* Anti-JCV antibody prevalence in a German MS cohort [abstract]. *Mult Scler* 2011; **10**(suppl): P520.
19. Achiron A, Chapman J, Dolev M, *et al.* Epidemiology of anti-JCV antibodies in Israeli multiple sclerosis population. *American Academy of Neurology* 2012 – 64th Annual Meeting. 2012; **78:** P02.138.
20. Calabresi PA, Giovannoni G, Confavreux C, *et al.* The incidence and significance of anti-natalizumab antibodies: results from AFFIRM and SENTINEL. *Neurology* 2007; **69:** 1391–1403.
21. Bloomgren G, Richman S, Hotermans C, *et al.* Risk of natalizumab-associated progressive multifocal leukoencephalopathy. *N Engl J Med* 2012; **366:** 1870–1880.
22. Krumbholz M, Meinl I, Kümpfel T, *et al.* Natalizumab disproportionately increases circulating pre-B and B cells in multiple sclerosis. *Neurology* 2008; **71:** 1350–1354.
23. Lee-Chang C, Zéphir H, Top I, *et al.* B-cell subsets up-regulate α4 integrin and accumulate in the cerebrospinal fluid in clinically isolated syndrome suggestive of multiple sclerosis onset. *Neurosci Lett* 2011; **487:** 273–277.
24. von Glehn F, Farias AS, de Oliveira AC, *et al.* Disappearance of cerebrospinal fluid oligoclonal bands after natalizumab treatment of multiple sclerosis patients. *Mult Scler* 2012; **18:** 1038–1041.
25. Villar LM, Garcia-Sanchez MI, Costa-Frossard L, *et al.* Immunological markers of optimal response to natalizumab in multiple sclerosis. *Arch Neurol* 2012; **69:** 191–197.
26. Mellergard J, Edström M, Vrethem M, *et al.* Natalizumab treatment in multiple sclerosis: marked decline of chemokines and cytokines in cerebrospinal fluid. *Mult Scler* 2010; **16:** 208–217.
27. Vagberg M, Kumlin U, Svenningsson A. Humoral immune response to influenza vaccine in natalizumab-treated MS patients. *Neurol Res* 2012; **34:** 730–733.