NOVEL STRATEGIES FOR THE IDENTIFICATION OF BIOMARKERS OF NON-HODGKIN LYMPHOMA: EVIDENCE FROM THE EUROPEAN PROSPECTIVE INVESTIGATION INTO CANCER (EPIC)

Rachel S. Kelly

May 2013

A Thesis Submitted in Fulfilment of Requirements for the Degree of Doctor of Philosophy of Imperial College London

Department of Epidemiology and Biostatistics
Imperial College London

ABSTRACT

Non-Hodgkin's Lymphomas (NHL) represent the eighth most common cancer in Western Europe. Yet despite their widespread prevalence and high mortality rate relatively little is known about the aetiology of these hematological malignancies. Consequently NHL represents an ideal candidate for the discovery of biomarkers lying along the causal pathway. Such biomarkers would allow the improved identification of risk factors and high risk individuals, as well as an enhanced understanding of lymphomageneisis. However, to date there has been little progress in determining validated predictive biomarkers of NHL.

This thesis attempts to address some of the issues that have previously hampered the study of NHL through novel strategies of biomarker identification utilising novel methodologies, technologies and statistical techniques. The thesis comprises a nested case-control study within the European Prospective investigation into Cancer (EPIC) cohort and is split into two parts: the 'validation of biomarkers' and the 'integration of biomarkers'. The most exciting finding was the identification of a novel biomarker for Follicular lymphoma based on the t(14;18) translocation which comprises a previously unknown pre-disease condition. Although no other predictive biomarkers were identified this work represents a 'proof-of-principle' for the use profile regression in the study of highly dimensional complex datasets, and the possibility of using mass-spectrometry derived metabolic profiles in the study of lymphoma. Part two of the thesis confirmed that the use of the '*meet-in-the-middle*' approach was a valuable and feasible method for studying the complete causal pathway from risk factor to disease.

Together these results highlight potential avenues for further study of NHL and confirm the utility of a number of novel strategies that can aid such work. Additionally it informs on some of the likely challenges that will be involved.

DECLARATION OF ORIGINALITY

I hereby declare that this thesis is my own work. All work included within that was not conducted personally be me has been acknowledged, cited and referenced as appropriate.

This was a multi-disciplinary thesis based on the statistical analysis and epidemiological interpretation of results produced from a variety of laboratory analyses and questionnaire data. All laboratory based work was conducted by others, as named in the thesis. The questionnaire and anthropometric data were collected within the European Prospective Investigation into Cancer (EPIC) cohort, and the Northern Sweden Health and Disease Study (NSHDS). The PI of EPIC is Professor Elio Riboli and the EPIC-Lymphoma working group is led by Professor Paolo Vineis. The NSHDS is led by Professor Göran Hallmans and Professor Beatrice Melin. The PI of the EnviroGenoMarkers Study is Professor Soterios Kyrtopoulos. The InterLymph Consortium is led by a coordinating committee with a rotating leadership. I am a member of the EPIC-Lymphoma working group, the EnviroGenoMarkers consortium and the InterLymph Consortium.

Chapters 1 (Introduction), 2 (Literature review) and 9 (Discussion) were researched planned and written solely by me. Chapter 3 (methods) was written with input and advice from the relevant laboratories and, in the case of the profile regression section, from my supervisor, Dr. Michail Papathomas.

In chapter 4 (Profile Regression as a Tool for Identifying susceptibility Biomarkers from Genetic Profiles) genotyping of the EPIC cases and controls was conducted by the National Cancer Institute's Cancer Genomics Research Laboratory, Bethesda, Maryland, USA, as part of the InterLymph Consortium's NHL GWAS. The quality control steps as applied to the EPIC cases, and all presented analyses were planned and conducted solely by me.

The body burden measurements and the metabolomics profiling presented in chapters 5, 6 and 8 were conducted as part of the EnviroGenoMarkers Project. The metals measurements, PCB/POPs measurements and the metabolomic profiling were performed in the laboratories of Thomas Lund (Department of Occupational and Environmental Medicine, Lund University Hospital, Sweden), Hannu Kiviranta, (Chemical Exposure Unit, National Institute for Health and Welfare, Kuopio, Finland) and Hector Keun (Department of Surgery and Cancer, Imperial College London) respectively. All presented statistical analyses were planned and conducted solely by me.

In chapter 7 screening for the t(14;18) translocation was conducted on cases and controls from EPIC in the laboratory of Bertrand Nadal (CNRS-INSERM, Université de la Méditerranée in Marseille, France). I was involved in the design, planning and execution of this project together with Bertrand

Nadal, Sandrine Roulland and my supervisor Paolo Vineis. All presented statistical analyses were conducted by me.

COPYRIGHT DECLARATION

ACKNOWLEDGEMENTS

I would firstly like to thank my supervisors Paolo Vineis, Marc Chadeau-Hyam and Michail Papathomas, for their invaluable help, guidance and support both in the writing of this thesis and throughout my time at Imperial College. I couldn't have asked for better supervisors and I am very grateful for everything they have done for me.

I would like to thank all of those involved in the EnviroGenoMarkers Project, in particular Soterios Kyrtopoulos, Thomas Lundh, Hannu Kiviranta, Hector Keun, Alexis Siskos, Anas Kamleh and Miquel Porta who have been incredibly helpful with their time and expertise. I am also very grateful to all those involved in both the EPIC-NHL and Interlymph GWAS projects, as well as to those from the Northern Sweden Health and Disease study who have assisted me on multiple occasions. I would like to thank Bertrand Nadal and Sandrine Roulland for involving me in their very exciting project, and for all their help and advice which has been instrumental in a proportion of this thesis.

I am grateful to the ESRC for funding my PhD and to everyone at Imperial College who has helped me during the three years of my studentship.

I would also like to thank all my friends at Imperial College, particularly Danielle and Karin who made the last three years so fun! As well as Harrietta, Alina, Mai, Pauline, Marc, everyone in room 171 and the DIDE crew. I also couldn't have done it without the support of all my other friends, especially the E-girls: Gem, Lex, Franks, Hols, Kaffy, Gosia and Faye. The MMT girls: Sarah, Vix, Debs and assorted antipodeans. The Warwick girls: Jessie B, Jess, Bry, Lisa and Em. All the netball girls, and everyone else who has been so wonderful, especially my fabulous housemate Lydia, Katie D, Leah, Kim and Alex.

Finally I want to thank my amazing family, my Mum and Dad who have supported me all the way. And my brother and sister, Paul and Helen for the invention of whiskey Sunday and for many 'educational' trips to Oxford.

Thank you.

CONTENTS

List of Figures

List of Tables

Definitions

2,4-D	2,4-Dichlorophenoxyacetic Acid
95% CI	95% Confidence Interval
AHR	Aryl Hydrocarbon Receptor
AIDS	Acquired Immune Deficiency Syndrome
AUC	Area Under the Curve
BDE-47	2,2',4,4'-Tetrabromodiphenyl Ether
BMI	Body Mass Index
Cd	Cadmium
CV	Coefficient of Variance
CYP	Cytochrome P-450 gene family
DDE	Dichlorodiphenyldichloroethylene
DDT	Dichlorodiphenyltrichloroethane
DNA	Deoxyribonucleic Acid
EBV	Epstein Barr Virus
EGM	EnviroGenoMarkers
EPIC	European Prospective Investigation into Cancer
FDR	False Discovery Rate
FFPE	Formalin-fixed Paraffin Embedded
GC-MS	Gas Chromatography-Mass Spectrometry
GST	Glutathione-S-Transferase gene family
GWAS	Genome Wide Association Study
GxE	Gene-Environment Interaction
GxG	Gene-Gene interaction
HCB	Hexachlorobenzene
HL	Hodgkin Lymphoma
HWE	Hardy-Weinberg Equilibrium
IARC	International Agency for Research on Cancer
IBD	Identity By Descent
ICD-0-3	International Classification of Diseases for Oncology - version 3
Ig	Immunoglobulin
IL	Interleukin
LOQ	Limit of Quantification
LCA	Latent Class Analysis
LD	Linkage Disequilibrium
m/z	Mass to Charge Ratio
MAF	Minor Allele Frequency
MBL	Monoclonal B-cell Lymphocytosis
MCMC	Markov Chain Monte Carlo
MDR	Multifactor Dimensionality Reduction
MDS	MultiDimensional Scaling
MGUS	Monoclonal Gammopathy of Undetermined Significance
MITM	Meet In The Middle
NAT	N-AcetylTransferase gene family
NCI-SEER	National Cancer Institute-Surveillance Epidemiology and End-Results study
NHL	Non-Hodgkin's Lymphoma
NMR	Nuclear Magnetic Resonance Spectroscopy
NPV	Negative Predictive Value
NSAID	Non-Steroidal Anti Inflammatory Drugs
NSHDS	Northern Sweden Health and Disease Study
OPLS-DA	Orthogonal Partial Least Squares Discriminant Analysis
OR	Odds Ratio
Pb	Lead

PCA	Principal Components Analysis
PCBs	Polychlorinated Biphenyls
PCR	Polymerase Chain Reaction
PLS	Partial Least Squares
POP	Persistent Organic Pollutant
PPV	Positive Predictive Value
PRoBE	Prospective-Specimen-Collection-Blinded-Evaluation
QC	Quality Control
REMARK	REporting recommendations for tumor MARKer prognostic studies
RNA	Ribonucleic Acid
ROC curve	Receiver Operator Characteristic Curve
rt	Retention Time
SCALE	Scandinavian Lymphoma Aetiology Study
SNP	Single Nucleotide Polymorphism
STROBE-ME	STrengthening the Reporting of OBservational studies in Epidemiology: Molecular Epidemiology
TCDD	2,3,7,8-Tetrachlorodibenzodioxin
TCE	Trichloroethylene
ttd	Time to Diagnosis
UPLC-MS	Ultra-high Performance Liquid Chromatography Mass Spectrometry
UV	Ultraviolet
VIP	Variable Importance in the Projection

B-cell NHL Subtypes

MALT lymphoma	Mucosa-Associated Lymphoid Tissue Lymphoma
DLBCL	Diffuse Large B-cell Lymphoma
BCLL	B-cell Chronic Lymphocytic Leukaemia
FL	Follicular Lymphoma
MM	Multiple Myeloma
MCL	Mantle Cell Lymphoma
LPL	Lymphoplasmacytic Lymphoma
MZL	Marginal Zone B-cell Lymphoma
HCL	Hairy Cell Leukaemia
BALL	B-cell Acute Lymphatic Leukemia
SLL	Small Lymphocytic Lymphoma
BNOS	B-cell NHL Not Otherwise Specified

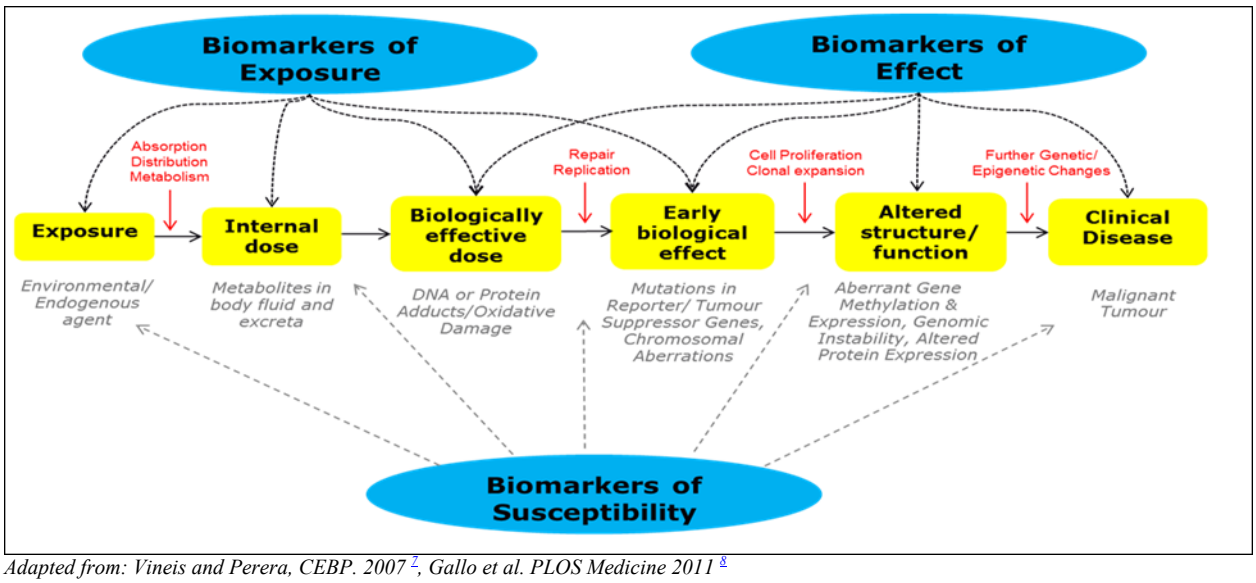
BACKGROUND, LITERATURE, AND METHODS

1 INTRODUCTION

Non-Hodgkin Lymphomas (NHL) represent the eighth most common cancer in Western Europe [1]. Yet despite their widespread prevalence and high mortality rate relatively little is known about the aetiology of these hematological malignancies, and they remain largely incurable [2]. The incidence of NHL, particularly the more highly aggressive subtypes, has been reported to be on the increase in the UK and most other Western countries since the 1980s and although this observed increase has recently begun to plateau, the evidence suggests worldwide incidence and mortality rates are still rising [3-5]. The causes of this increase have yet to be identified and are thought to be larger than can be explained by the diagnostic improvements and changes to the classification procedures which have occurred within the same time period [4,6]. Few definitive risk factors have been identified and NHL remains a poorly understood malignancy for which the elucidation of its aetiology is further complicated by the existence of multiple heterogeneous subtypes. These factors combine to make NHL a prime candidate for biomarker discovery.

Biomarkers are indicators in a biological system or sample, which include any objectively measureable characteristic that provides information on a physiological or pathological state. There are three major types of biomarkers (Figure 1): biomarkers of susceptibility, biomarkers of exposure and biomarkers of effect.

Figure 1: Schematic describing the different types of biomarker categories commonly employed in molecular epidemiology, their uses and specific examples



Adapted from: Vineis and Perera, CEBP. 2007 [7], Gallo et al. PLOS Medicine 2011 [8]

Biomarkers of susceptibility include multiple categories and are used to identify 'at risk' individuals where risk is either acquired (as measured by markers of previous disorders) or more usually, inherited (as measured by genetic markers) [8]. Biomarkers of exposure are used to capture the initiating events in the causal pathway and therefore are most commonly used to identify risk factors and to understand the casual mechanisms by which they may cause disease [9]. Finally, biomarkers of effect reflect the interaction between the risk factor and the individual [8] . Consequently they tend to be used in screening, diagnosis and disease progression monitoring [9].

This thesis attempts to address some of the issues that have previously hampered the study of NHL through the use of novel strategies of biomarker identification, utilising a nested case-control design within the European Prospective investigation into Cancer (EPIC) cohort. The thesis will be split into two parts: the 'validation of biomarkers' and the 'integration of biomarkers'. In part one various datasets within the EPIC-Lymphoma subcohort will be utilised in order to define potential biomarkers of susceptibility, of internal dose, of biologically effective dose and of early biological effect (figure 1). The clinical utility of these biomarkers will be assessed by validating their ability to predict future risk of NHL. Then, to provide a more global overview of the disease process, in the second part of the thesis two biomarker categories will be integrated in order to explore the extra information and the improved utility that can be obtained by considering them in combination. Together these analyses will provide an increased understanding of the mechanistic pathways involved in lymphomagenesis and the aetiology of NHL. In addition it will provide an overview of the current role of biomarkers in NHL, as well as possible research avenues for the future.

1.1 Histology

NHL comprises a heterogeneous collection of lymphoproliferative B or T-cell malignancies that typically present as a solid tumour of lymphoid cells [10]. The cells differentiate from hematopoietic stem cells within the bone marrow and mature into high affinity antibody producing cells through a number of stages involving genetic recombination and mutation. This inherent capacity to generate specific antibody-mediated immunity through immunoglobulin gene rearrangement leaves the cell open to mutations if the rearrangement goes awry [11]. The progressive accumulation of particular genetic lesions and translocations can deregulate certain oncogenes and disrupt homeostasis leading to proliferation, blocked differentiation and immortalisation, creating malignant T- or B-cells capable of clonal expansion. Further modulation by environmental, epigenetic, pathogenic or genetic factors may then lead to the development and progression of lymphoma [3,12]. Non-Hodgkin's lymphomas are distinguished from Hodgkin's lymphomas (HL) by the type of lymphocyte from which they arise; with Reed/Sternberg cells being the hallmark of HL [13].

The majority of NHL develop in lymph nodes, with primary extranodal disease, most commonly in the stomach, intestine or skin, accounting for only 20–30% of cases [4]. Lymphomas can be subdivided into two prognostic groups: indolent (slow growing) or aggressive (fast growing) which although they have a shorter natural history are frequently curable, unlike the indolent tumours which have a long median survival, but tend to be ultimately fatal [14].

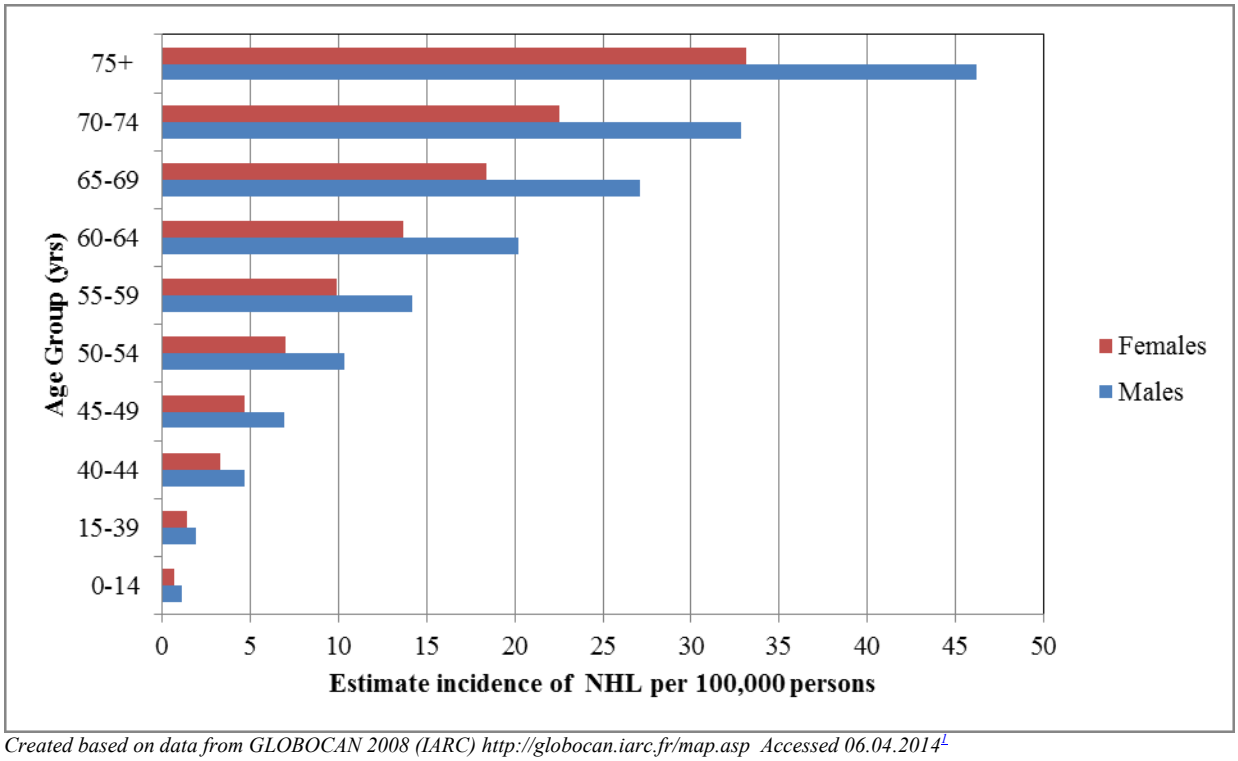
NHL was first evaluated as a separate entity in the late 1940s, and by the 1960s was characterized as either lymphosarcoma or reticulosarcoma [6]. Medical advances and increased understanding of the immune system, the bone marrow and of the genetic and cellular basis of malignant transformation provided the ability to subdivide these malignancies further [15]. A number of differing classification systems including the Rappaport formulation, the Working Formulation, the Kiel classification and the Revised European-American Lymphoma (REAL) classifications were subsequently introduced. These were superseded by the WHO classification of haematological malignancies which were most recently updated in 2008 [16]. These criteria marked a substantial change in the classification of haematological malignancies, representing an international consensus for the first time and addressing the issue of the lack of compatibility between the different classification systems that had been used previously [15]. The WHO classification refined and extended the definitions of existing haematological malignancies, incorporated clinical characteristics into disease criteria and identified new malignant conditions, based on biologically sound underlying principles to ensure clinical relevance, practicality and reproducibility [17,18]. Currently more than 50 different subtypes of B and T-cell NHL have been defined, differing in both clinicopathological and biological characteristics [18] and potentially in aetiology.

Recently Morton *et al.* [19] proposed a classification system for the epidemiologic analysis of lymphoma subtypes based on a hierarchical system which groups subtypes according to pathologic features including morphology, immunophenotype, genotype and stage of differentiation, and clinical features such as the site of occurrence. This was developed in order to address the challenges inherent in the epidemiological analysis of lymphoma. Evidence suggests etiological heterogeneity between the subtypes, but the various clinical and pathological schemes used over time and across the world complicate the study of individual subtypes. Standardising the groupings for epidemiologic research facilitates the comparison of subtype-specific literature, the harmonisation of cases diagnosed using differing systems and the pooling of cases for meta-analyses. All of these effects will further the elucidation of aetiology. The system, based on the WHO classification of lymphoid neoplasms [20] and the International Classification of Diseases – Oncology, Third edition (ICD-O03) [21], suggests that analyses are focused on the highest hierarchical group possible within the limit of the sample size of that study. This thesis will be focussed on hierarchical group four: peripheral B-cell neoplasms (from here on NHL will be taken to refer only to B-Cell NHL). The Morton *et al.* classification system was chosen for this thesis first to reflect the diversity of subtypes included, some of which pre-date the WHO 2001 classification. Second, within EPIC there is insufficient power to investigate each sample individually due to sample size. The Morton *et al.* system enables the grouping of pathologically and clinically similar neoplasms within EPIC, maximising the power and the likelihood of producing biologically relevant findings in this thesis, as well as thirdly allowing comparisons with the existing literature.

1.2 Incidence

The worldwide age-standardised incidence of NHL as a group (ICD-10 codes C82-85, C96) was estimated to be 5.1 cases per 100,000 persons in 2008. Risk increases with age from 1.7 cases per 100,000 in those aged 15-39 years to 28.1 in the 70-74 years age group and 40.2 in those aged 75 years and older [22]. The average age at diagnosis is between 60 and 70 years, although this differs markedly by subtype [2,15]. Precursor T- and B-cell malignancies are primarily diseases of children and young adults, while in adults most hematological malignancies arise from mature immunocompeptant B-cells. This is likely explained by the transition from an immune system rich in precursor cells in the young to one increasingly dominated by germinal centre and memory B-cells with increasing age. Overall incidence rates are roughly 1.5 times higher in males than females [15] (Figure 1.2a), with an age standardised rate of 6.0 cases per 100,000 compared to 4.2 in females. This gender-disparity has been hypothesised to relate to gender-specific differences in immune system regulation [15]. It has been observed to differ by subtype [15], but overall to increase with age, and it is particularly apparent over the age of 50 years [23].

Figure 1.2a: Estimated Incidence of NHL by sex and age group, per 100,000 persons



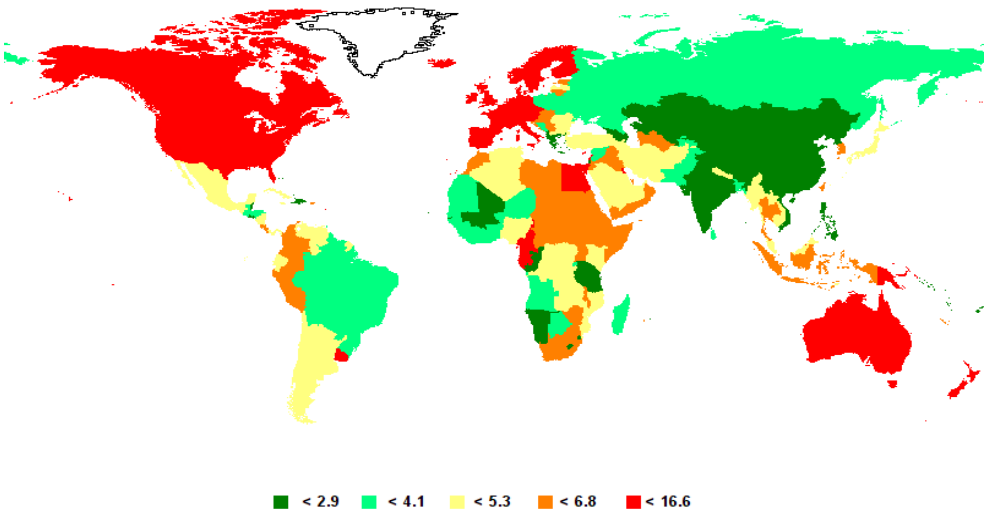
Created based on data from GLOBOCAN 2008 (IARC) http://globocan.iarc.fr/map.asp Accessed 06.04.2014[1]

NHL International Classification of Disease Codes (ICD) C82-85, C96. Includes HIV disease resulting in malignant neoplasms

Geography

In common with many cancers, overall incidence rates for NHL are higher in the Western world, namely North America, Europe and Oceania (Figure 1.2b). Again subtype plays a role, in particular diffuse and follicular lymphomas are very common in the Western world but rare in developing countries and the far-east, while Endemic Burkitts' lymphoma, which is related to infection with the Epstein Barr virus (EBV), occurs almost exclusively in Africa [1,4,24,25]. The extent to which these geographical differences represent true effects as oppose to a reflection of the better awareness and reporting of NHL in Western nations remains to be established. This is of particular interest as the burden of HIV/AIDs in much of the developing world would be expected to have a strong impact on prevalence. A proportion of the difference is likely to be attributable to the quality of the cancer statistics available from resource poor settings, which may not be entirely comparable to those from more developed nations. Although reporting can be problematic for all malignancies, it is particularly difficult for lymphomas due to the evolving classification, the poor standardisation of data, and the fact that unlike many other cancers, haematological neoplasms are diagnosed using multiple parameters. These include a combination of histology, cytology, immunophenotyping, cytogenetics, imaging and clinical data in addition to clinical and biological prognostic factors. This information can be difficult to access systematically and in enough detail to be able to implement the WHO classification [15]. These factors will undoubtedly impact the variation in global incidence, nevertheless given the strong association between age and risk, an even larger percentage is likely to be attributable to the differing age distributions between the developed and developing world, and the competing risks in developing nations.

Figure 1.2b: Estimated global age-standardised incidence rate of NHL per 100,000 persons: Males and females, all ages



GLOBOCAN 2008 (IARC) http://globocan.iarc.fr/map.asp Accessed 13.10.2013[1]

NHL International Classification of Disease Codes (ICD) C82-85, C96. Includes HIV disease resulting in malignant neoplasms

Incidence increase

The increase in incidence of NHL since the 1980s is well referenced in the literature [2] and has been observed to be particularly marked in developed nations. It has been reported in both males and females, across most ethnic and age groups and in nearly all subtypes, although it has been notably larger for aggressive NHL [4,26]. The rise has now begun to plateau, however the reasons for the changing trends remain unknown [3,27]. Interestingly, the incidence trend of NHL is not mirrored by that for HL or Leukaemia, which have remained relatively stable during the same time period despite the fact the three malignancies are believed to share a number of risk factors [28].

A large body of research has been dedicated to the investigation of the observed increase and to determining whether it represents a true effect or merely a data artefact. Changes in classification methods, advances in diagnostic technology and improvements in cancer registrations have all been postulated to play a role [5,29]. The impact of the WHO classification on incidence trends and disease burden should also not be underestimated. These criteria resulted in the inclusion in incidence data of previously unrecognized lymphoma types, such as mantle cell lymphoma, MALT lymphoma, and various T-cell lymphomas. Further, refinements in the histomorphological understanding of HL led to a large number of HL cases being reclassified as NHL [15].

Overall the weight of the evidence suggests that although such factors have inflated the incidence of NHL, the magnitude of its change is too large to be explained by data artefact alone [5,30]. This is supported by the fact that an increase was been reported in a variety of diverse nations with different diagnostic and reporting procedures [5,30]. In 1992 an international workshop concluded NHL constituted an "emerging epidemic" [31] implicating a pivotal role for an increase in causal environmental, lifestyle or other risk factors [4,5,26,32,33]. The subsequent levelling off of this increase has been attributed to cancer preventative measures and to changes in the prevalence of putative risk factors, such as restrictions in the use of certain chemicals, particularly in Western countries [34].

1.3 Aetiology

The specific mechanisms for the development of NHL have yet to be determined, however chronic antigenic stimulation, the induction of an immunosuppressive state or the disruption of normal cell proliferation are likely candidates [3]. It is similarly unknown what co-factors may interact with the normal regulation of lymphocytes to precipitate these malignant pathways, however the global rise in the incidence of NHL suggests that exposure to, or the prevalence of such factors may also be on the increase [33]. Genetic, viral, environmental and biological factors are all likely to play a role.

1.3.1 Immune dysfunction

Immunosuppression

NHL is a malignancy of the immune system and the most consistently and strongly associated risk factor is congenital or acquired immunodeficiency [35,36]. This is supported by the elevated incidence rates in immunosuppressed populations [2]. A quarter of patients with congenital immunodeficiency will develop a tumour at some point during their lifetime, of which NHL comprise about 50%. AIDS patients (excluding those in Africa) show an up to 100-fold higher risk of NHL than the general population [4]. The majority of AIDS-related NHL are high-grade tumors involving extra nodal sites, some of which are specific to HIV/AIDs patients and other which also manifest in HIV negative patients [37]. In fact, the onset of the AIDs epidemic in the 1980s was one of the postulated reasons for the observed increase in NHL incidence, however accumulated evidence suggests this is unlikely to the be case [32]. The immunosuppressive state is a result of the down regulation of the T-cell response to compensate for chronic antigenic stimulation [3]. This characteristic antigenic stimulation which induces immunosuppression also increases B-cell proliferation and consequently the probability of a random genetic mistake. Such mistakes are particularly prevalent in immunoglobulin gene rearrangement during lymphopoiesis [38], thereby increasing the risk of lymphomas. Risk increases with the degree of immunodeficiency and there is no evidence of a threshold [39].

The use of immunosuppresive drugs following organ transplantation or for the treatment of autoimmune disorders has also been associated with a dose-dependent increase in risk [4]. However it appears to be only a temporary effect with a high rate of tumour regression observed following the withdrawal of the drugs [3].

Autoimmunity

Autoimmune disorders themselves, which manifest in an imbalance of the immune system towards the T-helper 1-cell mediated response [40], have also been frequently associated with NHL risk independently of treatment. In particular Rheumatoid arthritis, systemic lupus erythematosus, Coeliac disease and Sjögren's syndrome have been implicated [40]. Some estimates suggest the risk of NHL

may be increased by up to 44 times in people with Sjögren's syndrome compared to the normal population [4,41]. The effect of autoimmunity appears to impact multiple subtypes, in particular DLBCL, FL and MZL, as well as NHL overall [32], and to be stronger in females than in males [26].

Allergy and Atopy

In comparison, the relationship between NHL and atopy is less well understood, and the literature less consistent [35,42] with positive [43,44], null [35,45] and negative associations [46,47] reported. Two conflicting hypotheses have been put forward to support the contradictory findings. The 'immune surveillance hypothesis' proposes that the atopic conditions may be protective against NHL due to the dominance of the humoral T-helper 2 response which heightens response to foreign antigens and thereby enhances the ability of the immune system to detect and eliminate malignant cells [46,48,49]. Conversely, ' The Antigenic stimulation hypothesis' suggests that atopy may in fact lead to an increased risk of malignancy [48] by creating a subchronic state of inflammation [42] which results in a higher possibility of random mutations occurring in dividing cells [35]. Epidemiological evidence relating to number of siblings, birth order and household crowding is inconsistent, and again appears to differ by subtype [35,50-52]. The relationship is further complicated by that fact that NHL itself has been associated with a reduction in atopic symptoms and in low levels of immunoglobulin subclasses including IgE [53]. Therefore the 'protective' effect, which has mainly been observed in case-control studies and only rarely in cohorts [35], may in fact be a result of reverse causation. Contrary to this, decreased levels of plasma cytokines such as Interleukin-2 and Tumor necrosis factor-α and down regulation of T helper-1 cytokines have been observed to be associated with future risk in prospective studies [27].

Infectious agents

Finally in addition to HIV which induces an immunosuppressive state though the depletion of CD4+ [37], a number of other infectious agents have also been implicated in risk [54]. This is thought to be mediated through their induction of cell proliferation, inflammatory cytokine release, stroma activation, and DNA damage [3,4]. Certain viruses are known to be able to directly transform lymphocytes. The most well-characterised is the relationship between EBV and Burkitts Lymphoma. Burkitt's tumour cells almost universally display a t(8;14) translocation which activates c-*myc* a potent oncogene [54]. EBV is also thought to be associated with NHL through immunosuppression and the resulting proliferation of transformed B cells normally controlled by T-cell-mediated immunity. Roughly half of HIV-associated DLBCL cases are EBV positive. EBV has further been hypothesized to increase risk through infectious mononucleosis, which has also been associated with NHL and of which EBV is a major cause [32,55].

More recently additional viruses and pathogens have been implicated in a number of NHL subtypes. HTLV-1 has been associated with the risk of adult T-cell Leukemia/lymphoma in a Japanese population [56], and H.pylori has reported to be causal for gastric MALT lymphoma due to chronic

inflammation [57]. Hepatitis B and C infection, herpes zoster, *Campylobacter jejuni*, *Chlamydia psittaci* and *Borrelia burgdorferi* have all also been implicated, although the evidence is inconsistent [32,5058]. Further the high prevalence of some of these infections relative to the incidence of NHL suggests they are interacting cofactors, and alone do not represent a sufficient cause [4]. In general the explorations of previous infection as a risk for future NHL is limited by the self-reported nature of the data and problems of recall bias, particularly among adults interviewed about their health in childhood, few of the existing studies use medical records for confirmation and for many infections and conditions differing definitions are used between different studies [55].

In summary, the overwhelming weight of the evidence supports immune dysfunction as the underlying basis of lymphomagenesis[27,59]. Nevertheless, although individuals with immunosuppressive and autoimmune disorders display a large increase in risk of NHL, in fact they account for only a small proportion of cases [2] . For the majority of cases, although pathogenesis may act through pathways of the immune system, immune dysfunction itself is not the cause.

1.3.2 Occupational Exposures

Epidemiological studies have noted some apparent associations between certain occupations and NHL. Meta-analyses and systematic reviews have most consistently found farmers, teachers, dry cleaners, butchers, car repairmen and workers in the wood and printing industries to be at an increased risk [60,61]. However to date, no workplace exposures have been conclusively identified as risk factors for NHL, and occupation is unlikely to represent a major risk factor in most populations [62].

For some of the implicated occupations, in particular the 'white collar' occupations, there is no obvious causal link with NHL. However, for others there is a biological rationale; the association with teaching, animal husbandry and farming is supportive of a viral or bacterial aetiology of NHL [32,60,61], while the relationship with farmers has also been used as evidence for a link with UV radiation exposure due the large amount of time farmers spend outdoors [3]. In fact, for the vast majority of observed associations the relationship with NHL is likely to be linked to exposure to the specific pesticides, solvents, oils, organic and inorganic dusts, aromatic amines, paints and varnishes associated with the particular occupation of interest [3,4,32,60,61].

1.3.3 Environmental pollutants

Despite evidence from occupational studies and plausible mechanisms, including genetic mutations and immunosuppressive effects [3], the role of environmental pollutants in NHL remains uncertain. Studies tend to be marked by small study populations with poorly characterized exposures leading to limited statistical power, poor quality of exposure measurements, minimal efforts to control for confounding, and lack of information on dose-response and temporality [3]. Occupational studies have

often been conducted in extremely high exposure settings, such as in the aftermath of an industrial disaster, and so are difficult to interpret in terms of actual population risk [63].

The Seveso disaster was an industrial accident in a chemical manufacturing plant in the Lombardy region of Italy. It resulted in the highest ever known human exposure to 2,3,7,8-Tetrachlorodibenzodioxin (TCDD) and has formed the basis of the evidence linking dioxins to NHL risk. Follow-up of the exposed population noted an almost doubling of the risk of lymphohemopoietic neoplasms [64], and these findings are supported by similar study of an industrial accident in a Dutch cohort [65] and by evidence from animal models [66]. Dioxin toxicity is mediated through binding to the aryl hydrocarbon receptor (AHR), a ligand-dependent transactivating factor which is expressed by major cell types of the immune system [67,68]. TCDD has been shown to target cells in the immune-haematopoietic system and consequently the observed association with NHL is thought to be mediated through its immunosuppressive effects [67,69].

A vast number of other chemicals with either known effects on the immune system or associations with other haematological malignancies, such as leukaemia, have similarly been explored. The dioxin-like properties of certain congeners is also, in part, the rationale behind the exploration of the polychlorinated biphenyls (PCBs) [70-72] which are mixture of synthetic chlorinated hydrocarbons produced directly or indirectly by multiple industrial and agricultural processes. Agricultural exposure to pesticides has been suggested as a further explanation for the increased risk of NHL observed in farmers. Of the most commonly used pesticides phenoxy herbicides and the herbicide 2,4-Dichlorophenoxyacetic acid (2,4-D) have been associated with the greatest risk however the lack of evidence for a dose–response relationship has raised doubts about the significance of these findings [3]. Multiple other pesticides which have demonstrated tumour initiating and/or promoting effects in animals have been studied [67,73-75], however data is scant and, like the study of PCBs, subject to problems of exposure measurements and classification, as exposure is frequently based on self-report or the use of job exposure matrices [3,4,32]. Duration of follow-up could also be an issue for epidemiological studies of NHL given that many subtypes represent indolent disease. Therefore although associations cannot be ruled out, to date neither PCBs, pesticides as a general group, nor specific chemicals have been defined as definitive risk factors [32,73,76-86]. Some of the strongest data relates to solvent exposure, in particular Benzene and Trichloroethylene (TCE). Yet despite a large number of studies the evidence is largely inconsistent, a causal association is yet to be established, and if it does exist appears to be restricted to FL and CLL [87-90].

1.3.4 Other environmental, lifestyle and anthropometric risk factors

A number of other environmental and lifestyle variables have been suggested as NHL risk factors. A history of blood transfusion has been associated with an up to four-fold increase in risk of NHL, which may be due to oncogenic virus transmission, transfusion-induced immunosuppression or

engraftment of malignant lymphoma cells from the donor [3,91]. A number of medications including statins, NSAIDs, antibiotics and antidepressants have been implicated by some studies [32], although it is difficult to separate their effects from those of the underlying condition they were prescribed for, particularly for antibiotics. Postmenopausal hormone therapy and an increasing number of pregnancies has been associated with a decreased risk, while long term (>5 years) use of hormonal contraception has been observed to increase risk [92-94]. In both instances these effects appeared to be specific to FL and DLBCL. The known carcinogenic effects of smoking, coupled with the evidence that tobacco can alter the immune response and that cigarettes contain substances known to be leukemogenic has led to smoking being suggested as a risk factor. The evidence supports an association but, again, suggests it is restricted to FL [95].The use of hair dyes, particularly before 1980 has been observed to increase the risk of FL and CLL/SLL [96]. Exposure to ionizing radiation, therapeutic radiation and ultraviolet radiation/sunlight [97,98] have all also been widely studied. However the data remain inconclusive [32].

Anthropometry

The global rise in obesity has been postulated as an explanation for the reported increase in NHL [99], given that obesity is also known to cause chronic low-grade inflammation and specific immune alterations [32]. Obesity has previously been associated with a number of other malignancies however it appears that the relationship with NHL, if real, may be restricted to DLBCL [100]. Height has been associated with overall NHL risk, but only in females [101].

Dietary factors

A high intake of fat, particularly animal and saturated fat, and of red meat has been suggested to increase risk of NHL, although previous studies have found it hard to disentangle the dietary effects from their strong correlation with obesity [99,102,103]. Conversely an inverse association with vegetables has been reported in a number of studies. This may be related to the high content of isothiocynates, which are hypothesised to induce apoptosis and growth arrest in preneoplastic and neoplastic cells [99]. Evidence is also suggestive but insufficient for a protective effect of wholegrains, fish, fruits, antioxidants and one-carbon nutrients [99,102]. Vitamin C intake has been reported to be inversely associated with FL risk [104]. Moderate physical activity is thought to similarly be associated with a reduced risk [99], and again it is hard to establish whether one, or all, of these associations are causal or whether they are merely proxies for a healthier lifestyle and therefore a lower risk of malignancy.

Perhaps at odds with the healthy lifestyle argument is the recent meta-analysis which provides quantitative evidence for a protective role of alcohol on NHL risk in both sexes and across all subtypes; the risk in alcohol drinkers is reduced by 15% compared to non-drinkers [105]. It has been suggested that alcohol may act through its immunomodulatory effect, the improvement of insulin sensitivity or via the antioxidants present in most alcoholic beverages. However at present such

mechanisms are merely postulations and there is a lack of a biological rationale for an association, as well as a risk of reverse causation if the onset of disease symptoms is associated with an increase in abstinence [105]. In general diet is a complicated risk factor to study given such possible confounding and the fluctuating nature of dietary patterns over time, and further studies are needed to investigate putative associations with NHL.

1.3.5 Genetics of NHL

A family history of any hematopoietic malignancy has been found to be associated with a 2-4 fold increased risk of NHL [106] suggesting a heritable component to risk. Aggregation of NHL in certain families has been observed, although this is rare [107]. In general, familial lymphomas account for around only 5% of all cases [3] and it is unclear to what extent such clustering is due to an increased awareness of the symptoms of this malignancy or to shared environmental factors. There does appear to be some non-uniformity in heritability by subtype, sex and the relationship to the affected relative. Siblings have been shown to confer the greatest risk, but overall the pattern of heritability remains poorly understood [106]. The primary congenital immunodeficiency disorders which are strongly associated with NHL such as Ataxia telangiectasia and Wiscott–Aldridge syndrome also impart a heritable component of NHL risk, although again such disorders are rare [3].

The role of a genetic component in risk has been supported by evidence from candidate gene studies considering germline mutations [108,109], and more recently by a number of genome-wide studies of NHL and its subtypes [110,111]. Due to the importance of immune dysfunction in aetiology, some of the most commonly explored pathways are those involved in the immune and inflammation response. Within these pathways studies have identified variants in genes which appear to be associated with NHL as a whole including toll-like receptor genes [112], *TRAF1, RIPK3, BAT2, MAP3K5, DUSP2, CREB1, B3GNT3, SELPLG, LSP1, FGG, ITGB3* [113], as well as those associated with particular subtypes. Genes involved in the TH1/Th2 balance have been shown to affect overall risk of NHL [114], but the findings are particularly strong for DLBCL with specific variants in *TNF* and *IL10* associated with a doubling of risk [115]. Similarly genetic variation in the major histocompatibility complex, which is within the 6p21.3 region, although found to be associated with all subtypes [116], has been observed to be of particular importance in FL [110,117], DLBCL [118] and CLL [119] susceptibility. Risk of CLL has additionally found to be associated with a further gene involved in the immune response, the transcription factor *IRF4* [120].

Variants in the one-carbon metabolism pathways genes such as *TYMS, MTHFR, MTR, BHMT, CBS, FPGS, FTHFD,* and *SHMT1* have also been implicated, and it is thought this may be mediated through the role of this pathway on immune function [2,121]. Genes involved in the oxidative stress pathway [122], DNA repair [123], the xenobiotic metabolism pathway [124] and a number of pro-apoptotic genes such as *BCL2L11*[125] have been similarly associated with risk. More recently, a new region of

interest on chromosome 9 has been identified. This region includes *SYK*, which encodes a kinase and is known to be of importance in B-cell development [116] . Furthermore, emerging evidence suggests a potentially important role for interactions between certain genetic variants and lifestyle and environmental variables in risk [126]. A number of gene-environment interactions, involving both the aforementioned and additional pathways such as vitamin D and leptin, have been reported in the literature (*see chapter 2 for details*).

However, studies considering a role for constitutional genetics in the aetiology of NHL are hampered by a lack of reproducibility, and therefore it is difficult to determine causal associations. This failure to replicate findings between studies has been attributed to inherent technical variability, heterogeneity between the studied populations and population stratification, issues of multiple testing and limited power, particularly when considering subtypes [2]. An International consortium of 22 lymphoma studies known as InterLymph is now addressing this by conducting the largest subtype-specific GWAS of NHL to date. The first paper to emerge from this consortium GWAS considered CLL and was able confirm a number of previously reported associations, as well as identifying nine novel loci. These loci were in close proximity to genes involved in apoptosis [87], supporting their biological plausibility, and also therefore the power of the InterLymph GWAS. The findings for DLBCL, FL and MZL are due to be published soon. Nevertheless, these studies are still limited by their focus on individual markers [2]. The most likely genetic basis of NHL seems to consist of multiple common variants each with only incremental individual effects influencing risk in combination [12].

Further information on the pathology of NHL and therefore potentially greater understanding of genetic susceptibility may also be explored by considering the tumour genome and the temporal sequence of somatic mutations in lymphoma tumours. Such studies have revealed that tumour phenotypes are a result of multiple oncogenic abnormalities acting synergistically [109]. Chromosomal translocations, including t(2;5)(p23;q35), t(3;22)(q27;q11), t(11;14)(q13;q32) and t(14;18)(q32;q21), are a genetic hallmark of NHL and have been shown to play a particularly important role in the pathology of lymphomagenesis through the inappropriate expression of certain crucial genes at reciprocal breakpoints [2]. The disruption of chromatin biology has also been suggested to be an important factor based on the observation that genes with roles in histone modification are mutated in some tumours [127]. While mutations in genes that play a vital role in the survival of B-cell, such as *BCR* and its co-receptors, are thought to enable lymphoid tumor cells to avoid apoptosis and have been associated with disease prognosis [128].

Again, while some of these mutations appear to pertain to multiple subtypes, others are specific. This has been observed to be particularly true for the translocations such as t(14;18) which is strongly associated with FL (*see chapter 7 for further details*). In fact, some of the most interesting findings to emerge from the study of tumour biology are those revealing the extensive genetic heterogeneity

between the subtypes, particularly between the germinal centre and non-germinal centre derived malignancies. Even within the subtypes gene-expression profiling and next generation based sequencing technologies have identified further molecular subgroups [129]. For example DLBCL has been shown to be comprised of at least two molecularly and clinically distinct diseases: a germinal center B cell-like (GCB) and an activated B cell-like (ABC) form. These two disorders arise from B cells at different stages of differentiation and are associated with specific genetic alterations, differentially expressed regulatory factors, different molecular signaling pathways, and different clinical outcomes [23,109,130]. Gene expression signatures tend to reflect the intrinsic properties of tumours cells as well as the tumour microenvironment and as such have provided a number of other novel insights into the pathobiology of human lymphomas, including predictions of response to therapy and of survival [109]. Such insights inform on aetiology and provide promising new therapeutic molecular targets for NHL [131]. Consequently their findings should not be overlooked, rather they should be utilized to help to refine the search for underlying genetic susceptibility markers.

1.3.6 Summary

In summary despite multiple studies, causal associations between environmental, lifestyle, dietary or genetic factors and NHL remain elusive. This may imply that causal associations do not exist with the risk factors explored to date, or alternatively it may be that genuine causal relationships have been obscured by bias or confounding in the study designs utilised.

A particular problem relates to the difficulty of accurately assessing environmental exposures and lifestyle factors. In NHL, a large number of the positive associtions with suspected risk factors arise from case-control studies. These are commonly cited dspite the fact the fact that such findings are subject to the problems of reverse causation as well as selection and recall bias [71]. Reverse causation in the case-control setting is also particularly problematic for putative NHL risk factors such as PCBs, as treatment or the disease itself has been reported to affect metabolism and blood concentrations of exogenous exposures [82,132]. Case-control studies additionally allow no measure of temporality. This may be a particular problem in occupational studies, which form a sizeable majorotiy of NHL research, where long-term exposure may be difficult to objectively quantify and where workers may be more aware of the harzardous nature of their exposure environments. Results from cohort studies can be considered as more definitive proof, however even here duration of follow-up is likely to be an issue given that many subtypes represent indolent disease and may be not be symptomatic for many years after exposure.

As discussed below, the heterogeneity of NHL is also problematic when considering aetiology as rather than representing a single disease NHL encompasses a heterogeneous group of malignancies varying in histological characterization and clinical manifestation [20], which may not share the same

aetiological pathway. Correct case-subtype classification is vital, but often lacking, particularly among older studies, given the multiple classification systems employed over time. Furthermore a large number of the putative risk factors are not specific to NHL and have been linked with other malignancies such as dioxins and prostate cancer [133], hormonal and reproductive factors and breast cancer [134], smoking and lung cancer [135], red meat consumption and obesity with colorectal cancer [136]. Similarly the 'protective' effect observed for certain dietary factors may simply be reflective of a lower overall risk of malignancy. Even for some of the strongest findings such as benzene, often more consistent evidence exists for these exposures with other haematological malignancies [137]. Hair dye usage appears to be one of the few risk factors specific to NHL [138]. Similarly, the weight of the evidence suggests NHL may be the most common malignancy in immunocompromised individuals, but alone these factors account for only a small proportion of disease.

Consequently the identification of validated risk factors for NHL is challenging. In fact, it is likely that, in common with most complex diseases, the majority of cases of NHL have a multifactorial aetiology, with each variable conferring only a small increase in risk. The incidence trend of NHL over recent decades suggests that exogenous factors including lifestyle and environment, play a particularly important role [4], but are associated with only small effect sizes [32] making their accurate measurement even more vital. However in the context of large epidemiological studies this can be extremely difficult; there is often a lack of consistency between studies in terms of how putative risk factors are measured. Furthermore, rather than being measured directly they are often inferred from occupation or area of residence. Finally, many studies consider single risk factors and even when confounding variables are taken into account, such univariate exposure is unlikely to reflect real world experience. Consequently in order to identify the factors contributing to risk, it is vital to have improved way of measuring them in the context of large epidemiological studies. The use of validated quantitative biomarkers could have the multi-beneficial impact of standardising risk factor assessment, improving definition of its relationship with NHL, and providing insights into potential mechanistic pathways.

1.4 Subtype heterogeneity

According to the WHO classification, which represents the international standard for clinical practice and research there are at least 50 different subtypes of NHL [16]. Even within the subtypes gene-expression profiling has revealed further subdivision, particularly for DLBCL and FL [117]. These different subtypes differ in genetics, morphology, immunophenotype, molecular signalling pathways and clinical outcomes, particularly regarding survival [2,17]. The incidence of NHL differs by subtype; the two most common subtypes are diffuse large B-cell lymphoma (DLBCL) and Follicular lymphoma (FL) which account for roughly 40% and 20% of all cases, respectively.

The heterogeneous nature of NHL complicates the study of lymphomagenesis. Differences in aetiology have not, to date, been well characterised, with both common and subtype-specific risk factors reported, and it is still not clear whether the existing subtype definitions are optimal for determining aetiology [19]. Lymphoid neoplasm subtypes all arise from the malignant transformation of normal lymphoid cells at various stages of differentiation, however what causes the original transformations is largely unknown and it could well be that there are common underlying risk factors for all subtypes. Similarly common susceptibility loci pertaining to all subtypes have been identified, although it is recognised these are less common than subtype specific loci [139].

In fact, some of the strongest evidence for causal risk factors to date pertain to particular subtypes, rather than NHL as a whole, and an increasing number of studies are demonstrating the importance of subtype specific studies. A large study [52] specifically designed to explore aetiological heterogeneity observed that late birth order and a high BMI only increased the risk of DLBCL, while autoimmune conditions and meat consumption appeared only to effect MZL and CLL respectively. A number of factors, such as HIV and Hepatitis C. infection, were also observed to increase the risk of multiple subtypes, and there were also common protective factors including alcohol consumption, sunlight, and consumption of cruciferous vegetables. In fact, some subtypes were found to be more closely related than others in terms of risk factors. They hypothesised that this related to the cell of origin which are thought to be germinal center or post–germinal center B cells for FL, DLBCL and MZL, while the cells of origin for CLL/SLL are not thought to have undergone germinal center transit [52]. In particular DLBCL and MZL both shared a number of risk factors that act through the immune response, which is consistent with these subtypes arising in the phases of B-cell differentiation that are most dependent on immune stimulation.

These findings may in part explain the observed differences in incidence rates by gender, age, country and time period. Understanding patterns of etiologic commonality and heterogeneity for non-Hodgkin lymphomas will additionally help to illuminate lymphomagenesis [52]. Consequently subtype

heterogeneity, should be taken into account, as much as possible in all studies of NHL. Given the discussed difficulties with defining subtypes in certain settings, again easily measurable biomarkers, which can help to distinguish subtypes and subtypes within subtypes, may prove useful here.

1.5 Biomarkers of NHL

NHL is a prime candidate for biomarker discovery given its high incidence, subtype heterogeneity and poorly understood aetiology with a lack of validated risk factors. Those factors that have been identified have little, if any, practical ability to predict disease risk of progression [2]. A wealth of emerging data is slowly beginning the address these issues, and some important successes have been reported [26]. Nevertheless, the risk factors identified to date can explain only a proportion of the incidence and progression of NHL and particularly for the genetic factors, replication and validation of findings is lacking [2].

It has been stated that the incorporation of objective molecular markers into epidemiological research can help to standardise risk and prognostic factors. This allows the pooling, interpretation and standardization of findings between multiple disparate studies and populations, ultimately increasing the ability to determine true effects [23]. Furthermore as stated, the diagnosis and classification of NHL, particularly in resource poor settings, is difficult, due to the multiple indices employed. Here again biomarkers could prove useful by representing a quicker, easier, and therefore most likely cheaper means of diagnosis [140]. In terms of the study of NHL, the use of biomarkers as a surrogate endpoint could decrease the required sample size, which is particularly important for the rare subtypes, as well as potentially circumnavigating ethical issues related to measuring true endpoints [140]. Consequently, biomarkers have multiple purposes and can take many forms [7], most of which would be advantageous to the study, prevention and treatment of NHL.

The different subgroups of biomarkers can be categorised as outlined in figure 1. Although there is some overlap, in general biomarkers of internal dose, biologically effective dose, early biological effect and altered structural function, can be considered as a temporal sequence of events between a risk factor and a disease outcome, and therefore in theory biomarkers can be identified at any stage along this causal pathway (although not every risk factor/disease relationship may follow this pattern). Additionally biomarkers of susceptibility can act at any of these stages. A number of biomarkers falling under each of these subgroups have been studied with relation to NHL, however substantial challenges remain, as discussed below.

1.5.1 Biomarkers of susceptibility

Biomarkers of susceptibility provide an indication of an individual's underlying susceptibility to a disease or its risk factors that may be inherited or acquired [8,141]. Given the role of immunodeficiency in NHL aetiology, immune disorders have been widely studied but as discussed, due to the small

proportion of cases arising in such susceptible individuals, these are normally treated separately from immunocompetant individuals. What is less well understood are more subtle changes in the immune system such as infection with EBV which has been demonstrated to interact with environmental exposures including organochlorines to modify risk [83].

Inherited genetic variants in the germline which predispose to risk through the modification of certain pathways, are likely to account for a larger proportion of cases. As discussed in section 1.3.5, the most likely candidate pathways include immune and inflammation response [112-114], one-carbon metabolism [2,121], oxidative stress [122], DNA repair [123], and xenobiotic metabolism [124]. However, the associations identified to date represent low penetrance variants, and a number may not be causal as studies have tended to be limited by problems of population stratification, improper control selection, genotyping errors, population size and failure to consider sex-, subtype-, or stage- specific effects [12,87]. Furthermore the majority of associations remain to be replicated [2].

Nevertheless, there is a vital need to consider genetic biomarkers of susceptibility in studies of risk factor-disease relationships as intra-population differences in such variants may explain some of the inconsistent associations observed for NHL identified to date. A number of such studies have been performed (*see chapter 2*). However it seems that the candidate gene approach taken to such studies may have reached its limitations and new methodologies considering interactions between multiple genes are required.

1.5.2 Biomarkers of internal dose

Dose represents the amount of a contaminant actually deposited within the body and so biomarkers of dose bear a quantitative relationship to previous and potentially cumulative exposure. Internal dose of suspected risk factors, often referred to simply as body burden, has been widely studied due to the suspected role of environmental pollutants in the aetiology of NHL, and the relative ease of accurate measurement. Among these, dioxins and PCBs are some of the most commonly investigated, but the results have been contradictory. This has been attributed to the use of case-control studies and the resulting potential for reverse causation bias, which is particularly important for the exploration of NHL, as the treatment or the disease itself has been reported to affect metabolism and blood concentrations of exogenous exposures [77,132].

1.5.3 Biomarkers of biologically effective dose

The biologically effective dose represents the dose at the site of toxic action. Biomarkers of biologically effective dose typically encompass interactive products, or interactions that change the status of the target molecule [141]. The most commonly used measures include DNA adducts and oxidative damage [142]. Such biomarkers are vital as they can provide an indicator of the potency of the

causal risk factor, established using biomarkers of internal dose. However this is a relatively unexplored area in NHL research and to date only oestrogen-DNA adducts have been investigated, with higher levels being observed in male NHL cases compared to controls [143]. Consequently novel approaches are required to fully exploit the potential of these biomarkers.

1.5.4 Biomarkers of early biological effect

Biomarkers of early effect indicate the presence and strength of a biological response to exposure to an environmental agent on the subcellular level [141]. By far the most widely studied in NHL aetiology is the t(14;18)(q32;q21) translocation, which is found in 70-90% of FL cases, 20-30% of DLBCL cases and 5-10% of the less common subtypes [20,144,145]. This chromosomal aberration increases cell proliferation and effectively immortalises lymphocytes [66,146], however it is also present in up to 80% [147,148] of healthy individuals, albeit at much lower frequencies [149]. This renders presence of the translocation alone an insufficient biomarker. Nevertheless its strong association with NHL and FL in particular, suggests this translocation could still be of some utility if exploited in a different way.

A number of other early effect biomarkers have also been explored in relation to NHL aetiology, including replicative index and mitotic index as measures of cell proliferation, and micronucleus frequency as an indicator of chromosomal damage. This has been determined mainly through their association with putative risk factors such as 2,4-D and Carbofuran [150-152]. Similarly the hypothesis that TCE and TCDD may induce NHL by suppressing the immune system is based on reported decreases in C4+ levels and major lymphocyte subsets, particularly sCD30 and sCD27 [153]. Consequently, although they provide some mechanistic insights, they are not specific to NHL and, as such, again are of little clinical utility. The problem of the inconsistency of the evidence for the association between NHL and most postulated risk factors makes their utility even more challenging. Therefore methods which integrate the risk factor, the biomarker and the disease are required to determine whether these relationships follow a temporal pathway, and therefore if the biomarker can be of predictive use.

1.5.5 Biomarkers of altered structure or function

Biomarkers of altered structure and function are the final stage in the pathway before clinical manifestation of disease. They are used to indicate subclinical changes in morphology or function, most commonly altered gene expression or a decrease in organ function are used [141]. Immune system function can be considered a good proxy in the investigation of lymphoma and a number of exciting results have been reported in this context. A recent study found that circulating immune markers including BCA-1, sTNFR2, and sVEGFR2 are associated with NHL risk well in advance of diagnosis [154]. While prediagnostic levels of IFNγ, TNF-α, sICAM-1, IL2 [27], IL10, sCD27 and sCD30 [155,156] have also been observed to be upregulated in NHL patients.

Similarly, the study of altered gene expression is perhaps currently one of the most promising areas in NHL aetiology. Accordingly a number of studies have reported differences in both epigenetic regulation [157] and gene expression [158] in NHL cases compared to healthy controls, which could represent potential candidate biomarkers. However, the majority of these studies, both of immune function and of altered gene expression tend to be case-control in design which may lead to issues of reverse causation, particularly for late stage biomarkers.

1.5.6 Summary

As with the risk factors themselves, multiple biomarkers have been explored to try and better characterise potential causal relationships between suspected risk factors and disease. Although there are some promising candidates, the majority of the available studies lack sufficient evidence to definitively identify biomarkers. There has also been a lack of a concentrated effort to replicate the findings in independent populations, which is the gold standard for determining utility.

1.6 The future of NHL research

Together these findings support investing in the study of biomarkers. Given that so much remains unknown regarding the mechanisms and aetiology of NHL, the issue of subtype heterogeneity, the difficulty of classifying disease and of standardising its classification, and the problem of recruiting large enough sample sizes, biomarkers, which could help to address all these issues, are paramount to the future of NHL research. This requires novel approaches in terms of methods, technologies and analysis. It also suggests a new way of thinking is needed, which should consider biomarkers not as single discrete entities, but as stages on a continuum between a risk factor and disease. These can be integrated in order to assist in the elucidation of aetiology and of mechanisms, in disease prediction and in the identification of targets for treatment and ultimately for prevention.

In this thesis, this will be performed using the methodologies and concepts detailed below.

1.6.1 Gene-environment interactions

Although well-established, gene-environment interaction models remain a relatively unexplored area in NHL research (see chapter 2: literature review for details). Gene-environment interaction or GxE is defined as "a different effect of an environmental exposure on disease risk in persons with different genotypes," or, alternatively, "a different effect of a genotype on disease risk in persons with different environmental exposures" [159]. Current GxE studies in NHL have all employed a candidate-gene approach (Chapter 2), which does not take into account the fact that multiple variants (and multiple environmental factors) may be jointly contributing to disease risk. Additionally, by only exploring genes from pathways with biological *a priori* hypotheses, there is the risk of missing novel variants and pathways which may be pivotal to lymphomageneisis. Furthermore, considering Figure 1, any stage of the causal pathway can be modified by host susceptibility factors and so the concept of GxE can also be expanded to include gene-biomarker interactions. New methodologies to deal with high order interaction analyses combining multiple risk factors, multiple genetic variants and potentially multiple biomarkers, are therefore required.

1.6.2 The *'meet–in-the-middle'* approach

The numerous epidemiological studies of NHL to date provided little mechanistic evidence for the aetiology or pathways of lymphomagenesis. What is required is a more holistic approach combining both the risk factor and disease in order to visualise the causal pathway. The "*meet-in-the-middle*" (*MITM*) approach [7] combines a retrospective search for biomarkers in those who develop disease, with a prospective search for associations with these biomarkers to past environmental exposures suspected to be risk factors. By linking exposure information with biological data measured on the same human biosample the mechanisms and events occurring along the theoretical continuum

between exposure and disease can potentially be identified, providing targets for both early diagnosis and therapy [7].

1.6.3 High throughput –omics technologies

The advent of the 'omics' era, together with improved bioinformatics and systems biology tools, has enabled a new generation of biomarker discovery allowing for the simultaneous quantitative analysis of tens of thousands of biological markers to differentiate between cases and controls. The combination of multiple –omics technologies including epigenomics, transcriptomics, proteomics and metabolomics provides a global picture of the functional state of an organism at the levels of DNA, RNA, proteins and metabolites. These measurements are co-informational, each shedding light on complementary biological process [160] and together they provide a holistic view of the effect of the environment on cellular pathways. In this way -omics technologies provide unique opportunities for the discovery of biomarkers of exposure to exogenous environmental compounds, diet, lifestyle factors, pathogens and anything else that perturbs the steady state of the human body. Consequently these technologies can also be used to identify biomarkers of the early effects of exposures that can be considered predictive for disease risk.

1.6.4 New statistical techniques

The advent of these new technologies introduces a range of new challenges. GxE utilising a GWAS approach and –omics analyses involve hundreds of thousands of potentially highly correlated variables, and no reasonable threshold for minimising false-positives has yet been agreed [161]. Similarly, to elucidate fully the role that a risk factor or interacting risk factors play in disease aetiology as is proposed in the *MITM* approach, it is necessary to determine the temporal sequence of initiation and the subsequent stages involved. Most biostatistical models such as multivariate regression are insufficiently able to take these levels of complexity into account [162]. Therefore to exploit these advances fully also requires novel statistical techniques [2].

A novel Bayesian methodology – profile regression – offers a potential solution [163,164165] (methods section 3.3.4 and chapter 4). Profile regression clusters individuals according to their genetic and environmental covariate pattern and simultaneously associates these patterns with the disease outcome, allowing for the identification and quantification of the effect of the genetic and environmental factors which jointly influence risk. Profile regression is a Bayesian technique which is optimally designed to deal with highly dimensional and highly correlated data, and which has previously been shown to be well suited to the analysis of genome-wide data in a study of lung cancer [165].

1.7 Conclusions

The end of the 1990s coincided with a reported decline or levelling off of the incidence of NHL in most Western countries [34], however despite encouraging developments in understanding the pathogenesis and mechanisms of disease [166] particularly from studies of tumours and NHL patients, determining aetiology and predicting risk remain challenging [167]. Immune dysfunction and related conditions continue to represent the most well defined risk factors while putative associations with occupational, environmental and dietary factors remain controversial. Similarly genetic studies have been promising but few causal variants have been identified to date. The limitations of traditional epidemiological studies have led to a shift in focus to more molecular epidemiology-based approaches characterised by the search for biomarkers of exposure and disease. However, to date this work has been hindered by the same factors as the conventional studies and again very little conclusive, validated evidence is available. What is agreed upon is that the aetiology of B-cell NHL is complex and largely unexplained. Consequently further research into the causal pathways of NHL is vital and methods of analyses which are able to take these levels of complexity into account must be employed

In the following chapters a number of these issues will be addressed and novel biomarkers on the causal pathway will be identified. These will help to better characterise risk-disease relationships and additionally to inform on disease mechanisms. This will be achieved utilising a nested case-control design within pre-existing cohorts with prediagnostic blood samples that can be used to precisely measure potential markers in a manner that allows temporal determination and is free from bias. The findings will then be integrated to build upon the current knowledge in this field to better understand the aetiology of this malignancy for which so little is currently known.

1.8 Objectives

The objectives of this thesis are as follows:

1. To identify and validate novel biomarkers of NHL including biomarkers of susceptibility, exposure and effect
2. To integrate two different biomarker categories using the *"meet-in-the-middle"* approach, to consider temporal causal pathways
3. To explore the utility of the identified biomarkers and the biomarker *"meet-in-the-middle"* approach in understanding NHL aetiology and in predicting risk in individuals
4. To consider how the findings of this thesis may inform future studies of NHL

2. GENE-ENVIRONMENT INTERACTION STUDIES IN NHL EPIDEMIOLOGY: A REVIEW OF THE LITERATURE

2.1 Introduction

As outlined in the previous chapter the aetiology of NHL remains largely unknown despite a multitude of studies investigating potential environmental, as well as genetic causes. Where putative associations have been observed the effect sizes tend to be small, rendering them of little value for clinical prediction. Consequently the population attributable risk for these factors is unclear.

It is now widely accepted that a majority of complex disorders including cancers arise, not from single gene-mutations or single risk factors but rather from the interplay between an individual's genetic-make up and their personal exposure and lifestyle histories. This renders elucidation of the aetiology of such disorders difficult. It is also thought to explain in part why, when considering environmental or genetic risk factors in isolation few significant associations have been observed. Those that have been identified, have been small in magnitude, as the observed associations tend to be shifted to the null [168]. It is hypothesised that by restricting the analysis of an environmental risk factor to those who are genetically susceptible, the power to identify associations will be increased [169]. Furthermore the identification of interactions also provides evidence on the pathways and mechanisms by which an environmental factor may be exerting its risk [170].

Consequently GxE studies have become popular in the study of complex diseases such as NHL. However, to date there are no existing literature reviews which consolidate all the available studies of GxE in the aetiology of B-cell NHL and which report the results of these studies in detail.

2.1.1 Objectives of this review

In order to fill this gap in the literature the objectives of this review are to identify, summarise and critique all the current published literature considering the interactions between specified genetic variants and specified environmental variables, including lifestyle, anthropometric, medical, exogenous and dietary factors, which affect the risk of B-cell NHL and its subtypes. The review was restricted to B-cell NHL in line with the recommendations by Morton *et al*. [19] regarding the proposed classification of lymphoid neoplasms for epidemiologic research, to reflect the higher prevalence of B-cell NHL in Western populations which comprise the population studied in this thesis, and in keeping with the overall focus of this thesis.

2.2 Search Strategy

PubMed, EMBASE and Web of Science were searched for peer-reviewed articles using the search term combinations described in figure 2.2 up until 14th April 2013. The population of interest was adult (aged 18 years and older) males and females from any country. The inclusion and exclusion criteria are described in table 2.2. The references of selected articles were also screened for potential inclusion, as were the references from any existing reviews of GxE in NHL [171].

Figure 2.2: Flowchart describing the results of the literature search

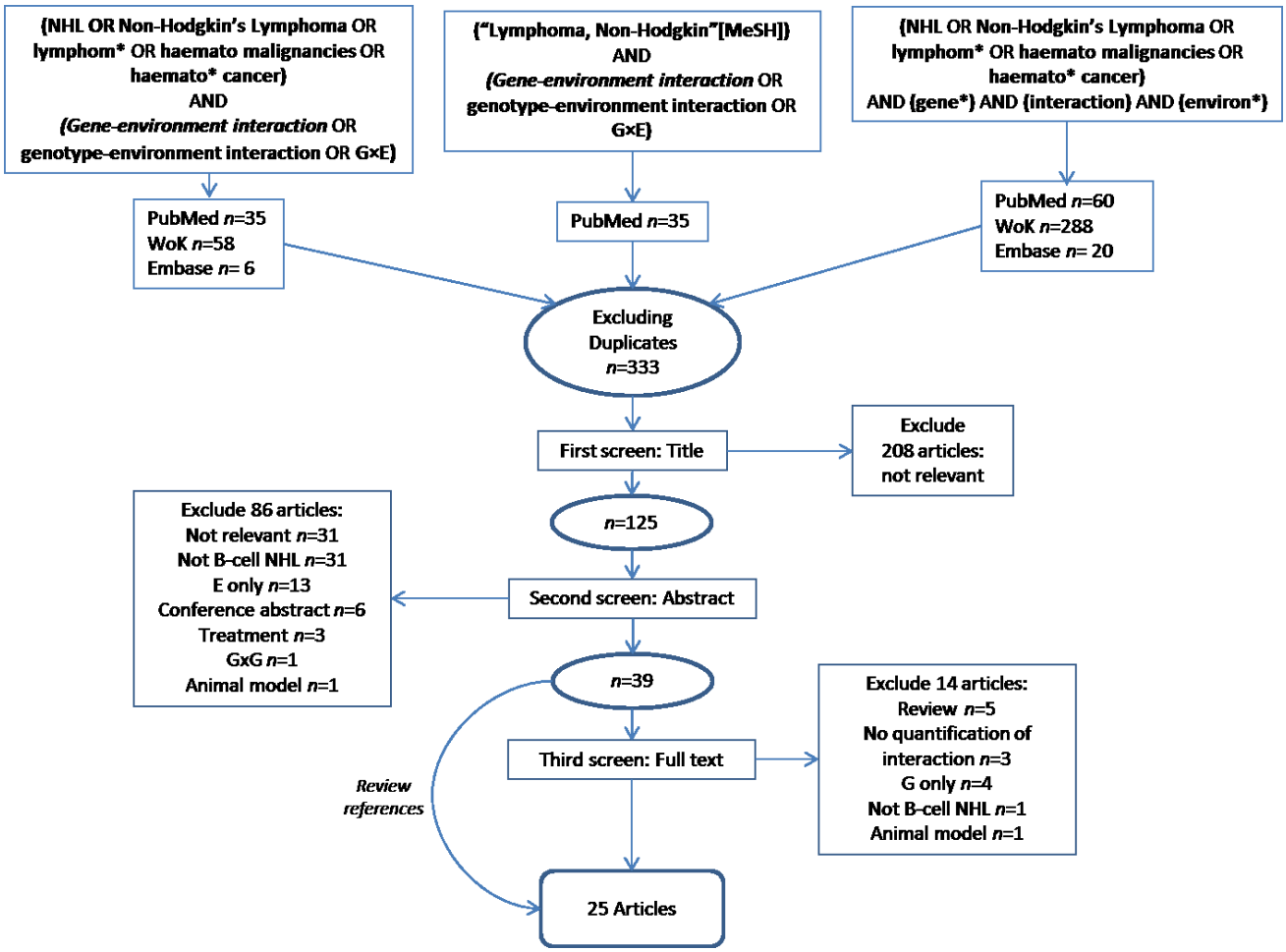


Table 2.2: Search Strategy Inclusion and Exclusion Criteria

Inclusion criteria	**Exclusion criteria**
• English-language articles published in per-reviewed journals • Articles published before April 2013 • B-cell NHL or a specified subtype as disease outcome • Specified polymorphism and specified environmental, lifestyle, dietary or anthropometric variable as an exposure • Quantified measure of interaction • Studies in humans	• T-cell NHL as the disease outcome • AIDs associated B-cell NHL cases • Articles in languages other than English • Articles not reporting original results • Articles considering mortality or treatment effects as outcome

2.3 Included articles

A total of 25 articles which met the inclusion criteria were selected following the screening process. These articles are described in table 2.3.

Table 2.3: Description of the Included Articles from the Literature Search

Study population	Year	Authors	No. cases	No. controls	Gene(s)	No. SNPs	Environmental variable	Subtype analyses
Population based Case-control study, UCSF2, San Francisco, USA (1988-1995) Age: 21-74	2004	Skibola C. *et al*	458	812	Leptin (LEP), Leptin-Receptor (LEPR)	3	Body Mass Index	All NHL, DLBCL, FL, Other
Population based Case-control Study (matched on sex and date of birth), England (1998-2001) Age: 18-64	2005	Willett E. *et al*	593	754	Leptin (LEP), Leptin-Receptor (LEPR)	4	Body Mass Index	All NHL, DLBCL, FL
Population based Case-control study (frequency-matched on age, sex, race, and registry) NCI-SEER, USA (1998-2000) Age 20-74	2010	Kane E. *et al*	471	468	Melanocortin 1 receptor	9	Sun exposure, eye colour, skin colour, hair colour, history of sun burn	All NHL, DLBCL, FL
	2006	Morton L. *et al*	1136	922	N-acetyltransferase 1 (NAT1) and 2 (NAT2)	10	Cigarette smoke, dietary heterocyclic amines	All NHL, DLBCL, FL, SLL, MZL, T-cell NHL
	2006	De Roos A. *et al*	566	422	11 genes involved in xenobiotic metabolism	15	Smoking status	Subtype specific data not shown
	2007	Lim U. *et al*	386	319	18 genes involved in one-carbon metabolism	30	107 food and beverage items	All NHL, DLBCL, FL, SLL
	2007	Morton L. *et al*	566	422	N-acetyltransferase 1 (NAT1) and 2 (NAT2)	10	Hair dye use	Subtype specific data not shown
	2007	Purdue M. *et al*	551	462	Vitamin D-receptor (VDR), Interleukin (IL4, IL10, IL12A, IL12B), Tumour necrosis factor (TNF)	8	UV exposure	All NHL, DLBCL, FL
	2007	Wang S. *et al*	1,172	982	Interleukin-10 (*IL10*), Tumour necrosis Factor (*TNF*)	2	Autoimmune conditions, Education, BMI, Diet, Height, Weight, Smoking status, Ethanol, b6, Sun exposure, Eye colour, Termite treatment, Alpha-chlordane, PCB180, Furans	All NHL, DLBCL, FL

Table 2.3 ***continued***

Study population	Year	Authors	No. cases	No. controls	Gene(s)	No. SNPs	Environmental variable	Subtype analyses
Population based Case-control study (frequency-matched on age, sex, race, and registry) NCI-SEER, USA (1998-2000) Age 20-74	2008	Keleman L. *et al*	371	311	17 genes involved in pathways of ROS neutralization and detoxification and DNA repair	28	Multiple dietary variables	All NHL
	2009	Colt J. *et al*	1321	1057	36 immune genes	61	Organochlorine exposure	All NHL
	2009	Gathany A. *et al*	990	828	IRF4 (MUM1)		Sun sensitivity	All NHL
Population-based Case-control study (matched on sex, age, and study region) , Germany (1999-2002) Age: 18-80	2008	Hoeft B. *et al*	710	710	Prostaglandin-endoperoxide synthase-2 (COX2), prostaglandin E synthase (PTGES), interleukin-1 alpha (IL1A), IL-1 beta (IL1B) IL-1 receptor antagonist (IL1RA)	10	NSAID use	All lymphoma, B-NHL, T-NHL, HL
Females only. Population based Case-control study (frequency-matched by 5 -year age groups), YALE Study, Connecticut, USA (1996-2000) Age 21-84	2009	Kilfoy B. *et al*	518	597	glutathione S-transferases (GSTs) and cytochrome P450s (CYPs)	5	Smoking status	All NHL, DLBCL, FL
	2009	Han X. *et al*	513	591	18 genes involved in the oxidative stress pathway	137	Fruit and vegetable intake	All NHL, DLBCL, FL, SLL/CLL
	2009	Zhang Y. *et al*	461	535	9 xenobiotic genes	19	Hair dye use	All NHL, DLBCL, FL, CLL/SLL, MZL, T-cell lymphoma
	2010	KilfoyB. *et al*	453 (NAT1), 456 (NAT2)	528 (NAT1), 533 (NAT2)	N-acetyltransferase genes NAT1 and NAT2	10	Smoking status	All NHL, DLBCL, T-cell NHL

Table 2.3 ***continued***

Study population	Year	Authors	No. cases	No. controls	Gene(s)	No. SNPs	Environmental variable	Subtype analyses
Females only. Population based Case-control study (frequency-matched by 5 -year age groups), YALE Study, Connecticut, USA (1996-2000) Age 21-84	2010	Li Y. *et al*	461	535	11 Cytochrome P450s (CYPs), glutathione S-transferases (GSTs) and N-acetyltransferases (NATs)	20	Alcohol consumption	All NHL, DLBCL, MZL, T-cell NHL
	2010	Barry K. *et al*	518	597	Cytochrome P450 2E1 (CYPE1), Epoxide Hydrolase 1 (EPHX1), NAD(P)H Dehydrogenase, Quinone 1 (NQ01), Myeloperoxidase (MPO)	6	Organic, chlorinated solvents (Benzene, chloroform, carbon tetrachloride, dichloromethane, dichloroethane, methyl chloride, trichloroethylene)	All NHL, DLBCL, FL, CLL, SLL, MZL, T-cell NHL
	2011	Chen Y. *et al*	785	868	Th1 and Th2 cytokine genes	5	Body Mass Index	All NHL, DLBCL, FL, SLL/CLL, MZBCL
	2012	Jiao J. *et al*	518	597	16 DNA repair genes	30	Chlorinated solvent, benzene and formaldehyde exposure	All NHL, DLBCL, FL
Population-based case-control study (frequency-matched by 10 -year age groups, sex and country) The Scandinavian lymphoma aetiology study (SCALE), Denmark and Sweden (1999-2002) Age: 18-75	2010	Ekström Smedby K *et al*	2448	1981	vitamin D receptor (VDR) protein	10	UV exposure	All NHL, DLBCL, FL, CLL
	2010	Fernberg P. *et al*	2449	1980	BCL2, CCND1, MYC, TNF, and IL10	11	History of autoimmune disorders	All NHL, DLBCL, FL, CLL, Mantel cell, T-cell NHL
Population based case-control study (matched on 5 year age group, sex and region) British Colombia, Canada (2000-2004) Age: 20-79	2010	Ng C. *et al*	422	459	AHR	5	Organochlorine exposure	All NHL, DLBCL, FL, Other B-cell, T-cell
Four studies from the USA; Nebraska, NCI-SEER, YALE Study, UCSF2. Two from Europe; SCALE, EpiLymph (Spain, Germany, Ireland, Czech Republic, France, Italy). One from Australia; New South Wales	2013	Gibson T. *et al*	5026	4630	N-acetyltransferase genes NAT1 and NAT2	10	Smoking status	All NHL, FL, DLBCL, ll/SLL, MZL, MCL, PTCL,MF/SS

2.4 Historical context

Many 'gene-environment interactions' have been known of since long before the advent of genotyping, although they may not have been identified as such. For example, the effect of sunlight exposure on fair as compared to dark skin, or the flushing response in certain individuals following alcohol consumption [169] . The formal study of the interaction between genes and the environment in the aetiology of disease has been suggested to extend back to as early as 1902 [169,172]. However the first studies which attempted to quantify interaction did not appear until the late 1960s as one of the fundamental concepts of the developing field of genetic epidemiology [159]. There has since been an explosion in the number of publications which cite 'gene-environment interaction' in the title, and this field now represents a major research theme which continues to grow with the development of newer and cheaper technologies for sequencing the genome.

Despite this, the study of GxE remains a relatively unexplored area in NHL aetiology, as evidenced by the small number of articles identified by this review (n=25). Nevertheless, the selected articles also reflect the general trend for an increase in studies of this type and suggest this is a growing field which will only expand. The first article was not published until 2004, and almost half were published from 2010 onwards.

2.5 Included populations

Although 25 unique articles were selected for review these in fact represent only seven independent study populations and one pooled study which encompasses four of these populations. This pooled study was included as it additionally contained unpublished results [126]. Two thirds (n=17) of the selected articles originated from two studies: the National Cancer Institute-Surveillance Epidemiology and End-Results (NCI-SEER) study [173] and the YALE study [174]. This may limit the generalizability of these findings to other populations and suggest further GxE analyses are hampered by the lack of available genotype data for NHL cases and suitable controls within large- scale epidemiological studies. Although there are a number of multi-study consortiums specifically considering NHL, notably InterLymph and EpiLymph, in common with a large number of epidemiological studies, many other studies of NHL lack the funds and infrastructure to collect biological samples.

Of the selected articles, all included very similar age ranges spanning from 18-84 years. They were all conducted in comparable time periods around the late nineties, early 2000s. Eight of the articles: those based within the YALE study, included only female participants while the rest considered both males and females. Where reported, the ratio of included males to females was similar except in the two articles from the SCALE study [175,176] and in Morton *et al.* 2007 [177] where a much higher percentage of males and of females, respectively, were included. Again when considering the generalizability of these studies it should be noted that they do not reflect the overall higher burden of NHL among males worldwide. In all studies cases and controls were matched by sex.

Although the included populations for the articles originated from the same few overall studies there was between-article heterogeneity in terms of the sample size (table 2.3), particularly within the NCI-SEER study. This was on account of the availability of environmental or genetic data for the specific variables investigated, or due to exclusions based on technical issues.

Ethnicity is of particular importance in studies of genetics, due to well-defined issues of population stratification [168,169]. However only four of the articles: those based within the English study and the Scandinavian SCALE study, utilised 'genetically homogenous' populations. The other studies contain a mixture of ethnic groups in their analyses, although these were not always specified or quantified. As well as the potential for ethnic differences arising through differing inherited allele frequencies, observed differences between different ethnic groups, may also be a reflection of cultural and societal differences, which will influence environment and lifestyle. This is also true of the different geographical populations investigated by the selected articles. Nineteen of the articles were based in North America including Canada, two in Scandinavia and the UK respectively and one in Germany. The pooled Study by Gibson *et al.* [126] additionally included a further North American population, six more European countries and an Australian study. It is of note that there were no studies conducted

outside Europe, North America and Australia, and all populations were either entirely or majority Caucasian. Given the recent findings of differences in genetic susceptibility to lymphoma among different ethnic groups [178,179], this may be an issue for the wider generalisability of these results. In any case the problems faced by main effects studies (i.e. the gene-disease or risk factor-disease relationship), in terms of ethnicity/race and geography, are potentially doubled in GxE studies. The currently available studies are most generalisable to a Western Caucasian population, and caution should be applied when considering their results in other populations and contexts.

2.6 Investigated environmental and lifestyle factors

Few definitive risk factors have been identified for NHL, which complicates the study of GxE as there are few obvious environmental variables to focus on. Although some studies have employed what they define as an 'Environment-wide Study design'[180], these are not entirely analogous to a GWAS as they cannot measure all possible environmental exposures and lifestyle factors, and so a targeted approach must be taken. The most well characterised risk factor for NHL is immune dysfunction [36], and accordingly two studies [106,176] investigated the interaction with a history of autoimmune conditions, which have been previously been associated with an increased risk of NHL [40].

For all other included factors (Table 2.3) there exists a body of literature considering their association with NHL, but the evidence is far from unequivocal. The most commonly investigated variable in the selected articles was smoking status (n=6), followed by UV/sun exposure (n=5), organochlorine exposure (n=5) and BMI (n=4). Despite multiple studies on the association with NHL, the literature has failed to reach a consensus for any of these factors. Interestingly UV exposure has actually been hypothesised to be protective against risk [181]. Three other articles, those by Han [182], Li [183] and Hoeft *et al*. [184] considered variables which are thought to be protective against NHL or which have shown inverse associations: fruit and vegetable consumption [99], alcohol intake [105] and NSAID (non-steroidal anti-inflammatory drug) use [185] respectively.

The main effect association between the investigated environmental factor and NHL was reported by 16 out of the 25 studies, either explicitly or in reference to previous work performed by the same group on the same population. The majority of these reported a significant association with the main effect. The exceptions were variables related to sun exposure [186], NSAID use [184], dietary mutagens [187] and one of the papers considering hair dye use [177]. As with hair dye use, for some factors investigated in more than one study conflicting results were found. Gibson *et al*.'s pooled paper [126] observed a significantly positive association between FL and ever smoking which was not reported by Kilfoy *et al* (2009 and 2010) nor Morton *et al.* (2006), although all three agreed there was a null relationship with NHL overall. Similarly Gathany *et al.* [188] reported that eye colour and number of hours spent in the sun were both inversely associated with risk, while Kane *et al.* [186] found no effect. In fact, such inconsistency in the literature may well be the result of GxE and therefore provides the very reason to conduct such analyses in the first place. The failure to report a main effect in the remaining studies complicates the interpretation of the interaction effect presented.

2.7 Investigated genes and pathways

All of the included studies have taken a candidate approach to the selection of genes based on known functions, biological processes or biological networks. Or based on the known mode of action of the environmental factor and therefore the pathways through which it may work [170,171].

The most commonly investigated single genes were Interleukin10 (*IL10*) and Tumour necrosis factor (*TNF*), both of which were investigated by the same five studies [176,189-192]. Three of these were based on the same overall population from the NCI-SEER study [189,190,192]. *IL10* and *TNF* are involved in the immune response as part of a regulatory feedback network that controls inflammation, and both have been frequently associated with NHL risk [189]. In fact genes of immune pathways are some of the most frequently explored with regard to lymphoma risk, and a number of significant findings have been reported [12]. Within this review in addition to *TNF* and *IL10* many other genes involved in the immune response were also explored including *Th1/Th2* [191], *COX2* [184], *IFNG* [192] and *IRF4* [188]. Most were chosen on the basis of previously reported associations in the literature, but in the few studies where main effects were reported, the only replicated associations were for *TNF* in Fernberg *et al.* [176] and *IRF4* in Gathany *et al*. [188]. Again, this may be taken as evidence for the role of GxE. Alternatively it may be reflective of the problems with the candidate gene approach and the lack of replication that has characterised such studies to date.

The most commonly explored pathway was xenobiotic metabolism, which was considered in ten articles [126,177,183,187,189,193-198]. Polymorphisms in this pathway resulting in a lowered capacity to detoxify and eliminate carcinogens from the body, have also previously been associated with risk of NHL [199]. Interestingly among these studies, more main effects were reported and replicated, compared to those investigating immune pathways.

In addition to genes involved in xenobiotic metabolism Keleman *et al.* [195] also studied a number of DNA repair genes, as did Jiao *et al.* [200]. Two studies considered Leptin (*LEP*) and Leptin receptor (*LEPR*) genes which are involved in body weight homeostasis [201,202] and two studies considered polymorphisms in the Vitamin D receptor gene [175,190]. The remaining articles considered genes involved in metabolism [203], one-carbon metabolism [204], oxidative stress [182] and pigmentation [186].

2.8 Methods

There are considerable methodological challenges, both including and extending beyond those for main effects studies, in the design and analysis of GxE studies [168]. This partly stems from the fact that, although many studies define themselves as GxE studies, they actually represent a wide range of statistical models that present different information in different ways, and which are often not comparable [169].

However, the included studies in this review were fairly homogenous: all employed a case-control design. Incident cases were identified and suitable unrelated, population-based controls selected, most commonly by population registries and random digit dialling. The only exceptions were the Irish, Czech and French EpiLymph studies from Gibson *et al.* which utilised hospital based controls [126]. This may introduce some of the well-defined bias associated with the use of such controls.

2.8.1 Sample size

There are a number of different approaches for calculating the required sample size to obtain optimum power in GxE studies [205]. However it is commonly accepted that the sample size of an interaction study needs to be at least four times the size required to calculate the main effect of each included variable with disease [205]. Allele frequency, the type of genetic model employed and the strength of the interaction will also affect sample size [206]. A high probability of misclassification of the risk factor will further increase the numbers required [169].

The samples sizes utilised in the articles range from nearly 700 [195] to over 4500 [175,176], and a total of 5026 cases and 4630 controls in Gibson's pooled study. Most authors recognised that the main limitation of their study design was a lack of statistical power due to sample size, although some made no mention of power and none reported a quantified measure of the actual power of their study to detect an interaction. This lack of statistical power was a particular problem for those studies reporting subtype analyses, and there is a strong probability that, especially for the weak interactions, and in those studies reporting multiple comparisons, some of the presented findings represent spurious associations [169].

The most effective way to increase power is by the pooling of studies as was done by Gibson *et al.*, but even here it was not feasible to recruit enough cases and controls to perform GxE studies with high power. Particularly when considering subtype-specific analyses, SNPs with low minor allele frequencies or difficult to classify environmental and lifestyle factors.

2.8.2 Statistical methods

Many models and methodologies have been proposed to assess GxE [207]. However all the selected articles in this review used the same statistical model, the inclusion a multiplicative term in the logistic regression model and the calculation of a p-value for interaction based on either the Wald's test or the likelihood ratio test. The only exceptions were Morton *et al.* (2006) and Kilfoy *et al.* (2010) who calculated the risk associated with hair dye use and smoking, respectively, stratified by *NAT1* and *NAT2* variants, but did present a quantified measure of interaction.

All the studies made assumptions about the underlying genetic model for their variant of interest. A dominant model was chosen by the majority of papers, most likely to simplify the interpretation of the interaction model through the use a binary 'G', as well as to retain power. Although the use of a dominant model may be justified for some of the investigated variants, none of the studies provided a biological justification for their choice. If another genetic model would be more appropriate then there is a risk of misleading odds ratio and interaction estimates, which have been shown to be sensitive to model assumption [208].

2.8.3 Environmental variable assessment methods

The lack of definitive risk factors identified for NHL to date has been suggested, in part, to reflect the difficulty of accurately measuring exposure, lifestyle and environmental factors in large scale epidemiological studies. This is a problem for any such study but the extent to which the included studies are subject to 'measurement error' varies according to the investigated variable and the study design employed. Blood measurements and clinically diagnosed medical conditions which, although still prone to measurement error, are in theory more objective than those such as sun exposure, which rely strongly on participant recall. Only eye colour, which was considered in two articles [186,189], as one of a range of variables, should be exempt.

Articles from the same 'parent' study by definition employed the same methodology to assess the environmental and lifestyle factors, and all used in-person or telephone interviews to complete the questionnaire data. Factors measured by more than one study tended to be ascertained in uniform ways. Nearly all relied on self-report, including height and weight from which BMI was calculated. Sun/UV exposure was calculated from self-reported time spent in the sun, taking geographical residence and use of sunbeds into account, and sun sensitivity was estimated from a variety of questions such as history of sunburn. Lim [204], Keleman [195], Morton (2007) [19] and Han *et al.* [182] used variations of food frequency questionnaires and calculated nutrient intake from these using standardised databases. Smokers were classified as never, former or current except in Kilfoy *et al.* (2009) [196] which included only two categories (smokers and non-smokers).

Only organochlorines were objectively measured in serum samples by Ng [193], Wang [189] and Colt *et al.* [192]. Colt additionally measured the levels in carpet dust from the homes of the participants. In Jiao *et al*. [200] and Barry *et al.* [203] a standardised job exposure matrix assessing occupational history, was used as a proxy for exposure.

2.8.4 Genotyping

DNA was extracted from blood or buccal cell samples in all studies, with the exception of the SCALE study [175,176] and Hoeft *et al*. [184] which used blood only. The majority of studies utilised Taqman (Applied Biosystems) for genotyping. Han *et al.* [182] used a custom-designed Illumina assay (Illumina, San Diego, CA), while Hoeft *et al.* [184] employed Pyrosequencing Restriction- Fragment-Length-Polymorphism (Biotage, Uppsala, Sweden) analysis for certain SNPs. The SCALE study[175,176] used Sequenom matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (Sequenom Inc, San Diego, CA) and in Kane *et al.* [186] *MC1R* genotypes were determined via direct sequencing (DxS Ltd., Manchester, UK).

2.9 Results

Eighteen of the 25 studies stated that they found evidence of at least one interaction between a defined SNP and a defined environmental or lifestyle factor in the aetiology of B-cell NHL or one of its subtypes. Some interactions were consistent across subtypes, while others were specific, although in actual fact all studies were underpowered to explore by subtype, particularly the rarer ones. Crucially though, the definition of success differed between the articles; for some a difference in risk estimate between different genotype strata was sufficient, in others a significant *p-value* for interaction was required. The confidence level required to denote significance also varied, and interactions reported as significant in some studies would be considered insignificant in others. A number of studies used the conventional level of statistical significance, a p-value of 0.05, to denote an interaction, despite the fact that most were considering multiple factors and genetic variants and some presented main effects as well as interactions. The majority of the studies did acknowledge this, although only nine corrected for multiple testing [175,182,191,193,195,198,200,203,204]. Additionally three studies argued that, although not robust for multiple testing, the consistency of their findings renders false-positives unlikely [177,190,192]. Table 2.9 describes the p-value for interaction in those studies where success was declared and the p-value was smaller than 0.05.

Jiao *et al.* [200] observed significant interactions between DNA repair genes and chlorinated solvents and benzene in the risk of NHL that remained significant after correction for multiple testing. This was also apparent when analysis was restricted to FL and DLBCL. Both Wang *et al.* [189] and Colt *et al.* [192] considered the joint effects of polymorphisms in *IL10* and *TNF* with organochlorine exposure in the NCI-SEER study. However, within the two differing subsets of the population different conclusions were drawn. Wang *et al.* stated they observed evidence of a joint effect of *IL10* with exposure to termite treatment/chlordanes; although the p-value was non-significant they observed consistent results across subtypes. Colt *et al.*, who explored a different polymorphism in *IL10* found no evidence of this but did report a significant interaction between PCB180 and *IL10* that was not observed by Wang *et al*. [189]. Colt *et al.* [192] additionally observed a significant interaction between *IL16* and PCB180 plasma levels and concluded that their results provide one of the 'first examples of potential gene-environment' interactions in NHL aetiology. Both Colt *et al*. [192] and Wang *et al*. [189] agreed that there was no significant interaction between *TNF* and organochlorines.

Ng *et al.* [193] considered organochlorines and variants in the Aryl hydrocarbon receptor gene (*AHR*). They did observe interactions with PCB118, oxychlordane and transnonachlor that were consistent across all subtypes, but not robust for their defined FDR threshold (0.01). Nevertheless, they conclude that the consistency of their findings does suggest a role for the *AHR* gene in mediating the risk of NHL with exposure to organochlorines. Similarly Barry *et al.*[203] reported a significant interaction between solvent exposure and *CYP2E1* and *EPHX1* in the risk of NHL.

Three studies considered dietary variables [182,195,204]; among the multiple genes investigated five were common between the articles by Han *et al.* [182] and Keleman *et al.* [195] both of whom considered vegetable consumption (*GPX1, MPO, NOS2A, NOS3* and *SOD2)*. Overall Keleman *et al.* concluded that, despite observing some interactions, these were likely false-positives, and that the examined genes in the pathways of reactive oxygen species neutralisation, detoxification and DNA repair were not modifiers of the association between vegetables or zinc intake and NHL risk. In contrast Han *et al*. noted a number of interactions with vegetable consumption mainly for genetic polymorphisms in *NOS1, MOP* and *SOD 3*. Most interactions were observed for DLBCL, but a number for FL, CLL and NHL overall were also reported.

Lim *et al.* [204] reported significant gene-nutrient interactions in the one-carbon metabolism pathway which may be aetiologically involved in lymphomagenesis. In particular they found a protective interaction between vitamin B6 with *FPGS, MTHFS* and *MTR*, and for methionine with *FTHFD, MTHFR* and *MTRR*.

Four studies reported on potential interactions with variables relating to UV exposure and sun sensitivity[175,186,188,190]. Among these, two [175,190] reported on polymorphisms in the vitamin D receptor. Purdue *et al.'s* [190] results suggested that the inverse association between UV exposure and NHL risk may be mediated by the vitamin D pathway, and this was particularly evident in FL. In contrast Smedby *et al.* [175] reported no evidence of interaction between *VDR* variants and UVR exposure on risk of overall NHL or B-cell lymphoma subtypes, although they did observe a potential interaction with T-cell risk. Both Kane [186] and Gathany *et al.*[188], neither of whom observed a significant interaction with sun exposure, considered a number of related variables including hair and eye colour. While Kane *et al.* reported that the joint effects of *MC1R* genotypes and skin colour influenced DLCL risk, Gathany *et al.* observed no significant interactions with *IRF4* for any of the investigated variables.

Morton *et al.* (2006) [187] concluded that there was an increased risk of smoking-associated NHL in *NAT2* intermediate/rapid-acetylators compared to *NAT2* slow-acetylators, but no interaction with *NAT1* and smoking for NHL overall. In contrast Kilfoy *et al.* (2010) [197] suggested that the association between NHL and smoking status may be modified by genetic variation in *NAT1* but not in *NAT2*, although they did not present p-values for interaction. Gibson's [126] pooled study included both these articles in addition to unpublished results on *NAT* polymorphisms and smoking from 5 other studies. They did not report significant interactions for either *NAT1* or *NAT2* and concluded that the positive association between risk of FL and smoking status is not modified by genetic variation in the N-acetyltransferase enzymes. However the exact variants explored differed slightly between the articles. Three other articles also considered smoking status [189,194,196]; Wang *et al.* [189] reported that *IL10* and

TNF genotypes had no effect on the relationship between smoking and NHL. Similarly Kilfoy *et al.* (2009) [196] did not report significant interactions for polymorphisms in GSTs or CYPs. De Roos *et al* [194], who explored a number of different genes of the xenobiotic metabolism pathway, including GSTs and CYPs, reported significant interactions for former smoking and *EPHX1* H139R and *PON1* Q192R, but concluded these had no realistic biological interpretation and were likely to represent false-positives.

Both Skibola *et al.* [201] and Willet *et al.* [202] considered the potential that genetic polymorphisms in Leptin or its receptor may be modifying the association between obesity and NHL. However neither reported evidence of a significant interaction. Conversely Chen *et al.* [191] reported that common genetic variation in Th1/Th2 pathway genes may modify the association between BMI and NHL risk. Wang *et al.* [189] found that risk differed by *TNF* genotype, although this was not supported by Fernberg *et al.*'s study [176].

Of the remaining articles Morton *et al.* (2007) [177] presented evidence suggesting that the risk of hair dye use may differ by genetic variation, although they do not quote an interaction statistic. Zhang *et al.* supported Morton's findings for *NAT2*, as well as observing interaction effects for hair dye use with *CYP2C9, CYP2E1, GSTM3 and GSTP1*. They did quote interaction statistics, and conceded they were not robust to correction for multiple testing. Hoeft *et al.* [184] reported evidence for the joint effects of NSAID use and *IL1RN* and *IL1B* on B-cell NHL risk. Finally Li *et al.* [183] reported interactions for multiple GSTs and NATs with alcohol intake on the risk of DLBCL, although again the reported p-values are unlikely to be robust for correction for multiple testing. These findings are summarised in table 2.9.

Table 2.9; Summary of the results of the selected publication reporting significant interactions with a p value<0.05

The table shows the interacting variant and environmental variable, the subtypes affected, the variant genotype and the direction of risk effect

Pathway	Genetic variant		Exposure	Subtypes	variant genotype	Risk
DNA Repair, Oxidative Stress, Xenobiotic Metabolism	***MGMT***	**rs12917**	**Chlorinated Solvents**	NHL**; DLBCL**; FL**	CT+TT	Increased risk with exposure
	NBS1	**rs1805794**	**Chlorinated Solvents**	NHL**	CG+CC	Increased risk with exposure
	BRCA2	**rs144848**	**Benzene**	NHL**	AC+CC	Decreased risk with exposure
	GSTP1	**rs1138272 (A114V)**	**Alcohol**$^{\infty}$	MZL*	CT or TT	Decreased risk with exposure
			Vegetables$^{\neq}$	NHL*	CC	Decreased risk with higher intake
		rs1695	**Alcohol**$^{\infty}$	MZL*	AG or GG	Decreased risk with higher intake
	GSTM3	**rs1799735**	**Alcohol**$^{\infty}$	DLBCL*	+- or --	Increased risk with higher intake
			Vegetables$^{\neq}$	NHL*	AGG/AGG	Decreased risk with higher intake
			Hair dye	NHL*	+- and --	Increased risk with use pre-1980
	OGG1	**S326C**	**Nutrients**	NHL*	CC	Decreased risk with higher Zinc intake
	XRCC3	**T241M**	**Vegetables**$^{\neq}$	NHL*	CC or CT	Decreased risk with higher intake
	AHR	**IVS1 + 4640G/A**	**Oxy-chlordane**	NHL*	GA or AA	Decreased risk in the highest quartile of exposure
			Trans-nonachlor	NHL*	GA or AA	Decreased risk in the highest quartile of exposure
			PCB118	NHL*	GA or AA	Decreased risk in the highest quartile of exposure
	NAT1		**smoking**	NHL**,$^{+}$	without *10	Increased risk in current smokers
	NAT2		**smoking**	NHL*,$^{+}$	intermediate-/rapid-acetylators	Higher risk in smokers
	MPO	**rs4401102**	**Vegetables**$^{\neq}$	NHL**,$^{+}$	CT + TT	Increased risk with higher intake
	NOS1	**rs2293054**	**Vegetables**$^{\neq}$	NHL**, FL***	AG + AA	Decreased risk with higher intake
		rs545654	**Vegetables**$^{\neq}$	DLBCL***	CT + TT	Increased risk with higher intake
		rs11068446	**Vegetables**$^{\neq}$	DLBCL**	CT + TT	Decreased risk with higher intake
		rs3782221	**Vegetables**$^{\neq}$	DLBCL**	AG + AA	Decreased risk with higher intake
		rs7298903	**Vegetables**$^{\neq}$	DLBCL**; FL***	CT + CC	Decreased risk with higher intake
		rs12424669	**Vegetables**$^{\neq}$	DLBCL**	CT + TT	Increased risk with higher intake
	EPHX1	**H139R**	**smoking**	NHL*	GG	Increased risk among non-smokers

Table 2.9; *cont.*

Pathway	Genetic variant		Exposure	Subtypes	variant genotype	Risk
DNA Repair, Oxidative Stress, Xenobiotic Metabolism	*PON1*	Q192R	smoking	NHL*	GG	Increased risk among non-smokers
	SOD3	rs2284659	Vegetables$^{\neq}$	CLL/SLL***	GT + TT	Increased risk with higher intake
	CYP1A1	rs1048943	Alcohol$^{\infty}$	DLBCL*	AG or GG	Increased risk with higher intake
	CYP2E1	rs2070673	Dichloromethane	NHL**	TT	Increased risk with exposure
			Carbon tetrachloride	NHL*	TT	Increased risk with exposure
			Methyl chloride	NHL*	TT	Increased risk with exposure
	CYP2C9	rs1799853	Alcohol$^{\infty}$	DLBCL*	CT or TT	Increased risk with higher intake
			hair dye use (pre-1980)	NHL*	CT + TT	Increased risk with use pre-1980
Immune response	*TNF*	rs1799724	BMI	NHL*; FL*	CC	Increased risk with BMI >25
	IL1RN	rs454078	NSAID use	NHL**	T	Decreased risk with use
	IL1B	rs16944	NSAID use	NHL**	A	Decreased risk with use
	IL1	haplo-type 3	NSAID use	NHL**	AACC	Decreased risk with use
		haplo-type 5	NSAID use	NHL***	TACC	Decreased risk with use
	IL4	Ex1-168C>T	TEQ	NHL*	CC	Increased risk with exposure
	IL7R	rs1494555	BMI	NHL*; MZL*; CLL/SLL*	AA	Increased risk with BMI >25
	IL12	rs568408	BMI	FL*	AG or AA	Increased risk with BMI >25
	IL13	rs20541	BMI	MZL*	AG or AA	Increased risk with BMI >25
	IL16	UTR, Ex871-A>G	PCB180	NHL**	AA	Increased risk with exposure
	IFNGR2	rs9808753	BMI	MZL**	AA	Increased risk with BMI >25

Table 9; *cont*

Pathway	Genetic variant		Exposure	Subtypes	variant genotype	Risk
One-Carbon metabolism	***FPGS***	**Ex15-263T>C**	**Vitamin B6 intake**	NHL**	CC	Decreased risk with high (>median) intake
	MTR	**Ex26-20A>G**	**Vitamin B6 intake**	NHL*	AA	Decreased risk with high (>median) intake
	FTHFD	**Ex10-40G>T**	**Methionine intake**	NHL*	GG	Decreased risk with high (>median) intake
	MTHFR	**Ex8-62A >C (1298A>C)**	**Methionine intake**	NHL**	CC	Decreased risk with high (>median) intake
	MTRR	**Ex5+36T>C**	**Methionine intake**	NHL*	TT	Decreased risk with high (>median) intake
Vitamin D	***Vitamin D***	**Ex11 + 32 T > C**	**UV exposure**	FL**	CC	Increased risk with < 7 h/week of exposure
Melanocortin	***MC1R***		**Skin colour**	DLBCL***	carrier of D84E, R151C, R160W or D294H	Increased risk in those with medium/dark skin

*Significance of the association; *<0.05, **<0.01, ***<0.001 [+]Not replicated in an independent study; \DLBCL – Diffuse Large B-cell Lymphoma; FL – Follicular lymphoma; MZL – Marginal zone lymphoma; MCL –Mantel cell lymphoma; CLL – Chronic Lymphocytic leukaemia (results for T-cell NHL not shown); ∞ includes beers, spirits and wine ≠ Any variables considering vegetable intake (cruciferous, leafing etc)*

2.9.1 Biological Plausibility and Replication

It is stated that a gene-environment interaction will only be accepted if it can be reproduced in two or more studies and if it is biologically plausible [169]. Although all the studies presented here are biologically plausible, indeed this was the impetus for conducting the studies, to date none of the positive findings have been replicated.

The existence of a statistical interaction between two factors does not necessarily inform understanding of the underlying biology of the joint effect of these factors on disease [168]. It can be difficult to determine biological plausibility, and in many studies it is often determined *post hoc* after the analysis has been completed to explain the findings. The potential plausibility of an interaction is further complicated by the many remaining 'grey areas' in the pathways and functions of genes, and by the subjective nature of the definition of plausibility [169].

One problem with the criterion of replication is that, as can be seen from this cross section of GxE studies, in such studies commonly multiple tests will be performed but 'current models of publication of individual studies favour suppression of negative results' [169,170]. This means replication of the interactions presented here may have been attempted but not found to be significant and therefore not published. In fact a number of the studies did agree on null findings, such as for BMI and *Leptin* [201,202] and *TNF* and organochlorines [189,192], but overall this is rare. The study of different SNPs within the same gene by different studies and the use of haplotype-tagging or linkage disequilibrium (LD) tagging SNPs to explore genetic associations further complicates the issue of replication and makes the identification of true causal associations even harder.

The lack of replications or attempted replications among the included studies meant that it was not possible to perform meta-analyses or to determine the heterogeneity in estimates between studies in this review.

2.10 Bias, confounding and limitations

All the included studies were case-control in design allowing detailed information, focussed on the variable of interest to be obtained and histology classification to be centralised. However, there are a number of unavoidable biases and limitations associated with GxE studies, in addition to the well-characterised problems faced by all case-control studies.

All participants were selected from larger parent studies, and were generally chosen based on the availability of both questionnaire data and DNA. If there is differential participation/availability by ethnicity, socioeconomic status or disease stage or grade, this may have an important impact on results. Selection bias was addressed in all studies by matching cases and controls on potential confounders, which tended to be uniform throughout the studies, in particular age and sex, and ethnicity where appropriate. However none reported on the grade or stage of disease at diagnosis, which may substantially affect results.

All studies, excluding those reporting on organochlorines, relied on self-reported information and the extent to which recall bias influenced the findings is difficult to know, but is likely to be sizeable given many of the studies are attempting to quantify lifetime or historical exposures and experiences. The case-control design may also introduce bias if diseased people choose to change their lifestyle following diagnosis or if the variable is altered by the pathology of the disease. The latter is of particular interest for those studies which measured serum concentrations of organochlorines [189,192,193] as there is some evidence to suggest that chemotherapy or NHL itself may affect body burden, and that therefore they may not be representative of exposure levels pre-diagnosis [132]. Measurements of risk factors are also subject to influence by both age and geography. The studied age range in the selected populations is very homogenous but encompasses most of adulthood which could be important for factors that may be strongly associated with age, such as BMI. Similarly in terms of geographical setting environmental and lifestyle factors also vary between populations. This is probably most pertinent for the studies considering UV exposure and other sun-related variables. However it may also be important for those studying chemical exposures where local restrictions, policies and usage patterns dictate the population exposure, and for diet which tends to be geographically heterogeneous.

Misclassified, noisy or biased information is likely to bias the risk estimate to the null [169], which may potentially explain the lack of strong findings in the included articles. All the studies used categorised variables, even when the underlying distribution was continuous. This is commonly done as it is much more straightforward to model and interpret joint effects, but it also loses information and introduces measurement error and further potential bias [209]. In contrast germline variation is static and discrete and therefore in theory can be measured without error [168,171]. However, there is still some potential for

bias when several polymorphisms in a gene contribute to altered function, or if the functional variants are not known and associations are identified through tagging SNPs [169]. Only Kane *et al*. [186] used direct sequencing and a number of the polymorphisms considered in the study are of unknown function. Due to these potential biases, which cannot be quantified all findings should be viewed with caution, although it should be noted that it has been shown that the presence of a biased main effect does not automatically imply a biased estimate of GxE [169].

As discussed, GxE studies are also problematic due to the issues of multiple testing which arises from models including multiple genes, multiple risk variables and multiple interactions. This is a particular issue in the selected studies, given that many of the articles were based within the same population, and some considered the same SNPs. Consequently even the within-paper FDRs, where these were calculated and applied, may not be sufficiently stringent. Furthermore it can be difficult to choose the optimal method of analysis, and to translate the statistical finding of an interaction into biology. Such limitations were touched on, in varying detail, in the majority of the included articles, although few sufficiently accounted for them in their conclusions.

2.11 Conclusions

Although the sequencing of the human genome has potentially been less successful than initially envisaged in terms of determining disease aetiology, it has provided the opportunity to explore environmental risk factors in the context of an individual's susceptibility through the study of GxE.

The 25 articles included here considered the joint effect of environmental or lifestyle factors with genetic variants in the risk of NHL. All the studies were similar in terms of design and methodology. In total, 18 of the studies reported that they found an interaction, and an even greater number concluded that their findings were of sufficient interest to warrant replication in larger studies. However, as discussed there are a number of ways of declaring a success in an interaction study, and the differing criteria employed by the different studies means that not all of the 'successes' can be considered equal.

This disparity is perhaps not unexpected given that, to date, there are few agreed guidelines or standards for the best way to conduct and report GxE. Interestingly none of the selected articles employ some of the more advanced statistical techniques available to explore interactions [207]. What is universally established is that to be confident that an observed interaction is a true effect it must be replicated in an independent population [169]. This was not the case for any of the findings presented here, which perhaps also is not unexpected given that are almost no examples of robust, replicated GxE for any disease. The best example is that between *NAT2*, smoking behaviour and bladder cancer risk [210], and a number of successes have also been noted in pharmacogenetics [169]. These interactions are characterised by genetic variants with clearly defined phenotypes and environmental variables that can be accurately and objectively measured, and it has been suggested that the study of GxE may only possible under such conditions [117]. The included studies here do not fulfil these criteria.

There has been some criticism of the popularity of GxE studies in the literature, which states that they are often not performed or interpreted correctly [211]. A statistical interaction does not necessarily correspond to a biological one. A biological interaction refers to an aspect of biological mechanism while a statistical interaction between two factors is a test of the fit of a particular model of joint action, and the extent of deviation from a model on a predefined scale [211]. Consequently the statistical model will almost always be a simplification of the underlying biology [209], and it has been argued that to date, a failure to grasp this has resulted in 'widespread over-interpretation of "interaction" in logistic regression models' [211]. Such an argument could be made for the included studies here. They used a simplistic model of interaction, in studies lacking sufficient power, yet the majority made fairly bold conclusions regarding their findings.

The ultimate goal of GxE studies can be said to be the development of individualised disease prevention or treatment [169]. The results from the studies to date are insufficient to do so. Although

some promising results are presented, which may help to inform or guide future studies and lines of investigation, until replication in an independent population can be demonstrated they cannot be considered definitive. There are a number of limitations and likely biases in the GxE studies available to date. The use of self-reported data, a candidate approach, dominant genetic models and simplistic interaction models, coupled with inadequate sample sizes, and a failure to fully account for multiple testing all affect the validity of the findings. Furthermore the use of different thresholds to determine a an interaction renders between study comparisons difficult, and among those studies investigating the same factors consistency was weak. Therefore, at worst these GxE studies may add to the confusion and uncertainty regarding NHL aetiology. Consequently it is recommended that future studies take a different approach to the consideration of GxE, and better reflect the ethnic and subtype heterogeneity which are likely to play an important role.

2.12 Recommendations

Although it is the approach chosen by the vast majority of the GxE studies for any disease, it could be argued that the main failing of the included studies is the candidate gene-approach taken. This is also the approach traditionally used for main genetic effect associations, but the success has been modest in these studies. This has been attributed to unfortunate choices of candidates and insufficiently stringent p-values. The same problems will therefore also plague GxE studies using this approach, and so other ways of identifying genes for study are required.

The advent of high-throughput genotyping technologies, has opened the possibility of exploring GxE at a genome wide level using an agnostic approach [171]. This approach is gaining popularity, but is currently limited by the lack of appropriate methods for modelling such a large number of interactions, by the extremely stringent significance thresholds that would be required for so many high-order interactions and by the fact that prior knowledge on biological processes and pathways is not taken into account [170]. It can be argued that in reality all disease arises from the interactions between our environment and our genome, and that therefore the benefit of GxE studies lies in determining the magnitude of these effects [209], identifying where they lie along the causal pathway, and how they relate to other casual pathways [171,209]. New methodologies which are able to deal with these issues are required.

This thesis will explore the use of one such methodology, profile regression, and apply it to GWAS and environmental data to determine whether it can offer new insights and understanding to the study of GxE, and in particular the roles such interactions may be playing in the aetiology of NHL.

3. METHODS

3.1 Study design

This thesis is based on participants from the European Prospective Investigation into Cancer (EPIC) cohort. Additionally, as part of the EnviroGenoMarkers (EGM) study a number of participants from the Northern Sweden Health and Disease study (NSHDS), have also been utilised in some analyses, where specified. NSHDS incorporates a proportion of EPIC-Sweden. The participants included in each chapter, and the crossover in participants between the chapters is described in table 3.1.2c at the end of this section.

3.1.1 EPIC

epic.iarc.fr

Ethics Statement

This thesis involves human genetic material, human biological samples and collection of personal data. All participants were adults and were healthy at recruitment. All the projects detailed in this thesis which include EPIC participants have been approved by ethical review boards of each participating centre and of the International Agency for research on Cancer (IARC). All participants provided written informed consent at recruitment using centre-specific forms which were administered in the local language, agreeing to have their health status followed for the rest of their lives.

Background

The European Prospective Investigation into Cancer (EPIC) is a multi-centre prospective cohort following over half a million healthy participants from middle age onwards. It was established to investigate the relationships between diet, nutritional status, lifestyle and environmental factors with the incidence of cancer. Detailed information on these variables in addition to anthropometric measurements and blood samples taken at recruitment is available for all EPIC participants. EPIC represents the largest study of diet and health ever undertaken and the goals of this project have now expanded to consider other risk factors and other chronic diseases [212].

Study subjects

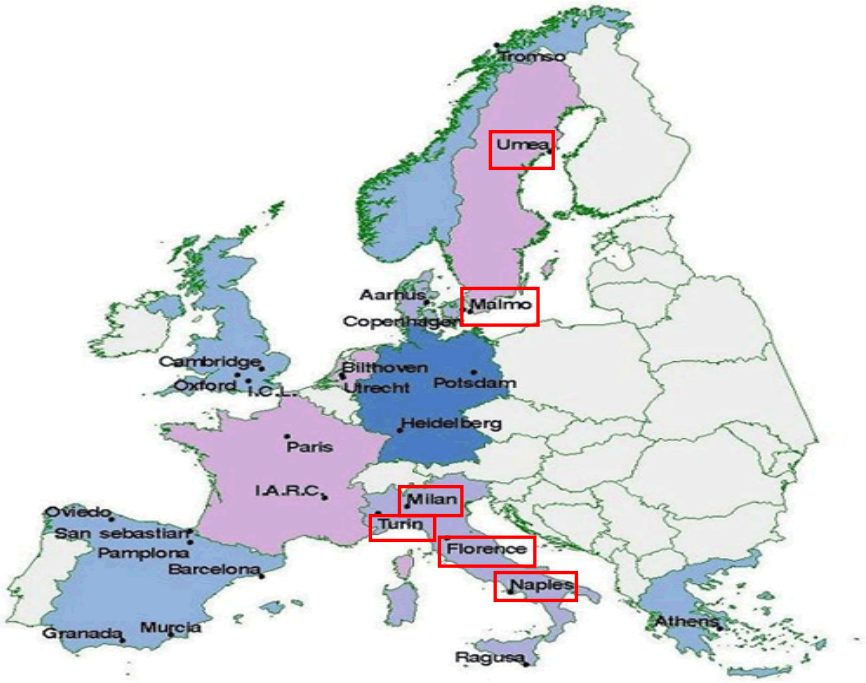
EPIC was originally initiated in 1992 in seven countries but it has now been expanded to include a total of 23 centres from ten European countries: Denmark, France, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and the United Kingdom (Figure 3.1.1.a). It is planned to follow-up the study participants for at least another ten years.

The choice of study population was country-specific determinant on practical considerations in that nation. Where follow up could be based on population cancer registries the general population was

targeted. Where cancer registries provide incomplete coverage, specific groups such as civil servants or members of health insurance plans were invited to participate. All countries included males and females aged 35-74 years with the exception of Norway, France and Utrecht (one of the two centres in the Netherlands) which included only females. Therefore women were over represented in the total epic cohort, accounting for ~70% of participants. No French cases were included in this thesis due to incomplete coding for lymphoid neoplasms at this time. Forty two percent of subjects were aged 49 years or younger at recruitment and more than 90% of participants were younger than 65 years. Potential study subjects were contacted via letter, phone or by personal contact dependent on the country.

Figure 3.1.1a: Map of the EPIC Collaborating Centres

The centres included in the EnviroGenoMarkers Project are indicated with a red box



Adapted from; http://epic.iarc.fr/centers/epicmap.php [213]

Questionnaire and dietary data and anthropometric measurements

Full details can be found in [214]. In brief, a standardized self-completed questionnaire was used to collect information on history of tobacco smoking, history of alcohol consumption, occupation, socioeconomic status, age at puberty and reproductive history, use of contraception and of hormonal drugs, history of previous illnesses and current physical activity. Height (cm), seated height (cm), weight (kg) and waist and hip circumference (cm) were measured using standardised methods and quality control checks for between- and within- observer variability.

To assess diet a 24-hour diet recall questionnaire was administered on 12 days over the course of the year. A built-in calibration study was also used based on precise measurements of single day's food intake in a subsample of 2% of the study population.

Blood samples

Blood samples were extracted in the centres at recruitment. Fourteen millilitres of blood was collected and aliquoted into plasma, serum, white blood cells and erythrocytes. They were stored in 0.5ml plastic straws in liquid nitrogen containers at -196°C.

Follow up for incident cancer cases

Cancer incidence data is obtained through population cancer registries, health insurance programs or though periodic personal contact dependent on the centre. Where possible, clinical data on anatomical localisation of the tumour, histology grade and stage has been collected for every cancer. Information on mortality is collected by record linkage with death certificate, health insurance or national health service files.

Lymphoma within EPIC

A total of 2116 incident cases of lymphoma have been identified to date of which 1921 have validated subtype information (table 3.1.1). Hodgkin Lymphomas (HL) comprise 94 cases. Of the Non-Hodgkin's Lymphomas 46 are of T-cell origin and 1757 are of B-cell origin. The diagnosis of cases in EPIC was initially based on the International Classification of Diseases for Oncology, second edition (ICD-O-2). These cases were then reclassified upon the publication in 2008 of the WHO document on classification of tumours of the hematopoietic and lymphoid tissues [16] by applying a conversion program available on the Surveillance, Epidemiology, and End Results Web page (http://seer.cancer.gov/tools/conversion/ICDO2-3manual.pdf) [215]. Those that could not be reclassified were termed "not otherwise specified" (NOS).

Multiple myeloma was the most common malignancy in this cohort accounting for 21% of cases (table 3.1.1), followed by CLL (17%), DLBCL (13%) and FL (12%). DLBCL and FL are commonly cited as the two most common NHL, however these estimates do not include MM which is recognised as the second most common blood cancer after NHL. For the purposes of this thesis MM is included as an NHL as discussed. Of the 1757 B-cell NHL cases 1524 (86.7%) had an available blood sample and could be matched to two controls who were lymphoma-free at the time of diagnosis in the case. Cases and controls were matched on study centre, age (+/- 1 year), sex and blood collection date (+/- 45 days). The characteristics of the eligible (n=1524) and non-eligible (n=233) participants are shown in appendix table 3.1.1. The non-eligible participants were more likely to be female and were slightly younger than the eligible participants. There were also some differences in country of origin, with a large percentage of Norwegians among the non-eligible cases.

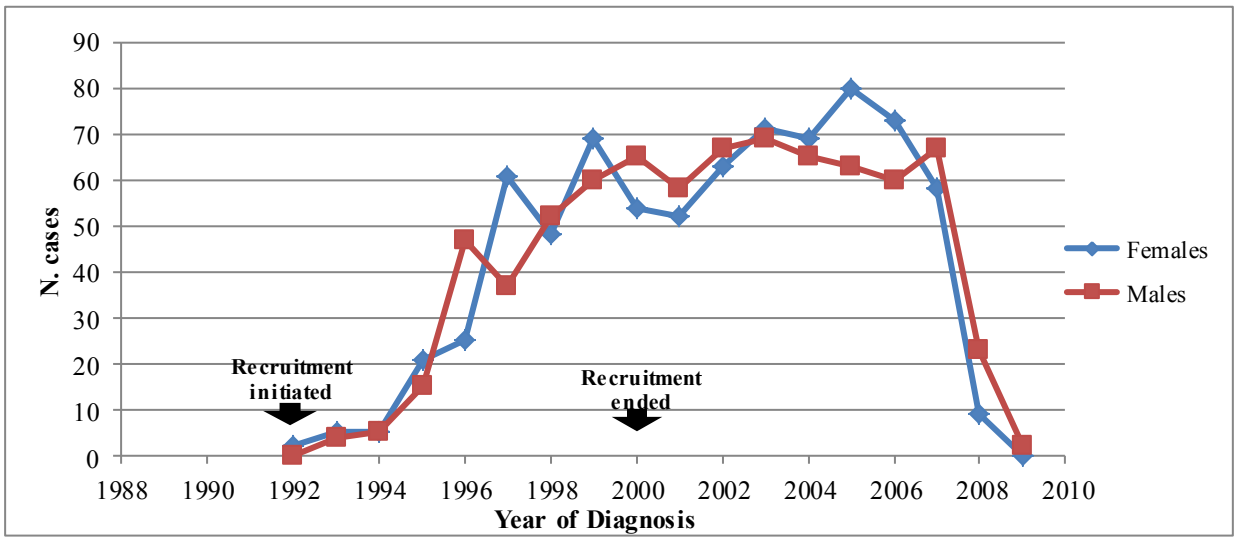
Despite the overall higher prevalence of NHL in males worldwide, due to the overrepresentation of females in the EPIC cohort of the 1524 cases 765 (50.2%) were female and 759 (49.8%) were male. The age distribution and the year of diagnosis were very similar between males and females. CLL, LL and HCL were more common in males, while FL and MZL were more common in females in this cohort (figures 3.1.1b ad c).

Table 3.1.1: Subtype Classification of 2116 Incident Lymphoma cases diagnosed within the EPIC cohort during follow-up

Subtype	*n.*	%
Non-Hodgkin's Lymphoma		
B-cell		
Multiple Myeloma	452	21.4%
Chronic Lymphatic Leukaemia	359	17.0%
Diffuse Large B-cell Lymphoma	284	13.4%
Follicular Lymphoma	243	11.5%
Lymphoplasmocytic Lymphoma	69	3.3%
Marginal Zone B-cell Lymphoma	22	1.0%
Hairy Cell Leukaemia	22	1.0%
B-cell Acute Lymphatic Leukaemia	21	1.0%
B-cell Lymphoma, Not Otherwise Specified	285	13.5%
T-cell		
Mycosis Fungoides	24	1.1%
T-cell Acute Lymphoblastic Leukaemia	3	0.1%
T-cell Lymphoma, Other	43	2.0%
Hodgkin's Lymphoma	94	4.4%
Histological Information Not Available	195	9.2%
Total	**2116**	100.0%

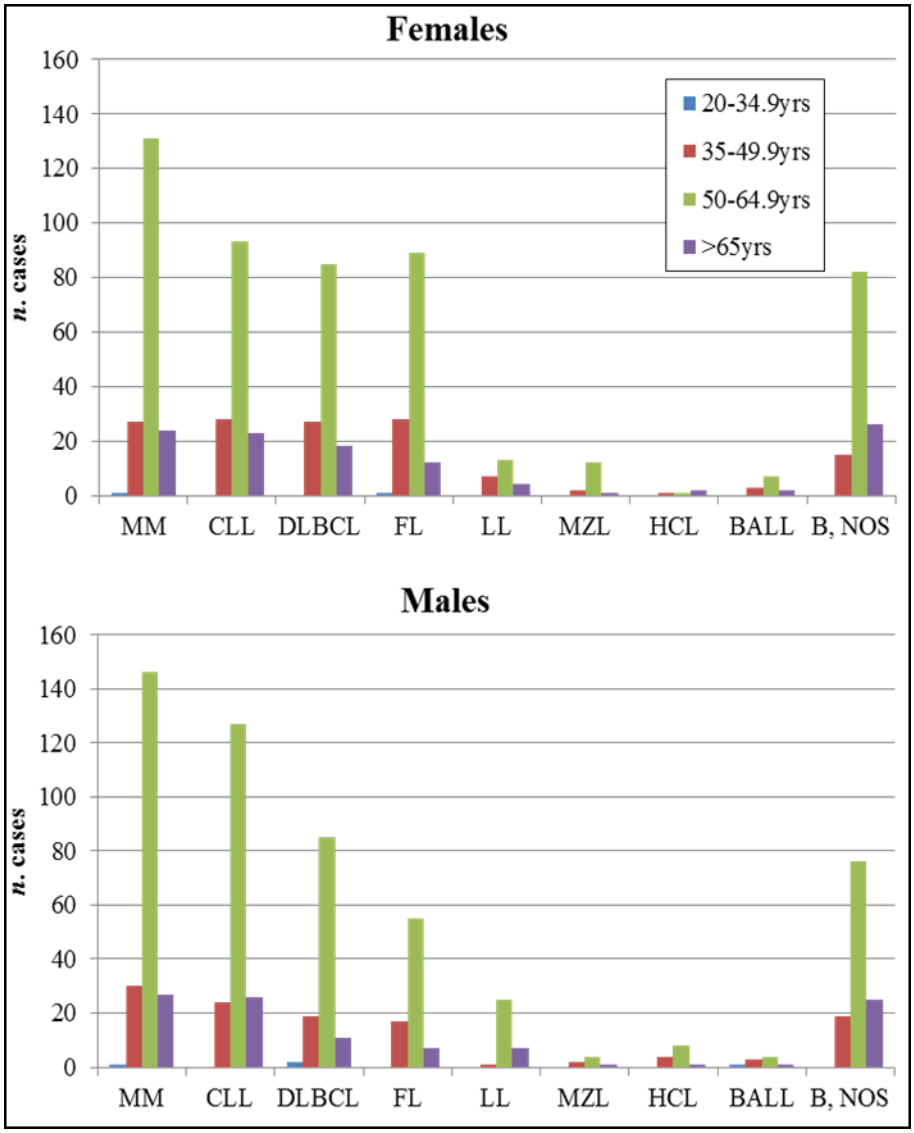
This does not include any cases from EPIC-France, who were not involved in the Lymphoma working group

Figure 3.1.1b Year of Diagnosis for 1524 incident NHL cases who had an available blood sample and could be matched to two lymphoma-free controls: Stratified by sex



Exact year of initiation and end of recruitment between 1992 and 2000 was dependent on centre

Figure 3.1.1c Number of Cases by Subtype and Age group in 1524 incident NHL cases who had an available blood sample and could be matched to two lymphoma-free controls: Stratified by sex



MM- Multiple myeloma, CLL – Chronic lymphatic leukaemia, DLBCL-Diffuse Large B-cell lymphoma, FL- Follicular lymphoma, LL-Lymphoplasmacytic lymphoma, MZL – Marginal Zone lymphoma, HCL – Hairy cell leukaemia, BALL- B-cell acute lymphatic leukaemia, B, NOS – B-cell NHL not otherwise specified.

3.1.2 The EnviroGenoMarkers study

www.envirogenomarkers.net

Ethics Statement

This study was approved by the committees on research ethics in Umea (Dnr 08-215M) and in Florence (ref 347/2009) in accordance with the Declaration of Helsinki of the World Medical Association. All participants provided written consent at recruitment agreeing to provide detailed

information on their dietary and lifestyle habits at recruitment and to have their health status followed for the rest of their lives.

Background

The EnviroGenoMarkers (EGM) Study [216] is a nested case-control study aiming at the development of a new generation of biomarkers to better characterise the relationship between exposure to environmental pollutants and adverse health effects.

Study Subjects

The EnviroGenoMarkers study is based on participants from two existing prospective cohort studies: EPIC-Italy and NSHDS (figure 3.1.1.3). All participants had blood samples collected prospectively at enrolment.

EPIC-Italy [217] is a subset of EPIC. A total of 47,749 participants aged 35-70 years were enrolled in five participating centres across Italy: Varese, Florence, Turin, Naples and Ragusa, between 1993 and 1998. Standardised procedures were used to identify newly diagnosed cases of cancer based on automated linkages to cancer and mortality registries, municipal population offices and hospital discharge systems. The only exception was Naples, where follow-up information was collected through periodic personal contact.

The NSHDS [218] includes participants from the Västerbotten Intervention program, the Västerbotten Mammary Screening Program and the Northern Sweden MONICA project. A total of 95,000 healthy individuals aged 40-60 were invited for inclusion in the project between 1990 and 2006. Subjects were asked to complete a self-administered questionnaire to collect demographic, medical and lifestyle information and a separate self-administered food frequency questionnaire at recruitment. Invasive cancers occurring among cohort members during the study period were identified by linkage with the Swedish Cancer Registry and the local Northern Sweden Cancer Registry. A proportion of the NSHDS participants included in the EGM study (61 cases and 56 controls) were also participants in the Swedish subset of EPIC.

Selection of cases and controls

Lymphoma cases were classified into subtypes according to the SEER ICD-0-3 morphology codes [21]. Incident cases occurring at least two years after blood draw were identified and a suitable cancer-free control was selected by incidence matching from the remaining populations for each case matched on sex, age (+/- 2.5 years), centre and date of blood collection (+/- 6 months). More than 95% of participants also had the same fasting status as their matched pair at blood extraction. Information from the two studies was integrated into a single database and calibrated. The baseline characteristics of cases and controls are described in table 3.1.2a. The time between sample extraction and diagnosis is shown in table 3.1.2b. The majority of participants were from Sweden (69%). There were no

significant differences between cases and controls in terms of baseline factors. Despite the fact that the incidence of NHL is greater in males than females, the population selected for this study included a slightly higher proportion of females than males. The mean age at recruitment was 53 years and the mean time to diagnosis was 73 months. This was similar between males and females, but longer in Sweden than in Italy, as the NSHDS included more recent follow up (up to 2013). The diagnosis by subtype is shown in figure 3.1.2, the most common diagnosis in all subgroups was MM. Roughly a quarter of the cases from NSHDS could not be assigned a subtype (BNOS).

To reduce the impact of technically-induced variation, matched pairs were analysed in the same batch on the same day. As matching was performed by sample date each case therefore had the same storage time as its matched control.

Table 3.1.2a: Baseline characteristics of cases and controls in the EnviroGenoMarkers study

Baseline variable		Cases (*n*=270) [*n* (%)]	Controls (*n*=270) [*n* (%)]	Difference (p-value)[α]
Cohort	**EPIC-Italy**	84 (31.1)	84 (31.1)	
	NSHDS	186 (68.9)	186 (68.9)	
Phase[∞]	**One**	100 (37.0)	100 (37.0)	
	Two	170 (63.0)	170 (63.0)	
Sex	**Male**	133 (49.3)	133 (49.3)	
	Female	137 (50.7)	137 (50.7)	
Mean Age	**(yrs)**	53.08	53.09	0.989
Mean Height	**(cm)**	169.55	168.24	0.120
Mean Weight	**(kg)**	76.48	75.09	0.267
BMI	**Underweight (<18.5 kg/m^2)**	1 (0.4)	0 (0.0)	
	Normal (18.5-24.9 kg/m^2)	104 (38.5)	109 (40.4)	
	Overweight (25-29.9 kg/m^2)	121 (44.8)	118 (43.7)	
	Obese (≥30 kg/m^2)	40 (14.8)	39 (14.4)	
	Unknown	4 (1.5)	4 (1.5)	0.883
Smoking status	**Never**	121 (44.8)	134 (49.6)	
	Former	90 (33.3)	72 (26.7)	
	Current	57 (21.1)	54 (20.0)	
	Unknown	2 (0.7)	10 (3.7)	0.269
Highest educational level	**None**	4 (1.5)	1 (0.4)	
	Primary	94 (34.8)	98 (36.3)	
	Technical/professional	68 (25.2)	54 (20.0)	
	Secondary	52 (19.3)	64 (23.7)	
	University/college	47 (17.4)	42 (15.6)	
	Unknown	5 (1.9)	11 (4.1)	0.202
Cambridge physical activity index	**Inactive**	80 (29.6)	75 (27.8)	
	Moderately inactive	106 (39.3)	95 (35.2)	
	Moderately active	69 (25.6)	76 (28.2)	
	Active	14 (5.2)	23 (8.5)	
	Unknown	1 (0.4)	1 (0.4)	0.510
Mean time to diagnosis	**(years)**	6.1 (range: 2.0, 16.0)		

[α] P-value for difference was calculated using the chi-squared test for categorical baseline variables and the student's t-test for continuous variables

[∞]In phase 1 50 case-control pairs were from NSHDS and 50 were from EPIC-Italy. In phase 2 136 pairs were from NSHDS and 34 were from EPIC-Italy

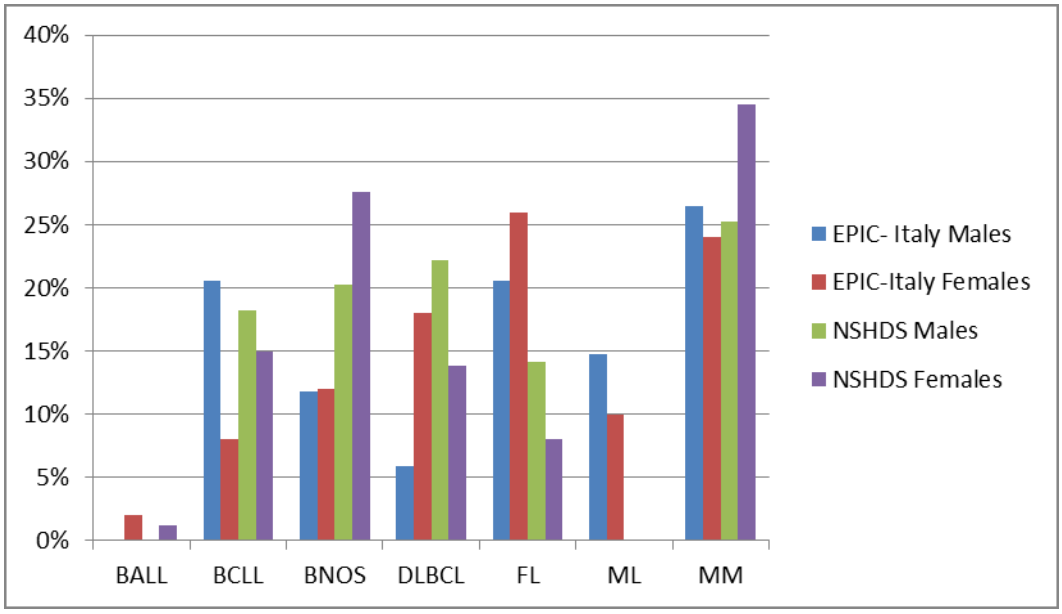
Table 3.1.2b; Time between sample extraction and NHL diagnosis in cases from the EnviroGenoMarkers study; mean median, range and number of cases by 5 year interval

Population	Median (months)	Mean (months)	Range	*n*. ttd<5 years (<60 months)	*n*. ttd ≥ 5 years (≥ 60 months)
Total	70.13	73.41	(24.03, 191.08)	121	149
Sweden	74.52	78.43	(24.03, 191.08)	70	116
Italy	54.62	60.25	(24.33, 119.15)	51	33
Men	66.16	73.70	(24.33, 191.08)	62	71
Women	71.74	73.11	(24.03, 190.92)	59	78

ttd – time to diagnosis

Figure 3.1.2 Classification of Included Incident B-cell NHL cases in the EnviroGenoMarkers Study: Stratified by Cohort and Sex

Percentage of the total cases classified into each subtype



BALL- B-cell Acute Lymphatic Leukemia , BCLL- B-cell Chronic Lymphocytic Leukaemia, BNOS - B-cell NHL Not Otherwise Specified, DLBCL - Diffuse Large B-cell Lymphoma, FL – Follicular Lymphoma, ML - Marginal Zone B-cell Lymphoma, MM- Multiple Myeloma

Schematic

The study was conducted in two phases. An initial pilot phase (phase 1) included 100 matched case control pairs (50 pairs from EPIC-Italy and 50 from NSHDS). Upon successful completion a second

roll-out phase (phase 2) was implemented including 34 case-control pairs from EPIC-Italy and 136 case-control pairs from NSHDS.

Table 3.1.2c Origin and number of included participants for each results chapter (4-8), and the between chapter crossover

POPULATION	CHAPTER	PARENT STUDY	DATA	*n.* participants	Cross Over between the Chapters (*n.*)				
					FOUR	FIVE	SIX	SEVEN	EIGHT
CASES									
	Four	EPIC	Genotyping performed as part of the InterLymph Consortium	**175**	**175**				
	Five	EGM	POPs and Heavy Metal Body Burden	**270**	25	**270**			
	Six	EPIC	Metabolomic Profiling	**210**	10	210	**210**		
	Seven	EPIC	Translocation Frequency and Prevalence	**165**	44	34	24	**165**	
	Eight	EGM	POPs and Heavy Metal Body Burden plus Metabolomic Profiling	**210**	10	210	210	24	**210**
CONTROLS									
	Four	EPIC	Genotyping performed as part of the InterLymph Consortium	**263**	**263**				
	Five	EGM	POPs and Heavy Metal Body Burden	**270**	53	**270**			
	Six	EPIC	Metabolomic Profiling	**210**	16	210	**210**		
	Seven	EPIC	Translocation Frequency and Prevalence	**336**	45	11	3	**336**	
	Eight	EGM	POPs and Heavy Metal Body Burden plus Metabolomic Profiling	**210**	16	210	210	3	**210**
TOTAL									
	Four	EPIC	Genotyping performed as part of the InterLymph Consortium	**438**	**438**				
	Five	EGM	POPs and Heavy Metal Body Burden	**540**	78	**540**			
	Six	EPIC	Metabolomic Profiling	**420**	26	420	**420**		
	Seven	EPIC	Translocation Frequency and Prevalence	**511**	89	45	27	**511**	
	Eight	EGM	POPs and Heavy Metal Body Burden plus Metabolomic Profiling	**420**	26	420	420	27	**420**

This table includes only those participants included in the final analysis (ie. excluded participants are not shown)

EGM incorporates both a subset of EPIC-Italy and a subset of the NSHDS

3.2 Technical (laboratory) methodologies

The following technical methodologies were conducted by other individuals, as specified.

3.2.1 Exposure assessment

Exposure to metals was assessed by measurements in stored erythrocytes, and exposure to persistent organic pollutants (POPs) was assessed using blood serum. Blood samples were collected without RNA preservative in citrate (EPIC-Italy) or heparin (NSHDS) tubes and processed by centrifugation on the same day of collection (EPIC-Italy) and within one hour of collection (NSHDS). Aliquots (0.5ml) were stored in liquid nitrogen tanks at -196°C in Italy and -80°C in NSHDS.

Cadmium and Lead

Included in chapters five and eight. Conducted as part of the EGM study by Thomas Lundh, Department of Occupational and Environmental Medicine, Lund University Hospital, Sweden.

Erythrocyte concentrations of Cadmium (Cd) and Lead (Pb) were determined by inductively coupled plasma-mass spectrometry (ICP-MS; Thermo X7, Thermo Elemental, Winsford, UK) in samples diluted with an alkaline solution according to Barany *et al.*[219]. The detection limit was calculated as 0.03 and 0.09 µg/L for Cd and Pb respectively based on the standard deviation (SD) of the blank multiplied by three. The analytical accuracy was checked against human blood reference material from Centre de Toxicologie du Quebec, International Comparison Program, Canada. The results (mean ± SD) obtained were for Cd (Lot. C0616 and C0912) 0.98 ± 0.04 and 5.0 ± 0.18 µg/L (n = 69) vs. recommended 1.0 ± 0.13 and 5.3 ± 0.43 µg/L, respectively and for Pb (Lot. L0909 and L0807) 23 ± 0.043 and 112 ± 3.9 µg/L (n = 69) vs. recommended 23 ± 1.1 and 112 ± 9.9 µg/L, respectively. All samples were prepared in duplicate and the method imprecision (calculated as the coefficient of variation for duplicate preparation measurements) was 5 and 3% for Cd and Pb respectively [216].

PCBs, DDE, HCB

Included in chapters five and eight. Conducted as part of the EGM study by Hannu Kiviranta, Chemical Exposure Unit, National Institute for Health and Welfare, Kuopio, Finland

The concentrations of the POPs (PCBs, DDE, DDT, HCB and BDE-47) in blood serum were determined using gas chromatography - mass spectrometry (GC-MS) analysis.

The quantification of POPs was performed by an Agilent 6890 gas chromatograph (GC) connected to a Waters Autospec Ultima high resolution mass spectrometer (HRMS). The GC column used was DB-5MS (J&W Scientific, 30m, ID 0.25 mm, 0.25 µm), and the splitless mode was used for injection. For the phase 2 EPIC-Italy samples a slightly different procedure was used involving an Agilent 7000B gas chromatography triple quadruple mass spectrometry (GC-MS/MS) system, with a DB-

5MS UI (J&W Scientific, 20m, ID 0.18 mm, 0.18 µm) GC column and the solvent vent mode. For quality control purposes in each batch of samples (43 batches with HRMS, 20 batches with MS/MS) 2 reagent blanks were additionally prepared and the average result of the blank samples subtracted from the results of the real samples. Furthermore two control samples of Standard Reference Material 1589a (PCBs, Pesticides, PBDEs, Dioxins/Furans in Human Serum) from the National Institute of Standards and Technology, were also included in each batch.

3.2.2 Metabolomic Profiling

Included in chapters six and eight. Conducted as part of the EGM study by Alexis Siskos and Anas Kamleh in the laboratory of Hector Keun, Department of Surgery and Cancer, imperial College London

Sample preparation

50 µL aliquots of plasma were thawed on ice then diluted with 50 µl of UPLC (Ultra-High Performance Liquid Chromatography) grade water (Romil LTD, Code H949, Cambridge , UK) and treated for protein precipitation using 300 µL of cold methanol (MeOH was pre-chilled at -20 C for 1 h). Samples were vortexed and kept in -80 C for a complete precipitation then centrifuged at 13000 rpm for 10 minutes. The supernatant was carefully removed to a clean eppendorf which was stored in -40 C until analysis day. A QC sample representing all samples was prepared by pooling 25 µl of each sample in a falcon tube. The pooled QC was aliquoted in 1.0 ml portion and stored at -40 C. A blank sample was prepared by replacing the plasma with 50 µl of pure water. On analysis day, 35 µl of each sample was added to 25 µl of an aqueous cocktail of internal standards (Hippuric acid-[D_2], Phenylalanine [^{13}C], Methionine [^{13}C] and Acetylcarnitine [D_3], Tryptamine-aplha,alpha,beta,beta-[D_4]).

All blanks, QCs and samples were loaded, in high recovery chromatography vials (Waters Corporation, Milford, MA, USA) into the UPLC-MS auto-sampler. Samples were run with UPLC-MS as matched pairs of cases and control, and QC was run every six injections (3 case/control pairs).

UPLC-MS experiment

Reversed-phased chromatographic separation of the plasma sample preparation was conducted using an Acquity UPLC system (Waters Corporation, Milford, MA, USA) on an Acquity HSS T3 C_{18} column 10mm x 2.1mm, 1.8 um (Waters) and a binary gradient elution comprising water + 0.1% formic acid (Sigma) and acetonitrile + 0.1% FA . Mass spectrometric analysis of the chromatographic eluent was performed using a quadrupole time-of-flight (QtoF-Ultima) spectrometer (Waters), with data collected in centroid mode in the m/z range 70-1000. Analysis was performed using positive ion mode electrospray ionization.

Data processing

Mass spectral profile data was converted to NetCDF format using the DataBridge software (Waters) before processing using the XCMS software package [220] running in the R computing environment. XCMS data processing typically comprised of peak picking (CentWave method, peakwidth=c(3,30),ppm=15,snthresh=12.5), peak grouping (group density, mzwid=0.07,bw=20), retention time alignment (rector method, f="s",p="m") and feature intensity filling and reporting (integrated peak intensity). The pooled QC samples, injected at intervals throughout the analytical run were used to normalise across batches and also to assess the quality of metabolite features.

Due to the untargeted nature of the methodology and the use of LCMS, only a proportion of metabolite features have been annotated. Unambiguous assignment requires the determination of retention time. This can only be determined by running authentic standards of the putative metabolites corresponding to the metabolite features and comparing the MS/MS fragmentation data. Each metabolite feature could potentially correspond to multiple metabolites, therefore to obtain the authentic standards and run the analyses for every possible metabolite for all the features identified in this project would take many years and was not possible during the scope of this thesis.

3.2.3 Screening for the t(14;18) translocation

Included in chapter seven. Conducted on a subset of the EPIC cohort by Bertrand Nadal, Sandrine Roulland and colleagues at CNRS-INSERM, Université de la Méditerranée in Marseille, France

DNA isolation

DNA samples were extracted from buffy coats using the QIAsymphony DNA Midi Kit or the QIAamp DNA FFPE tissue kit for Formalin-fixed paraffin embedded (FFPE) tumor biopsies (Qiagen, Crawley, UK).

Screening assay

Quantitative real-time PCR (Q-PCR) was conducted using an ABIPRISM 7500 system and Sequence Detection software (Applied Biosystems). The laboratory was blind to the case-control status of the samples. t(14;18) frequency within the BCL2 major breakpoint region (MBR) was measured using buffy coat-extracted DNA samples. Three PCR replicates were performed with at least 0.5-1 µg DNA for t(14;18), two replicates for the reference *GAPDH* gene and two for control cell lines. The absolute frequency of lymphocytes carrying the translocation was calculated using the standard curve method and normalized to the number of total lymphocytes in the samples. Frequency was defined as the number of circulating t(14;18) copies per 1,000,000 circulating lymphocytes. Prevalence ($t(14:18)^{+}$) was defined as more than 1 cell bearing the translocation per 1,000,000 circulating lymphocytes (≥1x10-6).

3.2.4 Genotyping

Included in chapter four. Conducted as part of the InterLymph NHL GWAS in a subset of the EPIC cohort at the National Cancer Institute's Cancer Genomics Research Laboratory, Bethesda, Maryland, USA

Genotyping was performed using Illumina OmniExpress at the NCI covering 71,5530 SNPs. Genotypes were called using Illumina GenomeStudio software, and quality-control duplicates (for the entire InterLymph GWAS) showed >99% concordance. Further details are provided in [87], and details of the quality control criteria applied to the EPIC population for this study are provided in chapter four.

3.3 Statistical methodologies

The statistical analysis plan employed for each chapter is outlined in the individual methods sections. Overarching statistical methodologies applied in more than one chapter are described here. All analyses were run using STATA 12.1 (Stata Corporation, College Station, TX), R version 3.0.0 or PLINK version 1.07 as specified. All tests are two tailed and the level of statistical significance was set at 0.05. However the potential for spurious results arising from multiple testing was also considered and where alternative significance levels were used these are indicated.

3.3.1 Unconditional and conditional logistic regression

The vast majority of association analyses within this thesis were based on a nested case-control design and so effect estimates were calculated as odds ratios using a logistic regression model. Logistic regression models the relationship between a dependent variable, in this case incidence of NHL and one or more independent variables which are chapter specific, using a generalized linear model with a binomial response and logit link function

The basic model considers occurrence of NHL as the binary dependent variable, the potential (chapter-specific) biomarker of interest as an independent variable and adjusts for the case-control study matching factors: age (+/- 1 year), sex, country and date of blood draw (+/- six months). In analyses based on the EGM study the basic model additionally includes study phase and processing batch.

As the parent study populations used a matched design where possible conditional logistic regression was utilised, which conditions out the matching factors from the analysis in order to eliminate unwanted nuisance parameters from the estimate. In practice, on the same data the basic and the conditional model will show similar odds ratios.

Additional included confounders for both the basic and the conditional model are described in the relevant chapters.

3.3.2 Association analysis for GWAS data

All genome-wide association analyses were conducted using PLINK version 1.07 utilizing all SNPs that passed the quality control thresholds (as detailed in chapter 4).

The association command *logistic*, which compares allele frequencies between cases and controls, was used to test the association with the defined phenotypes (NHL or translocation) adjusting for a defined set of covariates specified in a separate text file and with the *ci95* option to obtain 95% confidence intervals (95% CI) for the odds ratios. This command additionally provides the allele frequencies in cases and controls.

The *logistic* command reports a log-additive model so the resulting odds ratio describes the increase or decrease in risk associated with each extra copy of the variant (minor) allele.

3.3.3 Receiver Operator Characteristic (ROC) curve analysis

ROC curves are created by plotting the false-positive rate (1-specificity) against the true-positive rate (sensitivity) for a continuous predictor and a binary outcome.

In this thesis, ROC analysis was performed using three commands in STATA: *lroc* which plots the graph and computes the area under the curve (AUC, also known as the concordance index), *roctab* which provides the nonparametric estimation of the ROC curve and the asymptotic normal confidence intervals for the AUC, and *roccomp* which allows statistical comparison of multiple ROC curves.

3.3.4 Meta-analysis

Where multiple subgroups were compared heterogeneity in effect estimates between the groups was assessed using meta-analyses. The *metan* command in STATA was used to calculate the I^2 inconsistency matrix as a measure of the amount of heterogeneity not due to chance, and the corresponding p-value. The default fixed effects model was used, to account for the fact that the compared groups all came from the same parent population.

3.3.5 Profile regression

Profile regression is a novel Bayesian methodology specifically developed to analyse complex datasets in which multiple risk factors are likely to interact in the aetiology of common multifactorial diseases [163-165].

The benefits of profile regression include its ability to identify effects which are too small to be picked up as significant by standard logistic regression, and additionally to show the joint effect of covariates combined. In contrast to standard methodologies, it works better the more covariates there are, and the more correlated they are. Profile regression can handle over 1000 covariates and can include potential confounders as fixed effects in its estimation.

Profile regression acts to reduce dimensionality by using a profile formed of a sequence of covariate values, rather than individual risk factors, as its basic unit of inference. The covariate profile is used to cluster individuals into subpopulations, so that clusters are coherent with respect to risk factors among observations within clusters and dissimilar with respect to risk factors between clusters.

Clusters are determined via an assignment submodel, which evaluates the probability that an individual is assigned to particular cluster. Simultaneously a disease submodel links the profiles to an outcome of interest via an estimated association parameter. Both submodels are fitted jointly, in a

unified manner using Bayesian inference techniques and Monte Carlo Markov Chain (MCMC) sampling methods, so the allocation of individuals to clusters depends on both the covariate data in the assignment submodel and the outcome information in the disease submodel.

The MCMC method outputs a different clustering of data at each iteration of the sampler, with flexible clustering using the Dirichlet process. In Bayesian non-parametric models of data the prior and posterior distributions are not parametric distributions, but stochastic processes such as Dirichlet. The Dirichlet process is a probability distribution whose domain is itself a set of probability distributions which govern the relative chances of different outcomes. Therefore a draw from a Dirichlet process will return a theoretically infinite sequence of random variables. This makes it extremely flexible and an ideal candidate for clustering problems where the distinct number of clusters is unknown beforehand. In profile regression it allows the model to start with more clusters than needed and for the number of clusters to vary [221]. The best partitioning of the data i.e. the optimal number of clusters, is determined based on a score matrix, which calculates the percentage of times two people were in the same cluster [222]. The output of the MCMC sampler is exploited to assess parameter uncertainty and to compute the risk associated with specified profiles of interest. Each covariate is assigned a continuous selection weight (ρ) relating to its importance in the formation of clusters, relative to the outcome and all other covariates.

The script as run in R is shown in the appendix (Methods 3.4.4), using the PReMiuM package [223].

3.3.6 The *"meet-in-the-middle"* approach

The *"meet-in-the middle" (MITM)* approach [7] has previously been successfully applied to the study of diet and colon cancer using $^{(1)}$H NMR profiles from plasma samples collected before disease onset [224]. The principle is to use a nested case-control design to combine a retrospective search for biomarkers of disease development with a prospective search for associations with these biomarkers to past environmental exposures.

The *MITM* approach will be applied when combining biomarkers, and will be achieved using 3 steps to identify putative intermediate biomarkers:

STEP A: From disease to exposure

The 'top' (as defined by a biomarker specific threshold and significance level) disease-biomarker associations will be identified.

STEP B: From exposure to disease

The 'top' exposure-biomarker associations will be identified

STEP C: Parallel analysis

Biomarkers identified in both Step A and Step B (i.e. the common biomarkers) will be taken forward for further analyses to consider biological plausibility including pathway analysis and performance against biomarker criteria if appropriate.

3.4 Pathway analysis

SNP were mapped to genes using three different web-based programs (default settings):

Brain array SNP function Portal [225]
(http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx)
BRISK (Biology related Information Storage Kit) SNP to gene function
(http://genapha.icapture.ubc.ca/brisk/snpGene.do)
pfSNP: Potentially Functional SNP Search Engine [226]
(http://pfs.nus.edu.sg)

Genes which were identified by all three databases were then uploaded to the Database for Annotation, Visualization and Integrated Discovery (DAVID v6.7, http://david.abcc.ncifcrf.gov/) [227] for functional annotation. All available functional categories were considered, including Gene_Ontology, Pathways, and Protein_Interactions. A functional annotation chart was used to identify the most important pathways using defined criteria; the EASE (Expression Analysis Schematic Explorer) score was set to less than 0.01 and the minimum number of genes per group to 5. Pathways were considered of importance if their Bonferroni corrected p-value was less than $10e^{-3}$ and their fold enrichment greater than 1.5. These are established default values for this methodology.

3.5 Biomarker assessment

Although there exist a number of guidelines for the reporting of biomarkers including REMARK [228] and STROBE-ME [8]. There is less formal agreement on the best way of actually evaluating their utility in an epidemiological manner, and any rules that exist tend to be underdeveloped and/or not routinely applied [229].

Estimates of sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) are important measures of biomarker utility [230]. However, the most objective and statistically valid method for the evaluation of biomarker prediction is the use of ROC curves. The corresponding AUC can be interpreted as the probability that a classifier (in this case, biomarker) will rank a randomly chosen positive instance higher than a randomly chosen negative one; an AUC of 0.5 therefore is interpreted as having no practical utility [230]. In this thesis, these measures will be reported as standard where possible, in addition to adhering to the reporting guidelines as outlined by REMARK and STROBE-ME (the full guidelines can be found at [228] and [8] respectively).

As it will not always be possible to calculate AUC, sensitivity, specificity, PPV and NPV, and to enable a more holistic evaluation of biomarkers, a number of other guidelines and criteria will also be employed (Table 3.5). Standard epidemiological causation employs Bradford Hill's nine causation criteria [231], which can provide evidence on whether biomarkers lie along the causal pathway. Nevertheless, as biomarkers are not necessarily part of a causal association, the "Venice Criteria" [232] will additionally be employed. These criteria were developed for genetic studies as an alternative to the Bradford Hill Criteria, to reflect the fact that many of these guidelines are considered inappropriate for genetic associations, but they can also be well applied to molecular epidemiology studies. Finally, the themes addressed by Ransohoff in his paper 'Rule of evidence for cancer molecular-marker discovery and validation' [229] will be discussed. There is some crossover between these guidelines, particularly the importance of replicating findings and of accounting for potential bias and confounding.

The criteria for a successful screening test will also be discussed in this thesis and these are outlined in appendix box 3.5 [233].

Table 3.5; Three different sets of criteria for assessing biomarker utility and determining their association with an outcome of interest

Bradford Hill's Causation Criteria [231]	Venice Criteria [232]	Ransohoff's Rules of evidence for cancer molecular-marker discovery and validation [229]
Consistency/replication of the finding	Replication	Assessment of the reproducibility of the finding in an independent data set
Strength of the Association	Amount of Evidence	Protection from overfitting
Specificity of an association	Protection of the association from bias	Protection from bias
Temporality of exposure and effect		Protection from confounding
Analogy/consideration of alternative explanations		Generalisabilty
Biological gradient/dose response effect		Clinical usefulness
Biological plausibility of association		Consideration of cost vs. effort and benefit vs. harm
Coherence with existing knowledge		
Experimental evidence		

PART 1: VALIDATING BIOMARKERS

4.PROFILE REGRESSION AS A TOOL FOR IDENTIFYING SUSCEPTIBILTY BIOMARKERS FROM GENETIC PROFILES

BIOMARKER: SNP sets

BIOMARKER TYPE: Biomarker of Susceptibility

POPULATION: A subset of cases and controls from the EPIC cohort genotyped as part of the InterLymph consortium

4.1 Introduction

There is a wealth of data to support a genetic component to lymphoma aetiology [12], as evidenced by the increased risk observed for individuals with a family history and results from various genetic studies of both germline and somatic mutations. A number of compelling findings have emerged from studies of lymphoma tumours, however, in order to identify genetic biomarkers of susceptibility which can be used in prediction, constitutional genetics are of more interest. Initial candidate gene studies mainly focused on associations with genes involved in DNA integrity and methylation, the immune response, B-cell survival and growth and in the pathways of xenobiotic metabolism [108,123,124,187,234]. More recently genome-wide scans have identified variants involved in the inflammatory, folate, apoptotic, energy production and hormone metabolising pathways as also being of potential importance [12,110,111,124]. These findings have been discussed in more detail in chapter 1, section 1.3.5. Despite some success, to date both for the candidate and the genome-wide studies tend to have been characterised by modest effect sizes and low penetrance variants.

Consequently in an attempt to address this through the use of larger sample sizes there has been a shift towards consortium-led GWAS encompassing thousands of subjects from multiple studies. One such initiative is the InterLymph NHL GWAS [87], which contains over 11,000 NHL cases and 15,000 controls from 22 studies, including 9 prospective cohort studies, 8 population-based case-control studies and 5 clinic/hospital-based case-control studies. This represents the largest study to date enabling the investigation of individual subtypes with good power. InterLymph have previously reported a large number of significant associations between genetic variants and the risk of NHL and its subtypes using a candidate approach [110,114,115,125,126]. Many of which, particularly those involved in the immune response, have also been replicated in independent studies. Its more recent subtype-specific GWAS has already published strong findings for chronic lymphocytic leukaemia (CLL) [87]. Currently, the manuscripts for the remaining subtypes, including DLBCL, FL and MZL, are in preparation. Within the InterLymph NHL GWAS a total of 190 NHL cases and 265 controls from EPIC have been genotyped.

Although the emerging results from the consortium are very promising, it is hypothesised that a large proportion of genetic heritability remains unexplained. For CLL it is estimated that up to 46% of familial risk could be explained with common SNPS, while the identified variants explain only 17%. It is therefore likely that further SNPs, probably with only marginal effect sizes, remain to be discovered [87] . Although there is a counter-argument to this which states that not all that is heritable is necessarily genetic, and a large component of the 'missing heritability' may represent shared environmental and social factors or the inheritance of environment-sensitive epigenetic states [235,236].

Nevertheless, it has been observed that for most disorders GWAS can only detect a few common variants explaining only a fraction of the genetic component of risk [237]. In general if disorders are caused by rare genetic variants with minor allele frequencies (MAF) less than 1%, or by variants with effect sizes less than 10% then these are unlikely to be identified by standard GWAS methods. Additionally, these univariate methods are unable to elucidate how multiple variants may be working in combination to affect risk [87,238]. Consequently it has been proposed that GWAS may be better suited as a first preliminary step to prioritize information for additional analysis [237].

In particular three different ways of performing secondary analyses on the findings resulting from GWAS have been suggested: meta-analyses, pathway analyses and interaction analyses. Meta-analyses are now commonly used, mainly in systematic reviews, and the methods are well understood, although their application to GWAS data does introduce some novel challenges. These mainly relate to the heterogeneity between the contributing studies which, if too large, may obscure the association of interest. Pathway analysis is a way of integrating the findings from genetic studies into known molecular pathways and determining whether they are plausibly associated with the disorder of interest. This is a less well developed area than meta-analyses and it has been observed that findings can be very dependent on the various databases and methodologies employed [239]. Similarly the study of gene-gene interactions (GxG) is complex and currently it remains computationally and statistically implausible, if not impossible, to assess all the pairwise and higher order interactions involved in the context of a GWAS.

This chapter focusses on a novel methodology, profile regression, for the exploration of GxG following GWAS analyses, as a way of considering the findings in a different context and of providing a more global overview. Profile regression is also particularly well suited to the situation where the number of variables far exceeds the number of subjects, as in this dataset. Given the limited sample size available, the likelihood of obtaining interpretable results using standard GWAS analyses in this population is low.

Profile regression [163,165] is a dimension reduction technique that clusters individuals according to both their covariate pattern and an outcome of interest. The covariate profile in this instance is based on the genotype for each individual at each measured SNP, and the outcome is NHL. This allows for the

identification of the SNPs which combine to influence risk and the quantification of risk in each cluster. It is a Bayesian technique which incorporates flexible nonparametric clustering and a variable selection approach using a Dirichlet process to allow uncertainty to be taken into account. Each SNP is assigned a continuous selection weight which can be utilised to identify the variables most important for cluster formation and for determining outcome, relative to all other variables. Further details of the methodology are provided in methods chapter 3.3.5.

Here the results of profile regression analyses are presented for the strongest NHL associated SNPs in participants from the EPIC cohort.

4.2 Methods

4.2.1 Study population

The study population was drawn from the EPIC cohort (see methods 3.1.1 and table 3.1.2c). Blood samples from incident cases of NHL and two matched lymphoma-free controls, were sent to the Cancer Genomics Research Groups at the National Cancer Institute in Washington. Here, a subset of these cases (*n*=190) and controls (*n*=265) were selected for screening and the remainder saved for the taqman replication of promising findings from the total consortium.

4.2.2 Genotyping

Details of the genotyping and quality control (QC) procedure are in methods section 3.2.4. The data was received in its raw format so QC procedures were run independently on the EPIC dataset utilising the same thresholds applied by the InterLymph consortium in the pooled analysis[87].

4.2.3 Quality control in EPIC

First sample quality was checked. Samples with a call rate <93%, those with a mean heterozygosity <0.25 or >0.33 (based on the autosomal SNPs), those showing a discordant sex to that stated in the questionnaire (on the X chromosome >5% heterozygosity for males and <20% heterozygosity for females), unexpected duplicates (>99.9% concordance) and first-degree relatives (on the basis of identity-by-descent (IBD) with Pi-hat > 0.40) were excluded. The SNP QC criteria were then applied to the remaining samples. Monomorphic SNPs, SNPs with a call rate <93%, SNPs with a MAF <1% and those with a Hardy-Weinberg equilibrium p-value $<1x10^{-6}$ were excluded.

4.2.4 Association analyses

Association analyses were run including all individuals and SNPs that passed the QC criteria using PLINK: whole genome association analysis toolset [240]. The logistic command was used to produce an odds ratio, 95% confidence interval and p-value for each SNP, adjusted for age, sex, country and significant MDS (multidimensionality scaling) components. MDS is a dimensionality reduction technique that is a means of determining the levels of similarity between participants in a dataset. Components based on IBD were calculated using the *genome* command in PLINK.

SNPs were then ranked according to their p-value, with the smallest p-value ranked as one, and the strongest 5000 associations taken forward for profile regression analyses. To avoid simply picking up population substructure this list of 5000 SNPs was then thinned (meaning the number of SNPs were reduced) using purpose built code (Appendix 4.2.4). When the top 5000 SNPs are ordered by chromosome and then base-pair position, this code picks up the top SNPs for each region, according to p-value, by assuming that SNPs within 100kb of each other are in the same region and those more than 100kb apart are independent. In this way the top 1000 independent SNPs can be selected for profile regression

4.2.5 Profile regression

The top 1000 SNPs were analysed by profile regression adjusting for age, sex, country and significant MDS components by including these variables as fixed effects. SNPs were coded as 0, 1 or 2 depending on their minor allele count. Sex and country were included as binary and categorical variables, and age and MDS components were continuous. The MCMC output, in tandem with the model averaging approach, assigned individuals into clusters based on their covariate profile. Latent selection weights were used to identify the most important SNPs for the formation of these clusters. Results were obtained from 285,000 iterations, with a filter of one hundred after a burn-in sample of 285,000 iterations (in total 570,000 iterations). Burn-in refers to the process of discarding an initial proportion of the Markov Chain sample. This is performed to minimise the effect of the initial pre-convergence values on posterior inference, and to start the chain in, or near equilibrium. A filter of 100 was applied to the model, so that the simulated sample is recorded every 100^{th} iteration. This is performed to speed up subsequent computations and to reduce autocorrelation within the sample, because within the Markov chain high correlation between adjacent samples may be present. These values (the number of burn-ins and iterations) were chosen after multiple sensitivity analyses were run to determine the number of iterations required to produce consistent results. Further details on profile regression are detailed in the methods section 3.3.5.

4.3 Results

4.3.1 Study population and quality control

A total of 455 participants, 190 incident NHL cases and 265 controls from the EPIC cohort, were genotyped as part of the InterLymph NHL GWAS. These were selected by the genotyping lab from the 1524 B-cell NHL cases within EPIC with a blood sample and two matched controls. Selection was based on the numbers required for the subtype specific analyses of the whole consortia population. Those cases and controls not genotyped were saved for use in the replication stages of the promising signals emerging from the GWAS. Fourteen cases of T-cell NHL were excluded from further analysis. One case with a call rate <93% (90%) and two female controls with an X-chromosome heterozygosity >20% (21% and 24%) were also excluded. The remaining participants met all other sample QC thresholds and a total of 175 cases and 263 controls were included in the analysis. The baseline characteristics of these participants are shown in table 4.3.1. Before the QC thresholds were applied genotype data were available for 715,530 SNPs in the 438 remaining participants. After frequency and genotype pruning and the exclusion of heterozygous haploid genotypes and SNPs which failed Hardy-Weinberg Equilibrium (HWE), a total of 609,456 markers were included in the analysis.

Cases and controls displayed no differences with respect to gender age, BMI or smoking status (table 4.3.1). However, they did differ significantly by country with the majority of controls (76.8%) originating from Italy. This is despite the fact that Italy accounted for only 11% of EPIC cases overall (appendix table 3.1.1). Again, women were overrepresented in this study compared to global incidence and the age at diagnosis was younger than the European average. Almost three quarters of cases and of controls were overweight or obese and roughly half were former or current smokers. However due to the large number of different European countries included, and the fact that obesity and smoking prevalence have changed throughout the recruitment period, it is difficult to determine whether these values are higher than expected. The most common subtype was CLL (40.6%) followed by DLBCL and FL (both 26.3%). These were also the three most common subtypes in EPIC overall, excepting MM which was not included in the InterLymph GWAS.

Table 4.3.1: Baseline characteristics of EPIC NHL cases and controls with available genotype data from the InterLymph Consortium

Demographic characteristics		Controls	Cases	Difference (p-value)
N		263	175	
Age at blood draw in yrs, mean (range)		62.5 (44, 84)	61.4 (41, 83)	0.171
Sex - *n* (%)	Male	120 (45.6)	86 (49.1)	0.470
	Female	143 (54.4)	89 (50.9)	
BMI category - *n*(%)	Underweight	2 (0.8)	2 (1.1)	0.158
	Normal	118 (44.9)	61 (34.9)	
	Overweight	109 (41.4)	80 (45.7)	
	Obese	34 (12.9)	32 (18.3)	
Smoking status - *n*(%)	Never	106 (40.3)	81 (46.3)	0.061
	Former	77 (29.3)	61 (34.9)	
	Current	73 (27.8)	30 (17.1)	
	Unknown	7 (2.7)	3 (1.7)	
Country - *n*(%)	Italy	202 (76.8)	37 (21.1)	<0.001*
	Spain	10 (3.8)	24 (13.7)	
	United Kingdom	14 (5.3)	48 (27.4)	
	The Netherlands	15 (5.7)	26 (14.9)	
	Greece	6 (2.3)	13 (7.4)	
	Germany	15 (5.7)	26 (14.9)	
	Denmark	1 (0.4)	1 (0.6)	
Age at diagnosis in yrs, mean (range)			61.5 (42, 83)	
Subtype - *n*(%)	CLL		71 (40.6)	
	DLBCL		46 (26.3)	
	FL		46 (26.3)	
	MZL		8 (4.6)	
	MCL		3 (1.7)	
	B-cell NHL, NOS		1 (0.6)	

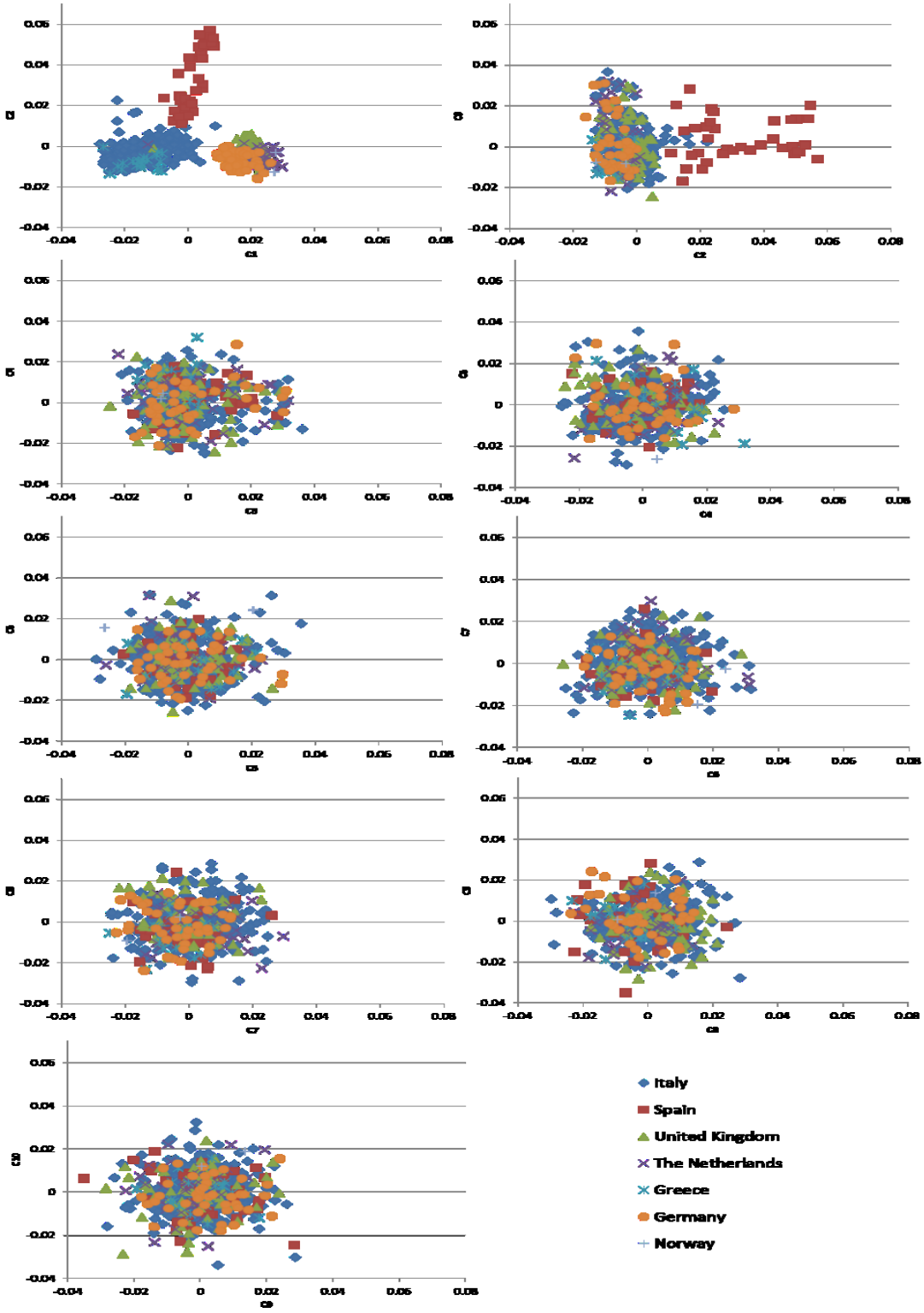
^α t-test for age, chi-squared test for all other analyses
**Significant at the 95% level*
CLL-Chronic Lymphocytic leakemia, DLBCL-Diffuse Large B-cell Lymphoma, FL-Follicular Lymphoma, MZL-Marginal Zone Lymphoma, MCL – Mantle Cell lymphoma, NOS- Not Otherwise Specified

4.3.2 Population stratification

To evaluate population substructure MDS components were calculated in PLINK which uses identical by state distance to perform hierarchical clustering. Plots for the first ten components are shown in figure 4.3.2. There was some evidence of clustering by country for the first two components. Additionally four components were associated with case-control status at the $p<0.1$ level (C1: $p<0.001$, C2: $p=0.093$, C8: $p=0.065$, C12: $p=0.073$), and were adjusted for in the association analysis.

Due to the large number of controls from Italy, country was also retained in the model, to further attempt to attenuate the potential effect of population stratification.

Figure 4.3.2: Plots of the first ten MDS components to evaluate population substructure in the total NHL EPIC genotyped population (cases=175, controls=263). Plots are coloured according to country of residence of the participants

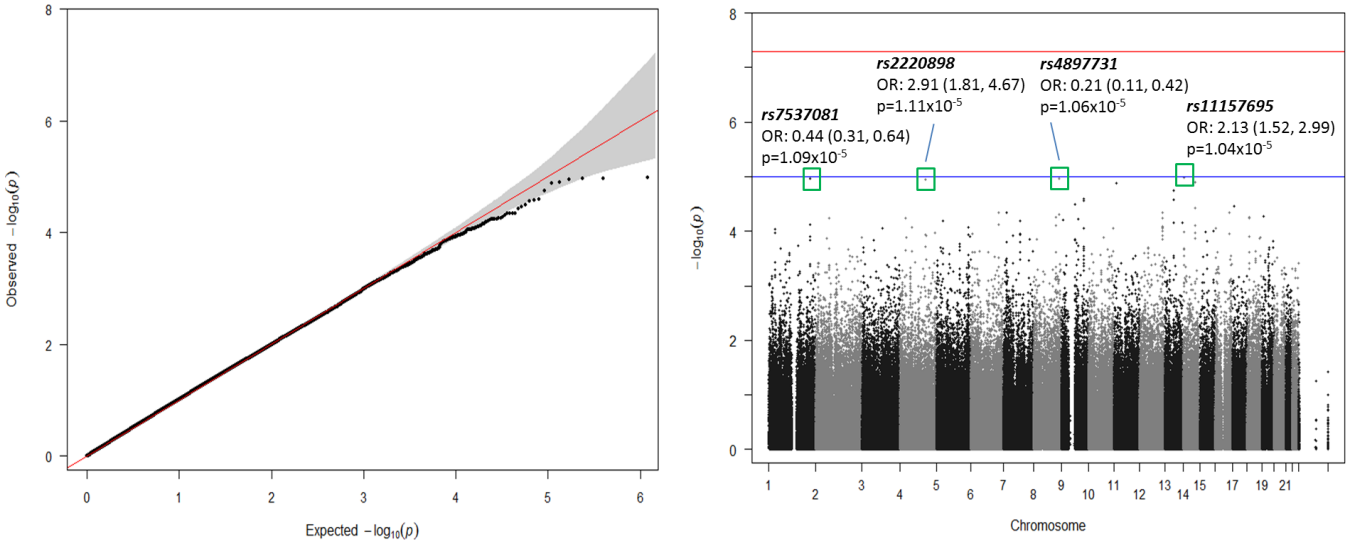


4.3.3 Association analysis

No SNPs achieved genome wide significance in this population ($p<5x10^{-8}$), although this was not unexpected given the modest sample size, and there was no evidence of an enrichment of SNPs with small values compared to the null distribution in the quantile-quantile plot (Figure 4.3.3). The SNPs that almost met the threshold for a 'suggestive association ($p<1x10^{-5}$) are highlighted on the Manhattan plot. In total 48 SNPs had a p-value $<10^{-4}$.

Figure 4.3.3: Quantile-Quantile (QQ) Plot and Manhattan Plot for the NHL-GWAS Association Analyses in the total NHL EPIC genotype population

In the QQ Plot the expected null distribution is shown as a diagonal red line. In the Manhattan Plot, the red horizontal line represents the genome-wide significance level ($5x10^{-8}$) and the blue line, the suggestive association level (10^{-5})



After thinning, a final 2263 SNPs, all of which are independent and represent the top SNP for each region were selected. The p values for these 2263 SNPs ranged from $1.04x10^{-5}$ to $8.47x10^{-3}$. The top 1000 of these SNPs (according to p-value) were taken forward for profile regression.

4.3.4 Profile regression

The SNPs were coded as 0, 1, and 2, according to the count of their minor allele. Latent selection weights were used for the detection of important genetic markers. Sensitivity analyses using different numbers of iterations were run to determine the number needed to obtain consistent results (appendix table 4.3.4). The final results are based on 285,000 iterations, after a burn-sample of 285,000. The MCMC output, in tandem with the model averaging approach assigned the subjects to two sub-populations (table 4.3.4a). There was a significant difference in risk between the two clusters. Cluster one contained 413 participants of whom 150 were NHL cases. Cluster 2 comprised 25 participants of

whom 25 were cases and membership of this group was associated with more than double the risk of cluster one according to the profile regression computed estimates of risk (OR: 2.29 (95% credible interval: 2.00, 2.44).

Table 4.3.4a; Assigned clusters for 438 NHL EPIC genotyped participants after 285,000 iterations of the sampler

Cluster	1	2
No. Participants	413	25
Risk (No. cases/no. participants in cluster)	0.363	1.00
Risk parameters computed by profile regression (95% CI)	0.371 (0.339, 0.402)	0.850 (0.678, 0.961)
Relative risk (95%CI)	1.00 (ref)	2.29 (2.00, 2.39)

CI – Credible Interval
Relative risk computed according to risk parameters computed by profile regression: RR in cluster2 =0.850/0.371

The estimates of cluster-specific risk are computed using post-processing of the whole MCMC output, in which the subjects may cluster slightly different at each iteration. At every iteration of the sampler, the risk parameters corresponding to the 25 subjects is averaged to obtain the distribution, therefore as some of the 25 participants may be in lower risk clusters at different iterations the computed risk is less than one.

The computed estimates should be robust to confounders, however the majority (n=22) of participants in the high risk cluster were from Italy, with the remaining cases from the United Kingdom (n=1), Spain (n=1) and Greece (n=1). Consequently there was a significant association between country of origin and allocation to the highest risk cluster (Chi-squared p=0.044). The effects of the confounder covariates are shown in table 4.3.4b.

Table 4.3.4b: Posterior estimates of the effects of the confounders included in the logistic regression model

Confounder	Posterior Median	95% Credible Interval
age	-0.017	(-0.045,0.009)
Italy	-2.095	(-3.921,-0.503)
Spain	1.117	(-0.514,2.875)
UK	1.713	(0.029,3.348)
The Netherlands	0.951	(-0.64,2.529)
Greece	1.055	(-0.647,2.786)
Germany	0.988	(-0.583,2.598)
Denmark	0.125	(-2.449,2.824)
sex	0.013	(-0.47,0.492)
c1	0.281	(-3.932,5.277)
c2	-0.026	(-4.843,3.819)
c8	-0.378	(-5.361,3.27)
c12	-0.926	(-6.035,3.473)

The sex ratio was similar in both clusters: 47% and 48% of the population were male in cluster 1 and 2 respectively (p=0.920). There were additionally no significant differences between the clusters in terms of MDS components 2 (p=0.267), 8 (p=0.693) or 12 (p=0.762), height (p=0.059), weight (p=0.054), BMI (p=0.417), educational level (p=0.152), activity level (p=0.700) or smoking status (p=0.629). Among the NHL cases there was no significant difference between the clusters in terms of time between blood draw and diagnosis (p=0.825). Individuals in the high risk cluster comprised diagnoses of all subtypes: CLL *n*=8, DLBCL *n*=4, FL *n*=8, MCL *n*=2, MZL *n*=3.

A total of 12 SNPs have a latent selection weight greater than 0.45 and two had a selection weight of 1, ranking them as the SNPs of greatest importance in the formation of clusters. Of the 12 SNPs three were located in chromosome 5, two were in chromosome 9, two were in chromosome 11 and there was one in each of chromosomes 1, 2, 6, 14 and 15. For 867 SNPs the selection probability was 0.

The posteriors medians of the latent selection weights for the top 1000 SNPs, ordered according to the strength of their association (ranked by p-value from 1-1000) with NHL, are shown in figures 4.3.4a and 4.3.4b. Of the SNPs with a variable selection weight >0.45 only two also had p-values in the top 100 in terms of association with NHL: rs10511476 on chromosome 9 which ranked 21st using classical methods ($p=6.19\text{x}10^{-5}$, posterior median: 0.46) and rs7402445 on chromosome 5 which ranked 32nd ($p=9.82\text{x}10^{-4}$, posterior median: 0.50). For the remaining SNPs with a weight >0.45 the p-values ranged from $1.7\text{x}10^{-4}$ to 0.0032.

Figure 4.3.4a; Posterior medians for the latent selection weights obtained using profile regression for the top 1000 SNPs associated with NHL in the logistic regression analyses

To compare the findings from the profile regression to the classical logistic regression method SNPs are ordered from 1-1000 based on the significance (p-value) of the association using the logistic regression model

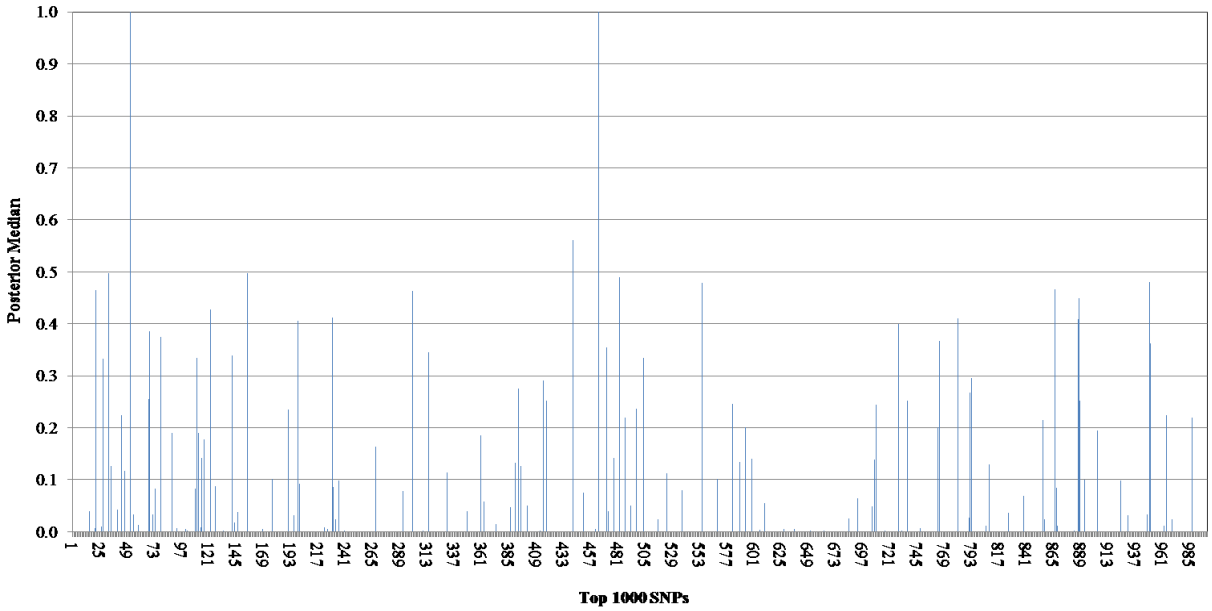


Figure 4.3.4b; Top 12 SNPs according to Posterior Median Latent Selection Weight obtained from the profile regression analyses and the associated p-values from the logistic regression analyses

The top 12 SNPs identified using profile regression are highlighted in black

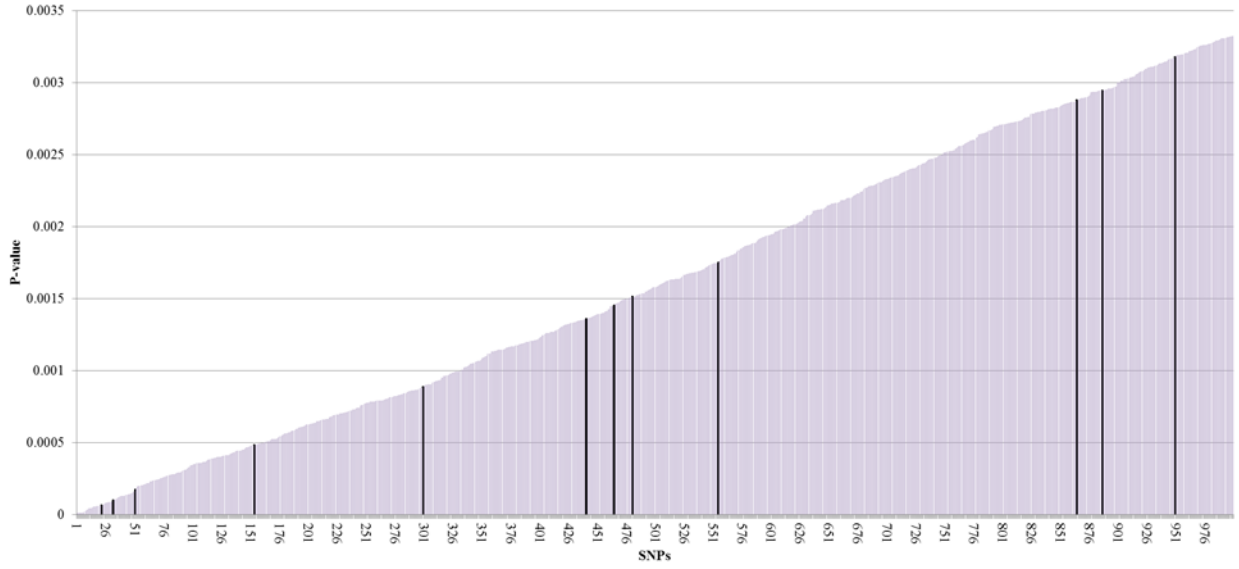
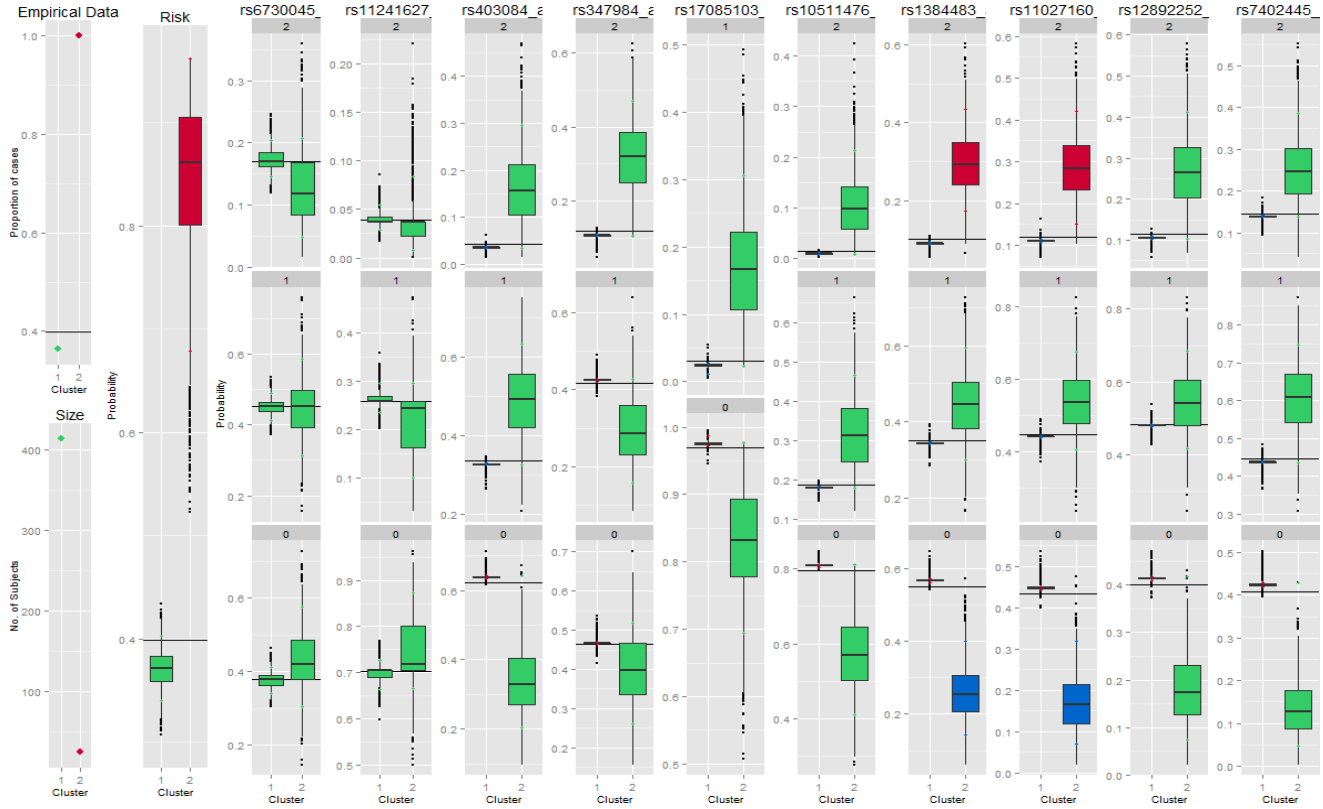


Figure 4.3.4c shows the top 12 SNPs and the associated genotypes which most contributed to the formation of the high risk cluster. The results show that individuals in the high risk cluster are significantly more likely to carry the rare homozygote for rs1384483 and rs11207160. Although the

differences were not significant for the remaining SNPs, individuals in the low risk cluster were more likely to carry the common homozygote, as would be expected.

Figure 4.3.4c: Profile of the top 12 SNPs influencing NHL risk cluster formation identified using the profile regression analyses



The bottom left hand plot displays the number of participants in the two clusters and the top left hand graph displays the proportion of these participants who were cases. The next plot shows the risk in the high risk (red) cluster relative to the average risk in the rest of the population. The remaining plots each represent one of the top 12 SNPs contributing to cluster formation, one column per SNP. The plots within the column represent the genotypes 0=common homozygote, 1=heterozygote, 2=rare homozygote. For each genotypes the boxplots show the likelihood of participants cluster 1 and cluster 2 (high risk) carrying this genotypes. Red indicates significantly more likely to, blue significantly less likely to, and green no significant difference.

There were no carriers of the rare homozygote for rs17085103

Biological plausibility of the findings

Of the top 12 SNPs (rs11241627, rs6730045, rs17085103, rs1384483, rs7402445, rs403084, rs11027160, rs12892252, rs1001606, rs10511476, rs347984 and rs12036302) none have previously been identified as being involved in the risk of NHL. One, rs17085103 which is found on chromosome 6, mapped to two genes: *OPRM* which is involved in immune function and *IPCEF1* which has previously been identified as being involved with carcinoma. However none of the other

top SNPs mapped to genes, and overall there was no strong evidence for the role of the top SNPs in lymphomagenesis.

All SNPs with a selection probability greater than 0 were then mapped to genes for pathway analysis using the methods described previously. To account for the thinning employed before profile regression the default settings in the SNP to gene translocation programmes were changed to include SNPs 100kb downstream and upstream. This resulted in a total of 52 genes. Of these 52 genes, only one, *DAPK1*, has previously been associated with lymphoma [241], however this represented a very low selection probability. Another gene, *E2F3,* has been previously associated with Chronic myeloid leukaemia, and a large number are associated with other malignancies. Similarly a number of the genes, notably *ELF1, VTCN1, PTMS* and *VEGFA* are involved in aspects of the immune response. However gene enrichment analysis using DAVID and the criteria outlined in the methods section (3.4) did not observe significant enrichment of any pathways. When less stringent criteria were employed the most important pathways were observed to be those involved in regulation of transcription, gene expression and biosynthetic and metabolic processes. One of the selected KEGG pathways was 'pathways in cancer', but the p-value was non-significant using the defined settings ($p=7.6x10^{-3}$). Additionally no immune related pathways were selected using this methodology.

Profile regression for prediction

These findings are not very promising, however the results of the ROC curve analysis did suggest that profile regression was able to better identify the SNPs providing the most discrimination between cases and controls. To explore this further three models containing different SNPs were created:

(i) The top 12 SNPs identified using profile regression,

(ii) The 133 SNPs with a selection probability>0, and

(iii) The remaining 867 SNPs with a selection probability of 0.

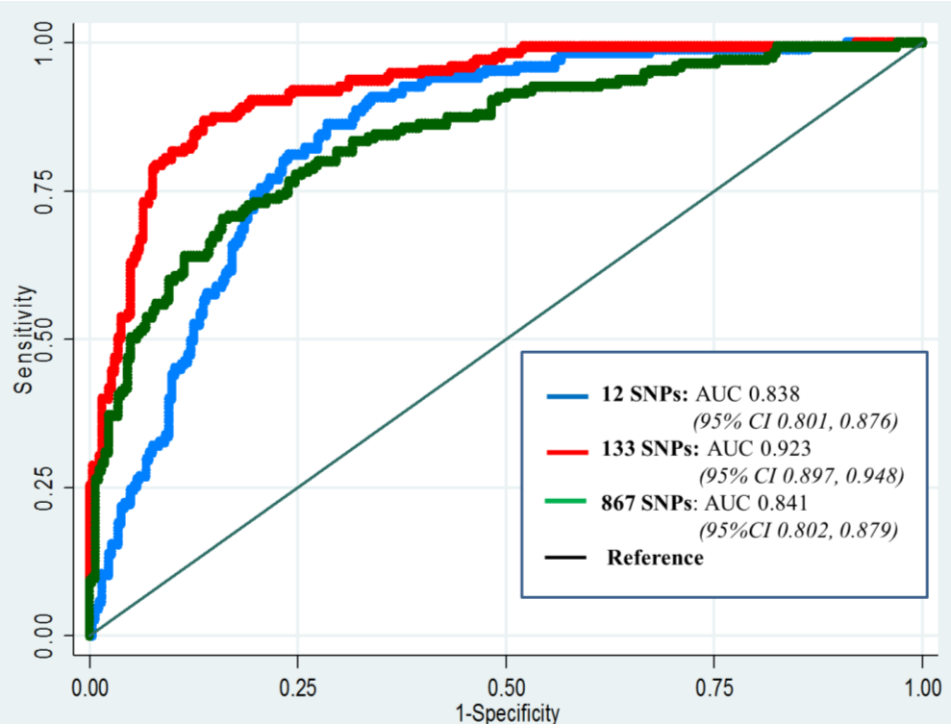
A continuous variable was created for each of these sets based on the sum of the rare alleles for the included SNPS, (i.e. for each SNP common homozygotes score 0, heterozygotes score 1 and rare homozygotes score 2). ROC curve analysis was then performed including the continuous variable as a predictor and adjusting for age, sex, country and the significant MDS components.

All three SNP sets performed better than a random predictor and the AUCs were high (figure 4.3.4d). This was expected given that all SNPs were preselected based on an association with NHL in this population. However, it is of note that the 133 SNPs selected by profile regression performed significantly ($p<0.0001$) better than the 867 with a selection probability of 0. This is of interest as these 133 did not represent the top associations when using the classical analysis. It suggests that profile regression is better at selecting the combination of SNPs that are working together. However,

it is still dependent on the pre-selection methods. The top 12 SNPs showed no improvement over those with a selection probability of 0, but this may just imply that 12 is too small a number to base prediction on. Proper validation of these findings requires replication in an independent cohort.

Figure 4.3.4d: Receiver Operator Curves and corresponding AUCs for three different SNP sets generated using profile regression analyses

ROC analyses was based on a continuous variable calculated as the sum of the rare alleles for the specified SNPs (12, 133 and 867) in each individual



4.4 Discussion

The data utilised in this chapter was generated as part of a larger initiative and alone the number of genotyped NHL cases and controls from EPIC was too small to conduct meaningful analyses using conventional methodologies. This is a particular problem in the study of NHL given that the few associations that have been identified for NHL to date, even amongst large studies, tend to be small in magnitude. This limitation was confirmed by the lack of SNPs reaching genome-wide or even suggestive levels of association significance in this population using standard methodologies. Additionally of the SNPs which almost reached suggestive significance, there was no strong evidence of any biologically plausible associations with NHL indicating these were merely chance findings.

Profile regression

A novel clustering approach that uses a Bayesian framework was chosen to investigate these data. This resulted in the formation of two distinct clusters, a high risk and a low risk cluster, for which the pattern of covariates in the high risk cluster was associated with a more than two-fold increased risk of NHL relative to the low risk cluster. As the covariates in this analysis were SNPs, clustering is based on genetic profile, with adjustment for potential confounders, and therefore these results suggest that the SNPs with high selection probabilities may help inform the genetic basis of NHL. However, as with the classical analysis, of the SNPs with the highest selection probabilities according to profile regression none had previously been associated with lymphoma.

Among the other SNPs, with lower selection probabilities there were potentially more promising results. In particular *DAPK1* was implicated and reduced expression of this kinase has previously been suggested to in part explain the heritable disposition to CLL [241]. This is of interest given that more than 40% of the cases included in the GWAS are CLL, however it should be noted that the selection probability of the mapping SNP was very low (rs3128495, ρ median=0.007). *DAPK1* falls in chromosome 9, which has previously been identified as of importance in NHL aetiology, along with the 6p21.3 chromosomal region [116]. Although no association with the specific variants noted in the literature are reported, it is of note that 9 associations with SNPs falling in chromosome 9 and seven in chromosome 6 were observed. Of these latter SNPs, two (including one mapping to *VEGFA)* fell in the 6p21 region.

The objective of this chapter was to identify novel SNPs which have not previously been identified in the literature. It is hypothesised that they have not been identified either due to their small effect sizes, because previous GWAS have not stratified by subtype or because their association with risk can only be identified when considered in the context of other interacting SNPs. Therefore it can be argued that the fact that the top SNPs identified here have never previously been associated with NHL may well have been expected. There is a lack of biological plausibility for the top SNPs. A previous study of

GxG in lung cancer utilising profile regression identified biologically plausible novel SNPs, in addition to confirming previously known associations[164]. However, a lack of biological plausibility for SNPs arising in GWAS is not uncommon, and does not necessarily preclude causal associations.

Pre-selection of SNPs

Profile regression does not require extremely large datasets and is suited to the analyses of data where the number of covariates exceeds the number of observations. In theory it can analyse a few thousand covariates (conditional on the number of subjects and number of levels of the categorical covariates). However sample size is a concern in this study, stemming from the underlying GWAS analysis that the selection of SNPs for profile regression is based on. As mentioned, GWAS invariably do require large sample sizes to overcome the issues of multiple testing and false-positive findings. The fact that the methodology relies on the pre-selection of a few thousand SNPs out of the whole genome represents the main issue with profile regression as a decision must be made on how best to select them. In this study the results of the association analyses presented in section 4.3.3 were used to obtain the top 5000 SNPs, and then thinned them to a total of 1000 SNPs. However it is known this methodology was limited in this population and furthermore that it did not pick up any of the SNPs previously associated with NHL in the literature. A recent article [179] identified 15 different SNPs that have previously been associated with NHL or its subtypes in GWAS, while the first InterLymph GWAS publication subsequently identified a further 7 reaching genome-wide significance for CLL [87]. Of these SNPs 20 passed QC criteria, but only 1 ranked in the top 5000 associations in the EPIC population, and so was eligible for inclusion in the profile regression analysis (rs735665, p=0.0044). After thinning, rs735665 was not selected for further analysis. So it is not known whether these SNPs identified in the literature would have influenced clustering. Although it should be noted that of the SNPs in the region surrounding rs735665 on chromosome 11 that were taken forward for profile regression, all had selection probabilities of zero, and there are no known SNPs in LD with rs735665. It would be of interest in future analyses to include in the SNP covariates those which have previously been associated with NHL in the literature, to determine whether they are also picked up using this methodology.

To counter this, it must also be considered that the role of genetic susceptibility in NHL is 'modest at best' [171] and overall there are a very small number of known associations for NHL. If working on a disorder with a more well-understood genetic basis, and with more confirmed associations there would be a greater chance of replicating at least one. Furthermore as the hypothesis was based on the ability to pick up joint effects, again these SNPs may not be expected to make it through the pre-selection stage if they don't also show a main effect, as standard methodologies are not optimised to detect such associations. Finally as mentioned, some argue that for most disorders the 'missing heritability' can be explained by environmental including epigenetic factors, and there may be no

more SNPs to identify [235,236]. Consequently even with a larger dataset the analysis would have been limited to pick up the SNPs previously reported in the literature.

The ability to detect true genetic associations in a GWAS is also dependent on the overall quality of the data. The DNA quality of real samples is usually less than of those used to benchmark the panels, and in the context of a GWAS even a small amount of systematic error could result in hundreds of false-positive signals [242] , which will then subsequently affect any downstream analysis such as profile regression. Both individual and marker QC were run on the data before initial association analyses using the same stringent criteria as employed by the InterLymph GWAS. This resulted in the exclusion of three individuals and more than 100,000 markers. Therefore the inability to detect significant associations does not stem from issues of data quality.

Pathway analysis

To further explore the findings of this study pathway analysis was applied. The term 'pathway analysis' can be defined in a variety of ways [239]. In general it refers to the practice of assigning genes to pathways on the basis of their functions; where a pathway represents a series of events that result in a particular end point or function. These end points include metabolic processes, biosynthesis, genetic information processing, cell signaling, immune responses and disease. The genes can then be tested in aggregate with computational tools and pathway databases to identify the most overrepresented or enriched pathways and therefore highlight the biological processes most pertinent to the investigated disorder [238]. It is also known as gene-enrichment analysis.

There are a number of benefits associated with performing pathway-based analyses, as compared to considering each gene individually. Primarily, the reduction in complexity by reducing dimensionality and the increase in explanatory power by incorporating information on the biological processes, components, or structures in which individual genes are known to be involved [239]. Disruptions in different genes may lead to the same disorder, and although affected individuals may share the same disrupted pathways, the mutated genes or variants within those pathways are likely to differ and this wouldn't be picked up using standard methods. Pathway analysis is able to determine if variants are working along the same pathway even when there is genetic heterogeneity between individuals [243]. By allowing for such heterogeneity, pathway analysis also permits the combination of multiple GWAS; different GWAS of the same disease often yield different results at the SNP level but the same findings at the pathway level, and of different ethnic groups [244]. Finally, it has been proven to help identify associations that are too weak to pick up under classical analysis [243].

In this population when all the SNPs with a selection probability greater than zero were mapped to genes, no significant enrichment of any pathways was observed. Although it should be noted this was based on a very small number of genes. A substantial number of the SNPs on GWAS platforms do not

explicitly tag particular genes, while some may map to different splice variants or different genes at the same chromosome location or to different genes in a different chromosomal location [226]. To attenuate this issue somewhat, only genes that were identified using at least two different databases were included. Consequently, the 133 selected SNPs only mapped to 52 known genes which could then be taken forward for analysis. This is a relatively small number, DAVID recommends ~100 to 2000 is optimal, as a larger gene list has a higher sensitivity to terms that are only slightly enriched and to more specific terms [227]. Although DAVID states that it could still be useful for small gene sets, the sensitivity in this set is unavoidably decreased towards broader more general terms, and highly enriched pathways. So in reality the pathway analysis was under powered to exploit in-depth meaning from the results. Additionally using this methodology it was not possible to assign greater weight to those genes mapping to more than one SNP, and the 81 SNPs not mapping to any known genes were discarded entirely.

Advantages and further applications of profile regression

A further reason for the lack of positive findings may also, yet again, reflect the subtype heterogeneity of NHL. All cases of B-cell NHL were considered as a single group to retain power. Genetic heterogeneity between subtypes has been proven in tumour based studies [23,109,131]. While the emerging results from the Interlymph GWAS consortium and other subtype specific GWAS [111,117], suggest there is also genetic heterogeneity between the subtypes in terms of susceptibility. Further it has been reported that susceptibility loci specific to a single subtype are more common than those relating to many [139]. The lack of strong findings to date may stem from the inability of previous studies to stratify analyses by subtypes, or the limited power due to small sample sizes when stratification was performed. The future subtype-specific GWAS papers from InterLymph will be able to better inform on this. In theory, you may expect profile regression to be able to pick up this heterogeneity. There was no evidence here of clustering by subtype based on the SNP profiles, but it would be of interest in a larger dataset to see whether this method resulted in a number of high risk clusters each representing a different subtype.

Such additional exploratory analyses represent further potential applications for profile regression, and despite the lack of success in this population, the available evidence suggests that this is a robust methodology which could be well-applied to an alternative dataset. For example the lung cancer dataset, where profile regression performed so well [164], was both larger and represented a single well defined outcome, unlike the heterogeneity of NHL.

However, even within the NHL dataset there were some interesting findings; the consistent clustering during sensitivity analysis and the consistent selection of the same SNPs contributing to the formation of these clusters suggests a genuine partitioning between the groups based on their genetic profile. The lack of biological plausibility may be a reflection of the fact that functional SNPs may be lost

when performing thinning. Thinning was performed to avoid picking up genetic substructure, as during initial runs many clusters based on SNPs in high LD or very close to each other on the genome were identified. The thinning script selected one SNP from each region on the basis of the most significant p-value, but did not take function into account, and the strongest finding does not necessarily represent the most informative. A thinning strategy which is able to take this complexity into account may provide more enlightening results. It is possible that the SNPs identified by profile regression are in LD or in haplotypes with the biologically important SNP and this hypothesis is somewhat supported by the results of the ROC analysis.

Utility of the identified profile as a susceptibility biomarker

For the three SNP sets the ROC curves were associated with very high AUCs in excess of 0.8 which would commonly rate as very good discriminatory ability. Rather than comparing against a gold standard, which is lacking in this instance, the ROC curves for each SNP set have been compared to each other. It is of interest that those with a selection probability>0 performed significantly better than those with a selection probability of 0. This is despite the fact that the majority of SNPs in this latter group had more significant p-values using the classical analysis. This suggests that further predictive ability is imparted through considering the 133 SNPs identified by profile regression in combination. Similarly regarding sensitivity and specificity, although not calculated here for this continuous biomarker, the relevant improvement in both for the 133 SNP set can be inferred from the ROC curve. While the 12 SNP set ultimately performs better at identifying true positives than the 867 SNP set, it also identifies more false-positives. Given the similarity in the AUCs careful consideration in this context would need to be given to which indices to optimise. Although it should be remembered there are also other important considerations and regardless of prediction ability a biomarker requiring information on a large number of variants is impractical.

These findings should be considered with caution. By validating within the same dataset as the results were produced there is a risk of overfitting and therefore in an over-inflation of predictive ability. Overfitting can occur when large numbers of potential predictors are used to discriminate among a small number of outcome events, and the larger the number of predictors the more likely overfitting is. The simplest way to check whether overfitting has occurred is to validate the findings in an independent dataset. Validation can be defined as "efforts made to confirm the accuracy, precision, or effectiveness of results", but if results are not reproducible then all other aspects of validity are irrelevant [229]. Due to the limited sample size the data could not be split into a training and validation set. This strategy means the validity of the multivariable results cannot be proven [229]. However although these results are of no use for prediction, they are still of interest to compare the three SNP sets to each other.

Technical bias was minimal in this study as all samples were genotyped in the same laboratory and stringent QC thresholds were applied. There was some potential for selection bias; all cases and controls came from the same population, but only a proportion was chosen for genotyping by the laboratory, with the selection criteria unknown. This is unlikely to have affected cases and controls differentially. There was also the potential for population stratification. Seven European populations were included in this study and although the case to control ratio was similar between most countries, it differed in Italy, and Italians comprised 76% of the controls in the whole study. This was accounted for this in analysis, by adjusting on both MDS components and country, as well as age and sex. Previous studies have shown that the inclusion of MDS correctly adjusts the estimates of risk associated with the clusters, and simulation studies prove that profile regression does not induce a false association to disease or false clustering based on the disease even in data sets where a genetic structure is clearly present. Therefore the possibility that the genetic profile of the cluster is a causal factor with regard to the increased risk cannot be ruled out. However, uncertainty remains about the effect of differing distributions of case-control status among distinct subpopulations for this methodology [164].

In terms of the other biomarker criteria (methods section 3.6) as mentioned many of Bradford Hill's criteria, such as temporality, are inappropriate for a genetics-based study. Among those that can be applied there was little evidence of biological plausibility, and related to this, little evidence of consistency or coherence with existing knowledge. This also pertains to the Venice and Ransohoff criteria regarding replication and amount of evidence. Again these issues could be addressed if replication in an independent cohort was possible. Given the relative weakness of the effect estimate, even if could be replicated it would be of virtually no clinical utility.

Limitations and strengths

In addition to those discussed there are a number of other limitations to this study including the exclusive use of a European Caucasian population which may limit its wider generalisability and external validity. Further, there was an overrepresentation of Italians and this did appear to affect the clustering, suggesting that country of origin is an important factor. As discussed in the results section, women were also overrepresented in relation to worldwide incidence, and there was a high proportion of overweight participants and of people with a history of smoking, given both of these factors have been associated with NHL in some studies [32], this may also have impacted findings.

Most of the other limitations are common to all genetic studies, for example the possibility that the true functional SNP of interest is in LD with that identified by the study. Allelic loss in cells may additionally prevent the genotyping assays from accurately distinguishing between homo- and heterozygote states [245].

A further common problem in genetic studies, the difficulty of estimating the combined effects of genetic polymorphisms, is addressed by the main strength of this study which is the use of profile regression, and this also circumvents the problem of multiple testing. There are now a number of methodologies available which aim to deal with this issue of GxG and which utilise clustering, including latent class analysis (LCA) and dimensionality reduction methods [246]. However, although these tools are useful for handling covariates in general, they do not routinely provide effect sizes or p-values, or account for possible confounding and in the case of MDR in particular, interpretation may be difficult[163,164]. In contrast profile regression can adjust for specified confounders, and although as a Bayesian methodology it doesn't generate p-values, it calculates a selection probability that provides an easily interpretable measures of the importance of each variable [238].

Conclusions

It has been stated that 'post-GWAS studies are likely to provide a clearer picture of the true role of common variants in common complex disorders'[238]. Here, a novel post-GWAS methodology in combination with pathway analysis, has been applied to attempt to better understand the genetic basis of NHL. Although it can be concluded that profile regression is well-suited to such analysis, ultimately as GWAS provide the 'etiological framework'[238], this study was underpowered to identify the genetic determinants of risk. If the GWAS is unable to accurately identify the top SNPs, then you need to be able to include more covariates in the profile regression analysis. However, despite the good computational performance and power of profile regression, after multiple sensitivity analysis it was not possible to include a larger number of SNPs in this particular analysis. It is currently not computationally possible, but it would be of interest in the future to consider a genuine genome-wide analysis using profile regression. Furthermore as profile regression provides a different interpretation of the data to logistic regression, ideally the two should be utilised simultaneously, to provide maximum information [163].

The PReMiuM package used to run these analyses can be utilised to make predictions in an independent cohort and assess the findings of the analyses [223]. An independent cohort was not available in this instance however this could represent possible future work. In theory if the findings could be replicated in an independent dataset, and the function of the associated SNPs could be ascertained, then the utility would lie in what this could tell us about mechanisms and particularly about the variants working in combination to affect risk. It may provide further evidence on the underlying genetic susceptibility for NHL, and help to infer how it may be interacting with environmental factors or with biomarkers on any stage of the causal pathway.

5. BIOMARKERS OF INTERNAL DOSE: EXPOSURE TO ENVIRONMENTAL POLLUTANTS AND FUTURE RISK OF NHL

BIOMARKER: Body burden of two heavy metals and ten persistent organic pollutants

BIOMARKER TYPE: Biomarker of internal dose

POPULATION: B-cell NHL cases and controls from the EnviroGenoMarkers Study

This chapter is based in part on the publication 'Blood Erythrocyte Concentrations of Cadmium and Lead and the Risk of B-Cell Non-Hodgkin's Lymphoma and Multiple Myeloma: A Nested Case-Control Study' by Kelly et al. (2013)[216]. Tables, figures and text have been modified accordingly.

5.1 Introduction

The epidemiological evidence to date suggests a largely environmental component to the aetiology of NHL. The reported increase in incidence of NHL had been noted to mirror the worldwide usage trends of a number of suspected chemical risk factors, with the plateauing of the increase observed to occur following the banning of many of these substances in much of the world [247]. An excess of cases has also been reported in those who are occupationally or residentially exposed to the suggested environmental risk factors [248]. While the relatively higher prevalence of NHL in males has been suggested to be related to the occupational nature of many of these environmental risk factors, and this is supported by the fact that incidence in women has been noted to rise with the increasing emergence of a female workforce in more traditionally male occupations such as industry [15,30]. However, this argument is countered by the fact that male excesses have also been observed in children [15]. Nevertheless, there exists a consistent evidence base for an important role of environmental factors in the aetiology of NHL.

However, assessing the association between environmental factors and NHL has been hampered by the difficulty of measuring exposure, which is often inferred from surrogate or proxy measures. Much of the evidence relating to the environmental epidemiology of NHL comes from case-control studies which are highly subject to recall bias and selection bias, or from investigations of high level accidental exposure [71]. The use of biomarkers of internal dose measured prospectively within a cohort study eliminates these issues, and provides an objective, accurate measure which is theoretically free from bias. In order to determine whether such a measure could be used to predict the risk of developing NHL in healthy persons, within the EnviroGenoMarkers study the body burden of twelve potential risk factors including six Polychlorinated biphenyls (PCB) congeners and two heavy metals, were measured in prospectively collected blood samples.

5.1.1 PCBs and NHL

PCBs have been commonly cited as NHL risk factors, but the findings have been inconsistent and a causal relationship is yet to be defined [63,70,249-251]. PCBs are a mixture of synthetic chlorinated hydrocarbons that are the products and incidental by-products of multiple industrial and agricultural processes. A PCB molecule consists of a pair of joined 6-carbon rings, with chlorines substituted at any one of the free 10-carbon positions, which can be twisted, relative to each other (non-coplanar), or aligned in the same plane (coplanar). There are 209 possible chlorine arrangements, or congeners, which display different properties [63] with commercial PCB mixtures typically containing 60 to 90 of these congeners [251]. PCBs have recently been upgraded to Class 1 carcinogens by IARC [252] and although once widely used, concerns about their health effects and toxicity has led to them being banned in most countries [66]. However, these compounds are exceptionally stable, chemically inert and highly soluble in lipids [251]. Consequently, they have extensively polluted the environment and are still found in all environmental media, including air, water, and soil which has led to bioaccumulation in the food chain. As such, diet represents a primary source of PCB exposure [63].

The evidence for an association with NHL is based on the observed tumorogenic and carcinogenic properties of PCBs and in particular the dioxin-like properties exhibited by the coplanar congeners [79,253]. The dioxin–like congeners share a biological mechanism, based on binding to the Ah-receptor [193], with 2,3,7,8-TCDD a compound which has previously been associated with an increased risk of lymphoid malignancies [254]. The relationship is further supported by the known immunosuppressive and inflammatory properties of PCBs [255], and by the temporal relationship between the incidence of NHL and the worldwide usage of PCBs.

5.1.2 Persistent Organic Pollutants (POPs) and NHL

In addition to PCBs, four other POPs hypothesised to have an effect on NHL risk were explored: Dichlorodiphenyldichloroethylene (DDE), Dichlorodiphenyltrichloroethane (DDT), Hexachlorobenzene (HCB) and 2,2',4,4'-Tetrabromodiphenyl ether (BDE-47) [76,78,81,83,86]. The biological rationale linking them to NHL is also based on their suspected carcinogenic and immunotoxic properties [85] and the literature linking them to NHL is similarly inconsistent. DDT (and DDE its main metabolite) is an insecticide which is now mostly banned in the western world, but due to its prolonged half-life and biologic persistence, which may extend over several decades, it remains a common contaminant in human adipose tissue [82]. HCB is a commonly used fungicide, that has also been found to be a contaminant in certain pesticides [82,256]. BDE-47 is widely used as a flame retardant in thermoplastics, textiles and building materials and its production is on the increase [257].

5.1.3 Cadmium, Lead and NHL

Non-essential metals such as cadmium (Cd) and lead (Pb) are ubiquitous persistent environmental toxicants which bioaccumulate in the human body. The global burden of Cd and Pb is high due to their widespread usage in industrial and manufacturing processes and through contaminated phosphate fertilisers [258-261]. Human exposure occurs primarily through the ingestion of contaminated food and drinking water, the inhalation of contaminated air and smoking [262-264]. Pb and Cd constitute some of the most widespread environmental pollutants and both are highly toxic in trace amounts [263], but have no known beneficial physiological role [263,265]. Consequently they represent an important public health burden [266,267].

Since 1993, Cd has been classified as a group 1 carcinogen by the International Agency for Research on Cancer (IARC) based on sufficient evidence from animal models and increased incidence of breast and lung cancer in humans [268]. Pb is classified as a possible human carcinogen (group2b), while its inorganic compounds are deemed probable carcinogens (group 2a) [269]. As with the investigated POPs, the strongest evidence for an association between Pb and Cd with NHL arises from the observation of elevated incidence in populations geographically or occupationally exposed to industrial sources [248], and the fact that both agents have been observed to modulate the immune system [259,270,271]. Furthermore, a primary target of Pb [272] and a secondary target of Cd is the hematopoietic system [272,273], making these metals plausible candidate risk factors for NHL.

Despite the available evidence and the numerous studies conducted to date, a causal relationship remains to be defined for all the investigated environmental contaminants with NHL.

5.2 Methods

5.2.1 Study subjects and exposure assessment

This study is based on all 270 incident cases of B-cell NHL and their matched controls from the EnviroGenoMarkers project. The study population and the methods used to assess body burden of six PCB congeners, DDE, DDT, HCB, BDE-47, Cd and Pb are outlined in the methods chapter (see methods chapter 3.1.2, tables 3.1.2a- c and figure 3.1.2 for details of the study population and section 3.2.1 for assessment of body burden).

5.2.2 Statistical analysis

All analyses were conducted on the total population as well as by cohort (EPIC-Italy and NSHDS) and by sex where appropriate. The statistical significance of the differences in baseline characteristics between cases and controls were determined using the Student's t-test and the chi-squared test for continuous and categorical outcomes respectively. Exposure concentrations were explored using non-parametric methods due to their typically non-Gaussian distributions. Cohort and sex specific analyses used the ranksum test while the case-control status analyses used the Wilcoxon signed rank test. Correlation between exposure levels was assessed using Pearson's correlation coefficient. Each pollutant was treated as a separate risk factor. PCBs mixtures based on the properties of the congeners [274] were additionally considered as risk factors: the dioxin-like congeners (118, 156), the non-dioxin like congeners (138, 153, 170, 180) and the immunotoxic congeners (118, 138, 156, 170). As was the sum total of all PCBs.

Cases were categorised into ordered quartiles based on the distribution of exposure in the control population. A further binary variable was created to assess whether blood concentration exceeding 'highly exposed' levels, defined as being above the 90th percentile of the control exposure distribution, was associated with greater risk by comparing it to those in the lowest quartile. For the remaining analyses, stratified by subtype, age group, BMI category and time to diagnosis, exposure levels were log transformed to normalise the distributions.

In the basic model, conditional logistic regression analyses were used to calculate odds ratios and 95% confidence intervals accounting for the matching factors. Tests for trend were performed using the Wald test statistic based on the continuous quartiles. Where all controls were included to maximise power, unconditional logistic regression adjusting for the matching variables was used. Confounding was assessed by additionally running an adjusted model to determine how the results differed. Potential confounders were identified as BMI (kg/m^2), height (cm), educational level (as a proxy for socioeconomic status), vegetable intake (g/day), dairy intake (g/day), protein intake (g/day), total fat intake (g/day) and alcohol consumption (g/day), based on both a significant ($p<0.05$)

association with cumulative POP concentration in the total population and a reported association with NHL in the literature [73,101] to ensure biological plausibility for the included confounders.

Heterogeneity between strata (subtype, time to diagnosis, age group and BMI category) was assessed using a meta-analytic model (3.3.4).

5.3 Results

The baseline characteristics of cases and controls are described in the methods section (3.1.2, table 3.1.2c). NSHDS accounted for 69% of the participants, 51% of participants were female, around 20% were current smokers and nearly 60% were overweight or obese according to their BMI. There were no significant differences between cases and controls for any of the investigated variables. A total of 270 incident cases (EPIC-Italy *n*=84, NSHDS *n*=186) and 270 matched controls were included. The median time from blood collection to diagnosis was 6.6 years in NSHDS and 5.0 in Epic-Italy (Table 3.1.2b) There was no association between time to NHL diagnosis and body burden for any of the pollutants amongst the cases (p-value>0.05).

The mean, median and range of exposure concentrations in cases and controls for each investigated pollutant are shown in table 5.3a. For BDE-47 and DDT more than 80% of the population had levels that could not be quantified and therefore these were excluded from further analyses. Additionally, one participant, a male control from the NSHDS was excluded from the PCB analyses as an outlier as his exposure levels were far above the 99th percentile of the distribution for all congeners.

There was no significant difference ($p<0.05$) between cases and controls in median exposure levels according to the Wilcoxon signed rank test for any of the investigated exposures (table 5.3a). However, stratified analyses by cohort and by sex showed significantly higher concentrations in female cases than in female controls for Cd (p=0.046) and PCB118 (p=0.027). While amongst males, PCB118 (p=0.024), PCB138 (p=0.036), PCB153 (p=0.033), PCB156 (p=0.017) and HCB (p=0.011) were significantly higher in controls (Appendix: table 5.3i).

Table 5.3a; Mean, median and range of exposure concentrations of six PCB congeners, HCB, DDT, DDE, BDE-47, Cadmium and Lead in cases and controls from the EnviroGenoMarkers study

Exposure	*LOQ*	Cases n=270				Controls n=270				Difference
		% <LOQ	Median	Mean	Range	% <LOQ	Median	Mean	Range	p-value[a]
PCB118	**5pg/ml**	0.00%	134.7	182	(8.5, 901.6)	0.00%	147.7	175.7	(11.6, 832.1)	0.991
PCB138	**5pg/ml**	0.00%	530	605.6	(11.0, 1810.1)	0.00%	562.3	636.4	(51.2, 2675.4)	0.276
PCB153	**4pg/ml**	0.00%	1007.8	1150.9	(32.5, 3308.0)	0.00%	1086	1205.5	(120.6, 4334.1)	0.239
PCB156	**2pg/ml**	0.00%	90.6	102	(19.8, 307.8)	0.00%	95.6	106.7	(15.5, 394.9)	0.186
PCB170	**4pg/ml**	0.00%	326.4	375.2	(62.7, 1082.3)	0.00%	357.6	389.4	(50.4, 1291.2)	0.284
PCB180	**3pg/ml**	0.00%	696.6	788.8	(139.7, 2708.8)	0.00%	727.8	807.9	(100.1, 2431.5)	0.284
HCB	**25pg/ml**	0.00%	257.8	398.1	(62.5, 3604.0)	0.00%	289.2	432.5	(64.5, 3882.2)	0.198
DDT	**200pg/ml**	85.90%	261.8	331.7	(206.6, 825.8)	86.70%	298.2	399.8	(218.8, 1636.9)	0.865
DDE	**5pg/ml**	0.00%	2478.7	4038.4	(16.4, 30992.8)	0.00%	2675.3	4200.1	(76.4, 23858.3)	0.374
BDE-47	**15pg/ml**	84.10%	23.5	34.2	(15.1, 133.4)	81.10%	25.8	95.1	(15.3, 2063.3)	0.657
Cadmium	**0.02µg/L**	0.00%	0.54	0.759	(0.098, 4.113)	0.00%	0.498	0.757	(0.099, 5.224)	0.398
Lead	**0.10µg/L**	0.00%	57.679	70.56	(15.423, 400.843)	0.00%	57.597	72.158	(11.199, 672.482)	0.768

LOQ – Limit of quantification
[a]Differences between median concentrations in cases and controls according to the Wilcoxon signed-rank test for paired samples
One participant with exposure levels above the 99th percentile was excluded from the PCB analysis

There was significant correlation between the various exposure levels in the total population, for all except Cd (table 5.3b). As expected, strong correlation amongst the PCBs was observed. Particularly amongst those PCBs with similar degrees of chlorination; the Pearson's ρ between PCB170 and PCB180 was 0.973, p<0.0001, which may in part explain the similar results observed for most of the PCB congeners in the subsequent analyses.

Table 5.3b: Pearson correlation coefficients (r^2) between exposure concentrations of six PCB congeners, HCB, DDE, Cadmium and Lead measured in participants from the EnviroGenoMarkers Study

	PCB118	**PCB138**	**PCB153**	**PCB156**	**PCB170**	**PCB180**	**HCB**	**DDE**	**Cadmium**	**Lead**
PCB118	1									
PCB138	*0.7747*	1								
PCB153	*0.7681*	*0.9681*	1							
PCB156	0.6544	*0.7938*	*0.8801*	1						
PCB170	0.5951	*0.8259*	*0.9088*	*0.9569*	1					
PCB180	0.6346	*0.8101*	*0.9032*	*0.9282*	*0.9723*	1				
HCB	0.6881	0.4544	0.4734	0.4199	0.3565	0.4454	1			
DDE	0.7327	0.6739	0.6336	0.3892	0.3864	0.4843	0.7317	1		
Cadmium	0.0364^{NS}	0.087	0.095	0.0837^{NS}	0.084^{NS}	0.108	0.116	0.170	1	
Lead	0.343	0.214	0.231	0.185	0.185	0.288	0.453	0.418	0.190	1

^{NS} Correlation is not significant (p>0.05)
r^2>0.7 are highlighted in italics to indicate high correlation
Log transformed exposure levels

Table 5.3c shows the association between increasing quartiles of exposure and risk of NHL, both under the basic conditional model, and when adjusting for potential confounders. Although under the basic model a number of significant inverse associations were observed there was no evidence of a dose-response trend. When adjusting on confounders, inverse associations with PCB138 and PCB153 survived and due to the high correlation so did those with non-dioxin like PCBs. Some suggestion of an inverse trend was also noted for the sum of total PCBs, HCB and DDE. Stratified analyses by gender (5.3d) indicated that these inverse associations were heavily influenced by males who displayed multiple dose-dependent inverse trends with the investigated POPs. This was not observed in females and there was some suggestion of an increased risk with high levels of exposure to Cd among women. Analyses stratified by cohort are shown in appendix 5.3ii. For the majority of exposures the inconsistent and non-significant results across the total population and in the subgroups indicated an essentially null relationship. These relationships were attenuated slightly by adjustment for confounders, but the conclusions were unchanged.

Table 5.3c: Association between NHL and quartiles of measured body burden of six PCB congeners, specified PCB functional groups, HCB, DDE, Cadmium and Lead in participants from the EnviroGenoMarkers Study

Expsoure		No. Ca	No. Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*
PCB118	Q1(8.46, 94.05)	80	68	1				1			
	Q2 (94.25, 148.04)	68	67	0.76	(0.44,1.32)	0.332		0.57	(0.31,1.08)	0.084	
	Q3 (148.37, 219.85)	42	66	0.45	(0.24,0.84)	0.012*		0.35	(0.17,0.72)	0.004*	
	Q4 (219.91, 901.63)	80	67	0.96	(0.49,1.87)	0.909	*0.571*	0.67	(0.3,1.49)	0.332	*0.168*
PCB138	Q1 (10.97, 396.03)	86	67	1				1			
	Q2 (399.73, 562.09)	60	67	0.67	(0.41,1.08)	0.102		0.53	(0.31,0.91)	0.021*	
	Q3 (562.64, 778.15)	63	67	0.63	(0.37,1.1)	0.104		0.50	(0.27,0.94)	0.031*	
	Q4 (778.44, 2675.35)	61	67	0.60	(0.33,1.08)	0.089	*0.093*	0.38	(0.18,0.78)	0.009*	*0.010**
PCB153	Q1 (32.47, 812.10)	88	68	1				1			
	Q2 (813.60, 1083.97)	59	66	0.60	(0.36,1.03)	0.064		0.56	(0.31,1.02)	0.058	
	Q3 (1088.49, 1489.97)	60	67	0.54	(0.3,0.98)	0.044*		0.47	(0.24,0.93)	0.031*	
	Q4 (1493.70, 4334.12)	63	67	0.55	(0.3,1.04)	0.066	*0.078*	0.37	(0.17,0.78)	0.009*	*0.011**
PCB156	Q1 (15.50, 69.69.62)	75	67	1				1			
	Q2 (69.95, 95.44)	71	67	0.88	(0.49,1.56)	0.654		0.9	(0.48,1.71)	0.754	
	Q3 (95.76, 126.68)	51	67	0.59	(0.31,1.11)	0.100		0.56	(0.27,1.16)	0.118	
	Q4 (126.87, 394.89)	73	67	0.86	(0.45,1.64)	0.644	*0.545*	0.69	(0.33,1.45)	0.326	*0.228*
PCB170	Q1 (50.39, 257.50)	73	67	1				1			
	Q2 (257.62, 358.81)	82	68	0.99	(0.59,1.66)	0.967		0.9	(0.51,1.6)	0.718	
	Q3 (359.10, 468.52)	43	66	0.48	(0.25,0.92)	0.028*		0.37	(0.17,0.79)	0.010*	
	Q4 (468.55, 1291.23)	72	67	0.83	(0.45,1.53)	0.55	*0.356*	0.59	(0.28,1.23)	0.160	*0.091*
PCB180	Q1 (100.08, 533.51)	72	67	1				1			
	Q2 (533.84, 730.03)	78	68	1.03	(0.61,1.73)	0.926		1	(0.55,1.81)	0.999	
	Q3 (730.81, 966.02)	46	66	0.57	(0.31,1.05)	0.071		0.45	(0.21,0.95)	0.035*	
	Q4 (966.55, 2708,83)	74	67	0.94	(0.52,1.69)	0.827	*0.595*	0.68	(0.32,1.42)	0.300	*0.168*
Dioxin-like PCBs (118, 156)	Q1 (36.16, 172.31)	82	67	1				1			
	Q2 (172.47, 247.91)	70	67	0.72	(0.42,1.22)	0.219		0.7	(0.39,1.27)	0.245	
	Q3 (248.24, 347.34)	39	67	0.41	(0.22,0.75)	0.004*		0.34	(0.17,0.68)	0.002*	
	Q4 (348.22, 1126.82)	79	67	0.82	(0.44,1.53)	0.539	*0.340*	0.63	(0.30,1.30)	0.209	*0.083*
Non-Dioxin like PCBs (138, 153, 170, 180)	Q1 (245.80, 2027.04)	81	67	1				1			
	Q2 (2027.72, 2696.84)	65	67	0.73	(0.43,1.23)	0.237		0.66	(0.37,1.17)	0.153	
	Q3 (2719.63, 3727.56)	62	67	0.63	(0.34,1.14)	0.127		0.54	(0.27,1.10)	0.091	
	Q4 (3747.581, 10380.82)	62	67	0.62	(0.34,1.16)	0.138	*0.144*	0.44	(0.21,0.92)	0.030*	*0.033**
Immunotoxic PCBs (118, 138, 156, 170)	Q1 (153.93, 851.71)	76	67	1				1			
	Q2 (857.09, 1162.29)	68	67	0.85	(0.5,1.44)	0.542		0.73	(0.4,1.32)	0.294	
	Q3 (1164.09, 1598.01)	58	67	0.69	(0.39,1.22)	0.202		0.57	(0.30,1.1)	0.093	
	Q4 (1606.86, 4907.36)	68	67	0.82	(0.44,1.50)	0.512	*0.431*	0.56	(0.27,1.16)	0.117	*0.092*
ΣPCBs	Q1 (329.28, 2192.24)	80	67	1				1			
	Q2 (2195.17, 3006.59)	72	67	0.82	(0.49,1.39)	0.46		0.73	(0.41,1.31)	0.29	
	Q3 (3018.09, 4016.41)	51	67	0.51	(0.28,0.94)	0.030*		0.43	(0.21,0.87)	0.019*	
	Q4 (4037.42, 11400.98)	67	67	0.68	(0.36,1.27)	0.229	*0.143*	0.46	(0.22,0.99)	0.047*	*0.026**

Table 5.3c: ***continued***

Exposure		No. Ca	No. Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*
HCB	Q1 (62.47, 181.87)	85	67	1				1			
	Q2 (184.78, 288.81)	62	67	0.58	(0.32,1.03)	0.064		0.52	(0.27,1.02)	0.057	
	Q3 (289.49, 474.86)	59	67	0.47	(0.24,0.94)	0.032*		0.28	(0.12,0.65)	0.003*	
	Q4 (477.00, 3882.19)	64	66	0.49	(0.22,1.07)	0.074	*0.061*	0.35	(0.14,0.87)	0.023*	*0.014**
DDE	Q1 (16.40, 1308.26)	81	67	1				1			
	Q2 (1310.23, 2666.60)	62	67	0.65	(0.37,1.14)	0.137		0.55	(0.28,1.05)	0.07	
	Q3 (2675.27, 5523.58)	68	68	0.67	(0.37,1.21)	0.186		0.40	(0.20,0.84)	0.015*	
	Q4 (5528.29, 30992.76)	59	66	0.55	(0.28,1.08)	0.082	*0.106*	0.34	(0.15,0.77)	0.010*	*0.008**
Cadmium	Q1 (0.098, 0.308)	70	68	1				1			
	Q2 (0.310, 0.498)	53	67	0.78	(0.46,1.30)	0.341		0.82	(0.46,1.45)	0.490	
	Q3 (0.490, 0.790)	74	66	1.12	(0.67,1.85)	0.668		1.12	(0.64,1.96)	0.679	
	Q4 (0.794, 5.224)	73	68	1.06	(0.63,1.77)	0.835	*0.487*	1.06	(0.59,1.92)	0.836	*0.555*
Lead	Q1 (11.199, 38.598)	70	67	1				1			
	Q2 (38.609, 57.594)	65	67	0.94	(0.56,1.56)	0.800		1.00	(0.55,1.79)	0.992	
	Q3 (57.597, 83.455)	64	67	0.92	(0.52,1.64)	0.782		0.80	(0.41,1.57)	0.522	
	Q4 (83.748, 672.482)	71	68	1.03	(0.53,1.99)	0.93	*0.957*	0.92	(0.44,1.94)	0.836	*0.724*

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence leve*

Table 5.3d Association between NHL and quartiles of measured body burden of six PCB congeners, specified PCB functional groups, HCB, DDE, Cadmium and Lead stratified by gender

POP by quartile		Males									Females								
		Ca/Co	OR	95% CI		P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*	Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*
PCB118	1	47/34	**1**					**1**				31/34	**1**				**1**			
	2	35/32	**0.6**	(0.27,1.32)	1.32	0.202		**0.69**	(0.28,1.72)	0.424		35/34	**1.28**	(0.59,2.81)	0.534		**1.09**	(0.46,2.6)	0.846	
	3	22/33	**0.32**	(0.13,0.78)	0.78	0.013*		**0.21**	(0.07,0.65)	0.007*		23/34	**0.89**	(0.37,2.12)	0.793		**0.69**	(0.24,1.96)	0.487	
	4	29/33	**0.39**	(0.15,0.98)	0.98	0.045*	*0.021**	**0.4**	(0.13,1.22)	0.109	*0.037*	48/34	**2.41**	(0.94,6.22)	0.068	*0.138*	**2.09**	(0.67,6.52)	0.203	*0.367*
PCB138	1	42/33	**1**					**1**				44/34	**1**				**1**			
	2	36/34	**0.69**	(0.34,1.38)	1.38	0.293		**0.53**	(0.22,1.25)	0.148		24/34	**0.58**	(0.29,1.13)	0.11		**0.38**	(0.17,0.87)	0.022	
	3	30/32	**0.55**	(0.24,1.27)	1.27	0.159		**0.55**	(0.2,1.54)	0.253		24/34	**0.55**	(0.25,1.19)	0.13		**0.4**	(0.16,1.02)	0.056	
	4	25/33	**0.41**	(0.17,1.03)	1.03	0.057	*0.052*	**0.25**	(0.08,0.81)	0.021*	*0.033**	45/34	**1.19**	(0.54,2.63)	0.67	*0.848*	**0.9**	(0.33,2.47)	0.845	*0.54*
PCB153	1	42/34	**1**					**1**				47/34	**1**				**1**			
	2	38/32	**0.73**	(0.32,1.64)	1.64	0.444		**0.59**	(0.21,1.68)	0.324		19/34	**0.39**	(0.17,0.86)	0.02		**0.4**	(0.17,0.97)	0.041	
	3	30/33	**0.5**	(0.21,1.19)	1.19	0.115		**0.52**	(0.18,1.55)	0.243		28/34	**0.61**	(0.28,1.3)	0.198		**0.55**	(0.23,1.31)	0.175	
	4	23/33	**0.35**	(0.13,0.92)	0.92	0.034*	*0.022**	**0.18**	(0.05,0.68)	0.011*	*0.014**	43/34	**0.89**	(0.4,1.98)	0.774	*1*	**0.68**	(0.26,1.83)	0.45	*0.537*
PCB156	1	43/33	**1**					**1**				40/34	**1**				**1**			
	2	39/33	**0.65**	(0.31,1.36)	1.36	0.257		**0.68**	(0.27,1.76)	0.431		27/34	**0.72**	(0.34,1.51)	0.388		**0.94**	(0.41,2.18)	0.889	
	3	25/34	**0.37**	(0.15,0.9)	0.9	0.029*		**0.34**	(0.11,1.01)	0.053		26/34	**0.69**	(0.31,1.55)	0.369		**0.8**	(0.3,2.12)	0.652	
	4	26/32	**0.42**	(0.17,1.02)	1.02	0.054	*0.032**	**0.31**	(0.1,1)	0.050*	*0.021**	44/34	**1.15**	(0.48,2.75)	0.746	*0.627*	**1.22**	(0.42,3.52)	0.717	*0.704*
PCB170	1	39/34	**1**					**1**				41/34	**1**				**1**			
	2	38/33	**0.86**	(0.41,1.79)	1.79	0.679		**0.8**	(0.31,2.01)	0.629		28/34	**0.71**	(0.34,1.46)	0.347		**0.87**	(0.39,1.96)	0.737	
	3	27/33	**0.56**	(0.23,1.33)	1.33	0.185		**0.55**	(0.19,1.61)	0.275		28/34	**0.61**	(0.25,1.49)	0.283		**0.68**	(0.23,1.96)	0.473	
	4	29/32	**0.64**	(0.27,1.54)	1.54	0.321	*0.217*	**0.38**	(0.12,1.17)	0.092	*0.066*	40/34	**0.92**	(0.4,2.14)	0.854	*1*	**0.94**	(0.33,2.68)	0.906	*0.905*
PCB180	1	34/33	**1**					**1**				38/34	**1**				**1**			
	2	44/33	**1.21**	(0.57,2.55)	2.55	0.616		**1.28**	(0.47,3.44)	0.632		37/34	**0.93**	(0.44,1.96)	0.857		**1.15**	(0.48,2.77)	0.75	
	3	30/33	**0.74**	(0.31,1.72)	1.72	0.479		**0.73**	(0.25,2.15)	0.572		20/34	**0.42**	(0.17,1.06)	0.065		**0.32**	(0.1,1.03)	0.056	
	4	25/33	**0.63**	(0.26,1.53)	1.53	0.312	*0.15*	**0.48**	(0.16,1.49)	0.205	*0.079*	42/34	**1.1**	(0.46,2.66)	0.829	*0.889*	**0.84**	(0.25,2.81)	0.779	*0.534*

Table 5.3d *continued*

POP by quartile		Males									Females								
		Ca/Co	OR	95% CI		P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*	Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*
Dioxin-like PCBs (118, 156)	1	44/33	**1**					**1**				38/34	**1**				**1**			
	2	42/33	**0.68**	(0.31,1.48)	1.478	0.331		**0.71**	(0.28,1.78)	0.463		29/34	**0.67**	(0.31,1.42)	0.293		**0.58**	(0.25,1.37)	0.214	
	3	18/33	**0.27**	(0.11,0.68)	0.684	0.006*		**0.17**	(0.05,0.59)	0.005*		20/34	**0.49**	(0.2,1.19)	0.115		**0.29**	(0.1,0.89)	0.03	
	4	29/33	**0.35**	(0.13,0.92)	0.917	0.033*	*0.011**	**0.29**	(0.09,0.94)	0.040*	*0.012**	50/34	**1.64**	(0.69,3.89)	0.265	*0.311*	**1.34**	(0.44,4.12)	0.607	*0.821*
Non-Dioxin like PCBs (138, 153, 170, 180)	1	34/33	**1**					**1**				45/34	**1**				**1**			
	2	47/33	**1.3**	(0.55,3.06)	3.062	0.544		**1.56**	(0.51,4.75)	0.432		24/34	**0.51**	(0.24,1.1)	0.088		**0.42**	(0.18,0.98)	0.045	
	3	26/33	**0.57**	(0.22,1.49)	1.493	0.252		**0.72**	(0.22,2.37)	0.584		28/34	**0.57**	(0.25,1.29)	0.178		**0.56**	(0.22,1.47)	0.24	
	4	26/33	**0.65**	(0.24,1.77)	1.769	0.396	*0.091*	**0.52**	(0.14,1.89)	0.317	*0.055*	40/34	**0.83**	(0.36,1.95)	0.674	*0.787*	**0.61**	(0.21,1.79)	0.37	*0.422*
Immunotoxic PCBs (118, 138, 156, 170)	1	33/33	**1**					**1**				42/34	**1**				**1**			
	2	44/33	**1.24**	(0.55,2.84)	2.839	0.602		**1.11**	(0.4,3.11)	0.842		23/34	**0.56**	(0.26,1.17)	0.122		**0.41**	(0.17,0.97)	0.043	
	3	30/33	**0.76**	(0.31,1.86)	1.859	0.549		**0.91**	(0.31,2.68)	0.866		27/34	**0.65**	(0.31,1.36)	0.254		**0.48**	(0.2,1.19)	0.113	
	4	26/33	**0.66**	(0.24,1.77)	1.773	0.406	*0.173*	**0.5**	(0.14,1.74)	0.274	*0.191*	45/34	**1.29**	(0.57,2.9)	0.546	*0.613*	**0.93**	(0.34,2.57)	0.892	*0.761*
ΣPCBs	1	35/33	**1**					**1**				46/34	**1**				**1**			
	2	46/33	**1.23**	(0.53,2.87)	2.872	0.632		**1.27**	(0.42,3.84)	0.675		23/34	**0.5**	(0.24,1.05)	0.066		**0.42**	(0.18,0.96)	0.039*	
	3	26/33	**0.55**	(0.22,1.39)	1.386	0.203		**0.7**	(0.22,2.22)	0.54		25/34	**0.47**	(0.2,1.1)	0.08		**0.36**	(0.13,1)	0.050*	
	4	26/33	**0.57**	(0.21,1.54)	1.544	0.271	*0.076*	**0.41**	(0.11,1.48)	0.174	*0.049**	43/34	**0.94**	(0.42,2.12)	0.881	*0.896*	**0.64**	(0.23,1.8)	0.4	*0.35*
HCB	1	54/34	**1**					**1**				32/33	**1**				**1**			
	2	33/33	**0.25**	(0.09,0.69)	0.69	0.007*		**0.25**	(0.08,0.77)	0.016*		31/34	**1.06**	(0.48,2.3)	0.892		**1.17**	(0.47,2.93)	0.733	
	3	16/33	**0.09**	(0.03,0.3)	0.3	<0.001*		**0.04**	(0.01,0.21)	<0.001*		42/34	**1.48**	(0.6,3.67)	0.398		**1.49**	(0.48,4.63)	0.488	
	4	30/32	**0.16**	(0.05,0.54)	0.54	0.003	*0.002**	**0.07**	(0.01,0.35)	0.001*	*0.001**	32/34	**1.14**	(0.38,3.48)	0.812	*0.647*	**1.26**	(0.34,4.67)	0.731	*0.688*
DDE	1	44/33	**1**					**1**				37/34	**1**				**1**			
	2	37/33	**0.64**	(0.29,1.41)	1.41	0.267		**0.46**	(0.16,1.26)	0.13		33/34	**0.82**	(0.37,1.83)	0.626		**0.53**	(0.19,1.45)	0.216	
	3	25/34	**0.37**	(0.15,0.9)	0.9	0.028*		**0.21**	(0.07,0.65)	0.007*		35/34	**0.83**	(0.36,1.93)	0.672		**0.4**	(0.13,1.22)	0.107	
	4	27/32	**0.39**	(0.15,1.02)	1.02	0.054	*0.031**	**0.16**	(0.04,0.6)	0.006*	*0.003**	32/34	**0.74**	(0.28,1.96)	0.54	*0.591*	**0.41**	(0.11,1.46)	0.167	*0.143*

Table 5.3d *continued*

POP by quartile		Males										Females									
		Ca/Co	OR	95% CI		P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*	Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*	
Cadmium	1	39/33	**1**					**1**				29/34	**1**				**1**				
	2	33/33	**0.78**	(0.38,1.61)	1.61	0.505		**0.76**	(0.33,1.74)	0.518		19/35	**0.55**	(0.24,1.23)	0.143		**0.63**	(0.26,1.51)	0.296		
	3	35/33	**0.93**	(0.48,1.81)	1.81	0.837		**1.07**	(0.49,2.35)	0.867		42/34	**1.48**	(0.73,3)	0.28		**1.39**	(0.64,3)	0.406		
	4	26/33	**0.57**	(0.25,1.31)	1.31	0.188	*0.322*	**0.58**	(0.21,1.58)	0.285	*0.565*	47/34	**1.68**	(0.87,3.25)	0.125	*0.042**	**1.75**	(0.82,3.7)	0.145	*0.074*	
Lead	1	33/33	**1**					**1**				39/35	**1**				**1**				
	2	29/33	**0.96**	(0.46,1.99)	1.99	0.909		**0.88**	(0.38,2.06)	0.772		33/34	**0.8**	(0.35,1.82)	0.603		**0.54**	(0.19,1.54)	0.252		
	3	40/33	**1.27**	(0.59,2.74)	2.74	0.543		**1.07**	(0.44,2.58)	0.885		31/34	**0.75**	(0.34,1.66)	0.48		**0.61**	(0.22,1.72)	0.355		
	4	31/33	**1.03**	(0.43,2.47)	2.47	0.956	*0.774*	**0.79**	(0.28,2.27)	0.663	*0.805*	34/34	**0.78**	(0.3,2.07)	0.624	*0.579*	**0.6**	(0.19,1.94)	0.395	*0.487*	

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence level*

5.3.1 Subtype-specific results

Information on histological subtype was available for 225 of 270 cases (83%, for details see Methods Figure 3.1.2) and the remaining 45 cases were classified as 'B-cell NHL, not otherwise specified'. The most common subtype was MM, which accounted for 28% of cases. The association between log transformed exposure levels and risk for the four largest subgroups are shown in table 5.3.1. There did appear to be some differences in risk estimates by subtype, particularly when confounders were taken into account. The significant inverse associations were limited to DLBCL. In fact the inverse associations with DLBCL, and to some extent MM, were quite striking, although they were based on relatively small numbers, and again multiple testing was not taken into account. When a meta-analyses of each exposure by subtype was conducted, the I^2 was 0% for all exposures except PCB170 (19.1%) and PCB180 (30.1%), and the p values were >0.1, indicating no significant heterogeneity between the subtypes. Subtype specific analyses stratified by cohort and sex were based on small numbers and no significant associations were revealed (Appendix: tables 5.3.1i and 5.3.1ii).

Table 5.3.1: Association between log-transformed body burden of six PCB congeners, specified PCB functional groups, HCB, DDE, Cadmium and Lead levels and NHL risk stratified by Subtype

Exposure	CLL (n=42)						DLBCL (n=45)					
	OR	95% CI	p-value	$OR^{(adj)}$	95% CI	p-value	OR	95% CI	p-value	$OR^{(adj)}$	95% CI	p-value
PCB118	**0.70**	**(0.31,1.58)**	0.398	**5.14**	**(0.56,47.08)**	0.147	**0.74**	**(0.27,2.04)**	0.561	**0.39**	**(0.06,2.68)**	0.335
PCB138	**0.63**	**(0.26,1.54)**	0.309	**1.33**	**(0.15,11.91)**	0.801	**0.37**	**(0.10,1.32)**	0.126	**0.33**	**(0.04,2.47)**	0.281
PCB153	**0.50**	**(0.16,1.52)**	0.221	**0.70**	**(0.06,7.77)**	0.772	**0.32**	**(0.08,1.26)**	0.103	**0.16**	**(0.01,1.84)**	0.140
PCB156	**0.57**	**(0.16,2.10)**	0.402	**1.47**	**(0.11,19.75)**	0.773	**0.25**	**(0.06,0.96)**	0.044*	**0.01**	**(0,0.53)**	0.023*
PCB170	**0.52**	**(0.16,1.66)**	0.267	**0.34**	**(0.03,3.61)**	0.372	**0.24**	**(0.06,1.02)**	0.053	**0.01**	**(0,0.47)**	0.021*
PCB180	**0.49**	**(0.14,1.77)**	0.278	**0.27**	**(0.02,4.15)**	0.351	**0.23**	**(0.05,0.97)**	0.045*	**0.01**	**(0,0.50)**	0.020*
Dioxin-like PCBs	**0.76**	**(0.43,1.35)**	0.346	**2.21**	**(0.56,8.67)**	0.255	**0.59**	**(0.29,1.19)**	0.138	**0.22**	**(0.04,1.09)**	0.063
Non-Dioxin like PCBs	**0.84**	**(0.62,1.13)**	0.247	**0.84**	**(0.44,1.62)**	0.604	**0.70**	**(0.48,1.02)**	0.061	**0.46**	**(0.21,1.00)**	0.051
Immunotoxic PCBs	**0.85**	**(0.63,1.14)**	0.282	**1.17**	**(0.60,2.29)**	0.651	**0.71**	**(0.48,1.04)**	0.080	**0.46**	**(0.21,1.01)**	0.054
ΣPCBs	**0.89**	**(0.73,1.09)**	0.259	**1.02**	**(0.65,1.60)**	0.924	**0.79**	**(0.61,1.02)**	0.067	**0.58**	**(0.34,0.99)**	0.047*
HCB	**0.42**	**(0.14,1.26)**	0.121	**1.12**	**(0.17,7.40)**	0.905	**0.90**	**(0.34,2.39)**	0.834	**0.41**	**(0.08,2.00)**	0.271
DDE	**0.83**	**(0.44,1.58)**	0.578	**1.99**	**(0.52,7.59)**	0.313	**0.88**	**(0.43,1.84)**	0.743	**1.17**	**(0.35,3.97)**	0.796
Cadmium	**0.88**	**(0.43,1.83)**	0.742	**0.60**	**(0.11,3.32)**	0.556	**1.15**	**(0.60,2.18)**	0.675	**1.85**	**(0.66,5.13)**	0.239
Lead	**0.77**	**(0.30,1.99)**	0.588	**0.43**	**(0.05,3.90)**	0.450	**0.76**	**(0.28,2.03)**	0.581	**0.70**	**(0.17,2.95)**	0.629

Results for the four largest subtype groups were included

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence level*

Table 5.3.1: ***continued***

Exposure	FL (n=39)						MM (n=76)					
	OR	95% CI	p-value	$OR^{b(adj)}$	95% CI	p-value	OR	95% CI	p-value	$OR^{b(adj)}$	95% CI	p-value
PCB118	**0.95**	**(0.41,2.21)**	0.904	**1.04**	**(0.34,3.13)**	0.947	**1.06**	**(0.55,2.07)**	0.856	**0.62**	**(0.23,1.63)**	0.330
PCB138	**0.91**	**(0.31,2.67)**	0.865	**1.17**	**(0.29,4.77)**	0.822	**0.74**	**(0.32,1.70)**	0.474	**0.45**	**(0.15,1.32)**	0.145
PCB153	**0.94**	**(0.29,3.04)**	0.918	**1.20**	**(0.27,5.31)**	0.814	**0.73**	**(0.29,1.85)**	0.512	**0.41**	**(0.12,1.42)**	0.159
PCB156	**0.99**	**(0.32,3.01)**	0.979	**1.12**	**(0.24,5.36)**	0.885	**0.77**	**(0.31,1.88)**	0.561	**0.47**	**(0.14,1.62)**	0.234
PCB170	**1.39**	**(0.43,4.51)**	0.580	**1.65**	**(0.33,8.24)**	0.540	**0.77**	**(0.30,2.00)**	0.588	**0.42**	**(0.11,1.66)**	0.216
PCB180	**1.43**	**(0.44,4.63)**	0.553	**1.52**	**(0.31,7.41)**	0.605	**0.90**	**(0.35,2.31)**	0.834	**0.56**	**(0.15,2.10)**	0.392
Dioxin-like PCBs	**0.97**	**(0.56,1.69)**	0.927	**1.04**	**(0.51,2.11)**	0.914	**0.97**	**(0.63,1.48)**	0.871	**0.70**	**(0.39,1.27)**	0.244
Non-Dioxin like PCBs	**1.03**	**(0.77,1.39)**	0.825	**1.08**	**(0.74,1.59)**	0.688	**0.93**	**(0.73,1.19)**	0.570	**0.81**	**(0.58,1.11)**	0.192
Immunotoxic PCBs	**1.01**	**(0.75,1.35)**	0.964	**1.05**	**(0.72,1.53)**	0.790	**0.95**	**(0.75,1.20)**	0.673	**0.81**	**(0.59,1.11)**	0.183
ΣPCBs	**1.01**	**(0.83,1.24)**	0.908	**1.04**	**(0.81,1.34)**	0.759	**0.97**	**(0.82,1.13)**	0.662	**0.87**	**(0.70,1.08)**	0.195
HCB	**0.82**	**(0.36,1.86)**	0.633	**0.83**	**(0.25,2.70)**	0.752	**0.70**	**(0.30,1.61)**	0.394	**0.31**	**(0.08,1.19)**	0.089
DDE	**0.77**	**(0.41,1.48)**	0.437	**0.93**	**(0.34,2.55)**	0.885	**0.66**	**(0.37,1.15)**	0.142	**0.52**	**(0.25,1.10)**	0.086
Cadmium	**1.59**	**(0.74,3.39)**	0.235	**2.03**	**(0.69,5.95)**	0.199	**1.05**	**(0.69,1.61)**	0.814	**1.47**	**(0.82,2.63)**	0.199
Lead	**1.37**	**(0.48,3.93)**	0.556	**1.17**	**(0.28,4.93)**	0.831	**1.12**	**(0.59,2.12)**	0.731	**1.20**	**(0.50,2.85)**	0.684

5.3.2 Stratification by potential confounders

In an attempt to explain the unexpected inverse associations noted for many of the POPs, analyses were stratified by age group, time to diagnosis and BMI category to determine whether there was evidence for confounding or effect modification. Age at recruitment (and therefore at blood draw; table 5.3.2a) did not appear to be modifying the findings, with no trend in risk observed and no significant heterogeneity in risk estimates between the age groups. Similarly, there was no evidence of an effect of time to diagnosis (table 5.3.2b). This remained true when confounding was taken into account. For both the age groups and time to diagnosis stratified analyses results from the meta-analysis indicated that the percentage of heterogeneity not due to chance (I^2) was 0% for all exposures, and all p-values were non-significant, when considering both the adjusted and the unadjusted models. When sex and cohort specific-analyses were run, the inverse associations were mainly confined to males, with females tending to show positive associations for most exposures (Appendix: tables 5.3.2i to 5.3.2iv). Nevertheless, none were found to be significant even without correction for multiple testing. Stratified analyses by subtype were not run due to small numbers, but this would be of interest for the future.

Table 5.3.2a; Association between log-transformed body burden of six PCB congeners, specified PCB functional groups, HCB, DDE, Cadmium and Lead levels and NHL risk stratified by age group at recruitment

Exposure	30-44 years (Cases n=40, Controls n=40)			45-59 years (Cases n=169, Controls n=158)			60-75 years (Cases n=61, Controls n=72)		
	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
PCB118	0.39	(0.11,1.35)	0.136	0.89	(0.58,1.37)	0.597	1.48	(0.69,3.17)	0.308
PCB138	0.39	(0.10,1.52)	0.176	0.80	(0.50,1.28)	0.347	0.98	(0.39,2.48)	0.968
PCB153	0.33	(0.07,1.51)	0.152	0.76	(0.45,1.30)	0.323	0.83	(0.31,2.22)	0.710
PCB156	0.15	(0.02,0.97)	0.047*	0.81	(0.44,1.52)	0.516	0.85	(0.31,2.33)	0.747
PCB170	0.31	(0.07,1.42)	0.130	0.85	(0.47,1.55)	0.596	0.79	(0.26,2.45)	0.686
PCB180	0.36	(0.08,1.64)	0.187	0.93	(0.51,1.70)	0.812	0.79	(0.25,2.48)	0.688
Dioxin-like PCBs	0.37	(0.14,0.99)	0.047*	0.91	(0.68,1.21)	0.516	1.14	(0.70,1.86)	0.600
Non-Dioxin like PCBs	0.75	(0.51,1.11)	0.149	0.95	(0.82,1.09)	0.455	0.96	(0.73,1.25)	0.762
Immunotoxic PCBs	0.64	(0.39,1.05)	0.076	0.95	(0.82,1.09)	0.448	1.02	(0.79,1.33)	0.858
ΣPCBs	0.77	(0.56,1.05)	0.097	0.96	(0.87,1.06)	0.460	1.00	(0.84,1.20)	0.992
HCB	0.47	(0.13,1.74)	0.257	0.76	(0.46,1.24)	0.267	0.62	(0.21,1.82)	0.389
DDE	0.80	(0.28,2.30)	0.685	0.85	(0.63,1.14)	0.270	0.86	(0.50,1.48)	0.579
Cadmium	0.95	(0.51,1.78)	0.878	1.35	(0.98,1.87)	0.064	0.80	(0.43,1.46)	0.457
Lead	0.64	(0.28,1.42)	0.272	1.36	(0.75,2.47)	0.304	0.65	(0.24,1.74)	0.394
	$OR^{(adj)}$	95% CI	p-value	$OR^{(adj)}$	95% CI	p-value	$OR^{(adj)}$	95% CI	p-value
PCB118	0.36	(0.04,3.04)	0.345	0.72	(0.43,1.23)	0.235	1.10	(0.4,3.01)	0.858
PCB138	0.43	(0.03,5.67)	0.521	0.59	(0.31,1.13)	0.110	0.70	(0.19,2.54)	0.584
PCB153	0.44	(0.02,9.11)	0.593	0.55	(0.27,1.15)	0.115	0.71	(0.2,2.53)	0.596
PCB156	0.21	(0.01,4.22)	0.310	0.70	(0.32,1.55)	0.382	0.90	(0.25,3.29)	0.874
PCB170	0.63	(0.05,8.89)	0.736	0.67	(0.31,1.44)	0.308	0.82	(0.18,3.77)	0.804
PCB180	0.97	(0.04,22.46)	0.983	0.76	(0.35,1.64)	0.483	0.87	(0.19,3.99)	0.860
Dioxin-like PCBs	0.34	(0.06,1.85)	0.211	0.81	(0.57,1.15)	0.244	1.01	(0.54,1.9)	0.972
Non-Dioxin like PCBs	0.85	(0.4,1.8)	0.68	0.88	(0.73,1.06)	0.18	0.93	(0.65,1.33)	0.69
Immunotoxic PCBs	0.66	(0.28,1.53)	0.33	0.88	(0.73,1.06)	0.18	0.97	(0.68,1.37)	0.85
ΣPCBs	0.80	(0.46,1.41)	0.446	0.92	(0.81,1.04)	0.185	0.97	(0.77,1.23)	0.801
HCB	0.45	(0.04,4.71)	0.506	0.69	(0.39,1.22)	0.198	0.30	(0.06,1.54)	0.150
DDE	1.16	(0.13,10.58)	0.898	0.65	(0.44,0.98)	0.038*	0.53	(0.24,1.16)	0.114
Cadmium	17.57	(1.5,205.32)	0.022*	1.25	(0.84,1.85)	0.271	0.95	(0.43,2.08)	0.900
Lead	0.89	(0.27,2.9)	0.849	1.06	(0.5,2.24)	0.879	0.66	(0.17,2.57)	0.548

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence level*

Table 5.3.2b; Association between log-transformed body burden of six PCB congeners, specified PCB functional groups, HCB, DDE, Cadmium and Lead levels and NHL risk stratified by time to diagnosis

Exposure	ttd< 5 years (n=108)			ttd> 5 years (n=149)			ttd< 5 years (n=108)			ttd> 5 years (n=149)		
	OR	95% CI	p-value	OR	95% CI	p-value	$OR^{(adj)}$	95% CI	p-value	$OR^{(adj)}$	95% CI	p-value
PCB118	0.82	(0.52,1.27)	0.368	0.95	(0.64,1.41)	0.808	0.75	(0.47,1.22)	0.253	0.88	(0.55,1.39)	0.576
PCB138	0.63	(0.39,1.01)	0.056	0.73	(0.46,1.14)	0.166	0.57	(0.34,0.95)	0.032*	0.63	(0.36,1.1)	0.105
PCB153	0.59	(0.34,1.02)	0.058	0.67	(0.39,1.13)	0.129	0.54	(0.3,0.98)	0.042*	0.58	(0.31,1.08)	0.084
PCB156	0.67	(0.36,1.27)	0.22	0.7	(0.4,1.22)	0.204	0.68	(0.34,1.33)	0.26	0.65	(0.35,1.23)	0.186
PCB170	0.61	(0.33,1.15)	0.125	0.71	(0.4,1.28)	0.258	0.59	(0.3,1.16)	0.129	0.62	(0.32,1.22)	0.169
PCB180	0.62	(0.33,1.17)	0.141	0.77	(0.43,1.39)	0.385	0.59	(0.29,1.18)	0.138	0.7	(0.35,1.39)	0.31
Dioxin-like PCBs	0.84	(0.63,1.13)	0.247	0.91	(0.7,1.17)	0.454	0.82	(0.6,1.12)	0.211	0.86	(0.65,1.16)	0.33
Non-Dioxin like PCBs	0.87	(0.75,1.01)	0.068	0.91	(0.79,1.05)	0.187	0.85	(0.73,1)	0.053	0.88	(0.75,1.04)	0.127
Immunotoxic PCBs	0.89	(0.76,1.03)	0.115	0.93	(0.81,1.06)	0.274	0.87	(0.74,1.02)	0.09	0.9	(0.77,1.05)	0.187
ΣPCBs	0.92	(0.83,1.02)	0.097	0.95	(0.86,1.04)	0.244	0.91	(0.81,1.01)	0.078	0.93	(0.83,1.03)	0.17
HCB	0.63	(0.38,1.06)	0.082	0.74	(0.46,1.21)	0.232	0.62	(0.36,1.07)	0.086	0.65	(0.37,1.14)	0.13
DDE	0.73	(0.54,0.99)	0.045*	0.79	(0.59,1.06)	0.118	0.65	(0.47,0.92)	0.013*	0.71	(0.51,1.01)	0.054
Cadmium	1.10	(0.81,1.5)	0.551	1.03	(0.78,1.35)	0.846	1.14	(0.8,1.61)	0.471	1.04	(0.76,1.44)	0.788
Lead	0.68	(0.38,1.21)	0.19	0.91	(0.56,1.46)	0.69	0.75	(0.4,1.39)	0.36	0.8	(0.47,1.39)	0.437

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence level*

ttd-time to diagnosis

Stratified analyses by BMI (table 5.3.2c), showed that risk estimates tended to decrease with BMI category, with the smallest ORs generally noted in obese participants. This was also observed in the stratified analyses with patients of normal weight more likely to show a positive association with NHL exposure in both cohorts and both genders (Appendix: tables 5.3.2v and 5.3.2vi). Effect modification by BMI may also in part explain the observed sex discrepancy, 69.6% of males are overweight or obese compared with only 50.2% of females ($p<0.0001$) and the inverse associations tended to be restricted to males. Although the results of the meta-analysis do not show significant heterogeneity between BMI strata, the p-values were much smaller than for subtype or age and the I^2 percentages were greater than 0%.

Table 5.3.2c; Association between log-transformed body burden of six PCB congeners, specified PCB functional groups, HCB, DDE, Cadmium and Lead levels and NHL risk stratified by BMI category

Exposure	Normal (Cases n=105, Controls n=109)			Overweight (Cases n=121, Controls n=118)			Obese (Cases n=40, Controls n=39)		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
PCB118	0.95	(0.43,2.13)	0.906	0.58	(0.28,1.22)	0.152	0.49	(0.06,3.72)	0.488
PCB138	0.77	(0.36,1.64)	0.497	0.16	(0.05,0.6)	0.006*	0.21	(0.02,2.87)	0.242
PCB153	0.74	(0.31,1.8)	0.511	0.17	(0.04,0.67)	0.011*	0.07	(0,2.91)	0.16
PCB156	0.78	(0.29,2.1)	0.621	0.35	(0.1,1.16)	0.085	0.07	(0,3.1)	0.17
PCB170	0.83	(0.31,2.27)	0.721	0.34	(0.1,1.13)	0.078	3.55E-03	(0,9.16)	0.159
PCB180	0.83	(0.29,2.34)	0.721	0.31	(0.09,1.04)	0.058	2.24E-03	(0,8.67)	0.148
Dioxin-like PCBs	0.92	(0.55,1.53)	0.739	0.63	(0.38,1.06)	0.083	0.49	(0.14,1.77)	0.278
Non-Dioxin like PCBs	0.94	(0.74,1.18)	0.58	0.66	(0.48,0.93)	0.016*	0.42	(0.13,1.34)	0.141
Immunotoxic PCBs	0.94	(0.73,1.21)	0.632	0.72	(0.53,0.97)	0.029*	0.60	(0.27,1.3)	0.194
ΣPCBs	0.96	(0.81,1.13)	0.615	0.78	(0.63,0.97)	0.023*	0.65	(0.36,1.19)	0.162
HCB	1.23	(0.45,3.36)	0.688	0.35	(0.12,1.03)	0.056	2.46E-03	(0,6.06)	0.132
DDE	0.83	(0.45,1.5)	0.529	0.28	(0.12,0.66)	0.003*	0.42	(0.06,3.23)	0.407
Cadmium	1.19	(0.7,2.02)	0.53	1.37	(0.76,2.45)	0.297	0.34	(0.02,4.77)	0.424
Lead	3.10	(0.89,10.8)	0.076	0.71	(0.3,1.65)	0.423	1.23	(0.18,8.44)	0.833
	$OR^{(adj)}$	**95% CI**	**p-value**	**$OR^{(adj)}$**	**95% CI**	**p-value**	**$OR^{(adj)}$**	**95% CI**	**p-value**
PCB118	0.68	(0.19,2.45)	0.551	0.49	(0.18,1.3)	0.153	Δ		
PCB138	0.53	(0.12,2.28)	0.393	0.08	(0.01,0.49)	0.007*			
PCB153	0.62	(0.12,3.25)	0.572	0.12	(0.02,0.68)	0.016*			
PCB156	1.06	(0.26,4.27)	0.934	0.28	(0.06,1.25)	0.095			
PCB170	1.44	(0.26,7.94)	0.678	0.22	(0.04,1.12)	0.069			
PCB180	1.44	(0.24,8.51)	0.685	0.19	(0.04,0.97)	0.046*			
Dioxin-like PCBs	0.88	(0.41,1.91)	0.749	0.56	(0.29,1.09)	0.090			
Non-Dioxin like PCBs	0.95	(0.63,1.45)	0.827	0.58	(0.37,0.91)	0.018*			
Immunotoxic PCBs	0.93	(0.61,1.41)	0.737	0.64	(0.43,0.96)	0.031*			
ΣPCBs	0.96	(0.72,1.28)	0.79	0.72	(0.55,0.96)	0.026*			
HCB	1.62	(0.4,6.65)	0.502	0.19	(0.04,0.96)	0.044*			
DDE	0.54	(0.22,1.35)	0.186	0.11	(0.02,0.53)	0.006*			
Cadmium	1.29	(0.54,3.09)	0.565	1.66	(0.74,3.71)	0.220			
Lead	8.08	(1.16,56.44)	0.035*	0.63	(0.2,2.02)	0.438			

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence level*

Δ *model would not converge*

5.3.3 Highly exposed individuals

The inverse associations also disappeared when risk was assessed in 'highly exposed' individuals i.e. when the risk in those with exposure levels above the 90th percentile of the control distribution was compared to those in the lowest quartile of exposure (table 5.3.3). Interestingly though, this was not observed for Cd which was found to be significantly associated with risk under almost every other model, particularly in females.

5.3.4 Summary

These findings have not been corrected for multiple testing. Given the large number of analyses explored including multiple strata and subcohorts, overall these results suggest a null association between body burden of six PBC congeners, HCB, DDE, Cd or Pb.

Table 5.3.3: Association between exposure and risk in 'highly exposed' individuals

High exposure is defined as exposure levels above the 90th of the control exposure distribution, the odds of disease are compared with those in the lowest quartile of exposure

Exposure	90th percentile control exposure distribution	Above 90th percentile		OR	95% CI	p-value	$OR^{(adj)}$	95% CI	p-value
		n cases (%)	*n* controls (%)						
PCB118	316.2	35(13.0)	28(10.4)	1.37	(0.76,2.45)	0.294	1.25	(0.65,2.41)	0.500
PCB138	1112.4	25(9.3)	29(10.7)	0.83	(0.45,1.52)	0.550	0.81	(0.40,1.64)	0.562
PCB153	1965.7	29(10.7)	29(10.7)	1.01	(0.56,1.81)	0.987	1.15	(0.58,2.26)	0.693
PCB156	176.4	23(8.5)	28(10.4)	0.79	(0.43,1.45)	0.447	0.89	(0.43,1.81)	0.739
PCB170	626.6	29(10.7)	29(10.7)	1.00	(0.57,1.78)	0.989	1.07	(0.55,2.06)	0.843
PCB180	1323.3	29(10.7)	28(10.4)	1.05	(0.58,1.88)	0.872	1.11	(0.57,2.14)	0.763
Dioxin-like PCBs	478.7	34(12.6)	29(10.7)	1.26	(0.70,2.26)	0.445	1.30	(0.67,2.53)	0.432
Non-Dioxin like PCBs	4937.2	29(10.7)	29(10.7)	1.00	(0.56,1.80)	0.988	1.18	(0.60,2.30)	0.626
Immunotoxic PCBs	872.8	33(12.2)	29(10.7)	1.19	(0.67,2.11)	0.549	1.16	(0.60,2.24)	0.665
ΣPCBs	5438.8	30(11.1)	28(10.4)	1.10	(0.61,1.97)	0.753	1.27	(0.65,2.51)	0.485
HCB	872.8	25(9.3)	29(10.7)	0.8	(0.41,1.55)	0.506	0.84	(0.42,1.68)	0.624
DDE	9787.7	27(10)	28(10.4)	0.95	(0.50,1.80)	0.878	1.01	(0.52,1.97)	0.978
Cadmium	1.7	22(8.1)	27(10.0)	0.79	(0.44,1.44)	0.444	0.77	(0.39,1.52)	0.458
Lead	125.3	30(11.1)	26(9.60)	1.22	(0.66,2.27)	0.528	1.20	(0.60,2.41)	0.612

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence level*

5.4 Discussion

In this chapter a prospective design was used to study the risk of NHL associated with internal dose of six PCB congeners, HCB, DDE, Cd and Pb under the hypothesis that an elevated body burden may increase future risk of this malignancy. This hypothesis was based on a body of epidemiological literature and experimental evidence. However these findings could not be replicated and instead a number of unexpected inverse associations with certain exposures are reported. Overall, due to non-correction for multiple testing, the evidence suggests a null relationship between the investigated exposures and NHL risk. Consequently it can be concluded that internal dose of these pollutants, as assessed by a single spot measurement in this population, has no utility in the prediction of NHL or the study of its aetiology.

Existing evidence from the literature

Although PCBs have been widely explored in relation to NHL, the evidence has been contradictory. While many studies report a positive association with the sum of PCB body burden, PCB mixtures (including dioxin-like, non-dioxin like and immunotoxic) and individual congeners [76-80], an equally large number have found no association [81-86]. Many of the positive findings come from case-control studies which are subject to the problems of reverse causation and other bias, or from investigations of high levels of accidental exposure [71], but even among highly exposed occupational cohorts the results have not been consistent [250]. Of the studies most directly comparable with this chapter, those utilising prediagnostic exposure measurements in the general population, two [80,275] reported an increased risk with total PCB exposure in both sexes, while a third [77] reported an increase in males. Two further studies [81,82] reported no significant associations for any of the investigated congeners. Even systematic reviews have been unable to confirm or refute an association. A recent meta-analysis reports a weight-adjusted odds ratio for all PCB congeners of 1.43 (95% CI 1.31, 1.55) and concludes that the literature supports a causal association between certain congeners and NHL risk [63] . Conversely, a separate review [250] states that it is unwarranted to conclude that PCB exposure is related to risk of NHL citing methodological limitations, congener-specific discrepancies in results, the lack of dose-dependent relationships and the likelihood of spurious associations arising due to multiple testing and chance.

The literature is similarly inconsistent for DDE and for HCB, which showed some of the strongest inverse associations in this population. Three case control-studies have reported a positive association between NHL and DDE. One was based on adipose tissue concentrations [86] and two measured levels in blood [76,78]. However, in a study of serum organochlorine levels in participants living near a municipal solid waste incinerator the relationship was borderline significant and the OR indicated only a marginal increase in risk [76]. Additionally a number of both case-control [79,85] and cohort [77,81,82] studies

have reported no association. Similarly, for those studies considering HCB; although one [78] observed a positive association with risk, two other studies report null findings [76,81]. Other work suggests that both DDE and HCB may act on risk through an interaction with antibodies to EBV early antigen (EA IgG) [83,276], but again these findings remain to be validated.

The putative association between heavy metals and NHL has been much less extensively studied, and the biological rationale for a relationship is based largely on animal models. In vitro studies indicate that both Cd and Pb have genotoxic properties [277,278] and can induce DNA damage via elevated levels of oxidative stress [263]. Additionally, there is evidence that, like POPs, they may interact with the immune system [259,267,270,271,279,280]. The epidemiological evidence is based mainly on residential proximity to sources of exposure such as oil refineries, however these also emit other suspected haematological carcinogens including dioxins, which may be confounding the relationship [248,281]. Only one previous study which directly explored the association with Cd in humans [282] reported an association between urinary Cd and NHL in males and an association with leukaemia in both males and females, a finding consistent with a case-control study in a Turkish population [272,282]. This study is the first to report on the risk of any haematological malignancy and internal dose of Pb.

Exposure assessment

In this study of a general, non-occupationally exposed cohort, exposure was assessed using erythrocyte concentrations of Pb and Cd and blood serum concentrations of eight POPs in samples taken from cancer-free individuals at recruitment to the cohort. This provides a point estimate of historical exposure free from recall bias. Internal doses of environmental pollutants vary by orders of magnitude from individual to individual. This can be attributed to various factors including, age, sex, area of residence, calendar year of sampling and route of exposure [283]. Global restrictions in the use of such pollutants has led to a reported decrease in human body burden over the last decade: for PCBs current exposure levels are estimated to be half those of 20-30 years ago[283], making it difficult to compare between populations and across periods. However, the observed exposure levels for both the POPs and the metals were within the range of those reported in studies of similar populations [284-287]. When considering the total population, most of the investigated exposures were observed at higher levels in the Italian cohort; only PCB170 was much higher in NSHDS. PCB118, HCB, DDE and Cd were significantly higher in females, while PCB170, PCB180 and Pb were significantly higher in males.

Sex specific effects and the influence of genetic susceptibility

One of the most striking findings of this study was the sex specific effects which were evident for nearly all explored analyses and models. In general, inverse associations were noted for males, while females tended to show positive, although non-significant associations, for most exposures. There was some evidence for an association between NHL and Cd in females, particularly at high exposure

levels. There is little epidemiological evidence reporting on the relationship between metals and NHL by sex although other Cd-associated health effects have previously been observed to be more common in females [288].

As discussed in the introduction, a number of studies have reported on genetic polymorphisms which may influence the uptake, metabolism and storage of these pollutants and which may be affecting the results [124,199,234,263,289-291]. Crucially a number of these genes have shown sex-specific differences in risk [194,234,292], and there has been an increasing interest in sex-specific susceptibility to carcinogens [293]. Sex-differences in the expression of xenobiotic metabolising enzymes including the cytochrome P450 system and the human glutathione S-transferases [294,295], and in candidate transporter genes for carcinogen export have all been reported [293]. A growing body of mechanistic, experimental and epidemiological evidence suggests that women may be more susceptible to the effect of carcinogens than males. This relates both to baseline differences in exposure levels to putative carcinogens as well hormonal, physiological, genetic and epigenetic effects [294,295]. Accordingly a number of studies have observed greater levels of markers of DNA damage including DNA adduct levels, sister chromatid exchanges and micronucleus frequency in female cancer patients compared to similarly exposed male cases [293]. DNA repair capacity has also been reported to be substantially lower in females, while expression of the oncogenic k-ras mutation has been found to be higher. It is thought these effects may be regulated by hormonal responses and that oestrogen in particular may play an important role [293-295]. Such variations may go some way to explaining these findings or alternatively the results may simply reflect the lower baseline risk of NHL in unexposed women [293]; NHL is 50% more common among men [3].

Potential confounding and effect modification

The heterogeneous nature of NHL may, in part, explain the differing findings between this study and previous work, as it is likely that the risk associated with POPs differs by subtype [86]. To take potential subtype heterogeneity into account, subtype specific analyses were conducted to the most detailed extent allowable by sample size, according to the hierarchies proposed by Morton *et al.* [19]. Despite the differing effect estimates observed between subtypes, most were non-significant and there was no evidence of subtype heterogeneity. However it should be noted that due to small numbers the power to detect subtype-specific associations was limited and the majority of rarer subtypes could not be analysed at all.

The differing composition of the PCB congeners included in different studies may also account for some of the observed discrepancy, particularly where a risk effect for the sum of PCBs has been reported. Individual congeners have been shown to have different biological activity, and therefore may produce health effects by different mechanisms [77,296]. In this study six PCB congeners were considered, all of which were non-coplanar and highly chlorinated, ranging from five (PCB118) to

seven (PCB170 and PCB180) substituted chlorine atoms, and can reasonably be considered as a group. Consequently, the sum of these six PCBs as well as dioxin-like congeners, non-dioxin-like congeners and immunotoxic congeners were also explored. However, due to the very high correlation between the included congeners in the respective groups, the mixtures merely reflected the constituent congeners and did not provide any further information on the relationship with NHL in this study.

In order to try and further disentangle the unexpected results observed, a number of additional analyses were performed. Previous studies of NHL and environmental exposures have noted an effect of 'time to diagnosis' i.e. the time between sample collection and diagnosis date, with stronger associations observed closer to diagnosis [81,82,249]. This is thought to be due to the pathophysiologic and metabolic changes, such as lipid mobilisation and weight loss, accompanying carcinogenesis which may affect the body burden of exposures [297]. To ensure that pre-diagnostic subclinical changes in cases who were diagnosed shortly after blood draw were not modifying the body burden of exogenous exposures [297], stratification by the time to diagnosis was performed. Results appeared consistent across strata. No effect of age at diagnosis was determined and adjustment for biologically plausible confounders did not alter the conclusions.

There was, however, some evidence of an effect of BMI, with the most strongly inverse associations noted in those who were overweight and obese. This may be of interest as organochlorines are stored in adipose tissue and therefore PCB levels in the blood have previously been reported to be inversely associated with weight gain, and overweight and obese individuals have been observed to have lower circulating levels [132,296]. In this analysis overweight/obese cases were compared to overweight/obese controls at baseline, therefore this would not explain the inverse associations between exposure and risk, unless BMI confounds the relationship between PCB levels and NHL through some unknown mechanism. Although no association between BMI and NHL was observed in this study it has been noted previously by others [99]. If BMI does act as a confounder this could explain the high number of inverse associations noted in this population which contained a high proportion (60%) of overweight and obese individuals. It would also explain why inverse associations were overrepresented in males, who had a significantly higher proportion of overweight and obese participants (69.6% compared to only 50.2% in females ($p<0.001$)). Furthermore it would explain why Cd and Pb which are not known to be lipophilic were not affected in the BMI stratified analyses.

Strengths and Limitations

The main strength of this study is its prospective design which protects against selection bias and it was determined that the case and control populations were similar with respect to baseline factors and additional confounders. This design also allows for the determination of a temporal and causal relationship between exposure and NHL, lacking in the majority of studies which tend to measure post-diagnostic or post-treatment levels. NHL or its treatment may affect metabolism and blood

concentrations of POPs [77,132], and chemotherapy has been observed to decrease PCB levels in the body by up to 30% [63]. The measures of exposure assessment were also free from recall bias, and were conducted using validated methodologies. The use of blood serum measurements has been shown to provide a reliable estimate of POP body burden due to their long half-lives [82]. Similarly Pb binds to erythrocytes and therefore body burden is optimally assessed by blood concentration [298]. Although urine is considered to be the best biomaterial to measure biomarkers of lifetime Cd exposure, blood Cd has been found to correlate highly with Cd concentrations in both urine and tissue and as such can be considered as a good estimate of body burden [267].

In this study, lipids were not adjusted for as it has been observed that direct standardisation can introduce bias [299]. Unadjusted volume-based concentrations of POPs, which have been found to correlate well with lipid based concentrations, are presented [300]. Furthermore, the possibility of 'reverse causation' and disease progression bias [297], which are the main reasons to adjust for lipids, (in order to take physiological pre-cancerous changes to normal metabolic process and lipid mobilisation into account), have been considered. The findings for BMI suggest that further exploration of these results using lipid adjusted concentrations may be of interest.

There were several other limitations to this study. In addition to lipids, information on the blood concentrations of beta-HCH and gamma-HCH, or of dioxins was not available. The high correlation between exposures, particularly the PCBs, restricted the ability to determine individual effects [301]. It is possible that there were further unmeasured confounders affecting the results, for example other organochlorines, metals or trace elements [275] which may bioaccumulate in parallel with the exposures of interest and exert interaction effects on the risk of NHL. Furthermore variables relating to the immune system, such as EBV antigens which have been shown to interact with organochlorines in NHL risk [83,276] can also not be accounted for. The estimate of body burden is based on a single one spot exposure measurement that is not necessarily representative of lifetime exposure. Finally, the generalisability of these findings to other populations should also be considered, given the relative overrepresentation of females, the excess of MM cases and the relatively high proportion of overweight participants and smokers. In particular, these findings should not be extrapolated into an occupational setting.

Conclusions

In conclusion, in this prospective nested case control study there is no consistent positive association with prediagnostic erythrocyte concentrations of Cd or Pb and risk of NHL. In fact significantly inverse associations were noted for some PCB congeners. A study in Finland has previously suggested that dioxins may act on the risk of soft-tissue sarcoma through a J-shaped dose response curve, and there in some support from animal models for such a relationship [302]. The results of the 'high exposure' analysis could considered to be supportive of a J-shaped curve, however again these results

were non-significant, and the results from the quartile analysis offer no support. It seems more likely the findings arose as a result of multiple testing among strongly correlated congeners or effect modification as discussed. Sex-specific susceptibility to xenobiotics and potential carcinogens is also likely to play a role in the findings. For the PCBs and POPs this adds to the existing body of literature and is in agreement with those studies which have previously reported a null association between these exposures and NHL [250]. For the metals, particularly Pb, this study represents among the first evidence in the field, and there is some indication that the relationship between NHL and Cd in females is worthy of further explorations in a larger cohort.

Overall an essentially null relationship between the investigated exposure and NHL is reported. Therefore internal dose of these exposures as measured in serum and erythrocytes cannot be considered as a reliable or useful biomarker of NHL, and has no utility in prediction.

6. METABOLIC PROFILING OF AN NHL CASE CONTROL STUDY TO IDENTIFY PREDICTIVE METABOLIC BIOMARKERS

BIOMARKER: Individual metabolite features or a defined metabolite profile

BIOMARKER TYPE: Biomarker of biologically effective dose

POPULATION: B-cell NHL cases and controls from the EnviroGenoMarkers Study

6.1 Introduction

Metabolomics is the study of the metabolome, which has been defined as 'the quantitative complement of all of the low molecular weight molecules present in cells in a particular physiological or developmental state' [303]. The term metabolomics was coined in 2002 by Fiehn *et al.* [304]. It is often used interchangeably with metabonomics which was introduced by Nicholson *et al.* in 1999 [305] Both methodologies share similar analytical and modelling procedures. The main difference is that while metabonomics measures global systemic change, metabolomics aims to characterise and quantify all the small molecules in a biological sample [306,307].

An individual's metabolome reflects their exposure experience, including dietary and lifestyle factors, as well as their genetics/genomics, and the interactions between these variables [308]. In this way it provides a down-stream measure of a whole system's activity that reflects and measures the genome, epigenome, transcriptome and proteome. This metabolic end-point is consequently thought to be more closely related to the phenotype than these preceding 'omes' and in individuals with disease the metabolome has also been shown to be reflective of the disease state. As such it potentially represents a rich source for biomarker identification [307,309], particularly in cancer epidemiology as cancer cells are known to possess highly unique metabolic signatures [310]. Accordingly, a number of studies have shown the ability of metabolomics to distinguish healthy tissue from tumour tissue and healthy controls from cases in a variety of cancers [311,312]. Particularly promising results concerning the use of metabolomics as a potential diagnostic tool have been reported for both prostate and breast cancer [310]. However, to date the majority of metabolomics studies have been limited to case-control designs, resulting in the identification of diagnostic biomarker profiles or biomarkers which assist with the classification of tumours by subtype or stage. Very few studies have employed a prospective design in an epidemiological context, those that have tend to have been characterised by small participant numbers [307] and none have explored NHL.

There is a strong biological rationale for the identification of prospective biomarkers predictive of disease onset from the metabolome [7] as metabolic changes initiated by a disease process precede the

overt phenotypic and functional changes, due to the effect of cellular regulation on small-molecule substrates [309]. Therefore changes in the metabolome should be apparent and detectable before the observable onset of clinical disease, allowing a critical window for prediction or early detection.

This chapter aims to harness this theory and apply metabolomics to the prediction of NHL risk, based on the nested case-control design from EGM study. A non-targeted approach and Ultra-high performance liquid chromatography Mass Spectrometry (UPLC-MS) were employed to identify individual metabolite features, fingerprints, profiles, or signatures that predict risk. Validation of the findings is then attempted in an independent cohort. For the purposes of this chapter these profiles are considered as biomarkers of biologically effective dose, as metabolomics reflects both the environment and the phenotype [310]. This represents the first study to report on the feasibility of using metabolomics in prospective serum samples to identify biomarkers for risk of NHL.

6.2 Methods

6.2.1 Study subjects

This chapter is based on the 540 participants from the EGM study (see methods chapter 3.1.2, tables 3.1.2a- c and figure 3.1.2 for details). For the purposes of validation the cohort was considered as two separate populations: NSHDS and EPIC-Italy

Discovery cohort

The discovery cohort was composed of the NSHDS component of the EGM study (phase 1 and 2).

Validation cohort

The validation cohort was drawn from phase 2 of the EPIC-Italy component of the EGM study. Participants from phase 1 of EPIC-Italy were not included as they were processed using a different analytical QC and were therefore not comparable.

6.2.2 Metabolic profiling

Prospectively collected blood samples from each participant underwent metabolic profiling using UPLC-MS as described in the methods 3.2.2. Each subpopulation (NSHDS phase 1, NSHDS phase 2 and EPIC-Italy phase 2) was run separately using the same analytical QCs. Unique metabolite features that could be identified in all three populations based on their mass to charge ratio (*m/z*) and retention time (*rt*) were selected for analysis.

6.2.3 Statistical analysis

The Student's t-test was used to assess differences in average metabolite intensities by sex and by cohort. Due to the highly dimensional and collinear nature of MS data, classical univariate, unsupervised multivariate and supervised multivariate models were employed to explore the relationship between metabolite profiles and risk of disease. All analyses were first performed on the discovery cohort then repeated in the validation cohort.

Classical methods

Conditional logistic regression was used to determine the association between each metabolite feature and disease risk, treating feature intensities as continuous explanatory variables. The strongest associations were further explored by considering dose-response relationships: intensities were categorised into quartiles based on their distribution in controls. Tests for trend were performed using the Wald's test statistic. Pearson's correlation coefficient was used to quantify the correlation among the metabolite features.

Selection of potential confounders

To account for potential confounding, additional variables were included in the models. These were associated both with the metabolite features and with NHL, either in this population or in the literature. The associations between each covariate and every metabolite feature were computed using univariate linear regression where metabolite intensity was considered as the continuous outcome variable.

Time to diagnosis

To determine whether metabolite profiles, and potential associations differed as a function of time from blood draw to clinical onset, time to diagnosis (ttd) stratification was implemented by ttd <5 and ttd>5 years. Stratified analyses included all controls and adjusted for the matching factors: age, sex, metabolite batch and phase.

Multivariate projection methods

The data was imported into SIMCA-P software (SIMCA-P 12.0, Umetrics, Umeå, Sweden) to run both unsupervised and supervised multivariate projection analyses. The data were log-transformed and *Pareto* scaled prior to analysis. PCA was used to describe associations and patterns among the variables, considering sex, cohort, phase, age group and metabolomic processing batch in addition to case-control status. The QCs were also represented on the score plot to determine if they were evenly distributed throughout the samples and check for analytical drift. Strong outliers were identified using Hotelling's T2 range plot and DmodX (distance to model). Once identified, they were excluded from further analysis. For each component, R^2 was computed to determine the goodness-of-fit of the model and Q^2 to determine its predictive power. To guard against model over-fitting, a cross validation procedure (SIMCA's default 7-fold internal cross-validation) was implemented.

Partial least squares (PLS) and Orthogonal-partial least squares projection to latent structures-discriminant analysis (OPLS-DA) were carried out to discriminate the metabolic patterns between NHL cases and controls. The R^2 and Q^2 measures were inferred using cross validation. As a further measure of internal validation, a permutation test was conducted to compare the real Q^2 to a generated Q^2_{max} based on 999 permuted models. The variable importance in the projection (VIP) values of all the metabolite features and the S-plot from the cross-validated OPLS-DA model were used to select those which best discriminated cases from controls. The retained features were then taken forward for the ROC curve analysis.

Stratified analyses were also performed for the four largest subtype groups: MM, DLBCL, CLL and FL and their matched controls.

ROC curve analysis

The classical and multivariate projection methods resulted in the formation of potentially predictive metabolite feature sets which were then taken forward for logistic regression. For each exposure-disease combination, logistic models using PCA summary of the data subset were run and their predictive abilities were assessed by drawing ROC curves, and summarised using the AUC. The model with the highest AUC was identified and replication in the validation cohort was attempted.

6.3 Results

6.3.1 Dataset

After the exclusion of analytical failures (together with their matched pair), of the 186 case-control pairs (table 3.1.2a) a total of 183 pairs (366 participants) from the NSHDS component of the EGM study were included in the discovery cohort. Of the 34 phase 2 case-control pairs from EPIC-Italy 27 pairs (54 participants) were eligible for inclusion in the validation cohort (appendix table 6.3.1).

In this subset of cases from the EnviroGenoMarkers study, there was a higher proportion of NSHDS participants (87.1%) than in the total population (68.9%) due to the exclusion of EPIC-Italy phase 1 participants. There were also slightly more males (51.4% compared to 49.3%) and the average age was marginally younger in the subpopulation (52.8 years compared to 53.1 years). There was a borderline significant difference in height between cases and controls (171.1cm compared to 169.3cm, p=0.046), however there were no other significant differences and the distributions of the other investigated characteristics were very similar to the total EnviroGenoMarkers population, therefore this subpopulation can be considered representative.

A total of 2016 metabolite features were identified in NSHDS and 1959 in EPIC-Italy. In total 902 unique metabolite features were common between both sets of profiles based on their *m/*z and *rt*. Of these, 152 were excluded from further analysis as their coefficient of variation (CV) was >30% of the cohort-specific QC in at least one of the populations. For the remaining 750 features, 244 (32.5%) had a CV less than 15% in both populations and the CV% were similar between NSHDS and EPIC-Italy (appendix figure 6.3.1i). There were high levels of correlation between feature intensity (appendix figure 6.3.1ii). The QCs revealed no evidence of analytical drift. According to the test for skewness, 682 features (90.9%) displayed a non-normal distribution ($p<0.001$), so all features were log transformed to obtain normality.

A total of 198 (26.4%) features displayed differential intensities by sex ($p<0.05$), of which the majority (n=131) were higher in males. Five hundred and fifteen (68.7%) differed by cohort, with the majority of these (n=362) displaying higher intensities in NSHDS. Subsequent principal components analysis (PCA) revealed strong discrimination by cohort along the first two components, which explained more than 35% of the variation. This also resulted in some clustering by phase and processing batch. There was no obvious clustering by sex or age group (appendix figure 6.3.1iii).

6.3.2 Association between disease risk and metabolite profile in the discovery dataset

Classical methods

Using a conditional logistic regression model (model 1) a total of 13 metabolite features were associated with disease risk with a p-value <0.05. Ten were upregulated in cases and three were downregulated. None of these associations would survive any realistic multiple testing correction. Potential confounders were identified and two further adjusted models were run (details in appendix section 6.3.2i). Using the same arbitrary significance level (0.05) nine associations between metabolite features and NHL risk in model 2, and nine in model 3 were found. The number of up and downregulated features in the three models and the crossover between them is shown in table 6.3.2. There was relatively high overlap between the identified features; 3 were upregulated in all three models and 1 was down regulated in all three. In total, 19 unique metabolite features were associated with NHL in at least one model.

For 9 of these 19 strongest metabolite features, an evident dose-response relationship was seen in at least one of the models (appendix table 6.3.2ii). While the pairwise correlation coefficients (ρ) tended to be significant across the 19x19 pairs, their intensity was relatively low (appendix table 6.3.2.1iii): only four pairs showed a ρ >0.5 and none had a ρ >0.7. The comparison of the *m/z* and *rt* of these 19 candidates indicated that they represented distinct features, and their CV% ranged from 11.2 to 29.7% (mean; 20.2%, median 18.3%).

Table 6.3.2: The number metabolite features significantly (p<0.05) associated with NHL risk and the direction of their effect using three different modes.

The number of features that were common between the three models is shown in the columns

Direction of effect			Model 1 Conditional	Model 2 (conditional + 10 covars[a])	Model 3 (conditional + 5 covars[b])
Upregulated	**Model 1**	**(n=10)**	10		
	Model 2	**(n=5)**	4	4	
	Model 3	**(n=6)**	4	4	6
Downregulated	**Model 1**	**(n=3)**	3		
	Model 2	**(n=4)**	2	4	
	Model 3	**(n=3)**	1	2	3

Associations were determined using a conditional logistic regression model, models 2 and 3 were additionally adjusted for the following covariates:

[a]BMI, educational level, alcohol consumption, vegetable consumption, dairy product consumption, protein consumption, total fat consumption, energy intake, fruit consumption and fish consumption

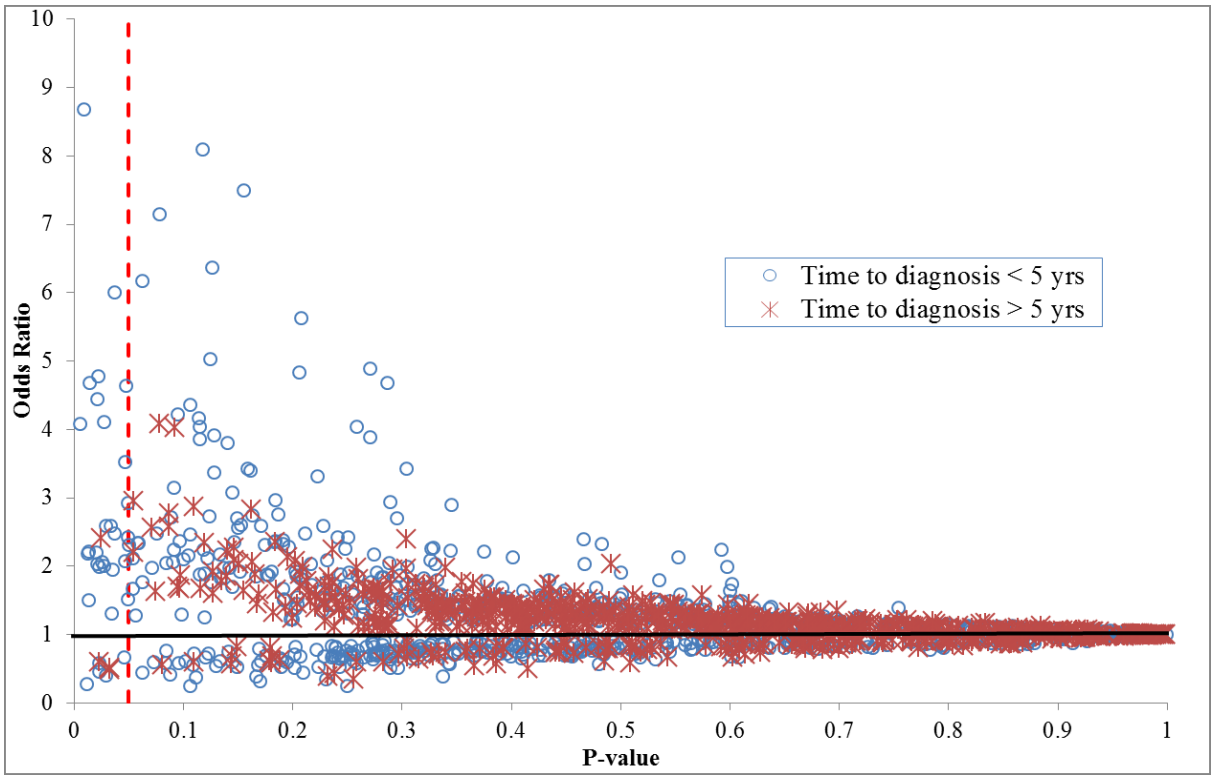
[b]BMI, educational level, alcohol consumption, total fat and energy consumption

Time to diagnosis

For model 1, 32 associations (p<0.05) between metabolite features and disease risk were identified in those who were diagnosed within five years of blood draw (ttd< 5 years *n*=70 cases), but only 6 in those diagnosed more than 5 years later (ttd> 5years *n*=113 cases). No metabolite features were common between the two subgroups. As expected, effect size estimates tended to be larger in the group diagnosed closer to blood draw (figure 6.3.2a). The same pattern was observed for model 2 (< 5 years *n features*= 25, ≥5 years *n features*=8) and model 3 (< 5 years *n features*=16, ≥5 years *n features*=10). In the ttd<5years subpopulation, across all models 13 unique features were identified. Of these, two were common with the 19 selected in the full population.

Figure 6.3.2a Significant (p<0.05) associations between metabolite features and NHL risk for model 1 stratified by time between blood draw and NHL diagnosis <5 years and >5 years)

The black line represents an OR of 1, and the dashed red line indicates a p-value of 0.05



The Odds ratio represents the increase in odds of incident NHL per unit increase in metabolite feature intensity

Replication of findings in the Validation cohort

Analyses for all 750 metabolite features were run in EPIC-Italy. Despite a non-stringent significance threshold of p<0.05, none of the 19 putative associations from the NSHDS full population were replicated in EPIC-Italy. Linear regression was used to consider the relationship between the p-value rankings of the features in the two populations. There was no significant correlation for any of the models (p>0.1) and the selected features from NSHDS did not rank among the top 50 in EPIC-Italy in

any case. In addition to this, under both models more than half of all the metabolite features showed discordant directions of effect between the two cohorts.

Only one of the 13 features identified in the NSHDS ttd<5 years subpopulation was found in EPIC-Italy ($p<0.05$), but the direction of effect was discordant and so cannot be considered replicated.

The *mz* and *r/t* of the top selected features from NSHDS and EPIC-Italy were compared. Again there was very little evidence that similar features were being identified within the two cohorts (appendix figure 6.3.2i)

Multivariate projection methods

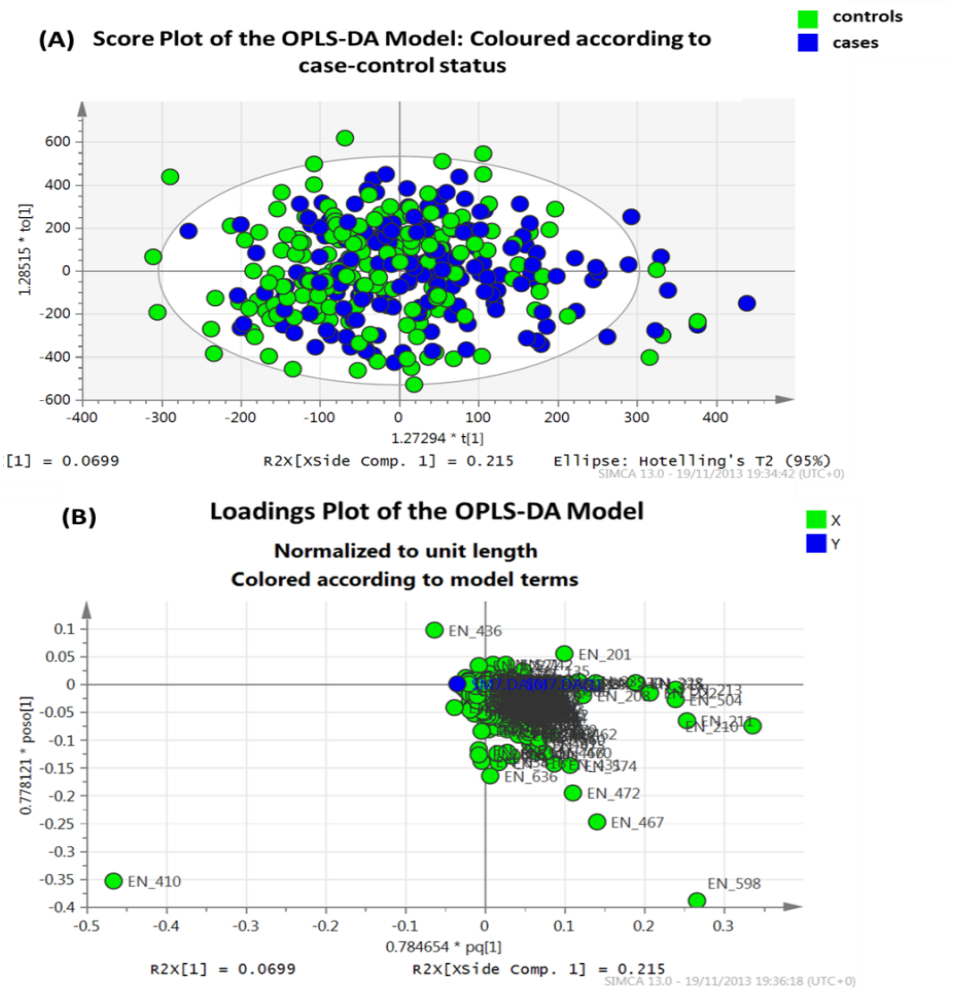
PCA was applied to the discovery set to identify major trends and any important outliers. This resulted in the exclusion of 17 outliers: eight cases and nine controls. These outliers comprised males and females, all batches, both phases and a number of different subtypes. The score plot for the first two components in the remaining 349 participants is shown in the appendix figure 6.3.2ii. There was no evidence of any discrimination between cases and controls.

To determine whether the model did have any predictive ability, and to improve interpretation, an OPLS-DA model was then built containing one predictive component and one orthogonal component (to explain the systematic variation not related to disease status). The model was found to explain 28.5% of the total variation in metabolite profiles: 7.0% corresponding to disease risk and 21.5% to the variation unrelated to risk. The remaining 71.5% can be interpreted as noise. Internal validation of this model revealed a negative Q^2 (Q^2Y=-0.104) indicating the model should not be used for prediction purposes. Permutation tests performed on the corresponding PLS model confirmed the lack of predictability as the Q^2_{max}, exceeded that obtained by the real model.

The score and loadings plots for the OPLS-DA model are shown in figure 6.3.2b. In accordance with the lack of predictability of the model these provide little evidence of separation between cases and controls from the predictive components. Nevertheless a number of features of potential interest were observed in the loadings plot. Their putative association with disease was further supported by the S-plot and the Variable of Importance in the Projection (VIP) plot (appendix: figure 6.3.2iii). Metabolites with VIP scores >1 generally represent those carrying the most information to discriminate classes [313]. A total of 54 features showed a VIP value >1 and confidence intervals that did not include 0 (model 4). Five (9.3%) of these 54 were common with the features identified by the classical analysis in the full population (this is defined as model 5).

Figure 6.3.2b: Results from the OPLS-Discriminant analysis in the discovery set

(A) Score Plot; each data point represents a participant. The x-axis describes the between group, or predictive, variation and the y axis the within group variation (B) Loadings plot: each data point represents a metabolite feature allowing the identification of those features most influential in the formation of the predictive component.



The feature names have been shortened to EN_XXX for the purposes of these figures for easier visualisation

Models were built for each subtype separately: MM, DLBCL, CLL and FL, to account for NHL heterogeneity. For each subtype only the matched controls were included to minimise nuisance variation from age and sex. Again these models were found to confer no predictive ability (appendix: table 6.3.2iv).

Time to diagnosis

An OLPS-DA model was built in the ttd<5 years subpopulation. The model performed slightly better in terms of predictive ability: 8.7% of the total variation in the metabolite profile was attributable to disease risk and 17.6% to systematic variation, after the exclusion of nine outliers. Again internal

cross-validation revealed a negative Q^2 (Q^2Y=-0.035), and therefore a lack of utility for prediction purposes.

There was higher overlap in results from the logistic regression analysis for this subgroup. Thirty features with a significant VIP value >1 were also selected in the classical analyses (model 6).

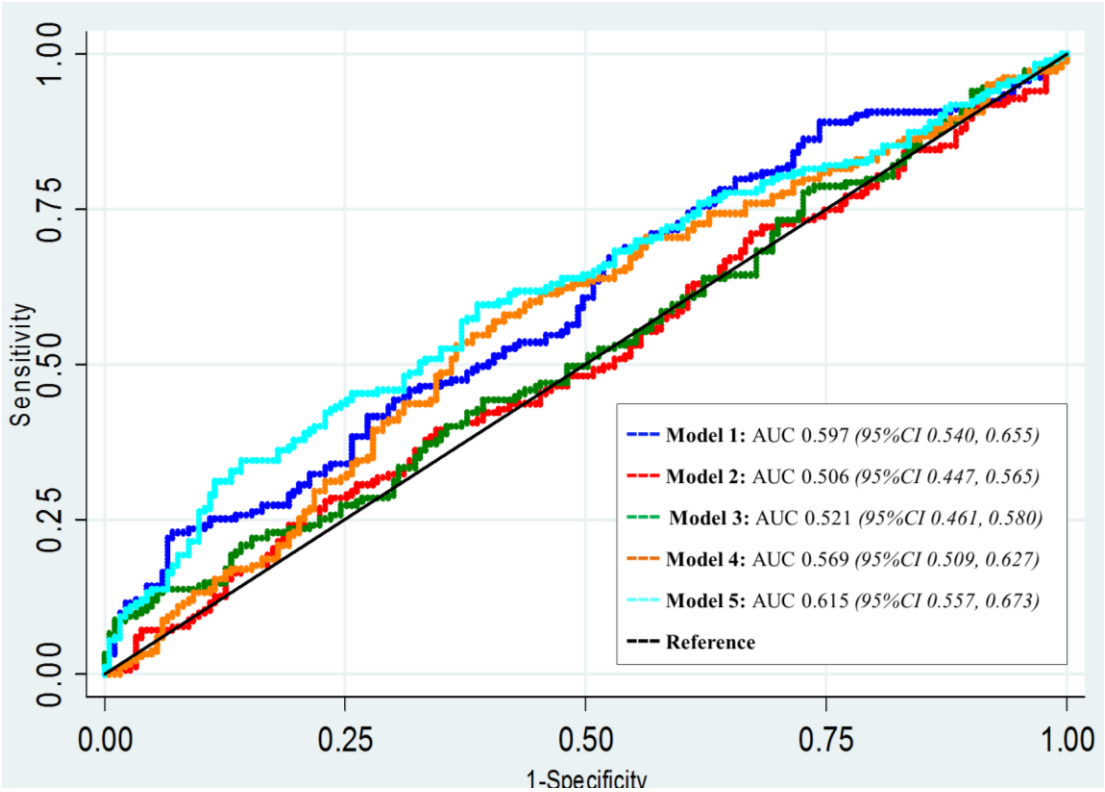
6.3.3 Metabolite profile as a predictive biomarker of NHL

The metabolite features included in each of six the models are shown in the appendix table 6.3.3.

To evaluate the ability of the metabolite feature sets to discriminate individuals who will develop NHL, ROC curves were determined using the scores for the first principal component as a continuous predictor. In all models, the first component explained more than 20% of the variation in the data, and specifically for model 4, 59% was explained (appendix table 6.3.3). The ROC curves and corresponding AUCs for models 1-5 in the full discovery population are shown in figure 6.3.3a.

Figure 6.3.3a: ROC curves describing the AUC and 95% CI for five different metabolite feature sets run in the total discovery cohort

ROC Curves are shown relative a random predictor (black line). ROC analyses were performed on continuous variables computed from principal component analyses of the specified metabolite features



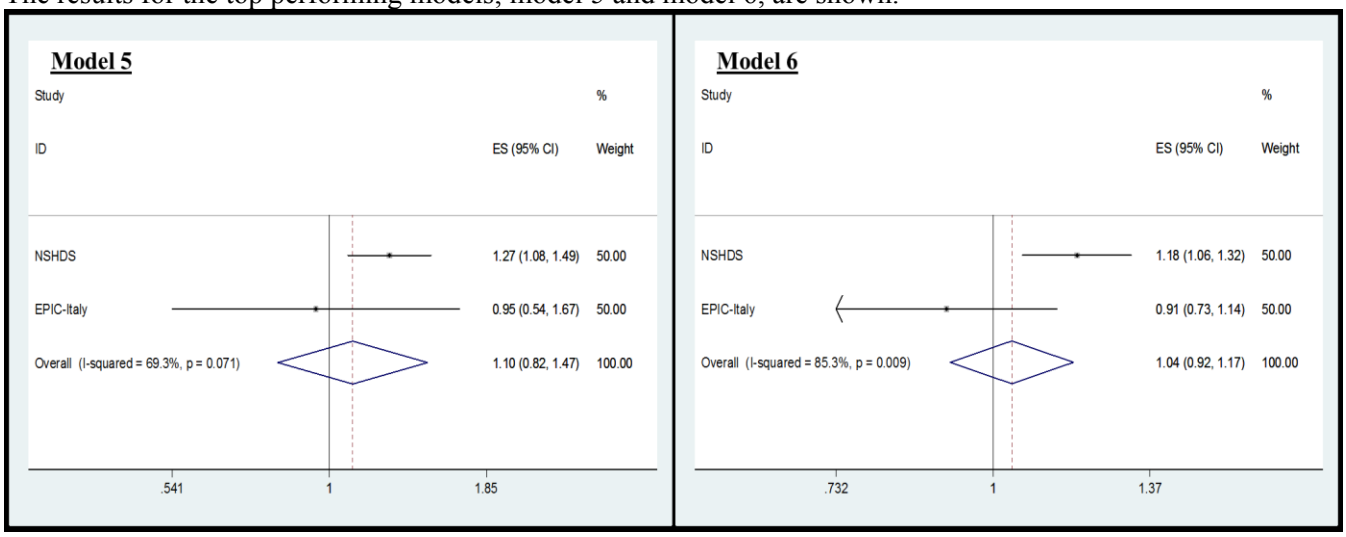
Model 6 is not included here as it pertains only to a subset (ttd<5 years) of the population

The AUCs were low for all 5 models, and the confidence intervals provided little evidence that they perform better than a random predictor at identifying individuals who will develop NHL. The highest AUC was for model 5 (AUC 0.615, 95% CI 0.557, 0.673), which incorporated the findings from both the classical and multivariate projection analyses. There was no evidence that this metabolite dataset could predict risk in an independent population, as there was no association between the computed PCA score and risk for these metabolite features in EPIC-Italy (OR: 0.95, 95% CI 0.54, 1.53). This was also true of the other models.

In model 6 (ttd<5 years subpopulation) the AUC was 0.679 (95%CI 0.603, 0.754), but results could not be replicated in EPIC-Italy. For both model 5 (as the 'best' performing model in the full population) and model 6 (representing the ttd<5 years subpopulation) there was significant heterogeneity in the effect estimates between the discovery and validation cohorts, and the pooled estimate suggested no significant effect (Figure 6.3.3b). These findings are perhaps not surprising, as even in the two best performing models, the AUCs of <0.7 and associated confidence intervals would not generally be considered to have strong predictive ability.

Figure 6.3.3b: Forest plot showing significant heterogeneity in the effect estimates between the discovery (NSHDS) and the validation (EPIC-Italy) cohorts for the first principal component score and disease risk

The results for the top performing models; model 5 and model 6, are shown.



ES – Odds Ratio

6.4 Discussion

Metabolomics is an interdisciplinary "omics" science that combines pattern recognition approaches and bioinformatics with epidemiology, analytical biochemistry and biology to provide a global quantitative assessment of endogenous metabolites within a biological system [310]. Metabolomic profiling is now being increasingly used in cancer biomarker development, owing to its down-stream nature and consequent close relationship to disease phenotype [307]. Accordingly, it is most commonly applied to the classification of diseased individuals, and there is increasing evidence that the diagnosis of a large number of cancer types can be assisted through the use of metabolomic profiles. The strongest findings to date relate to breast, prostate, brain [310] and colorectal cancer [314], with a large number of other cancer sites also investigated [307]. Such tests are most commonly used on tumour tissue, but a role for other biological samples is currently developing, probably most successfully for blood serum and ovarian cancer [315]. It is cheaper and easier to perform metabolic profiling on bodily fluids and consequently this opens up the potential for the use of metabolomics as a predictive screening tool [310]. As part of the EnviroGenoMarkers study, the feasibility of using prediagnostic blood samples to predict future risk of NHL, based on metabolite profiles characterised using mass spectrometry was investigated.

Summary of findings

Using logistic regression models, a total of 19 metabolite features were significantly associated with risk of NHL. This represented less than 5% of all the metabolite features explored, and therefore there is a strong possibility of spurious findings in this population, which is supported by the fact that none of the associations were robust to correction for multiple testing. Confounding was also an issue given the known influence of external lifestyle and dietary factors on metabolite profiles [316]. However, a number of the strongest associations were consistent across the various models and additionally showed evidence of a dose-response relationship. This hints at a true association for these features, and therefore a possible role in prediction. Encouragingly when stratification by time to diagnosis was performed, a larger number of significant associations of greater magnitude were observed in the subgroup of cases diagnosed closer to blood draw. This provides some evidence of a biological plausibility as it would be expected that the metabolite profile in blood samples extracted closer to diagnosis would more accurately reflect the disease phenotype.

The multivariate projection methods did not show strong results and internal validation suggested no predictive utility. More promising results were found when the analysis was restricted to those cases diagnosed within five years of blood draw, and there was also greater overlap in the features identified by the two analytical strategies in this subgroup. Overall, however, due to fact that the findings were not robust to correction for multiple testing, the poor performance using the multivariate methods and

the lack of replication in an independent cohort, no definitive conclusions about the utility of metabolomic profiles as predictors of future risk of NHL can be made.

The lack of strong and replicated positive findings are perhaps not unexpected. To date no studies have identified predictive cancer biomarkers from the metabolome through the use of mass spectrometry and a prospective cohort design [307]. Although a small study within EPIC utilising NMR did identify eight spectral regions which were associated with future colon cancer risk over an average follow up time of seven years [224]. The feasibility of conducting metabolomics studies using a prospective design has been additionally proven in prognostic studies, successful metabolomics based classifications for recurrence of both hepatocellular carcinoma [317] and breast cancer [318] have been developed. However such studies are not really comparable to this one, and operate on a much shorter time scale. Therefore, the question remains as to whether it is possible to predict the risk of developing a malignancy up to fifteen years later using an individual's metabolic profile.

Assessment of methodologies

In this study, 750 metabolite features representing potential biomarkers, or constituents of a biomarker profile were investigated. These features represent only a fraction of the possible metabolites that could have been detected. The precise numbers of human metabolites is unknown, although it is estimated to lie within the range of 10^4 to 10^5, and is somewhat dependent on the different analytical and detection methods used [307]. To date, the Human Metabolome Database contains about 15,000 entries [319]. Unfortunately, there is not a single analytical methodology that can, even closely, cover the whole range of the metabolome. The number of features measured in this study is within the range of what may be expected using our methods [320]. These are currently unannotated, and typically one or more of these features may represent the same metabolite, or its adduct, isotope or fragment. This was explored through the plots of *m/z* versus *rt* for the significant associations in the classical analyses, and there were some groupings identified. However, further inferences cannot be made without annotation.

The correct annotation of the identified metabolomic features represents one of the biggest challenges in mass spectrometry-based metabolomics. Although retention time or ion mobility can be descriptive of the physical and chemical properties of a molecule, putative annotation, absolute identification and quantification can only be made in the presence of an authentic standard, and unambiguous annotation from database searches is not currently possible [310,321-323]. This is more of an issue for MS than for the other most commonly used methodology, Nuclear magnetic resonance spectroscopy (^{1}H-NMR). Furthermore, MS will only work on ionisable molecules and it is biased towards those that ionise most efficiently. In this study only the positive ionisation mode was used, potentially missing metabolites that only ionise in the negative mode. Ion suppression can also be a problem with MS as can adduct

formation and the more intensive sample preparation required, which can increase sample-to-sample variability [307,310,321,322].

UPLC-MS was chosen on the basis of its increased sensitivity to the order of 10 µmol/L which leads to wider class detection [307], its robustness, high resolution and ability to identify multiple metabolites at low concentrations with high specificity [160,310,321]. Due to its increased sensitivity, LC-MS analysis was feasible with the limited sample volume available for the current study, (50 µL), for which ^{1}H-NMR analysis would have been impossible. Liquid chromatography (LC) was chosen over gas chromatography (GC) mainly due to the fact that as there is no limit on molecular weight a larger number of metabolites can be detected [307]. Due to the hypothesis-free nature of this study an untargeted approach, analogous to a GWAS, was taken to pick up as many metabolites as possible, rather than a targeted approach to identify a few pre-defined metabolite subsets or pathways [307].

Sample variability

Sample handling can influence metabolite profiles. Metabolic pathways are highly susceptible to the exogenous environment so within the same study method standardisation is vital [310]. This may in part explain the dramatic differences seen between EPIC-Italy and NSHDS in this study. Although the metabolic profiling was carried out under controlled standardised conditions for all samples, the original sample extraction was conducted differently in the two cohorts. Previous work within the EGM cohort has found the anticoagulants used and the time between blood collection and freezing (benchtime) may both affect the metabolome [324]. This nuisance variation should not affect our final analyses however: benchtime was only observed to have an effect if greater than 8 hours and although exact benchtimes are not available for NSHDS, it was <1 hour for all samples. For all EPIC-Italy samples it was <6 hours. Within the cohorts the same anticoagulant was used for all samples. However the anticoagulant induced variation may have been a further factor in the inability to replicate between cohorts.

It could also be argued that the difference between the cohorts represents a batch effect. Although given the strong within-cohort batch concordance observed this seems unlikely. It seems more probable that exogenous differences including the sample collection as well as diet and environmental exposures between Italy and Sweden explain the metabolic distinction between the cohorts [308]. In accordance with the literature [325], multiple strong associations between baseline characteristics and metabolite features were observed. It has also been shown that there is a genetic component to the metabolome [326] which could further explain the geographical differences observed.

Statistical techniques and prediction

Regardless of the laboratory techniques used, complex analytical methods are required to deal with the multivariate, highly collinear noisy datasets produced [224,306]. Multiple testing is a particular problem: in this study none of the candidate associations would survive any realistic multiple testing correction. Commonly metabolomics employs pattern recognition or clustering techniques to determine differences between cases and controls [310]. Unsupervised methods, such as PCA, measure the innate variation in data sets, whereas supervised methodologies, including OPLS-DA are also able to take the outcome into account in the formation of the clusters, and to analyse systematic (orthogonal) variation that is not related to the outcome separately, thereby improving interpretability [310]. This is particularly important in metabolomics where variation may arise from environmental influences or from experimental variation [325] disturbing the multivariate modelling and causing imprecise predictions and lower model robustness [327]. As these methodologies cannot naturally account for the effect of confounders, they are therefore considered in combination with classical logistic regression. Within this project neither methodology provided reproducible predictive metabolite features.

Additionally results from OPLS-DA showed poor predictive power for the orthogonal component, explaining only a small proportion of such variation, and it was noted that the discriminant analyses were highly sensitive to the selection of outliers for exclusion. This may suggest more refined analyses are required to identify the true sources of variation in this dataset. One possibility is the use of a lower CV% cut-off for feature inclusion. A threshold of <15% has been suggested in the literature [230]. However under this more stringent threshold a large number of the features included in models 1-6 would have been excluded from the discovery set. Although again, the ttd<5 years group performed better in this respect.

The ability to stratify by time to diagnosis is one of the major strengths of this study, as was the existence of a validation cohort. A recent review reported that only 17% of published MS-based studies even attempted to replicate their findings in an independent cohort, so it is difficult to estimate the number of spurious associations that may be reported in the literature [307]. The lack of internal validation was a cause for concern. To ensure that the data were meaningful, an OPLS-DA was built for sex. This showed good separation by sex and decent predictive ability for epidemiological data (R^2 0.283, Q^2 0.257) as expected, suggesting the low Q^2 for the disease model was not due to data quality. An obvious limitation of the study was the lack of annotation for the features, which means that biological plausibility of the putative markers cannot currently be explored. To make the most of the data, existing literature can be used to suggest a number of possible metabolites that the unannotated candidates may represent if they do indeed lie on the carcinogenesis pathway.

Potential pathways

Two NHL metabolomics case-control studies have been conducted on human subjects to date. One reported an inverse association with hypoxanthine (HX) [328] measured in urine, the other found increased pyruvate and glutamate and decreased isoleucine concentrations in the serum of CLL patients compared to controls [329]. These findings may help inform on the metabolite features identified in this study. However, it should be noted in addition to their case control design and the use of urine by one, the studied populations were not entirely comparable to the population investigated here. In contrast to this population both contained more males than females (>60%).The populations were older on average; the mean age was 57 years in Yoo *et al.*'s [328] study and 71 years in the population studied by MacIntyre *et al*. [329]. Both papers were also much more homogeneous in terms of subtypes; MacIntyre [329] included only CLL cases, and Yoo *et al.*'s population [328] was predominantly (73%) DLBCL. In addition both papers had information on stage which was lacking from this study, although they reported no information on height, weight, BMI, smoking status, or any other variables which may be affecting the metabolite profile and which were available for the EGM population.

Studies of lymphoma in mice additionally suggest a role for altered levels of xanthine, adenosine and xanthosine monophosphate (XMP), together with higher levels of inosine and inosine monophosphate (IMP). Work in cell lines suggests decreased tryptophan and proline levels may also be expected [330]. To date the strongest findings in metabolomics tend to be related to those malignancies which are intricately involved in metabolism, such as pancreatic cancer. However cell line studies have shown that neoplastic lymphoma cells exhibit distinct metabolic profiles compared to normal lymphoctyes [330], and that a ^{1}H-NMR-based metabolomic approach can better identify patients with poor prognosis on the basis of their metabolic profile [329]. Therefore the evidence suggests we may well be able to obtain biologically meaningful results regarding lymphoma from the findings.

Only annotation of the identified metabolites will allow the further exploration of these hypotheses, and the determination of whether statistical significance translates to biological significance in this population. In particular the use of enrichment set analysis could be employed, which may additionally allow for the better comparison of the two datasets. Annotation of these features is ongoing.

Subtype specific results

Interrogation of the PCA score plots did not reveal any clustering by subtype (appendix figure 6.3.1ii) and this was supported by the negative predictive value of the OPLS-DA subtype specific models. However it has been suggested that the complexity of the genetic and epigenetic interplay in the various subtypes is likely to be translated to metabolic regulation [331]. As mentioned the positive findings from both MacIntyre *et al.* [329] and from Yoo *et al.* [328] pertained largely or completely to single

subtypes. Therefore it is suggested that subtype should also be considered further in future, larger studies.

Potential biomarker utility

For an annotated set of features these findings may be able to impart important mechanistic information, and there may indeed be some potential for validation of the findings once the true metabolites are known. This would allow consideration of a number of the biomarker criteria discussed in methods 3.5: biological plausibility, coherence with existing knowledge and amount of evidence. In common with the biomarkers discussed in the previous chapters it is possible to determine a temporal relationship, and in this chapter a number of associations of relatively high magnitude with a dose-response effect were observed. However careful thought must be given to practicality vs. prediction when measuring multiple metabolites. Furthermore, given the low AUCs reported and the current lack of replication, with its subsequent risk of overfitting, at present there is no utility for metabolomics profiles in the assessment of future NHL risk based on these findings. In common with all MS-based metabolomics studies there is also potential for bias, particularly through the pre-processing steps, including the import and conversion of raw data files, the detection of signals (peak picking), the assignment of single ions to the same metabolite (deconvolution) and the integration and alignment of the chromatographic peaks. Currently no standards exist for these different steps; a recent review of mass-spectrometry based metabolomics in cancer research found more than 10 different software programs were used for pre-processing in just over 100 studies [307], and there is no single methodology which is currently able to measure all the metabolites in a sample [316,323]. Agreement on the optimal protocol to maximise reproducibility, reliability and sensitivity and to minimise analytical drift in order to make the existing methodologies comparable is urgently required [323].

The generalizability of this cohort to the wider population is also an important consideration for potential biomarker utility. Although, the gender ratio has been slightly shifted towards males, in all other respects is it essentially the same as the population discussed in chapter five, and therefore subject to the same generalisability issues. Further, this population is even more heavily dominated by Swedish participants.

Conclusions

With the advancement of high-throughput technologies, metabolomics is emerging as a powerful tool that has the ability to sub-classify cancer phenotypes and to separate cases from controls on a large scale [309,316,323]. It was hypothesised that this may also be possible in pre-diagnostic blood samples. The results displayed little predictive ability on the basis of the logistic regression, and none when multivariate methods were employed. However, given that carcinogenesis represents a continuous

spectrum rather than a discrete separation between healthy and diseased [7] this is perhaps not surprising; the cases developed cancer at different times after blood sample extraction, so were at different stages of the carcinogenesis continuum at enrolment. Therefore given the dynamic nature of the metabolome one would perhaps not expect them all to exhibit a profile so similar it could be used to distinguish them from controls. This hypothesis is further supported by the fact that stronger associations were seen in the subgroup of cases diagnosed closer to blood draw. An interesting facet of the findings was the strong discrimination between two geographically distinct cohorts, as well as by sex.

Here differences in metabolite intensities between cases and controls have been utilised to attempt to disentangle the mechanistic underpinnings of disease. These differences can either be considered as a reflection of changes in the environment or in the genome, epigenome, transcriptome, or proteome, or alternatively as the drivers behind pathogenesis. These questions can begin to be answered by integrating other information into the study of the metabolome and this is explored further in chapter 8 utilising the EGM exposure data and the '*meet-in-the-middle*' approach.

7. t(14;18) TRANSLOCATION: A PREDICTIVE BLOOD BIOMARKER FOR FOLLICULAR LYMPHOMA

BIOMARKER: Prevalence and frequency of the t(14:18) translocation

BIOMARKER TYPE: Biomarker of early biological effect

POPULATION: Incident Follicular Lymphoma cases and control from the EPIC cohort

This chapter is based in part on the publication 't(14;18) Translocation: A predictive blood biomarker for Follicular Lymphoma' by Roulland S, Kelly RS et al.[332] Tables, figures and text have been modified accordingly.

7.1 Introduction

Follicular Lymphoma (FL), a lymphoma of follicle centre B-cells, represents a particularly attractive model to study early biological effects in lymphomagenesis as it is characterised by an indolent asympotomatic course and up to 90% of cases share the same genetic hallmark, the t(14;18) (q32;q21) translocation [333]. The t(14;18) translocation juxtaposes the BCL2 (*B-Cell leukaemia/Lymphoma 2) proto-onco*gene on chromosome 18q21 near to the immunoglobulin heavy chain (IGH) locus on chromosome 14q32. This rearrangement subjects *BCL2* to the control of the IGH-E enhancer leading to overexpression of BCL2 protein, which inhibits apoptosis, thereby increasing cell survival and uncontrolled cell proliferation in the germinal centres [147]. This effectively 'immortalises' the lymphocyte, so although with this translocation alone they are not neoplastic, any t(14;18)$^+$ cells[1] that develop subsequent oncogenic mutations are unlikely to be eliminated through routine apoptotic mechanisms [334].

A large proportion (50-70%) of healthy individuals also harbour low levels of circulating t(14;18)$^+$ cells [335,336], yet will never develop FL and the relationship between the translocation and progression to disease remains unclear. In rare cases, apparently tumour-free "healthy" individuals carry unusually high frequencies (10-100 times greater than the average population) [337]. It has previously been demonstrated that a high frequency of cells carrying the translocation constitute an expanding clonal population of atypical B-cells, which are issued from the germinal centre, and share genotypic and phenotypic features exclusively seen in FL [338,339]. Nevertheless, a key gap in understanding is whether such t(14;18)$^+$ cells with "FL-like" features in healthy individuals constitute clonally-related FL precursors, and if high t(14;18) frequencies in blood represent a suitable predictive biomarker of FL development.

[1] From here on t(14;18)$^+$ will be used to denote prevalence/positivity of the translocation (more than 1 cell bearing the translocation per 1,000,000 circulating lymphocytes (≥1x10-6))

7.2 Methods

7.2.1 Study population

Discovery cohort

The study population was drawn from the EPIC cohort (methods 3.1.1 and table 3.1.2c). Eligible cases were those participants who were diagnosed with FL during follow-up, according to the ICOD-3 (9690/3, 9691/3, 9695/3, 9698/3) and who had an archived cryopreserved buffy coat sample collected at enrolment. Each case had two matched controls (methods 3.1.2.7). Cases and controls with insufficient DNA quality or equivocal ICD03 codes were excluded (the number of FL cases are smaller than reported in table 3.1.1.7 as no Scandinavian cases were eligible for inclusion in the discovery cohort).

Validation cohort

The validation cohort was drawn from the NSHDS (methods 3.1.2) according to the eligibility criteria used for the EPIC cohort.

The baseline characteristics of the discovery and validation cohorts are shown in table 7.3.1a.

7.2.2 Assessment of t(14;18) prevalence and frequency

Details in the methods section 3.2.3.

7.2.3 Statistical analysis

Frequency analyses were conducted on both the total population and within the $t(14;18)^+$ subgroup to obtain effect estimates both for a general population and for a population known to be positive for the translocation. Differences in t(14;18) prevalence and frequency were analysed using the chi-squared and Mann–Whitney U tests, respectively. Correlations between frequency and time-to-diagnosis were performed using Spearman's rank tests. Unconditional logistic regression was applied to determine the association between t(14;18) frequency and FL, adjusting for the matching factors: age, sex, time-since-blood-draw and country of residence. An unconditional model was used to maximize power, as after exclusion of poor quality DNA samples, there was no longer a 1:2 ratio for all cases and controls, and therefore utilizing matched analysis would have resulted in further exclusions. Frequency was then dichotomized based on six pre-defined thresholds ($5x10^{-6}$, $1x10^{-5}$, $5x10^{-5}$, $1x10^{-4}$, $5x10^{-4}$, $1x10^{-3}$), and the risk of FL associated with a frequency above these thresholds calculated. These thresholds were chosen to be equally spaced on the basis of the frequency distribution in t(14;18)+ participants. ROC curve analysis was used to test the overall ability of t(14;18) frequency to predict FL. An optimal predictive t(14;18) frequency threshold was determined based on a context-specific trade-off between the AUC calculated from the ROC curve analysis, sensitivity, positive predictive value (PPV) and negative predictive value (NPV).

This optimal (in the context of this biomarker) threshold was then applied to the validation cohort to confirm its ability to predict FL risk in an independent cohort. Meta-analysis of the two cohorts was used to assess heterogeneity and provide a summary risk estimate. Stratified analyses by time to diagnosis were performed on the pooled dataset. Cases were categorized into four groups based on the time between blood draw and diagnosis: <5 years, 5-10 years, 10-15 years and >15 years. For each strata unconditional logistic regression was run adjusting for the matching factors using all controls as baseline. Heterogeneity between the strata was assessed using meta-analyses.

7.3 Results

The baseline characteristics of cases and controls in the discovery and validation cohort are shown in table 7.3.1a. For the discovery cohort, a total of 100 incident cases of FL with good-quality archived prediagnostic blood collected at enrolment were identified. Time to FL diagnosis ranged from 2 months to 13 years (mean: 6.4 years). The control population comprised 218 participants who did not develop lymphoma during follow-up (from the date of their blood draw to the present time). The mean age at diagnosis was 61.6 years in the combined cohort, which is in line with the reported median age of 60 years at diagnosis for this malignancy [340]. As with the previous chapters, females were over-represented, particularly in the discovery cohort, however it has previously been reported that FL may not exhibit the same male-bias as other lymphoma subtypes [15].

7.3.1 t(14;18) Prevalence and Frequency in cases compared to controls

Overall, the prevalence of $t(14;18)^+$ was significantly higher in incident FL cases compared to the controls (56.0% *vs.* 28.9%, $p<10^{-3}$) in the discovery cohort. Similarly, the mean and median frequencies were significantly higher, both when considering the total population and the $t(14;18)^+$ participants only, and this was evident in both sexes (Table 7.3.1b and Figure 7.3.1). Although t(14;18) frequencies varied over a 4-log range (from undetectable to one every 50 cells), the highest frequency quartile in prediagnostic FL reached levels unseen in the EPIC control group and never observed previously in large screenings of healthy subjects in other studies [337].

Table 7.3.1a. Demographic Characteristics and t(14;18) Analysis According to Study Cohort

	EPIC discovery Cohort			NSHDS Validation Cohort			Total Cohort		
Demographic and clinical characteristics	**Subjects who did not develop FL (controls)**	**Subjects who developed FL (incident FL)**	***P* value^**	**Subjects who did not develop FL (controls)**	**Subjects who developed FL (incident FL)**	***P value^***	**Subjects who did not develop FL (controls)**	**Subjects who developed FL (incident FL)**	***P* value^**
N	218	100		128	65		346	165	
Age at blood draw in yrs, mean (range)	54.8 (27.3 - 77.7)	54.1 (27.2 - 77.2)	*NS*	55.8 (30.0 - 73.70)	52.9 (30.3 - 72.0)	*0.030*	55.2 (27.3 - 77.6)	53.6 (27.2-772)	*NS*
Age at FL diagnosis in yrs, mean (range)	n.a	60.3 (37.7 - 79.2)		n.a	63.1 (38.4 - 81.7)		n.a	61.6 (37.7 -81.7)	
Sex - no (%) Male	66 (30.3)	32 (32.0)	*NS*	65 (50.8)	32 (49.2)	*NS*	131 (37.9)	64 (38.8)	*NS*
Female	152 (69.7)	68 (68.0)		63 (49.2)	33 (50.8)		215 (62.1)	101 (61.2)	
Calendar years of enrolment	1993-2002	1993-2002		1990-2010	1990-2010		1990-2010	1993-2010	
Calendar years of FL diagnosis	n.a	1995 - 2007		n.a	1994-2013		n.a	1994-2013	
Time from blood collection to diagnosis in months, mean (range)	n.a	76.9 (2.7 - 161.3)		n.a	124.0 (5.9 - 244.5)		n.a	91 (2.7-244.5)	

NS - non-significant at the 5% confidence level

^p-value for difference between controls and incident cases; t-test for age at blood draw, chi-squared test for sex

Table 7.3.1b t(14;18) Characteristics according to Study cohort and gender

Population	t(14;18) characteristics	EPIC discovery Cohort			NSHDS Validation Cohort			Total Cohort		
		Subjects who did not develop FL (controls)	Subjects who developed FL (incident FL)	*P value^*	Subjects who did not develop FL (controls)	Subjects who developed FL (incident FL)	*P value^*	Subjects who did not develop FL (controls)	Subjects who developed FL (incident FL)	*P value^*
Total	Prevalence (%)	63/218 (28.9)	56/100 (56.0)	*<0.0001*	52/128 (40.6)	46/65 (70.8)	*<0.0001*	115/346 (33.2)	102/165 (61.8)	*<0.0001*
	Range (per 10^{-5})	0.1 - 45.6	0.1 - 1860		0.1-6.7	0.1 - 2129.5		0.1 - 45.6	0.1 - 2129.5	
	Mean frequency (per 10^{-5})	0.7	91.9	*<0.0001*	0.4	98.3	*0.003*	0.6	94.1	*<0.0001*
	Mean frequency in t(14;18)+ samples (per 10^{-5})	2.3	162.1	*<0.01*	0.9	138.9	*0.026*	1.7	151.5	*0.0001*
	Median frequency (x10-5)	0.1	0.2	*<0.0001*	0.1	0.3	*<0.0001*	0.1	0.2	*<0.0001*
	Median frequency in $t(14;18)^+$ samples (per 10^{-5})	0.2	3.8	*<0.0001*	0.2	1.6	*<0.0001*	0.2	1.8	*<0.0001*
	Prevalence of samples above 10^{-4} (%)	4/218 (1.8)	21/100 (21)	*<0.0001*	0/128 (0)	12/65 (18.5)	*<0.0001*	4/346 (1.2)	33/165 (20.4)	*<0.0001*
Males	Prevalence (%)	15/66 (22.7)	21/32 (65.6)	*<0.0001*	30/65 (46.2)	22/32 (68.8)	*0.036*	45/131 (34.4)	43/64 (67.2)	*<0.0001*
	Range - 10^{-5}	0.1 - 13.4	0.1 -1860		0.1 - 4.8	0.1 - 1069.4		0.1 - 13.4	0.1 - 1860	
	Mean frequency (per 10^{-5})	0.4	134	*0.005*	0.5	62.4	*0.017*	0.4	98.2	*0.0003*
	Mean frequency in t(14;18)+ samples (per 10^{-5})	1.4	204.2	*0.098^{NS}*	0.9	90.8	*0.05*	1.1	146.2	*0.0094*
	Median frequency (per 10^{-5})	0.1	0.5	*<0.0001*	0.1	0.3	*0.002*	0.1	0.3	*<0.0001*
	Median frequency in $t(14;18)^+$ samples (per 10^{-5})	0.2	5.8	*0.002*	0.2	1.3	*0.01*	0.2	2.7	*<0.0001*
	Prevalence of samples above 10^{-4} (%)	1/66 (1.5)	10/32 (31.3)	*<0.0001*	0/63 (0.0)	7/32 (21.9)	*0.001*	1/131 (0.8)	17/64 (26.6)	*<0.0001*

Table 7.3.1b *continued*

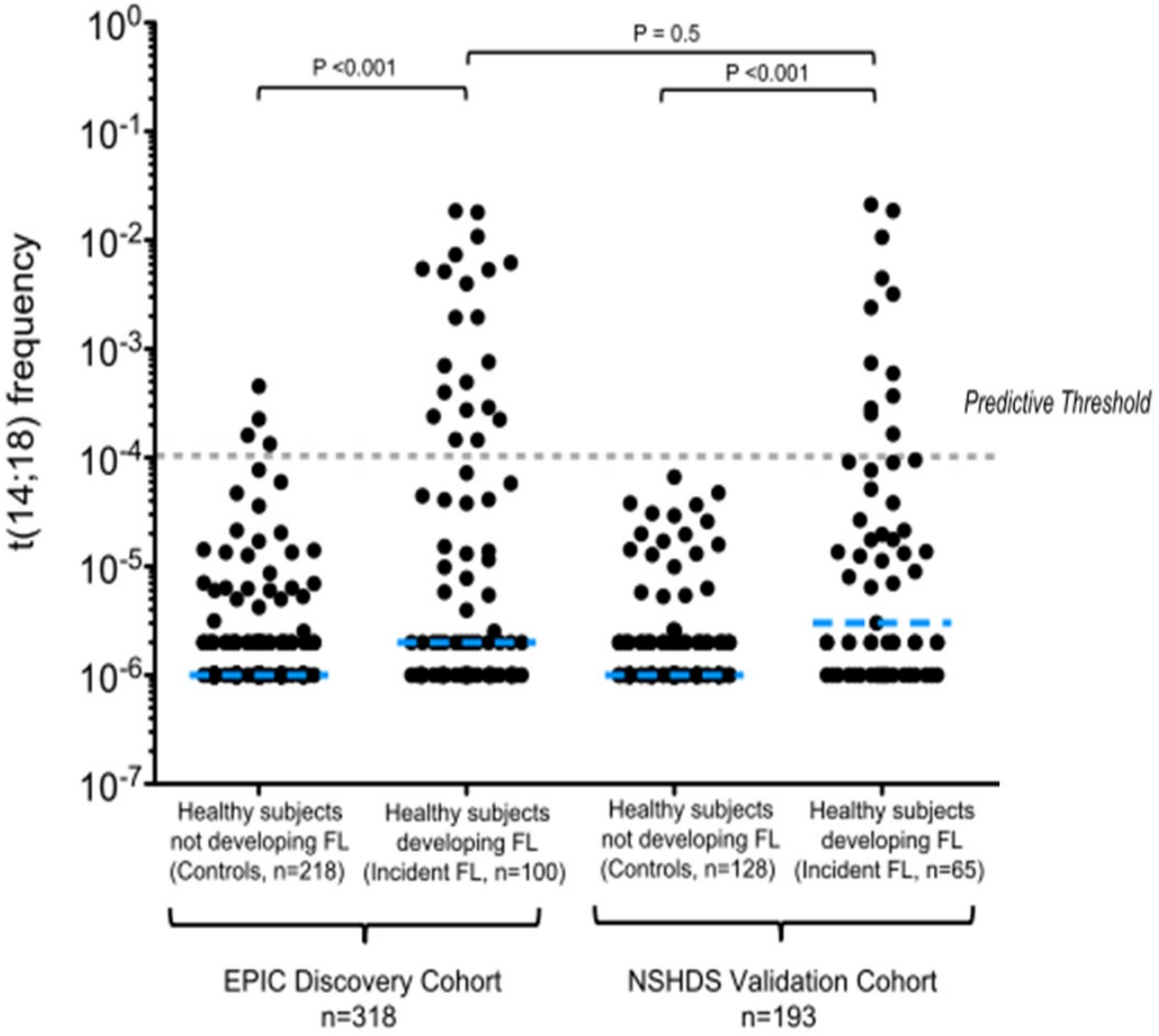
Population	t(14;18) characteristics	EPIC discovery Cohort			NSHDS Validation Cohort			Total Cohort		
		Subjects who did not develop FL (controls)	Subjects who developed FL (incident FL)	*P value^*	Subjects who did not develop FL (controls)	Subjects who developed FL (incident FL)	*P value^*	Subjects who did not develop FL (controls)	Subjects who developed FL (incident FL)	*P value^*
Females	Prevalence (%)	48/152 (31.6)	35/68 (51.5)	*0.005*	22/63 (34.9)	24/33 (72.7)	*<0.0001*	70/215 (32.6)	59/101 (58.4)	*<0.0001*
	Range (per 10^{-5})	0.1 - 45.6	0.1 - 1810		0.1 - 6.7	0.1 - 2129.5		0.1 - 45.6	0.1 - 2129.5	
	Mean frequency (per 10^{-5})	0.9	71.2	*0.001*	0.4	133	*0.032*	0.8	92	*0.0002*
	Mean frequency in t(14;18)$^{+}$ samples (per 10^{-5})	2.6	136	*0.0112*	0.1	183	*0.137^{NS}*	2.1	155.5	*0.005*
	Median frequency (x10-5)	0.1	0.2	*0.0005*	1	0.6	*<0.0001*	0.1	0.2	*<0.0001*
	Median frequency in t(14;18)+ samples (per 10^{-5})	0.3	1.3	*0.0098*	2	2	*0.002*	0.2	1.4	*0.0001*
	Prevalence of samples above 10^{-4} (%)	3/152 (2.0)	11/68 (16.2)	*<0.0001*	0/63 (0.0)	5/33 (15.2)	*<0.0001*	3/215 (3.4)	16/101 (15.8)	*<0.0001*

NS - non-significant at the 5% confidence level

^p-value for difference between controls and incident cases; t-test for mean frequency, chi-squared test for prevalence, Mann-Whitney U test for median frequency

Figure 7.3.1. t(14;18) frequency distribution in incident cases and healthy controls from in the discovery and validation cohorts

(Adapted from *Roulland et al.* [332]) this figure shows t(14;18) frequencies in blood from healthy individuals who subsequently developed FL (incident FL cases) as compared to healthy individuals who did not (controls) in the EPIC discovery cohort and the NSHDS validation cohort. The dashed horizontal lines indicate the cutoff value of $1x10^{-4}$. The dashed blue line represents the median frequency. The negative threshold value was defined as half the detection limit of the Q-PCR assay (e.g. $1x10^{-6}$).



7.3.2 t(14;18) frequency as a predictive biomarker of NHL

To evaluate the potential of t(14;18) frequency to discriminate individuals who will develop FL, a ROC curve was determined from the logistic model. The AUC was computed to be 0.744 (95% CI 0.654, 0.834) when carried-out on $t(14;18)^+$ samples, indicating that t(14;18) frequency constitutes a good predictive marker, allowing the discrimination of individuals who will develop FL from controls

(Figure 7.3.2). Considering the AUC together with sensitivity, specificity, PPV and NPV it was found that the optimal cut-off frequency value to obtain good predictive ability for the risk of FL development (OR: 9.88, sensitivity: 93.7%, PPV: 84.0%), whilst maintaining a reasonable specificity (38.2%) was 10^{-4} (table 7.3.2). Unconditional logistic regression adjusting for the matching factors suggested an almost ten-fold increase in risk of NHL associated with being above this threshold in $t(14;18)^{+}$ individuals (OR; 9.88; 95% CI, 2.98-32.70; $P=1.77x10^{-4}$). If this threshold was applied to all individuals regardless of t(14;18) prevalence, the risk increased by more than 15-fold (OR; 15.52; 95% CI, 5.12-47.11; $P=1.29x10^{-6}$).

Figure 7.3.2: The receiver-operating-characteristic (ROC) curves for t(14;18) frequency in the discovery cohort ($t(14;18)^{+}$ samples)

The solid diagonal line shows the performance of a random predictor.

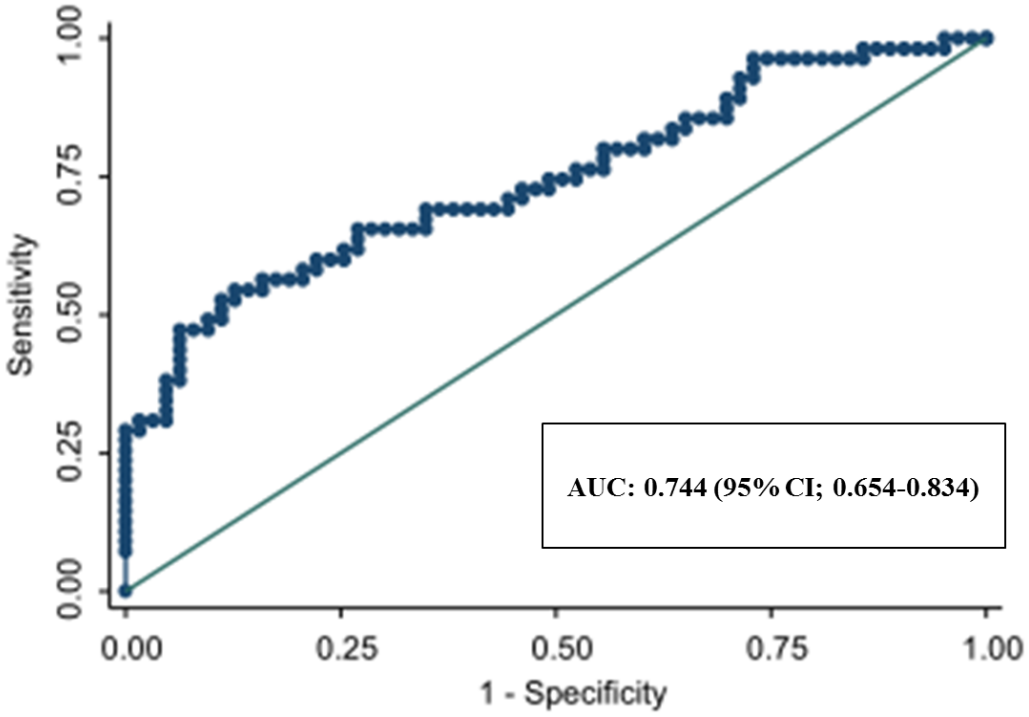


Table 7.3.2: Ability of varying thresholds of t(14;18) frequency to predict risk of FL among $t(14;18)^+$ individuals in the discovery cohort

t(14;18) frequency threshold	Sensitivity	Specificity	PPV	NPV	OR≠ (95% CI)
$5x10^{-6}$	67.3%	60.3%	59.7%	67.9%	2.97 (1.36, 6.45)
$1x10^{-5}$	60.0%	74.6%	67.3%	68.1%	4.35 (1.94, 9.77)
$5x10^{-5}$	43.6%	90.5%	80.0%	64.8%	7.43 (2.66, 20.73)
$1x10^{-4}$	38.2%	93.7%	84.0%	63.4%	9.88 (2.98, 32.70)
$5x10^{-4}$	23.6%	100.0%	100.0%	60.0%	ł
$1x10^{-3}$	20.0%	100.0%	100.0%	58.9%	ł
Continuous variable					1.01 (1.00, 1.01)

≠ OR adjusting for age at screening, sex, country and days between blood draw and frequency screening
ł No controls were above this threshold so an odds ratio could not be computed

7.3.3 Validating the utility of t(14;18) frequency as a predictive biomarker

These initial findings were confirmed in an independent validation cohort. In this cohort $t(14;18)^+$ prevalence and mean and median frequency were all significantly higher ($P<0.05$) in those participants who went on to develop FL (Table 7.3.1). The AUC in this population was 0.796 (0.708-0.883), and those participants with a frequency above $5x10^{-5}$ had a greater than forty fold increase in risk of FL (OR 41.82 95%CI 4.25, 411.75, p=0.001) (appendix table 7.3.3 and figure 7.3.3). No controls displayed a frequency $>1x10^{-4}$, so an odds ratio could not be computed. However, using this threshold 12 out of 65 participants (18.5%) who went on to develop FL were correctly identified.

Meta-analysis of the discovery and validation phases found no heterogeneity in the effect estimates between the two cohorts at the thresholds where an odds ratio could be computed in both cohorts; at $5x10^{-6}$ the p-value for heterogeneity was 0.509, at $1x10^{-5}$ p=0.631 and at $5x10^{-5}$ p=0.177 (figure 7.3.3i).

When both the discovery and the validation cohorts were combined, the AUC was 0.764 (0.700-0.827) (Fig. 7.3.3b). Table 7.3.3 describes the predictive ability of the varying thresholds in the pooled cohort and confirms that a threshold of $1x10^{-4}$ is associated with a high predictive ability, in terms of odds ratio, sensitivity and PPV, whilst maintaining a specificity greater than 30%.

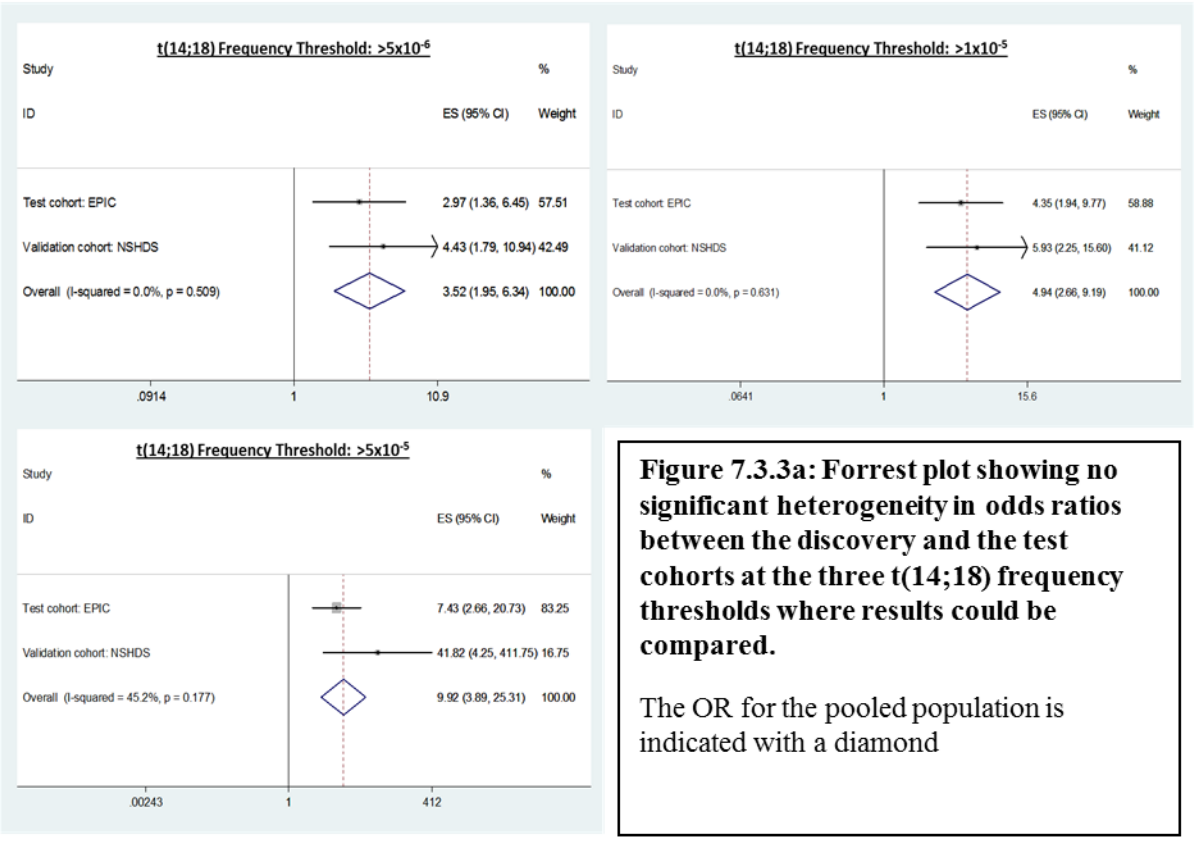


Figure 7.3.3a: Forrest plot showing no significant heterogeneity in odds ratios between the discovery and the test cohorts at the three t(14;18) frequency thresholds where results could be compared.

The OR for the pooled population is indicated with a diamond

Figure 7.3.3: The receiver-operating-characteristic (ROC) curves for t(14;18) frequency in the validation cohort (t(14;18)$^{+}$ samples)

The solid diagonal line shows the performance of a random predictor.

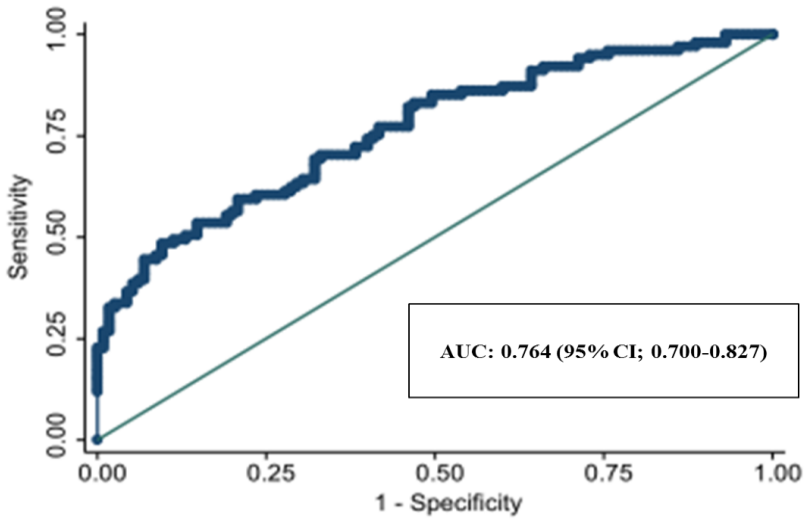


Table 7.3.3: Ability of varying thresholds of t(14;18) frequency to predict risk of FL among t(14;18)$^+$ individuals in the pooled cohort (discovery and validation cohorts combined)

t(14;18) frequency threshold	Sensitivity	Specificity	PPV	NPV	OR$^{\neq}$ (95% CI)
5x10^{-6}	68.3%	61.7%	61.1%	68.9%	3.48 (1.96, 6.02)
1x10^{-5}	60.4%	73.9%	67.0%	68.0%	4.84 (2.64, 8.90)
5x10^{-5}	40.6%	93.9%	85.4%	64.3%	11.53 (4.74, 28.05)
1x10^{-4}	32.7%	96.5%	89.2%	62.0%	14.95 (4.94, 45.25)
5x10^{-4}	20.8%	100.0%	100.0%	59.0%	Ł
1x10^{-3}	16.8%	100.0%	100.0%	57.8%	Ł
Continuous variable					1.01 (1.00, 1.01)

PPV – positive predictive value
NPV – Negative predictive value
AUC – area under the curve
≠ OR adjusting for age at screening, sex, country and days between blood draw and frequency screening
Ł No controls were above this threshold so an odds ratio could not be computed

Overall, within the pooled cohort a t(14;18) frequency >1x10^{-4} was associated with a 15-fold increase in risk among t(14;18)$^+$ individuals (OR: 14.95; 95% CI, 4.94-45.25; P=1.7x10^{-6}) and a 23 fold increase in risk when considering both t(14;18)$^+$ and t(14;18)$^-$ participants[2] together (OR: 23.17; 95% CI, 9.98-67.31; P=7.7x10^{-9}) .

7.3.4 Stratification by time to diagnosis

To determine whether increased t(14;18) frequency might be preferentially observed in pre-diagnostic samples collected close to diagnosis, analysis on the pooled cohort (n=511) was additionally stratified by time to diagnosis which ranged from 2 months to 20 years (mean 7.9 years)[3]. Overall, median time to diagnosis did not differ according to t(14;18)$^+$ prevalence (p=0.48). Among t(14;18)$^+$ cases, there was no significant correlation between t(14;18) frequency and time to diagnosis (r_s=-0.06, p=0.563).

When analysis was stratified by 5-year ttd intervals, the risk of NHL associated with a frequency >10^{-4} was significant in all intervals in the total population, but not in cases diagnosed >15 years after enrolment (Table 7.3.4). There was no heterogeneity of effect estimates across ttd strata.

[2] From here on t(14;18)$^-$ is defined as participants with no cells carrying the translocation
[3] Time to diagnosis (ttd) was different in the pooled cohort as compared to the discovery cohort, due to the inclusion of extra participants

Table 7.3.4; Risk of FL associated with a frequency >1x10^{-4} stratified by time to diagnosis in the pooled cohort (discovery and validation cohorts combined)

Results for the total population and for t(14;18)$^{+}$ participants are shown separately

	All participants (n=511) *p for heterogeneity= 0.837^{NS}*				t(14;18)$^{+}$ participants (n=217) *p for heterogeneity= 0.665^{NS}*			
Time to diagnosis	**No. cases**	**No. controls**	**OR$^{\neq}$ (95% CI)**	**p-value**	**No. cases**	**No. controls**	**OR$^{\neq}$ (95% CI)**	**p-value**
< 5 years	49	346	23.84 (7.06, 80.51)	3.26x10^{-07}	30	115	17.17 (4.48, 65.81)	3.35x10^{-05}
5-10 years	63	346	30.11 (9.58, 94.68)	5.70x10^{-09}	36	115	22.87 (6.70, 78.08)	5.89x10^{-07}
10 - 15 years	38	346	15.34 (3.68, 63.91)	1.77x10^{-04}	25	115	8.40 (1.92, 36.73)	4.68x10^{-03}
>15 years	15	346	16.07 (1.24, 208.67)	0.034	11	115	5.85 (0.42, 80.94)	0.187^{NS}

≠ OR adjusting for age at screening, sex, country and days between blood draw and frequency screening
p for heterogeneity assess whether there was a significant difference between time strata
NS – non significant at the 95% level

7.3.5 t(14;18)$^{+}$ clones as disease precursors

To confirm that t(14;18)$^{+}$ clones present in the prediagnostic samples constituted the precursors which progressed to overt FL, biopsy samples for all included cases were requested from the corresponding EPIC centres. Due to centre-specific ethical restrictions and the logistical difficulties of obtaining tumours samples stored in various non-centralised locations, only 3 good-quality archival formalin fixed paraffin embedded FL biopsies from EPIC subjects were recovered. Their t(14;18) clonotypic breakpoints were determined, and compared to the ones found in the corresponding prediagnostic samples. All FL biopsies were found to be t(14;18)$^{+}$, including one mcr variant, and corresponded to the t(14;18)$^{+}$ prediagnostic samples. In all cases, the same t(14;18) breakpoint was found before and after FL outcome, demonstrating that t(14;18)$^{+}$ clones in blood from these healthy individuals constituted *bona-fide* precursors, subsequently progressing to FL. These findings were independent of frequency, as one of the precursor clones was detected in a case with a very low t(14;18) frequency (1/500.000 cells), and of time to diagnosis, which ranged from 4.9-10.6 years in the three participants with biopsy samples.

7.4 Discussion

The strong association between FL and high frequencies of t(14;18) is well documented, and the presence at low frequency of $t(14;18)^+$ clones in the blood from healthy individuals is equally well known [333,336,337,341]. What is not understood is the relationship between the two and if, and how, those $t(14;18)^+$ individuals who go on to develop FL can be distinguished from those who don't. The timeline for disease progression still remains poorly characterized [341-343]. However a recent report of synchronous FL development, occurring from the same $t(14;18)^+$ precursor 9 years after bone marrow transplantation in both the donor and the recipient [343], demonstrated that commitment to malignant transformation can occur nearly a decade before disease onset.

The literature suggests that this translocation may represent a useful biomarker in the study of NHL. The prospective design of the EPIC cohort provides a valuable opportunity to investigate the association of this chromosomal aberration with FL onset. In this study it was demonstrated that $t(14;18)^+$ prevalence was significantly associated with future risk of FL in a cohort of people who were healthy at blood sampling, and that there is a dose-dependent relationship between translocation frequency and risk. These findings were used to identify a frequency threshold that had strong predictive ability in terms of specificity and positive predictive value, but that maintained reasonable sensitivity. The utility of this threshold was then confirmed in an independent cohort. Above this threshold, t(14;18) frequencies reaching one in every 10,000 blood cells, a greater than 23-fold greater risk of FL was observed in this population.

7.4.1 Assessment of t(14;18) threshold as a predictive biomarker of FL risk

As discussed, the presence of the translocation is insufficient as a biomarker. Similarly a continuous frequency distribution is not appropriate for clinical usage and a cut-off point somewhere along this distribution must be chosen. This is by its nature somewhat abitrary, and indeed the investigated cut-off points chosen here, were selected to be equally spaced throughout the distribution in this specific population. This raises questions regarding the validity of using these particular thresholds. Of the thresholds, one was then chosen as 'optimal' based on its ability to maximise the indices of interest in this population. These are discussed in more detail below. However, despite the arbitrary nature of the chosen threshold, it was possible to demonstrate that it performed equally well in an independent population.

ROC curves

In both the discovery and pooled cohorts AUCs in excess of 0.7 were observed with confidence intervals ranging from at least 0.65 at the lower bound to greater than 0.8 at the upper bound. As this constitutes the first predictive biomarker of FL there is no gold standard for comparison. However these AUCs would be ranked as fair to good under most commonly accepted definitions.

The ROC curves and corresponding AUCs were presented for $t(14;18)^+$ individuals, rather than for the total population to demonstrate what the specific frequency cut-off identified in this study provides in terms of predictive ability above and beyond classifying participants as positive or negative for the translocation. The odds ratios have also been presented for the $t(14;18)^+$ individuals only as including $t(14;18)^-$ individuals would inflate the effect estimates. Instead the presented effect estimate provides a conservative, but still strong, estimate of the association of this specific threshold with the outcome of interest.

Sensitivity and Specificity

Due to the heterogeneous nature of most cancers, it is unlikely a single biomarker will be able to detect all cases of a particular subtype with high sensitivity and specificity [344]. There does exist a measure called the Youden index, which can be used to determine the point on the ROC curve that maximises both sensitivity and specificity. The drawback of this index is that it affords equal weight to sensitivity and specificity, which in a real world setting is rarely appropriate. Rather in most instances a trade off must be made between sensitivity and specificity, that is optimal for the usage of the that particular biomarker. In this study for the defined threshold, lower sensitivity is accepted as a trade-off for increased specificity. This is optimal for the potential usage of this biomarker, that is, to predict progression to disease in a target population with the aim of administering preventative therapy to those at highest risk. Although this biomarker performs poorly at predicting all cases of FL that eventually developed within these cohorts, with sensitivities in the two populations below 40%, it is very good at identifying a certain subgroup of those who will go on to develop disease. Furthermore its false-positive rate is extremely low, hence its high positive predictive value.

Within a population-wide setting even a low false-positive rate would translate to vast numbers of wrongly diagnosed people. Given the monetary and emotional costs of chemotherapy, and the fact that there are other ways of diagnosing FL in its latent state, it is better to avoid unnecessary or overtreatment in individuals for whom there is a high degree of certainty they will not develop the disease. Hence a high specificity was favoured.

Bradford Hill Causation Criteria

In terms of Bradford Hill's causation criteria the biomarker performed well. A strong association that displayed a biological gradient was observed. This finding is coherent and consistent with the current literature and with independent analyses of different populations conducted in Bertrand Nadal's lab. Additionally the nature of the nested case-control design allowed the determination of temporality of

association. It is known that $t(14;18)^+$ prevalence is not specific, as these translocations have been observed both in other NHL subtypes and more importantly in healthy individuals. Within this study the chosen frequency threshold showed specificity for FL development of 94% in the discovery cohort and 97% in the pooled cohort. Additional experiments in healthy populations and in those who go on to develop other malignancies associated with this translocation, such as DLBCL, would be necessary to further define the specificity of this threshold for risk of FL. It should also be noted that the absence of specificity does not negate a causal relationship.

Biological plausibility of the association has been discussed in detail. This is further strengthened by the finding that in three participants for whom post-diagnosis biopsy samples were available, the $t(14;18)^+$ cells in the biopsy sample were clonally related to those in the prediagnostic blood sample in the same individual [332]. Funding is available to collect biopsy samples from more of the included cases in order to expand upon these findings and offer further support. Negotiations with the sites to obtain these biopsies are ongoing. The fact progression to disease was observed in some people below the threshold may argue against plausibility, but it is not unsurprising, since it is known there are also a proportion of FL cases, who display only low frequencies of the translocation or lack it altogether. This may suggest differing pathways of development and favour further subdivision of the subtypes.

Potential confounders such as age and country of residence were controlled for in this study – but the possibility that a further factor, such an environmental exposure may be causing both the clonal expansion of the translocation as well as lymphomagenesis through an unrelated mechanism cannot be ruled out. What also remains to be fully determined is Bradford Hill's eighth criteria 'experiment': whether the condition can be prevented or altered with intervention to the causal factor. Studies have shown decreasing frequencies of t(14;18) in some but not all patients following treatment with the drug Rituximab [345]. This needs to be further clarified but could potentially have a significant implication for usage of this biomarker.

Venice-Criteria

In terms of the Venice criteria, this biomarker would score well. Although a sample size of over 1000 (cases and controls combined assuming a 1:1 ratio) is recommended to achieve large scale-evidence, this generally refers to genetic studies which have issues of power and multiple testing not relevant in this study of a single biomarker. Furthermore, although this represents the first prospective study, there is already a large amount of data confirming a strong association between t(14;18) frequency and FL.

Almost uniquely, this biomarker has been replicated in an independent cohort. To date cancer biomarker research has been characterised by 'inflated expectations, followed by disappointment

when the results cannot be reproduced' [229]. Within this study meta-analysis of the two independent cohorts found no significant heterogeneity.

The third Venice Criteria is protection from bias. Bias was minimal in this study: the biomarker was measured in an objective manner using a recognised and proven screening assay in the same lab under the same conditions using prediagnostic blood samples. All participants came from the same cohort and were healthy at recruitment. There is some potential for selection bias as participants were excluded on the basis of unavailable or poor DNA quality DNA samples, but this is unlikely to be related to disease status. Cases and controls were adjusted for potential confounding factors such as age which may be related to t(14;18) frequency and it can be concluded that bias is unlikely to be affecting the presence of an association, although it may have some effect on its magnitude

7.4.2 Strengths and Limitations

The main strength of this study, resides in its nested-case control design within an established cohort which allows the establishment of a temporal relationship and eliminates recall and selection bias. This design is in accordance with the PRoBE (prospective-specimen-collection-blinded-evaluation) design, which has been proposed to optimize the biomarker validation studies [346]. The cases and controls are from the same at-risk population and fulfil the same eligibility criteria.

The ability to investigate t(14;18) frequency by time to diagnosis represents a further strength. Interestingly, no association between frequency and time to diagnosis was observed and it was determined that the threshold showed high predictive ability both close to and up to 15 years before diagnosis. This is particularly important for translation to a clinical setting [344]. The main limitation of this study, when considering the usage of a biomarker in a population-wide setting, is its relatively small sample size which led to wide confidence intervals and a consequent lack of precision in the estimates of effect size. Further the European Caucasian population may limit its wider generalizability and external validity. However, as discussed in the results section, the baseline characteristics were in keeping with what may be expected in terms of gender and age at diagnosis. Furthermore it is of note that the findings were replicated in an independent, geographically distinct cohort.

The lack of information on grade of disease represents a limitation that may be particularly important in the context of FL. A close relationship between grade 3 FLs and progression to the more aggressive DLBCL lymphoma has been observed [340], while it has been noted that the prevalence of the translocation may be uncommon among such high grade cases [17]. This would have important implications for the clinical relevance of this biomarker.

7.4.3 Clinical utility

Based on these performance indices, the translation of the biomarker into clinical practice can be considered. The ultimate goal of biomarker discovery is to create a predictive model with optimal sensitivity and specificity that can be utilized in clinical practice, most commonly in the form of a screening test. The criteria for a useful screening test are outlined in the methods section 3.5 and appendix box 3.5. This biomarker would perform well against a number of these criteria including the testing method; translocation frequency is detectable in blood using a simple, reproducible, low-cost assay.

However, any screening test for FL would fail to meet all these criteria, particularly those regarding the disease itself. The natural history is not well understood and although NHL as a collective group represent an important health problem, the relative risk of FL remains low and population wide-screening would not be cost-effective. Furthermore, despite the advent of effective therapies which have significantly improved clinical outcome, FL remains incurable and patients continue to die from the disease following resistance to treatments or transformation into a more aggressive DLBCL [347]. When considering the economics of the potential clinical translocation of this biomarker it must also be noted that, in comparison to many malignancies, the survival estimates for FL are good. For those with limited disease median survival is 19 years, and up to 80% of FL cases survive for ten years or more [340]. Nevertheless, the majority of cases have advanced disease at diagnosis, and as mentioned FL often transforms into the more lethal DLBCL [340], meaning that there is an argument for investing in early detection.

The findings suggest that this biomarker could be considered a pre-disease condition analogous to monoclonal B-cell lymphocytosis, which has been shown to be a precursor for CLL [348], or monoclonal gammopathy of undetermined significance which is a precursor for MM [349]. The identification of such a condition would enable the identification of high risk groups for monitoring. It has also been suggested that such *in situ* disease can be used as an intermediate outcome to increase sample size in future studies [171].

7.4.4 Biomarker as a tool for understanding biological processes

One of the major failings of this biomarker as a clinical tool is that, as observed in this cohort, a large proportion of individuals who will develop FL do not display elevated t(14;18) levels before FL onset, and thus cannot be detected. This 'failing' in fact suggests two possible diverse pathways of lymphomageneis in FL: one in which progression to FL systematically occurs from clonally-related t(14;18)$^{+}$ *bona fide* lymphoma precursors (as confirmed in the study of prediagnostic/tumour sample

pairs, details in [332]), and one in which there is no clonal expansion of this precursor, or it is absent altogether. The literature support the existence of different molecular subtypes [350], and the utility of this biomarker may lie in its ability to distinguish between these two forms of FL. In other words it can be argued that this biomarker does not predict risk of FL, but rather predicts risk of a certain subset of FL associated with a very high frequency of t(14;18) and which may be aetiologically, clinically, pathologically and genetically different from other FL subtypes. This adds further support for calls in the literature to reclassify FL based on translocation status as more clinical, epidemiological and genetic data become available [17].

Alternatively the findings may suggest a greater than previously suspected role for alternative BCL2 breakpoints in aetiology. Among the t(14;18)$^-$ participants who went on to develop FL, to exclude the possibility that they represented false-negatives, a refined nested PCR assay was run to screen for three additional breakpoint clusters in a subset of participants (details in [332]). These alternative breakpoints are each involved in <10% of FL cases and together with the initial screen cover ~70% of BCL2 breakpoints [350,351]. Among 30 screened t(14;18)$^-$ EPIC participants who subsequently developed FL, 4 were identified with alternative *BCL2/J_H* translocations (two *mcr/J_H* and two *3'MBR/J_H*) and frequency distributions similar to cases with *MBR/J_H* junctions.

Either way, the identification of committed FL precursors in prediagnostic blood samples suggests a mechanistic continuum between pre-clinical and clinical phases of FL progression in t(14;18) associated lymphomagenesis [352]. The identification of these cells is the first step in their isolation and geno-phenotypic characterisation. This will provide important insights into the factors driving FL progression [353,354], suggest future therapeutic avenues, and help to target the identification of future biomarkers which could be utilized together with this frequency threshold in a multivariate model.

This finding is important in the wider field of biomarker discovery. It has recently been stated that current design of using diagnostic samples for discovery and then attempting to validate the findings in prediagnostic samples, has been unsuccessful and that such markers have little value for early detection [355]. Here, the findings from numerous case-control studies of translocations and NHL were validated using a cohort design, suggesting that this method can be successful and potentially paving the way for future similar studies.

7.4.5 Conclusion

In conclusion there is a strong association between the defined t(14;18) frequency threshold (>1x10^{-4}) and risk of FL, which is apparent up to 15 years before diagnosis. Crucially these findings have been validated in an independent cohort, achieving consistent results. In this way this threshold fulfils most of the 'Rules of Evidence' for molecular markers proposed by Ransohoff [229]. Although tailored to

high through-put discovery based –omics studies, these rules can also be considered in the context of this study. The investigation of a single biomarker is free-from many of the problems of an –omics based approach, most notably the dangers of overfitting and multiple testing. However by validating the findings in an independent cohort, the possibility of such issues influencing the results have been eliminated anyway. Similarly, the possibility of residual bias and confounding is minimised meeting two more of Ransohoff' rules. Although generalisability is an important issue, sufficient evidence is presented to warrant the further investigation of this biomarker in larger and more diverse populations.

Ransohoff's final rules pertain to clinical usefulness in terms of benefit, harm, cost and effort [229]. Due to the relatively low prevalence of FL in the population and the lack of effective preventative or curative therapies for those in whom FL could be predicted there is currently no direct clinical application for such a biomarker.

The biomarker is useful in a number of other ways. First it provides information on mechanisms of lymphomagenesis. Second, given its known association with a number of risk factors, it can help to establish or refute their causal relationship with NHL. Smoking [148], UV radiation [356] and exposure to both pesticides and dioxins [357] have all been associated with translocation frequency or prevalence. If such associations can be established, then this biomarker may be of use in a screening capacity in high risk exposed populations, such as farmers working with certain pesticides. Third, it may represent a precursor condition that can be managed accordingly. Finally it may provide further insights to enable the division of FL into more appropriate subtypes. It could build upon the proposed subdivision between $t(14;18)^{+}$ and $t(14;18)^{-}$ [334], introducing a further group: $t(14;18)^{HF(high\ frequency)}$, which may more accurately allow the discrimination of committed precursors from otherwise inoffensive $t(14;18)^{+}$ clones. Given the lack of rationale for population wide screening for FL such insights into etiology and mechanisms are arguably more useful than prediction in this instance. As such this frequency threshold can be described as a useful biomarker of early biological effect.

PART 2: INTEGRATING BIOMARKERS

8. INTEGRATING BIOMARKERS: THE "*MEET-IN-THE-MIDDLE*"APPROACH

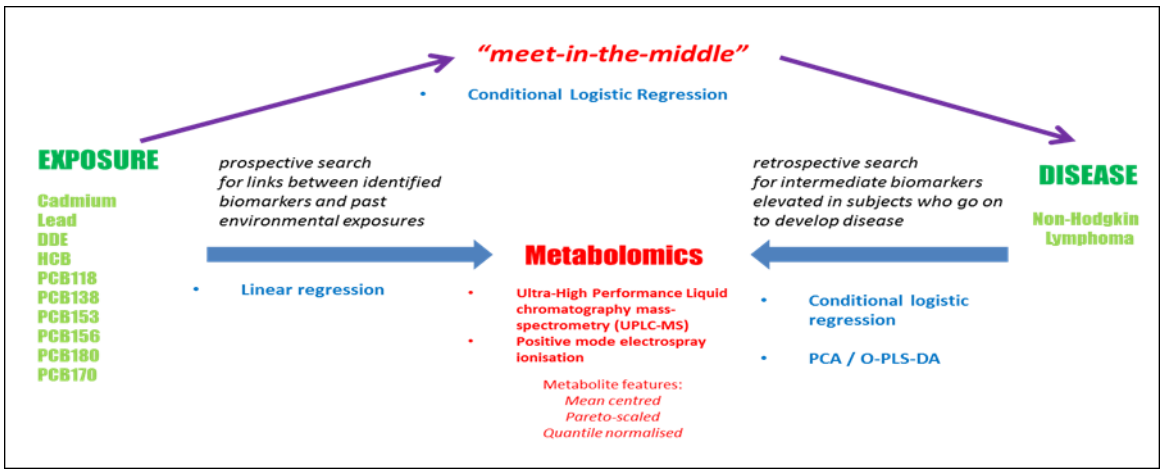
POPULATION: B-cell NHL cases and controls from the EnviroGenoMarkers Study

8.1 Introduction

One way of integrating biomarkers is to use the *MITM* approach. This concept relies on two important findings that have emerged from the development of the field of molecular epidemiology. First, the fact that exposure to certain pollutants can be directly correlated to biomarkers of biologically effective dose and early biological effect [7]. For example styrene and DNA adducts [358]. Second, that cancer risk has been associated with the levels of such biomarkers [359]. Therefore, the identification of elevated biomarkers levels both in those who are exposed to certain pollutants and in those who develop disease has the two-fold benefit of strengthening the causal link with the exposure, while also providing information on the causal pathway and the potential mechanisms through which this exposure may be acting [7].

The application of this theory to NHL is problematic as biomarkers related to exposure and those related to disease tend to have been studied separately (see chapter one; Introduction). More importantly, none of the biomarkers identified to-date are specific to either NHL (with the possible exception of the translocation threshold) or its putative risk factors, making the definition of a causal pathway difficult. The aim of the *MITM* approach is to search on the same human biosample (i)- prospectively for biomarkers of past environmental exposures, and (ii)- retrospectively for biomarkers associated with future disease risk.

Figure 8.1 Schematic describing the "*meet-in-the-middle*" approach for molecular epidemiology



In this way the *MITM* approach can identify intermediate biomarkers that lie along the causal pathway [7] i.e. a biomarker from any of the categories falling between exposure to a risk factor and disease as outlined in figure 1 (chapter 1). Here this principle is applied to the metabolite profiles from the EGM study in order to search for intermediate biomarkers associated with both environmental pollutants and with NHL (figure 8.1).

Omics are particularly well suited to this approach as their methods are hypothesis-free and they allow biological and mechanistic interpretation. This allows for the identification of novel candidates that do not necessarily have a biologically plausible a *priori* relationship [7]. There is a particularly strong rationale for the link between a metabolite profile and disease (see chapter 6) [310], and there is wealth of data in preclinical models that demonstrate the effect of exogenous exposures on metabolite levels [320]. Furthermore the concept has already been shown to provide very promising results in a metabolomics profiling study of dietary fibre intake and risk of colon cancer [224].

This chapter represents the first study to report on metabolomics profiles following exposure to PCBs. The results are of particular interest, given the number of unexpected inverse associations reported in chapter 5 for these pollutants with risk of NHL.

8.2 Methods

8.2.1 Study subjects, exposure assessment and metabolite profiling

This chapter is based on the same EGM discovery and validation cohorts described in chapter 6 (table 3.1.2c, appendix table 6.3.1). All participants had information on the body burden of six PCB congeners, DDE, DDT, HCB, BDE-47, Cd and Pb, as described in methods 3.2.1. Metabolomic profiling of cases and controls was performed as described in methods 3.2.2 and chapter 6.

8.2.2 Disease vs. metabolite profile

The identification of individual metabolites and metabolite profiles associated with NHL is described in chapter 6.

8.2.3 Exposure vs. metabolite profile

Linear regression was used to assess the relationship between log-transformed exposure levels and metabolite intensities in the discovery cohort. The model was adjusted for the EGM matching factors: age, sex, phase, as well as metabolite batch (matched pairs underwent profiling in the same batch) and fasting status (>95% cases had the same fasting status as their matched control).

8.2.4 Identification of potential *MITM* candidates

The parallel analyses conducted in 8.2.2 and 8.2.3 resulted in two lists of the metabolite features most strongly associated with disease risk and exposure concentration, respectively. These features were selected on the basis of their strength of association (as measured by their p-value) and cut-off was arbitrarily set to 0.05. The resulting lists were subsequently considered as potential *MITM* biomarkers.

Biological plausibility was explored by considering the direction of effect. Validation of the candidates was attempted by replicating the findings in EPIC-Italy.

8.3 Results

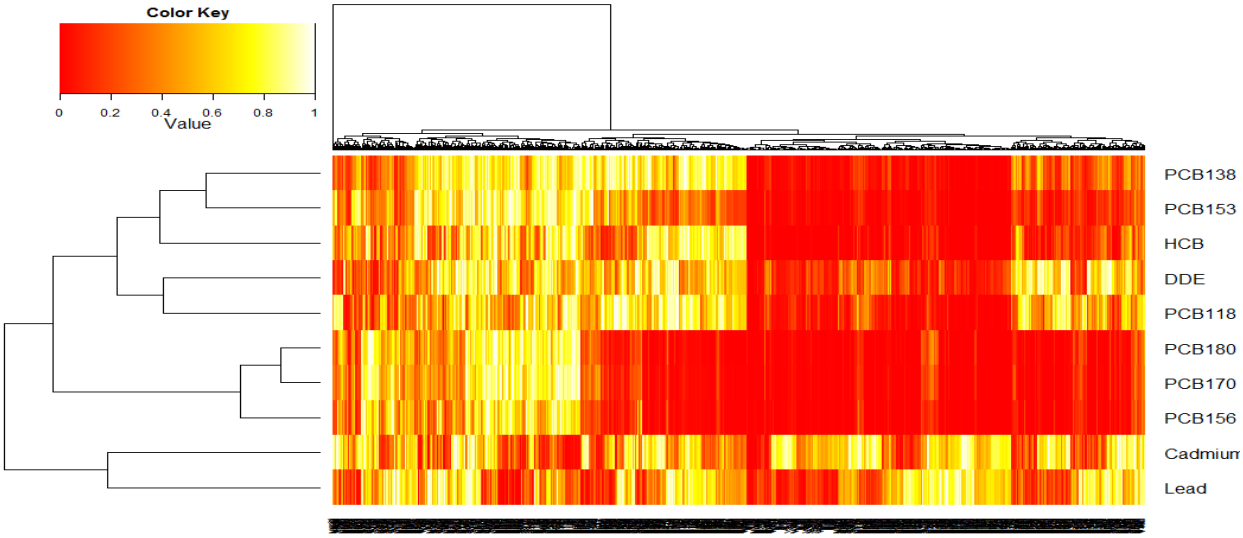
8.3.1 Disease vs. metabolite profile

Within the discovery cohort, six potentially predictive metabolite sets were selected, representing a total of 90 different metabolite features (chapter 6, results and appendix 6.3.3). Of these, 60 were unique to the total population (models 1-5), 22 were unique to the ttd <5 years subpopulation (model 6), and 8 overlapping features were found irrespective of the model used. The candidates from models 1-5 and from model 6 will be considered separately.

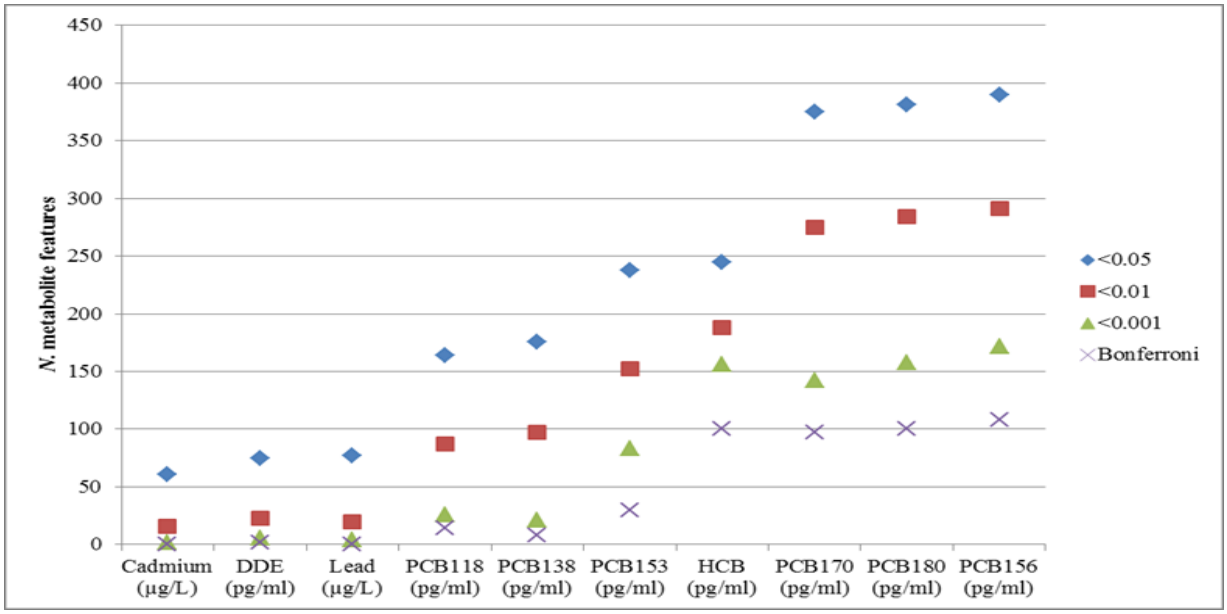
8.3.2 Exposure vs. metabolite profile

In keeping with previous analyses, metabolite features with a CV>30% in either the discovery or the validation cohorts were excluded, as were exposures with >10% of samples below the LOQ. Consequently, 750 metabolite features and 10 exposures were included in the analyses. The associations between the POPs, PCBs and metals with metabolite features are summarised in figure 8.3.2. At the stringent Bonferroni level, a large number of associations were observed for the more highly chlorinated PCB congeners and for HCB, which is also highly chlorinated. Fewer associations were found for metals and DDE. Figure 8.3.2 plots the correlation between the p-values obtained for each exposure-metabolite feature combination. It shows clear clustering across both metals as compared to the POPs, and a cluster among PCB170, 180 and 156 within the POPs. There is also evidence of clustering by metabolite feature suggesting the existence of related features.

Figure 8.3.2: Associations between measured body burden of six PCB congeners, HCB, DDE, Cadmium and Lead with intensity levels of 750 metabolite features in the discovery cohort according to significance level



Significance was computed using univariate linear regression for each exposure-metabolite pair. In the heat map the p-value for every exposure-metabolite association is presented coloured according to p-value (see colour key) the y-axis dendogram shows the clustering between the exposures and the x-axis dendogram the clustering between metabolite features. The metabolite feature names are represented along the bottom x-axis



8.3.3 Potential "*Meet-in-the-Middle*" Candidates

Figure 8.3.3a shows the effect size estimate as a function of the p-value for each of the 90 disease-related metabolite features defined in section 8.3.1. All PCBs showed high numbers of low p-value associations, particularly PCB153, 156, 170 and 180 which also tended to show the stronger effect sizes. HCB was again observed to show similar results to these more highly chlorinated congeners, whilst very few associations were noted for DDE.

Figure 8.3.3a: Association between exposure and metabolite feature intensity for 90 metabolite features identified as being associated with NHL

The β coefficients and p-values from the linear regression model are plotted to determine the strength and the significance of the association. The dashed red line indicates a p-value of 0.05. The 90 features were identified as being associated with disease risk in chapter six. They are plotted as different series depending on which disease-metabolite model they were originally identified in. full details of these models are given in appendix table 6.3.3)

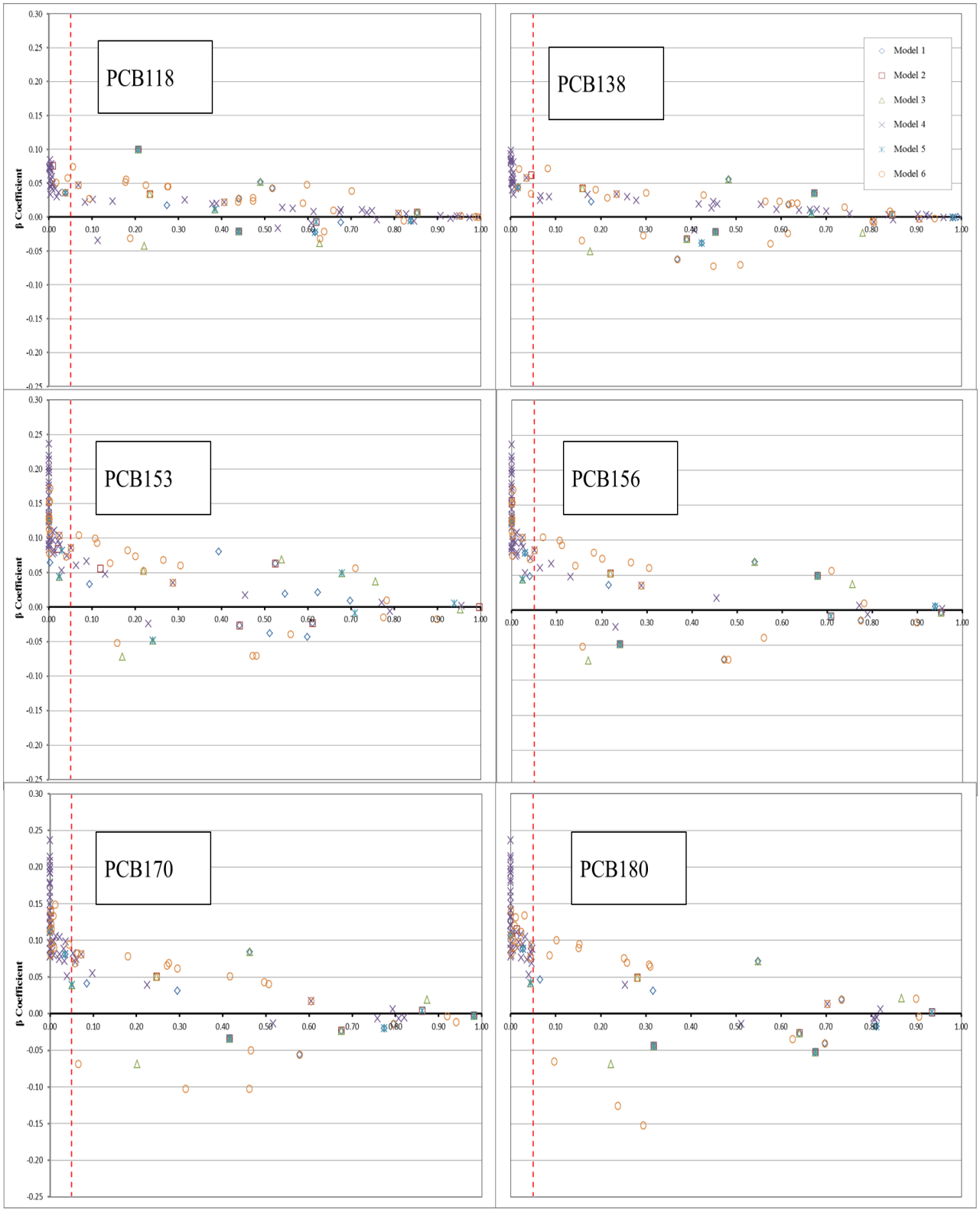
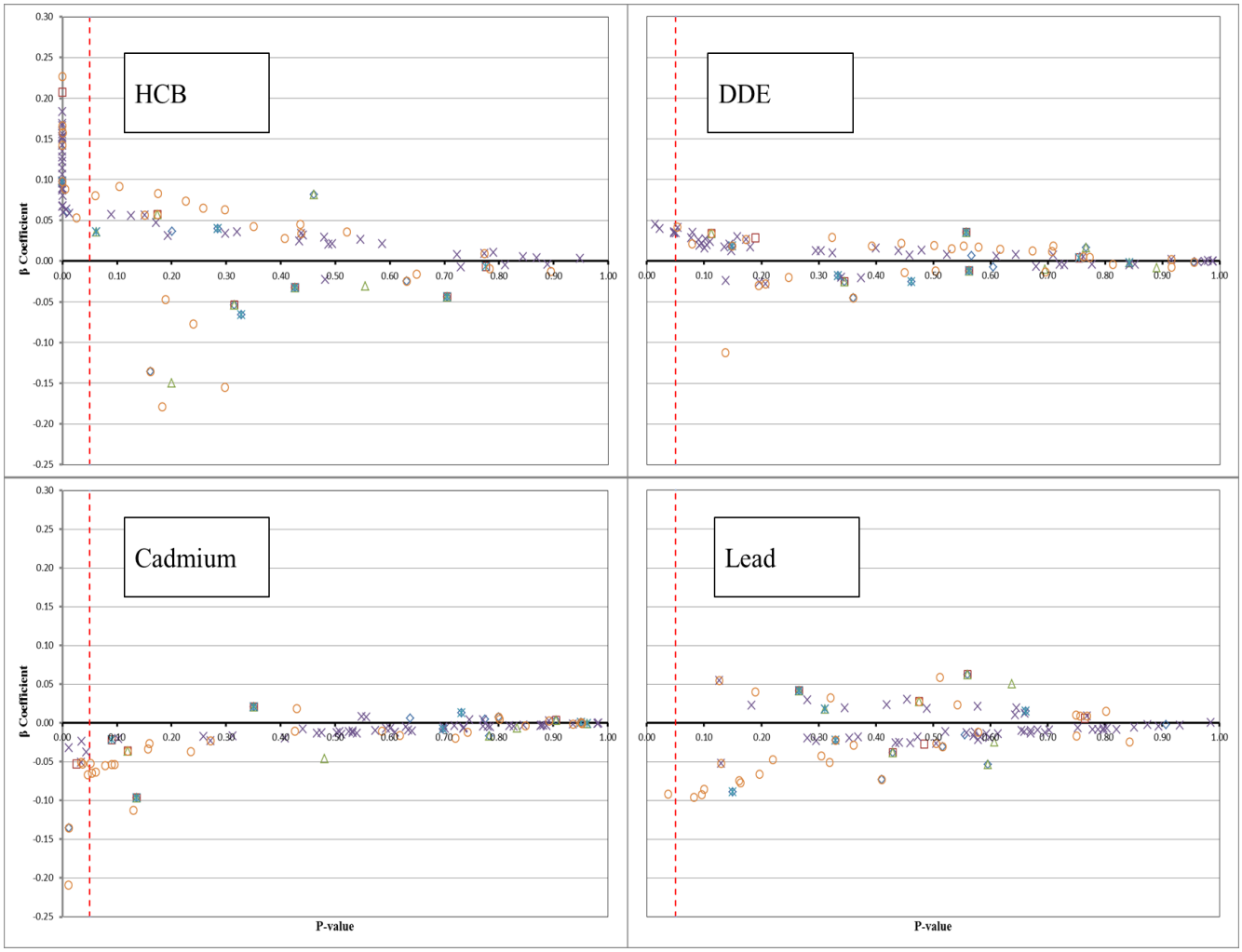


Figure 8.3.3a *continued*



To explore these results further, the direction of effect was determined for those metabolites found to be associated with the investigated pollutants and with disease (table 8.3.3). The β coefficients and p-values from the linear regression model for each of the 90 features are plotted. This provides a measure of the correlation between features intensity and pollutant concentration. Interestingly for the POPs, the intensity levels of all the *MITM* candidates were positively associated with both exposure and disease. For the heavy metals, intensity was negatively associated with exposure levels but positively associated with disease risk. There was a strong overlap between the candidates found for POPs and between the metals. In total, 62 different metabolite features can be considered as potential *MITM* candidates: 46 that were unique to the pooled population, 11 that were unique to the ttd<5 years subpopulation, and five that were common to both.

Table 8.3.3 Direction of the association with exposure concentration and with disease risk for the metabolite features identified as potential *MITM* middle candidates

Feature-disease associations selected in models 1-5 (associated ($p<0.05$) with disease in the total population) and model 6 (associated with disease in ttd<5 years) were considered separately. The *MITM* candidates were stratified into (i) those for which feature intensity was positively associated with disease and with exposure to the investigated pollutant (ii) those for which feature intensity was negatively associated with disease and with exposure to the investigated pollutant and (iii) where direction of effect was discordant between exposure and disease

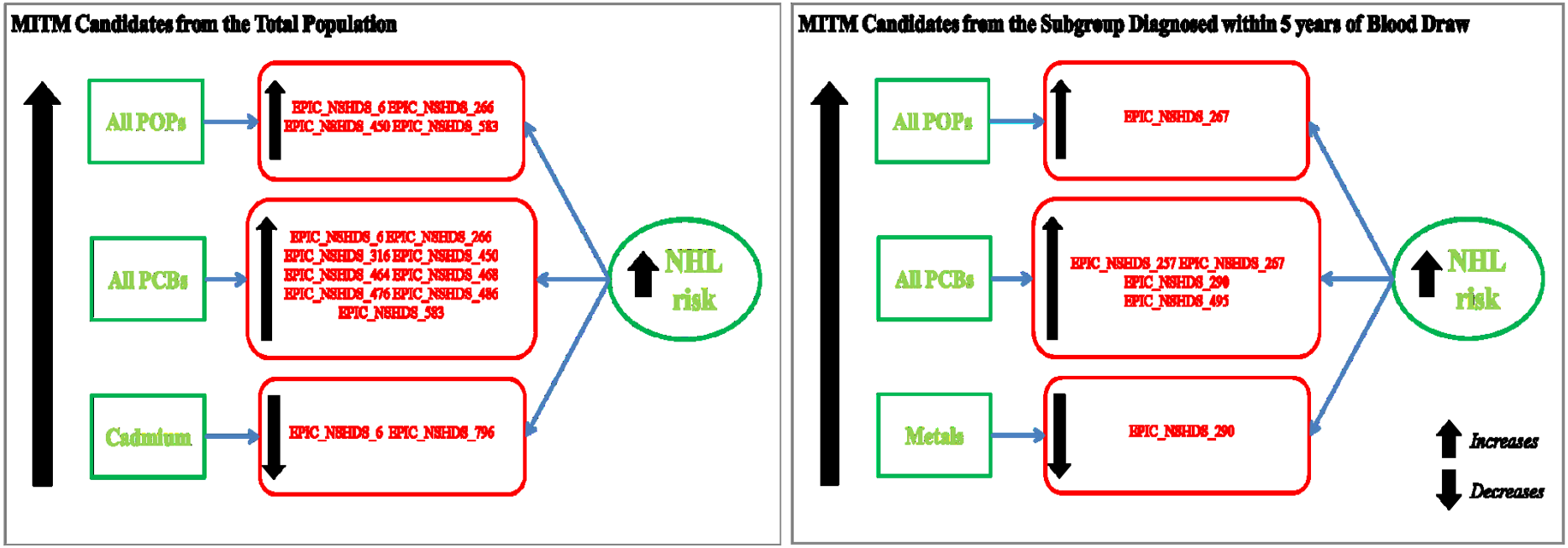
		***n*. metabolite features associated with disease and exposure**			
Model	**Exposure**	**Total**	**(i) Positive association with disease and exposure**	**(ii) Negative association with disease and exposure**	**(iii) Discordant direction of effect**
1- 5 *(associated with disease in total population)*	PCB118	23	23	0	0
	PCB138	24	24	0	0
	PCB153	45	45	0	0
	PCB156	46	46	0	0
	PCB170	43	43	0	0
	PCB180	47	47	0	0
	HCB	28	28	0	0
	DDE	6	6	0	0
	Cadmium	6	0	0	6
	Lead	0	0	0	0
6 *(associated with disease in those diagnosed within 5 years of blood draw)*	PCB118	3	3	0	0
	PCB138	4	4	0	0
	PCB153	12	12	0	0
	PCB156	12	12	0	0
	PCB170	10	10	0	0
	PCB180	12	12	0	0
	HCB	7	7	0	0
	DDE	1	1	0	0
	Cadmium	5	0	0	5
	Lead	2	0	0	2

One feature EPIC_NSHDS_244 was excluded from the model because the direction of effect with NHL was discordant among models 1-5

For these 62 features, for validation purposes, the analyses were repeated in EPIC-Italy. It comprised a much smaller sample size, and there were fewer significant associations for all the exposures. Specifically, none reached the Bonferroni threshold (appendix figure 8.3.3), as detailed in chapter six, and there was no overlap with the NSHDS results. Due to this, the replication criteria were loosened, and only the direction of effect (both with the exposure and disease) considered as an indication of consistency across both populations. These consistent features are listed in Figure 8.3.3b. For clarity, and due to the strong overlap in results across PCBs, these were considered as a single group. All POPS (PCBs, DDE, HCB) and metals were considered as two further groups. The largest number of potential candidates (n=9) are for the putative relationship between PCBs and NHL. This figure

indicates that increased intensity for the features is associated with both higher levels of PCBs and an increased risk of future disease. For the metals the results suggest that increased intensity is associated with a decreased metal erythrocyte concentration, but an increased risk of future disease.

Figure 8.3.3b: Metabolite features identified as potential *"Meet-in-the-Middle"* Candidates in the Discovery Cohort and which could be Validated in EPIC cohort



For the total model only cadmium is included as there were no candidates associated with lead and disease
None of the associations were significant in EPIC-Italy 'validation' was based on concordance of direction of effect

8.4 Discussion

Introduction

Metabolite profiles can reasonably be considered as either biomarkers of biologically effective dose or of early biological effect relating to certain exposures or to particular diseases, depending on the particular metabolites observed [224]. As such, these profiles may provide candidates for intermediate biomarkers lying on the causal pathway between the exposure and disease that could inform on mechanisms. Here this "*meet-in-the-middle*" approach has been applied to the previously explored PCB/POP/metal relationship with NHL (chapter 5) to determine if it can gain biological/mechanistic information on these results, and potentially draw conclusions on putative casual pathways.

MITM candidates

The metabolite profiles showing the strongest association with future risk of NHL risk were presented in chapter six. These findings lacked replication and the significance levels were not sufficiently strong to rule out false-positive findings. It was speculated that this failure to identify definitive associations may relate to the time-lag between blood draw and disease diagnosis, and stronger findings were observed in the population restricted to those diagnosed closer to blood draw. Accordingly, a much larger number of associations were observed for the exposure-metabolite relationships, even using stringent multiple testing correction strategies. This was to be expected, given that the exposure levels and the metabolite profile represent the same time frame, and the metabolome is known to reflect current experience. There was high concordance between the PCBs in terms of the metabolite feature associations, and for PCB170 and 180 an almost identical set were selected. Fewer associations were noted for the metals, possibly suggesting that they interact with metabolism to a lesser extent (or through a more complex mechanism) than the organochlorines.

There was an overlap of 62 biomarker features (table 8.3.3) that were associated with both exposure to one the investigated pollutants and disease. Of these, 15 features were considered to be validated in EPIC-Italy (11 from total population and 4 from ttd<5 yrs) using a loose criterion (due to the smaller sample size in EPIC-Italy). Although, this represents less than half, potential biological plausibility is strengthened by the fact that there were clear differences between the PCBs as compared to the metals. In summary, the results suggested that increased POP (PCBs, HCB, DDE) exposure levels were associated with an increased intensity of metabolite features that were also associated with an increased risk of disease. Conversely, an increased body burden of Cd and Pb appeared to be associated with decreased intensity in NHL associated metabolite features.

The causal pathway

These results are intriguing given the findings reported in Chapter 5. There was no evidence of a positive association between exposure and disease, and for PCBs, some evidence of an inverse association was observed. Therefore the positive associations between the '*MITM* candidates' with both exposure and disease seem somewhat counter-intuitive. There are a number of possible explanations for these findings:

- Either the metabolite-disease or the metabolite-exposure relationships represent chance associations.
- The disease associated metabolite profiles may be associated with another unmeasured NHL risk factor rather than exposure to environmental pollutants.
- The exposure-associated metabolite profiles could relate to a different health outcome not investigated here (or not to any outcome at all).
- There is no association between POPs and NHL, and the apparent disease-metabolite relationship may actually result from confounding by a further factor associated with both POPs and disease which may be reflected in the metabolome
- There is a strong positive association between POPs and NHL, but this was not identified in chapter 5 because body burden is a sub-optimal measure of exposure history. Changes to the metabolome are more sensitive to the effects of PCBs, so can act as an intermediate step linking them to the disease. This would be supported by the body of literature which has reported a positive association between PCBs and risk of NHL.
- The inverse association with metals could be explained if the observed metabolite profiles represent early stages in the causal pathway. Metal exposure may, through interaction or negative feedback, lead to an up regulation of later stage biomarkers, which are associated with an increased risk of disease.

To determine which, if any, of these hypotheses may explain the results, annotation of the metabolite features is required in order to identify the pathways and mechanisms involved.

Potential mechanisms

A number of studies have investigated metabolic profiles following PCB exposure in animal models [360,361]. Lipid metabolism has been shown to be particularly affected with lysophosphatidylcholine (16:0), tyrosine, PC(18:4/18:1) and lysoPE(18:2/0:0), all found to be significantly altered. Tryptophan metabolism and phenylalanine metabolism are also thought to be affected and interestingly sex-specific differences in results have been noted. This is the first study to consider perturbations to the metabolome with PCB exposure in humans and the first in any medium for HCB and DDE. A recent publication has reported that the early effects of Cd exposure in humans can be detected by metabolic profiling using NMR and urine samples [362]. They found a positive association between citrate levels and Cd exposure but an inverse association with 3-hydroxyisovalerate, dimethylglycine, creatinine, creatine and 4-deoxy-erythronic acid, which may be of particular interest. Animal studies of the

metabolite profiles following Cd exposure have reported a role for carboxylic acids, amines, sulfated compounds, glucuronides, and glycosides in addition to citrate [363]. Encouragingly, a number of the pathways in which these metabolites are involved have been previously implicated in NHL and other haematological malignancies [364-366]. Therefore, if such metabolites were included in the annotated findings this would considerably strengthen the evidence for a relationship with disease.

Strengths and limitations

Currently the biological pathways relating to these results can only be speculated upon. As in chapter six, the lack of annotation, represents a significant limitation of this study. This study was also limited by the factors stated in chapter five relating to the reliance on a single spot measure as an assessment of body burden. This is even more pertinent in this context as the metabolome is also known to fluctuate over time, suggesting this study could be considerably improved by the use of repeat measurements [7]. Only classical, rather than multivariate projection methods such as PLS were utilised in this study. This strategy was chosen is in accordance with the previous *MITM* colon cancer study [224], and to allow for the easy and intuitive adjustment of matching factors.

Subtype heterogeneity was not assessed in this chapter despite the fact that, as discussed in chapter six, metabolite profiles may be subtype specific. This issue is magnified with the use of the *'meet-in-the-middle'* approach, as the potential for the association between the disease and the exposure to also vary by subtype was observed in chapter five. Nevertheless, this study was underpowered to run subtype specific analyses and therefore, given a lack of association between the exposure and the disease had already been demonstrated; these analyses were conducted as a proof of principal of the approach.

There were also a number of strengths specific to this study, mainly stemming from its novelty, and the variety of pollutants included. This generated particularly interesting findings regarding the differences in the metabolomics profile with exposure to POPs compared to metals. The clustering of PCBs by degrees of chlorination and the similarity of findings for HCB with the lower chlorinated congeners are also of interest. This study additionally overcomes one of the challenges discussed in the introduction regarding the identification of NHL biomarkers, as it links the risk factor and the disease to potential intermediate biomarkers on the same biosample. Further, the prospective design allows the assessment of both the POPs and the metabolite profile free from the influence of inherent metabolic changes due to the disease itself [224]. Unlike in studies observing positive associations between PCBs and NHL [249], it was not possible to adjust on lipids in EGM. Since these studies indicated lipid metabolism was affected by PCB exposure [360,361], it may be worthy of further exploration within EGM. Finally, given that metabolomics has been stated to be at a developmental

stage with respect to cancer epidemiology, with the majority of studies including <100 patients [307], this analysis represents one of the largest to-date.

As discussed in chapters five and six, the generalisability of this population must be considered in the analysis of these findings. Given that the source of exposure to POPs and metals are likely to vary between populations and by geography, the excess of Swedish participants in this population must be taken into account. This is especially pertinent as a strong division by geographical cohort was observed in the PCA analysis in chapter six.

Biomarker utility

There is little clinical or predictive utility that can be garnered from this study. However the preliminary findings suggest that metabolite profiles may represent useful biomarkers to assess exposure to or experience of suggested NHL risk factors. This rests on the association between disease and the risk factor. If true, this indicates that the methods used in chapter five were inadequate to assess lifetime exposure such as would be expected to increase NHL risk. It may be that the metabolome provides a better measure.

If the exposure-metabolite profile associations are assumed to be true then the metabolome would fulfil a substantial proportion of criteria for a 'good' biomarker, including importantly replication in an independent cohort. Despite the high correlation between the PCBs the differences between the similar structural groups, such as the more highly chlorinated PCB congeners as well as the metals, suggests some specificity. There is also coherence with existing knowledge in terms of the well-known perturbations to metabolite profiles caused by exogenous exposures. The most common use of metabolomics to date has probably been in the field of toxicology. Consequently in future analyses it would be of interest to attempt building predictive exposure models similar to those illustrated in chapter six. These may have utility as biomarkers in biomonitoring studies or environmental risk assessment [367].

Regarding the other biomarker criteria, coherence cannot be considered until the annotations are complete, and this will also inform on biological plausibility. Bias cannot be ruled out for the exposure-metabolite associations, only the matching factors have been adjusted for, although an attempt has been made to control for laboratory or sample induced 'nuisance variation' by adjusting on processing batch and fasting status. If lipid adjustment was in some way biasing the results between PCBs and disease, it may also be expected to affect the association between PCBs and metabolite profiles. Furthermore, as with the subtypes-specific analyses, sex-stratified analyses were not run to retain power, to avoid issues of multiple testing and because, these analyses were conducted as a proof of principal. However, the results of chapter five suggest sex-stratified analyses may be of interest, as do the findings from the metabolomics and PCB exposure studies in animal models [360,361].

Nevertheless there is sufficient evidence to warrant the further exploration of the highlighted candidates as a means of gaining a better biological understanding of both PCB exposure and lymphomagenesis, and the potential relationship between them. In this way, the *MITM* approach provides a promising strategy to which the agnostic –omics methodology, which is able to simultaneously and objectively measure many hundreds of potential biomarkers, is particularly well suited. These biomarkers can then be considered as a profile or set, allowing better consideration of pathways than is available with the more traditionally employed exploration of single biomarkers. Finally these profiles may represent any stage of disease pathogenesis [367]. If intervention is the goal then it may be of more interest to search for biomarkers in those diagnosed more than five years after blood draw, as these will have the potential to represent early stage disease biomarkers. This was not possible here as very few features were identified for the ttd>5 years group and none were validated as *MITM* candidates.

Conclusions

This is the first study to attempt to gain insight into the biology underlying suspected relationship between POPs and NHL through the identification of intermediate biomarkers in the potential causal pathway using the *MITM* approach. This was a hypothesis-generating study and these primary results, once annotated and validated in a new dataset, will provide the starting point for future analyses.

Metabolomic profiles are particularly useful as they directly mirror a cell's function [307] and consequently provide considerable insight into dysregulated metabolic pathways, inherent disease development, and pathophysiological alterations caused by external stressors [367]. Therefore, it was hypothesised metabolomics may provide a rich resource for biomarkers lying between a risk factor and disease. The benefits of identifying such biomarkers would be three-fold: the identification of biomarkers of (i)- POPs, (ii)- and of NHL, and (iii)- the improved characterisation of their uncertain association. Based on these results, the strongest utility for metabolic profiling in this context probably lies in the first of these three. While these results are only suggestive, the number of common metabolites between the PCBs, POPs and the metals, also suggest potential common pathways associated with exposure that are worthy of further investigation. Although not conclusive, the results do go some way to providing a more comprehensive understanding of the pathways of NHL. A number of *MITM* biomarkers have been identified, however whether they belong to the causal pathway between exposure and disease can only be determined with further studies.

9 DISCUSSION

B-cell Non-Hodgkin lymphomas represent the most frequent hematopoietic cancers of adulthood, and an increase in incidence has been reported in most countries during the past 30 years, which is now beginning to abate. Although a proportion of this increase can be attributed to changes in classification and reporting of NHL within this time period, the evidence suggests these findings represent more than a mere data artefact. However, the risk factors driving the incidence trend of NHL over the last decades remain elusive and the mechanisms of malignancy poorly understood. The more that is known about this malignancy, the greater the importance of subtype heterogeneity is revealed to be, further complicating its study. The aim of this thesis was to identify novel biomarkers of the pathways of lymphomagenesis which could provide further insights into the aetiology of NHL. The generation of such biomarkers could have multi-faceted benefits for research including: improved risk factor assessment, determination of biologically important dosage of exposure, identification of individuals at high risk, and improved mechanistic insights into the lymphomagenesis pathway. This would subsequently impart translational potential to provide novel targets for prediction, prevention and therapy, and ultimately to advance the control of NHL.

9.1 Rationale and Aims

The rationale behind these aims was that novel approaches for the study of both NHL risk factors and NHL biomarkers were required, necessary as a means to try and address the relative lack of success achieved with the currently adopted strategies, as discussed in the introduction. As well as to further inform on the reported increases in NHL incidence over the previous decades. It has been stated that only if the underlying disease patterns are understood can hypotheses regarding aetiology and pathogenesis be accurately tested [15].

The thesis was split into two parts, the first of which investigated and attempted to validate four different types of biomarkers occurring at different stages along the causal pathway: biomarkers of susceptibility, internal dose, biologically effective dose and early biological effect. These were chosen on the basis of a known or suspected relationship with NHL. In the second part of the thesis the additional utility which could be conferred by integrating different biomarker types using the *MITM* approach was explored. The included population was based on a number of different studies which contained a subpopulation of EPIC NHL cases and controls. Not all EPIC NHL cases were included in the thesis, although exploration of the various subpopulations in each chapter suggested they were representative of EPIC overall. As the original projects for which these subpopulations were selected were not planned around the aims of this thesis there were some sample size limitations. Consequently a largely hypothesis-free approach was taken to determine the potential role of the biomarkers in NHL research and the feasibility of integrating them.

9.2 Summary of findings

This thesis employed novel techniques and methods and represents a number of firsts in the study of NHL. These include the potential causal relationship with Cd and Pb exposure and the determination of the predictive ability of metabolomic profiles in prediagnostic blood samples. It is additionally the first to determine a predictive threshold for the t(14;18) translocation in Follicular lymphoma risk. Finally it is among the initial few studies considering the use of the '*meet-in-the-middle'* approach and the utility of profile regression for GWAS data in any malignancy. The findings have been discussed in detail in the relevant chapters and are summarised below.

Biomarkers of susceptibility to NHL were explored using EPIC NHL cases and controls, genotyped as part of the InterLymph NHL GWAS consortium, which has reported exciting findings and novel susceptibility loci for a number of NHL subtypes, including those involved in apoptosis with CLL [87]. These findings were not replicated in the EPIC subpopulation, nor were any novel genetic variants identified. However this was not unexpected and it can be concluded that this was a function of a sample size rather than the novel statistical methodology, profile regression, which was used. In fact profile regression was found to outperform the conventional association analyses in terms of predictive ability in this population, and there was some weak evidence for the identification of biologically plausible genes such as *DAPK1*. These findings are therefore supportive of the benefits of this methodology for analysing highly dimensional, complex data [164], and pave the way for future studies.

Prediagnostic concentrations of PCBs, POPs and metals were considered as biomarkers of exposure. Intriguing results regarding the role of gender and obesity, and for the DLBCL subtype are worthy of exploration in future studies. However, overall the evidence suggested a null association between POPs and NHL, reflecting the ambiguity in the existing literature. A possible explanation for the null findings with POPs, metals and NHL risk may be that the blood measures used do not accurately reflect long-term exposure. There was some support for this theory on the basis of the metabolomics results, and the findings indicate that metabolomic profiles may be a useful method of exposure assessment or environmental monitoring. The tentative evidence that metabolomics may additionally have a role in disease prediction means that a metabolic profile can reasonably be considered as an intermediate biomarker along the causal pathway between risk factors and disease, using the theory of the *MITM* approach. In general the metabolomics results were promising. Within the discovery cohort there was evidence of biological plausibility for an association with disease risk including dose-response, robustness to adjustment for confounding, and the fact that the strongest associations were identified in those diagnosed closest to blood draw. It was not possible to replicate these findings in a validation cohort, although this may relate to underlying differences between the cohorts primarily due to anti-coagulant induced bias. Conversely for the POPs analysis coherence of results between

structurally similar pollutants supported biological plausibility and a large number of findings were replicated.

Finally the most compelling results were those related to frequency of the t(14;18) translocation as a biomarker of risk for FL. The strength, and crucially the replication, of this finding suggests an exciting potential for eventual translation into a predictive methodology. Consequently a patent based on these results has been applied for.

9.2.1 t(14;18) frequency as a candidate biomarker for risk prediction

The ultimate goal of biomarker discovery is the translation into a clinical setting. The most likely candidate in this study is the t(14;18) translocation. The results suggest that the identified high frequency threshold actually represents a previously unknown pre-disease state that will progress into overt malignancy. This would comprise the third such condition for NHL after chronic lymphocytic leukaemia/monoclonal B-cell lymphocytosis (CLL/MBL), and multiple myeloma/monoclonal gammopathy of undetermined significance (MM/MGUS) [348,349]. Both MBL and MGUS have been found to confer significant benefit in the identification of patients at high risk of CLL and MM, respectively, who can then be managed accordingly [368-370]. It can be hoped that as the first example of a pre-disease condition for FL, it may confer similar utility in the management of this subtype. Alternatively it may allow further subdivision of FL on the basis of t(14;18) frequency, which could ultimately prove useful in determining aetiology or clinical management. However, any application of this biomarker can only be speculated at presently. Further studies in larger and more diverse populations and with additional validation are needed before any kind of clinical translation can be considered.

9.3 Strengths and Weaknesses

Some caution must be applied when considering the findings in this thesis, due to the limitations discussed in the relevant chapters, which stem mainly from sample size. Power was computed at the outset of the EGM study, although the strength of the emerging findings now suggest the sample size chosen may have been insufficient. For this reason, and due to the hypothesis generating nature of this thesis, power has not been calculated here for the individual chapters. The power of the individual analyses was further limited where subtype-specific analyses were conducted. In fact, subtype heterogeneity represents an important limitation of any study considering NHL as a group. This is discussed further in section 9.5 below.

There were a number of other limitations, misclassification of disease or subtype status can often represent a significant issue in studies of NHL. This can be a particular problem in prospective cohorts which encompass a number of the different classification systems used over the years [26], however within the EPIC cohort procedures have been implemented to address this [215]. With regards to the EPIC NHL data, there was a lack of information on grading and staging of disease with only 21% of cases having this information available, and no information on family history. However the fact that this population arose from a large well-characterised cohort with inbuilt quality control, calibration and validation procedures represents one of the major strengths of this thesis and minimises many of the potential problems inherent in epidemiological studies.

As a cohort study the baseline characteristics of the participants are free of the recall bias which has plagued many of the case-control studies comprising the majority of the epidemiological literature for NHL. Nevertheless although the information was collected at baseline, it still relies on self-reported data. In fact, as discussed, the inconsistencies in the NHL literature may well arise from the inaccurate measurement of many risk factors which have been based on self-reported data. In this study the use of self-reported information for smoking status, physical activity index and education level could have introduced bias. For all these variables what is of interest is lifetime experience, which may be particularly difficult for people to report accurately.

There was a possibility of selection bias, as analyses were performed on the available data, rather than specifically selected populations. However, again such bias was likely to be minimal as the majority of the analyses were on matched case-control pairs. The nesting within EPIC raises issues of the generalisability of the findings, both in terms the genetic and metabolomic data and also the data on the suspected risk factors which are subject to geographical preferences and regulations. Consequently this may need to be addressed in the forward translation of these findings.

Nevertheless, the underlying focus of this thesis was to generate hypotheses and assess the feasibility of novel strategies for biomarker development. Therefore its main strengths lie in the inferences which can be gleaned from these analyses and the way in which they can inform future work. These are discussed in further detail below.

9.4 Recommendations and implications for biomarker research

Prospective cohort studies are vital for the study of risk factors and predictive biomarkers

Although a number of the Bradford Hill criteria have been rendered redundant in this era of molecular epidemiology, what remains true is that to establish a causal relationship exposure to, or experience of, a risk factor must precede disease. Similarly a predictive biomarker must be shown to be present and measurable before disease onset so it can be determined that is not influenced by the pathophysiological changes due to the disease itself [224]. Indeed, as mentioned, the prospective design employed represents one of the main strengths of this thesis. Although case-control studies may provide useful hypotheses, the comparison of the results presented here to the existing case-control based literature and the consideration of the inherent bias associated with such studies leads to the conclusion that a predictive biomarker can only be accurately assessed in the context of a cohort study. To date there has been a failure to address this in the NHL literature. There is still a lack of large-scale population based cohort studies. They are hampered by cost and the relative rarity of the individual subtypes. More are needed and, crucially, they must be constructed in a way that allows for molecular epidemiological analysis.

Accurate measurement of exposure is vital

Given the suspected environmental component to NHL aetiology, one of the biggest challenges in its study is accurately measuring the risk factor. Even for data that is based on measurements rather than self-report current, methods for measuring external exposure, such as through residential proxies, JEM or measure from the surrounding environment, tend to lack accuracy. Consequently they may not be reflective of the dosage actually experienced by an individual. This underlines the need for internal molecular biomarkers with increased validity.

Here, blood concentration was used as a proxy for lifetime exposure, but the results suggest that the metabolomic profiles may represent a more accurate picture of the effect of such exposures. The effect of exposure can then be better correlated with actual disease risk. This is in agreement with the literature which observes that omics technologies provide a theoretically objective and informative measure of the long-term exposure to environmental pollutants based on the resulting perturbations of cellular pathways and networks [224,367]. Other omics technologies, such as epigenetics, are likely to be better suited for this purpose than metabolomics and should play a pivotal role in future studies

Factors occurring in the early stages of the causal pathway may be too far removed from the disease to act as biomarkers in diseases with long latency; the most promising biomarkers are late stage biomarkers

Early stage biomarkers may be too far upstream of the event of interest to act in a predictive manner. This was evidenced by the fact that within this thesis the strongest associations by far were noted for a

late stage biomarker - translocation frequency. For early stage biomarkers the relationship with disease, which may not occur for many years, is likely to be distorted by the intervening stages and pathways and the potential bias and confounding these may introduce into the association. This could again explain the null relationship with the POPs/metals and is also supported by the fact that for the metabolomics markers the strength of the associations increase closer to diagnosis. This suggests that metabolomics profiles can act as both early and late stage biomarkers, with the utility of the early stage biomarkers lying in their improved assessment of risk factors and in the inferences they can provide on mechanisms. Annotation of the metabolite profiles will help disentangle this further.

Integrating biomarkers provides increased utility and interpretability

It can be concluded that additional utility was imparted through the integration of different biomarker categories to provide a more global view of the disease process using the MITM approach.

The results of chapter five, when considered in isolation, may lead to the conclusion that PCB exposure was protective against NHL. The inclusion of the metabolic profile enables the consideration of the alternative hypotheses that the blood-based measure of POPs does not provide an accurate assessment of exposure. Further work may consider gene-environment interaction. Similarly the profile regression analyses emphasised the importance of not considering genetics in isolation given the differences in profile regression results when environmental factors were included in the analysis. Furthermore, integrating additional biomarkers such as genetic variants or environmental exposures into the consideration of t(14;18) translocation frequency helps to improve the understanding of the causes and drivers of this translocation and consider its translation into a clinically useful screening strategy in high risk groups. However, within this thesis the numbers are too small to draw definitive conclusions.

The role of *a priori* hypotheses should not be overlooked

In order to integrate biomarkers in a biologically meaningful way it is vital to choose them carefully, based on known biology, pathways and mechanisms. In an era where there is a general trend to focus on the use of large scale hypothesis-free studies using next generational high throughput techniques, the findings from this study suggest that hypothesis-driven analyses should not be overlooked, even though they have been criticised by some [169,371]. The most successful analysis was that conducted with a biological *a priori*, namely the translocation study. The GWAS approach did not yield biologically meaningful findings, and it is difficult to draw definitive inferences on the metabolic profiling without annotation and validation. Candidate studies may be particularly appropriate when, as in this thesis, sample size is limited.

Omics focussed studies are likely to play an increasingly important role

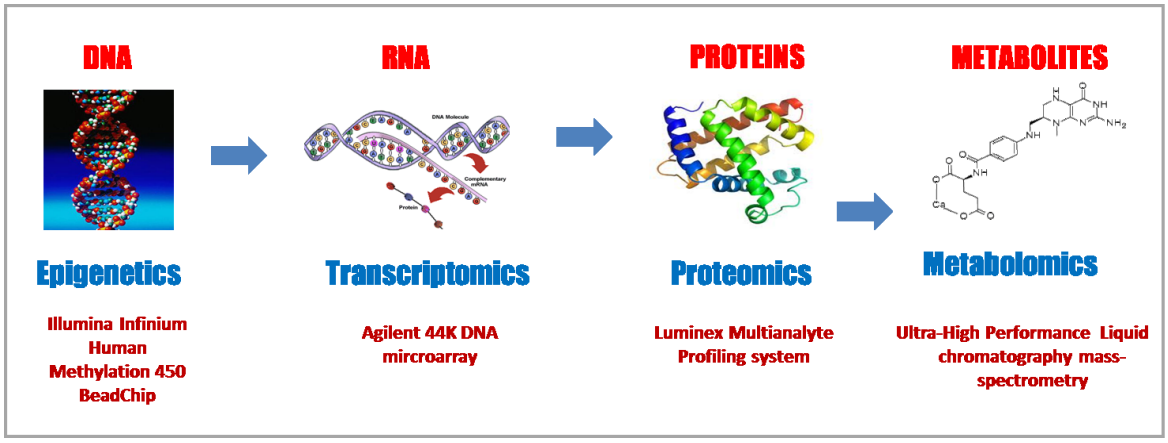
If a population is large and contains a sufficient number of cases then omics, which tend to be hypothesis-free, can provide almost unprecedented opportunity to search for biomarkers of exposure

or disease at a population-wide level. By considering many hundreds or thousands of variables simultaneously a temporal picture of disease evolution can be constructed based on the interacting pathways and networks identified [367].

In addition to metabolomics, other omics technologies were conducted within the EGM cohort (figure 10.4) of which the transcriptomic analyses produced particularly interesting results. Over 700 candidate genes showing altered expression years before clinical onset were observed for Chronic Lymphocytic Leukaemia. These genes were involved in cell signalling networks and immune system regulation pathways, providing plausible new mechanistic insights [372]. Encouragingly, like the metabolomics results, they also showed a clear distinction in findings by time to diagnosis.

Consequently, the results to date suggest an important future role for omics technologies, which as has been shown, can be complemented by the *MITM* approach. This has also been applied to the transcriptomics data with very encouraging results (*Castagne and Kelly et al. under preparation*[4]). As for the more traditional biomarkers, it should also be noted that the greatest utility will likely be garnered by integrating the different hierarchical omics levels (figure 9.4) to potentially allow the characterisation of the entire malignant phenotype [310].

Figure 9.4 Methodologies and '-omics' technologies included in the EnviroGenoMarkers Study



Novel Statistical methodologies are vital

Such developments also introduce new challenges which are only just beginning to be addressed. Most omics markers have not yet been validated, few longitudinal studies have been conducted, standardized protocols are lacking and uncertainty surrounds their applicability and role in the causal paradigm [367]. In addition their highly dimensional nature raises major challenges in terms of statistics

[4] Castagné R, Kelly RS, Vineis P, Kyrtopoulos S, Vermeulen R, Chadeau-Hyam on behalf of the EnviroGenoMarkers project consortium Integration of gene expression profiling and meet-in-the-middle strategies to dissect transcriptomic changes after environmental pollutant exposures in a B-Cell Chronic Lymphocytic leukaemia cohort study *(under preparation)*

and study design, with complex biostatistical methodologies and high computing capabilities required to deal with these issues [238]. These challenges extend even beyond those facing researchers when they first tried to characterise the human genome, due to the additional temporal component of these downstream –omes as well as issues regarding their stability, variation, plasticity and the ways in which they relate to an individual's lifecourse [367]. Large changes in the philosophies and methods of statistical analyses applied to epidemiological data are required to cope with these issues and keep pace with the technological advances [169].

It is vital to make biological as well as statistical meaning from such data [239]. This requires a multi-disciplinary systems biology approach incorporating pre-existing knowledge and taking the multistage evolution of disease into account [168,367]. In fact it has been argued that the as the field develops the focus should be on approaches that improve the understanding of the biology rather than on the development of new statistical methods [238]. Pathway analysis is the most popular method for converting high throughput data into biologically meaningful results [239], as demonstrated in chapters 4 and 8, and there additionally exist similar methods for metabolomics which could be applied to annotated data [230]. However pathway analysis is also entirely reliant on the underlying databases utilized. The existing databases are not completely comparable in terms of nomenclature and most are in varying stages of completion; a large number of genes remain unannotated and for those that are, many annotations are of a low quality and potentially inaccurate. Consequently a further required development is improved mapping techniques and high resolution annotation knowledge bases which are specific to the methodologies and technologies used [169,239]. Only this can allow the understanding of whether the identified biomarker does in fact belong to the causal pathway, and the role it plays.

Replication holds the key

Throughout the thesis three criteria sets were used to assess the explored biomarkers: Bradford-Hill, Ransohof and the Venice criteria, with ROC curves, sensitivity and specificity additionally presented where appropriate. In fact, it is argued by Ransohof, that the most important of the biomarker criteria is replication, and without this the other factors are meaningless [373]. This is well illustrated by the contrast in the confidence that can be placed in the findings of the metabolomics analysis relative to those of the translocations where near identical results were observed in the discovery and validation cohorts. It should also be noted that it is possible that the failure to replicate the findings in the metabolomics validation cohort stemmed from the comparability of the samples which were processed with a different anti-coagulant introducing substantial nuisance variation into the results. Therefore this underlines the importance, not only of validating, but of validating in a suitable population.

In addition to 'epidemiological validation' of a finding through replication in an independent cohort, is it also vital to show that a proposed biomarker displays 'technical validation' in terms of intrinsic

measurement error and analytic sensitivity which can arise from within subject variability. The best way to minimise such variability is through the use of repeat samples [7]. These would be of particular value for all the biomarkers explored here by allowing the visualisation of the temporal progression of the disease pathway, and potentially of 'critical windows' of exposure. Therefore, although not available in EPIC, where possible repeat sampling should always be built into such studies

Clinical prediction is still not a reality in most instances

In the introduction to this thesis, it was stated that biomarkers that could help to identify individuals at an increased risk of NHL would be sought, with a view to eventual translation into clinical practice, such as screening strategies. Currently however, despite the strong findings reported here as well as very promising findings from the EGM transcriptomics study, translating their potential predictive ability into a clinical capacity is currently not feasible. In fact much of this stems from the nature of lymphoma itself rather than from failings with the predictive test. With reference to the discussion on FL in chapter 7 (7.3.4) NHL in general is not well suited to population-wide screening regardless of the strength of the association with the identified biomarker.

In fact very few conditions do meet all these criteria, and the number of population-wide screening programs, in any nation or health system is incredibly small. More common are targeted screening tests in certain sub-populations, such as those exposed to known causal agents. For NHL this may include measuring translocation frequency in groups such as pesticide sprayers or workers exposed to particular chemical, although definitive validation of such risk factors would first be required.

Ethical implications must be taken into account

Thus the search for biomarkers will generate vast amounts of new information, some of which may represent plausible candidates for forward translation into clinical practice, and some of which will represent incidental findings. Both have important ethical, legal, financial and physiological implications [374,375]. Genetic screening for cancer susceptibility has become an accepted part of oncologic care particularly for individuals with an inherited disposition to cancers of the breast, ovary, colon, stomach and uterus [376]. However current guidelines are less than definitive and only pertain to high penetrance mutations. The identification of predictive biomarkers from other technologies, particularly as incidental findings, potentially opens 'Pandora's box' regarding issues of ownership, dissemination, regulation and informed consent [374-376]. This may prove particularly problematic for omics as there are still so many unknowns in these types of analyses and the development of new biomarker criteria must reflect this.

9.5 NHL and Subtype Heterogeneity

The points discussed here can well be applied to the study of NHL and indeed the lessons learnt are already beginning to be assimilated into ongoing studies. However NHL remains a challenging malignancy to study. The single biggest factor behind this is its heterogeneity. In fact it could be argued that NHL doesn't represent one disease at all, but multiple. Unfortunately due to small sample sizes, subtype specific analyses were limited here. Although as a group NHL represents the eighth most common malignancy in Western Europe, each individual subtype has relatively low incidence [22]. Among a cohort of more than half a million people even the most common subtype - Multiple myeloma (MM) - comprised only 452 cases. The two subtypes generally reported as being the most common among the NHLs, DLBCL and FL comprised 284 and 243 cases respectively, and in total 1757 incident cases of NHL have been identified in the EPIC cohort to date (This total does not include French cases as France was not involved in the Lymphoma working group). In fact MM is often considered a distinct entity from the other B-cell NHL due to its pathophysiological differences, and is excluded from analyses. However under the most commonly used classifications scheme today [17,19] MM is classed as a B-cell NHL (hierarchical group 4) and therefore was considered with the other B-cell subtypes here.

Here, there was no evidence of statistically significant heterogeneity by subtype for either the PCB/POP/Metals analysis or the metabolomics. Conversely, the most promising findings were those restricted to one subtype, FL, and similarly the aforementioned EGM transcriptomics findings pertain only to CLL [372]. Therefore when exploring aetiology and biomarkers these results support the utility of considering NHL both as a cohesive group and by subtype, numbers permitting, particularly when taking a hypothesis free approach, as was done in this thesis. Furthermore to reflect the gender-disparities by subtype, ideally subtype analyses should also be gender specific [23]. Finally, the additional subclassification of the subtypes based on the findings emerging from genetic and epigenetic studies also needs to be considered [18]. The high risk biomarker for FL identified here, adds to this discussion.

9.6 Concluding comments

In conclusion, through addressing the aims set out in the introduction a number of interesting results have emerged, as well as a number of challenges. It was stated that in the introduction that NHL was a prime candidate for biomarker development. Although this remains true, the results suggest that obtaining biomarkers that are accurate and of utility for all subtypes may not be feasible. One of the major conclusions of this thesis is the increasingly important role that subtype heterogeneity plays in all aspects of NHL research.

In terms of the results, a novel biomarker to inform on the risk of one of the most common NHL subtypes, FL, in the form of a t(14;18) translocation frequency threshold has been identified. Although currently unsuitable for clinical usage, it allows further insights into the mechanisms of this malignancy, and additionally represents a previously unknown pre-disease state.

The results also emphasise again the relevance of gene-environment interactions in the aetiology of NHL and suggest that future work should also consider gene-biomarker interactions and the importance of the interacting omics levels including the genome. If undiscovered risk factors for NHL do in fact exist, it may be that such integrated analyses are the only way to identify them. Integration of different technologies was shown to improve utility and encouraging proof of principle results are presented for the *MITM* approach, suggesting that it is a feasible methodology that could reasonably be applied to the other omics technologies. Similarly profile regression could well play an important role in the future of molecular epidemiology. Finally this thesis confirms the importance of omics technologies in the future of NHL research but also underlines the continued importance of targeted studies and suggest the most fruitful identification of biomarkers and the most complete picture of the disease process is likely to be achieved through the integration of the two. The subsequent translation of such biomarkers into clinical practice could have huge implications for this highly prevalent malignancy.

REFERENCES

1. Ferlay J SH, Bray F, Forman D, Mathers C, Parkin DM. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 In: IARC; 2013.
2. Zhang Y, Dai Y, Zheng T, Ma S. Risk Factors of Non-Hodgkin Lymphoma. Expert Opin Med Diagn 2011;5:539-50.
3. Fisher SG, Fisher RI. The epidemiology of non-Hodgkin's lymphoma. Oncogene 2004;23:6524-34.
4. Müller AS, Ihorst G, Mertelsmann R, Engelhardt M. Epidemiology of non-Hodgkin's lymphoma (NHL): trends, geographic distribution, and etiology. Ann Hematol 2005;84:1-12.
5. Holford TR, Zheng T, Mayne ST, McKay LA. Time trends of non-Hodgkin's lymphoma: are they real? What do they mean? Cancer Res 1992;52:5443s-6s.
6. Groves FD, Linet MS, Travis LB, Devesa SS. Cancer Surveillance Series: Non-Hodgkin's Lymphoma Incidence by Histologic Subtype in the United States From 1978 Through 1995. Journal of the National Cancer Institute 2000;92:1240-51.
7. Vineis P, Perera F. Molecular Epidemiology and Biomarkers in Etiologic Cancer Research: The New in Light of the Old. Cancer Epidemiology Biomarkers & Prevention 2007;16:1954-65.
8. Gallo V, Egger M, McCormack V, et al. STrengthening the Reporting of OBservational studies in Epidemiology – Molecular Epidemiology (STROBE-ME): An Extension of the STROBE Statement. PLoS Med 2011;8:e1001117.
9. Mayeux R. Biomarkers: Potential uses and limitations. Neurotherapeutics 2004;1:182-8.
10. Mead GM. ABC of clinical haematology: Malignant lymphomas and chronic lymphocytic leukaemia. BMJ 1997;314:1103.
11. Bauer S. B-Cell differentiation and risks for neoplastic transformation. In: Marti GV, R; ed. Determining the role of environmental exposures as risk factors for B-cell lymphoproliferative disorders. Georgia: U.S Department of Health and Human Services; 1997:51-68.
12. Skibola CF, Curry JD, Nieters A. Genetic susceptibility to lymphoma. Haematologica 2007;92:960-9.
13. Kuppers R, Hansmann ML. The Hodgkin and Reed/Sternberg cell. Int J Biochem Cell Biol 2005;37:511-7.
14. Cabanillas F. Can we cure indolent lymphomas? Clinical Cancer Research 1997;3:2655-9.
15. Smith A, Howell D, Patmore R, Jack A, Roman E. Incidence of haematological malignancy by sub-type: a report from the Haematological Malignancy Research Network. Br J Cancer 2011;105:1684-92.
16. Swerdlow SH, Campo, E., Harris, N.L., Jaffe, E.S., Pileri, S.A., Stein, H., Thiele, J., Vardiman, J.W, ed. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. fourth ed. Lyon: IARC; 2008.
17. Campo E, Swerdlow SH, Harris NL, Pileri S, Stein H, Jaffe ES. The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. Blood 2011;117:5019-32.
18. Jaffe ES. The 2008 WHO classification of lymphomas: implications for clinical practice and translational research. ASH Education Program Book 2009;2009:523-31.
19. Morton LM, Turner JJ, Cerhan JR, et al. Proposed classification of lymphoid neoplasms for epidemiologic research from the Pathology Working Group of the International Lymphoma Epidemiology Consortium (InterLymph). Blood 2007;110:695-708.
20. WHO. Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues. In: Jaffe ES HN, Stein H, Vardiman JW, , ed. World Health Organization Classification of Tumours. Lyon: IARC Press; 2001.
21. Fritz A PC, Jack A, Shanmugaratnam k, Sobin L, Parkin DM. International Classification of Diseases for Oncology (ICD-O). Geneva; 2000.
22. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F.

GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet].Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr
23. Smith A, Roman E, Howell D, et al. The Haematological Malignancy Research Network (HMRN): a new information strategy for population based epidemiology and health service research. British Journal of Haematology 2010;148:739-53.
24. Anderson JR, Armitage JO, Weisenburger DD. Epidemiology of the non-Hodgkin's lymphomas: distributions of the major subtypes differ by geographic locations. Non-Hodgkin's Lymphoma Classification Project. Ann Oncol 1998;9:717-20.
25. Boffetta P. I. Epidemiology of adult non-Hodgkin lymphoma. Annals of Oncology 2011;22:iv27-iv31.
26. Roman E, Smith AG. Epidemiology of lymphomas. Histopathology 2011;58:4-14.
27. Saberi Hosnijeh F, Krop EJM, Scoccianti C, et al. Plasma Cytokines and Future Risk of Non-Hodgkin Lymphoma (NHL): A Case-Control Study Nested in the Italian European Prospective Investigation into Cancer and Nutrition. Cancer Epidemiology Biomarkers & Prevention 2010;19:1577-84.
28. Vineis P. Incidence and time trends for lymphomas, leukemias and myelomas: Hypothesis generation. Leukemia Research 1996;20:285-90.
29. Wheeler DC, De Roos AJ, Cerhan JR, et al. Spatial-temporal analysis of non-Hodgkin lymphoma in the NCI-SEER NHL case-control study. Environ Health 2011;10:63.
30. Han YY, Dinse GE, Davis DL. Temporal and demographic patterns of non-Hodgkin's lymphoma incidence in Pennsylvania. Int J Occup Environ Health 2010;16:75-84.
31. Levine PH, Hoover R. The emerging epidemic of non-Hodgkin's lymphoma: current knowledge regarding etiological factors. Cancer Epidemiology Biomarkers & Prevention 1992;1:515-7.
32. Bassig BA, Lan Q, Rothman N, Zhang Y, Zheng T. Current understanding of lifestyle and environmental factors and risk of non-hodgkin lymphoma: an epidemiological update. J Cancer Epidemiol 2012;2012:978930.
33. Adamson P, Bray F, Costantini AS, Tao M-H, Weiderpass E, Roman E. Time trends in the registration of Hodgkin and non-Hodgkin lymphomas in Europe. European Journal of Cancer;43:391-401.
34. Hardell L, Eriksson M. Is the decline of the increasing incidence of non-Hodgkin lymphoma in Sweden and other countries a result of cancer preventive measures? Environ Health Perspect 2003;111:1704-6.
35. Grulich AE, Vajdic CM, Falster MO, et al. Birth Order and Risk of Non-Hodgkin Lymphoma—True Association or Bias? American Journal of Epidemiology 2010;172:621-30.
36. Kinlen L. Immunosuppressive Therapy and Acquired Immunological Disorders. Cancer Research 1992;52:5474s-6s.
37. Carbone A, Gloghini A. AIDS-related lymphomas: from pathogenesis to pathology. Br J Haematol 2005;130:662-70.
38. Filipovich AH, Mathur A, Kamat D, Shapiro RS. Primary Immunodeficiencies: Genetic Risk Factors for Lymphoma. Cancer Research 1992;52:5465s-7s.
39. Grulich AE, Vajdic CM, Cozen W. Altered Immunity as a Risk Factor for Non-Hodgkin Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2007;16:405-8.
40. Anderson LA, Gadalla S, Morton LM, et al. Population-based study of autoimmune conditions and the risk of specific lymphoid malignancies. International Journal of Cancer 2009;125:398-405.
41. Kane E, Newton R, Roman E. Non-Hodgkin lymphoma and gluten-sensitive enteropathy: estimate of risk using meta-analyses. Cancer causes & control: 2011;22:1435-44.
42. Wang H, Diepgen TL. Is atopy a protective or a risk factor for cancer? A review of epidemiological studies. Allergy 2005;60:1098-111.
43. Mills PK, Beeson WL, Fraser GE, Phillips RL. Allergy and Cancer: Organ Site-Specific Results from the Adventist Health Study. American Journal of Epidemiology 1992;136:287-95.
44. Eriksson NE, Holmen A, Hogstedt B, Mikoczy Z, Hagmar L. A prospective study of cancer incidence in a cohort examined for allergy. Allergy 1995;50:718-22.

45. Briggs NC, Levine RS, Brann EA. Allergies and risk of non-Hodgkin's lymphoma by subtype. Cancer Epidemiol Biomarkers Prev 2002;11:401-7.
46. Vajdic CM, Falster MO, de Sanjose S, et al. Atopic Disease and Risk of Non–Hodgkin Lymphoma: An InterLymph Pooled Analysis. Cancer Research 2009;69:6482-9.
47. Cozen W, Cerhan JR, Martinez-Maza O, et al. The effect of atopy, childhood crowding, and other immune-related factors on non-Hodgkin lymphoma risk. Cancer causes & control: 2007;18:821-31.
48. Ellison-Loschmann L, Benavente Y, Douwes J, et al. Immunoglobulin E Levels and Risk of Lymphoma in a Case-Control Study in Spain. Cancer Epidemiology Biomarkers & Prevention 2007;16:1492-8.
49. Söderberg K, Hagmar L, Schwartzbaum J, Feychting M. Allergic conditions and risk of hematological malignancies in adults: a cohort study. BMC Public Health 2004;4:1-6.
50. Becker N, Fortuny J, Alvaro T, et al. Medical history and risk of lymphoma: results of a European case–control study (EPILYMPH). J Cancer Res Clin Oncol 2009;135:1099-107.
51. Smedby KE, Hjalgrim H, Chang ET, et al. Childhood Social Environment and Risk of Non–Hodgkin Lymphoma in Adults. Cancer Research 2007;67:11074-82.
52. Morton LM, Wang SS, Cozen W, et al. Etiologic heterogeneity among non-Hodgkin lymphoma subtypes. Blood 2008;112:5150-60.
53. McCormick DP, Ammann AJ, Ishizaka K, Miller DG, Hong R. A study of allergy in patients with malignant lymphoma and chronic lymphocytic leukemia. Cancer 1971;27:93-9.
54. Engels EA. Infectious Agents as Causes of Non-Hodgkin Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2007;16:401-4.
55. Becker N, Falster MO, Vajdic CM, et al. Self-reported history of infections and the risk of non-Hodgkin lymphoma: An InterLymph pooled analysis. International Journal of Cancer 2012;131:2342-8.
56. Arisawa K, Soda M, Endo S, et al. Evaluation of adult T-cell leukemia/lymphoma incidence and its impact on non-Hodgkin lymphoma incidence in southwestern Japan. International Journal of Cancer 2000;85:319-24.
57. Sagaert X, Van Cutsem E, De Hertogh G, Geboes K, Tousseyn T. Gastric MALT lymphoma: a model of chronic inflammation-induced tumor development. Nat Rev Gastroenterol Hepatol 2010;7:336-46.
58. Du MQ. MALT lymphoma : recent advances in aetiology and molecular genetics. J Clin Exp Hematop 2007;47:31-42.
59. Lan Q, Zheng T, Rothman N, et al. Cytokine polymorphisms in the Th1/Th2 pathway and susceptibility to non-Hodgkin lymphoma. Blood 2006;107:4101-8.
60. Neasham D, Sifi A, Nielsen KR, et al. Occupation and risk of lymphoma: a multicentre prospective cohort study (EPIC). Occupational and Environmental Medicine 2011;68:77-81.
61. Boffetta P, de Vocht F. Occupation and the Risk of Non-Hodgkin Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2007;16:369-72.
62. Schenk M, Purdue MP, Colt JS, et al. Occupation/industry and risk of non-Hodgkin's lymphoma in the United States. Occup Environ Med 2009;66:23-31.
63. Freedman M, Kohles S. Plasma Levels of Polychlorinated Biphenyls, Non-Hodgkin Lymphoma, and Causation. Journal of Environmental and Public Health 2012;2012:15.
64. Bertazzi PA, Consonni D, Bachetti S, et al. Health Effects of Dioxin Exposure: A 20-Year Mortality Study. American Journal of Epidemiology 2001;153:1031-44.
65. Hooived M, Heederik DJJ, Kogevinas M, et al. Second Follow-up of a Dutch Cohort Occupationally Exposed to Phenoxy Herbicides, Chlorophenols, and Contaminants. American Journal of Epidemiology 1998;147:891-9.
66. Baccarelli A, Hirt C, Pesatori AC, et al. t(14;18) translocations in lymphocytes of healthy dioxin-exposed individuals from Seveso, Italy. Carcinogenesis 2006;27:2001-7.
67. Saberi Hosnijeh F, Lenters V, Boers D, et al. Changes in lymphocyte subsets in workers exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD). Occupational and Environmental Medicine 2012;69:781-6.
68. Dutta SK, Mitra PS, Ghosh S, et al. Differential gene expression and a functional analysis of PCB-exposed children: understanding disease and disorder development. Environ Int 2012;40:143-54.

69. Cole P, Trichopoulos D, Pastides H, Starr T, Mandel JS. Dioxin and cancer: a critical review. Regulatory Toxicology and Pharmacology 2003;38:378-88.
70. Freeman MK, SS. Plasma Levels of Polychlorinated Biphenyls, Non-Hodgkin Lymphoma, and Causation. Journal of Environmental and Public Health 2012;2012:15.
71. Alcock RE, Behnisch PA, Jones KC, Hagenmaier H. Dioxin-like PCBs in the environment - human exposure and the significance of sources. Chemosphere 1998;37:1457-72.
72. Boers D, Portengen L, Turner WE, Bueno-de-Mesquita HB, Heederik D, Vermeulen R. Plasma dioxin levels and cause-specific mortality in an occupational cohort of workers exposed to chlorophenoxy herbicides, chlorophenols and contaminants. Occupational and Environmental Medicine 2012;69:113-8.
73. Alexander DD, Mink PJ, Adami H-O, et al. The non-Hodgkin lymphomas: A review of the epidemiologic literature. International Journal of Cancer 2007;120:1-39.
74. Zheng T, Hoar Zahm S, P Cantor K, D Weisenburger D, Zhang Y, Blair A. Agricultural Exposure to Carbamate Pesticides and Risk of Non-Hodgkin Lymphoma. Journal of Occupational and Environmental Medicine 2001;43:641-9.
75. Zahm SH, Blair A. Pesticides and non-Hodgkin's lymphoma. Cancer Res 1992;52:5485s-8s.
76. Viel J-F, Floret N, Deconinck E, Focant J-F, De Pauw E, Cahn J-Y. Increased risk of non-Hodgkin lymphoma and serum organochlorine concentrations among neighbors of a municipal solid waste incinerator. Environment International 2011;37:449-53.
77. Bertrand KA, Spiegelman D, Aster JC, et al. Plasma Organochlorine Levels and Risk of Non-Hodgkin Lymphoma in a Cohort of Men. Epidemiology 2010;21:172-80
78. Spinelli JJ, Ng CH, Weber J-P, et al. Organochlorines and risk of non-Hodgkin lymphoma. International Journal of Cancer 2007;121:2767-75.
79. De Roos AJ, Hartge P, Lubin JH, et al. Persistent Organochlorine Chemicals in Plasma and Risk of Non-Hodgkin's Lymphoma. Cancer Research 2005;65:11214-26.
80. Rothman N, Cantor KP, Blair A, et al. A nested case-control study of non-Hodgkin lymphoma and serum organochlorine residues. Lancet 1997;350:240-4.
81. Brauner EV, Sorensen M, Gaudreau E, et al. A prospective study of organochlorines in adipose tissue and risk of nonHodgkin lymphoma. Environ Health Perspect 2012;120:105-11.
82. Laden F, Bertrand KA, Altshul L, Aster JC, Korrick SA, Sagiv SK. Plasma Organochlorine Levels and Risk of Non–Hodgkin Lymphoma in the Nurses' Health Study. Cancer Epidemiology Biomarkers & Prevention 2010;19:1381-4.
83. Hardell K, Carlberg M, Hardell L, et al. Concentrations of organohalogen compounds and titres of antibodies to Epstein-Barr virus antigens and the risk for non-Hodgkin lymphoma. Oncology reports 2009;21:1567-76.
84. Cantor KP, Strickland PT, Brock JW, et al. Risk of non-Hodgkin's lymphoma and prediagnostic serum organochlorines: beta-hexachlorocyclohexane, chlordane/heptachlor-related compounds, dieldrin, and hexachlorobenzene. Environ Health Perspect 2003;111:179-83.
85. Cocco P, Brennan P, Ibba A, et al. Plasma polychlorobiphenyl and organochlorine pesticide level and risk of major lymphoma subtypes. Occupational and Environmental Medicine 2008;65:132-40.
86. Quintana PJ, Delfino RJ, Korrick S, et al. Adipose tissue levels of organochlorine pesticides and polychlorinated biphenyls and risk of non-Hodgkin's lymphoma. Environ Health Perspect 2004;112:854-61.
87. Berndt SI, Skibola CF, Joseph V, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat Genet 2013;45:868-76.
88. Karami S, Bassig B, Stewart PA, et al. Occupational trichloroethylene exposure and risk of lymphatic and haematopoietic cancers: a meta-analysis. Occupational and Environmental Medicine 2013;70:591-9.
89. Kane EV, Newton R. Benzene and the risk of non-Hodgkin lymphoma: A review and meta-analysis of the literature. Cancer Epidemiology 2010;34:7-12.
90. Cocco P, t'Mannetje A, Fadda D, et al. Occupational exposure to solvents and risk of lymphoma subtypes: results from the Epilymph case–control study. Occupational and Environmental Medicine 2010;67:341-7.

91. Castillo JJ, Dalia S, Pascual SK. Association between red blood cell transfusions and development of non-Hodgkin lymphoma: a meta-analysis of observational studies. Blood 2010;116:2897-907.
92. Kane EV, Roman E, Becker N, et al. Menstrual and reproductive factors, and hormonal contraception use: associations with non-Hodgkin lymphoma in a pooled analysis of InterLymph case–control studies. Annals of Oncology 2012;23:2362-74.
93. Kane EV, Bernstein L, Bracci PM, et al. Postmenopausal hormone therapy and non-Hodgkin lymphoma: a pooled analysis of InterLymph case-control studies. Ann Oncol 2013;24:433-41.
94. Costas L, Casabonne D, Benavente Y, et al. Reproductive factors and lymphoid neoplasms in Europe: findings from the EpiLymph case-control study. Cancer Causes Control 2012;23:195-206.
95. Nieters A, Rohrmann S, Becker N, et al. Smoking and Lymphoma Risk in the European Prospective Investigation into Cancer and Nutrition. American Journal of Epidemiology 2008;167:1081-9.
96. Zhang Y, Sanjose SD, Bracci PM, et al. Personal Use of Hair Dye and the Risk of Certain Subtypes of Non-Hodgkin Lymphoma. American Journal of Epidemiology 2008;167:1321-31.
97. Chang ET, Canchola AJ, Cockburn M, et al. Adulthood residential ultraviolet radiation, sun sensitivity, dietary vitamin D, and risk of lymphoid malignancies in the California Teachers Study. Blood 2011;118:1591-9.
98. Karipidis KK, Benke G, Sim MR, et al. Non-Hodgkin lymphoma and occupational radiation exposure assessed using local data. Occupational Medicine 2009;59:437-9.
99. Skibola CF. Obesity, diet and risk of non-Hodgkin lymphoma. Cancer Epidemiol Biomarkers Prev 2007;16:392-5.
100. Larsson SC, Wolk A. Body mass index and risk of non-Hodgkin's and Hodgkin's lymphoma: a meta-analysis of prospective studies. Eur J Cancer 2011;47:2422-30.
101. Britton JA, Khan AE, Rohrmann S, et al. Anthropometric characteristics and non-Hodgkin's lymphoma and multiple myeloma risk in the European Prospective Investigation into Cancer and Nutrition (EPIC). Haematologica 2008;93:1666-77.
102. Cross AJ, Lim U. The role of dietary factors in the epidemiology of non-Hodgkin's lymphoma. Leuk Lymphoma 2006;47:2477-87.
103. Cross AJ, Ward MH, Schenk M, et al. Meat and meat-mutagen intake and risk of non-Hodgkin lymphoma: results from a NCI-SEER case–control study. Carcinogenesis 2006;27:293-7.
104. Thompson CA, Habermann TM, Wang AH, et al. Antioxidant intake from fruits, vegetables and other sources and risk of non-Hodgkin's lymphoma: the Iowa Women's Health Study. International Journal of Cancer 2010;126:992-1003.
105. Tramacere I, Pelucchi C, Bonifazi M, et al. Alcohol drinking and non-Hodgkin lymphoma risk: a systematic review and a meta-analysis. Ann Oncol 2012;23:2791-8.
106. Wang SS, Slager SL, Brennan P, et al. Family history of hematopoietic malignancies and risk of non-Hodgkin lymphoma (NHL): a pooled analysis of 10 211 cases and 11 905 controls from the International Lymphoma Epidemiology Consortium (InterLymph). Blood 2007;109:3479-88.
107. Linet MS, Pottern LM. Familial aggregation of hematopoietic malignancies and risk of non-Hodgkin's lymphoma. Cancer Res 1992;52:5468s-73s.
108. Cerhan JR, Novak AJ, Fredericksen ZS, et al. Risk of non-Hodgkin lymphoma in association with germline variation in complement genes. Br J Haematol 2009;145:614-23.
109. Staudt LM, Dave S. The biology of human lymphoid malignancies revealed by gene expression profiling. Adv Immunol 2005;87:163-208.
110. Skibola CF, Bracci PM, Halperin E, et al. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. Nat Genet 2009;41:873-5.
111. Di Bernardo MC, Crowther-Swanepoel D, Broderick P, et al. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. Nat Genet 2008;40:1204-10.
112. Nieters A, Beckmann L, Deeg E, Becker N. Gene polymorphisms in Toll-like receptors, interleukin-10, and interleukin-10 receptor alpha and lymphoma risk. Genes Immun 2006;7:615-24.
113. Cerhan JR, Ansell SM, Fredericksen ZS, et al. Genetic variation in 1253 immune and inflammation genes and risk of non-Hodgkin lymphoma. Blood 2007;110:4455-63.

114. Skibola CF, Bracci PM, Nieters A, et al. Tumor Necrosis Factor (TNF) and Lymphotoxin-α (LTA) Polymorphisms and Risk of Non-Hodgkin Lymphoma in the InterLymph Consortium. American Journal of Epidemiology 2010;171:267-76.
115. Rothman N, Skibola CF, Wang SS, et al. Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. The Lancet Oncology 2006;7:27-38.
116. Wang SS, Menashe I, Cerhan JR, et al. Variations in chromosomes 9 and 6p21.3 with risk of non-Hodgkin lymphoma. Cancer Epidemiol Biomarkers Prev 2011;20:42-9.
117. Conde L, Halperin E, Akers NK, et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat Genet 2010;42:661-4.
118. Smedby KE, Foo JN, Skibola CF, et al. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. PLoS Genet 2011;7:e1001378.
119. Slager SL, Rabe KG, Achenbach SJ, et al. Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. Blood 2011;117:1911-6.
120. Crowther-Swanepoel D, Broderick P, Di Bernardo MC, et al. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. Nat Genet 2010;42:132-6.
121. Kim HN, Lee I-K, Kim Y-K, et al. Association between folate-metabolizing pathway polymorphism and non-Hodgkin lymphoma. British Journal of Haematology 2008;140:287-94.
122. Lightfoot T, Skibola C, Smith A, et al. Polymorphisms in the oxidative stress genes, superoxide dismutase, glutathione peroxidase and catalase and risk of non-Hodgkin's lymphoma. Haematologica 2006;91:1222-7.
123. Shen M, Menashe I, Morton LM, et al. Polymorphisms in DNA repair genes and risk of non-Hodgkin lymphoma in a pooled analysis of three studies. British Journal of Haematology 2010;151:239-44.
124. Gra OA, Glotov AS, Nikitin EA, et al. Polymorphisms in xenobiotic-metabolizing genes and the risk of chronic lymphocytic leukemia and non-Hodgkin's lymphoma in adult Russian patients. Am J Hematol 2008;83:279-87.
125. Nieters A, Conde L, Slager SL, et al. PRRC2A and BCL2L11 gene variants influence risk of non-Hodgkin lymphoma: results from the InterLymph consortium. Blood 2012;120:4645-8.
126. Gibson T, Smedby K, Skibola C, et al. Smoking, variation in N-acetyltransferase 1 (NAT1) and 2 (NAT2), and risk of non-Hodgkin lymphoma: a pooled analysis within the InterLymph consortium. Cancer causes & control: 2013;24:125-34.
127. Morin RD, Mendez-Lago M, Mungall AJ, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature 2011;476:298-303.
128. Doménech E, Gómez-López G, Gzlez-Peña D, et al. New Mutations in Chronic Lymphocytic Leukemia Identified by Target Enrichment and Deep Sequencing. PLoS One 2012;7:e38158.
129. Morin RD, Mungall K, Pleasance E, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood 2013;122:1256-65.
130. Scott DW, Mungall KL, Ben-Neriah S, et al. TBL1XR1/TP63: a novel recurrent gene fusion in B-cell non-Hodgkin lymphoma. Blood 2012;119:4949-52.
131. Morin RD, Gascoyne RD. Newly Identified Mechanisms in B-Cell Non-Hodgkin Lymphomas Uncovered by Next-Generation Sequencing. Seminars in hematology 2013;50:303-13.
132. Baris D, Kwak LW, Rothman N, et al. Blood levels of organochlorines before and after chemotherapy among non-Hodgkin's lymphoma patients. Cancer Epidemiol Biomarkers Prev 2000;9:193-7.
133. Leng L, Chen X, Li CP, Luo XY, Tang NJ. 2,3,7,8-Tetrachlorodibezo-p-dioxin exposure and prostate cancer: a meta-analysis of cohort studies. Public Health 2014;128:207-13.
134. Anothaisintawee T, Wiratkapun C, Lerdsitthichai P, et al. Risk Factors of Breast Cancer: A Systematic Review and Meta-Analysis. Asia-Pacific Journal of Public Health 2013;25:368-87.
135. Lee P, Forey B, Coombs K. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. BMC Cancer 2012;12:385.
136. Johnson C, Wei C, Ensor J, et al. Meta-analyses of colorectal cancer risk factors. Cancer causes & control: 2013;24:1207-22.

137. Khalade A, Jaakkola M, Pukkala E, Jaakkola J. Exposure to benzene at work and the risk of leukemia: a systematic review and meta-analysis. Environmental Health 2010;9:31.
138. Saitta P, Cook CE, Messina JL, et al. Is there a true concern regarding the use of hair dye and malignancy development?: a review of the epidemiological evidence relating personal hair dye use to the risk of malignancy. J Clin Aesthet Dermatol 2013;6:39-46.
139. Vijai J, Kirchhoff T, Schrader KA, et al. Susceptibility loci associated with specific and shared subtypes of lymphoid malignancies. PLoS Genet 2013;9:e1003220.
140. Aronson JK. Biomarkers and surrogate endpoints. Br J Clin Pharmacol 2005;59:491-4.
141. Links JM, Kensler TW, Groopman JD. Biomarkers and mechanistic approaches in environmental epidemiology. Annu Rev Public Health 1995;16:83-103.
142. Perera F, Brenner D, Jeffrey A, et al. DNA adducts and related biomarkers in populations exposed to environmental carcinogens. Environ Health Perspect 1992;98:133-7.
143. Gaikwad NW, Yang L, Weisenburger DD, et al. Urinary biomarkers suggest that estrogen-DNA adducts may play a role in the aetiology of non-Hodgkin lymphoma. Biomarkers 2009;14:502-12.
144. Janz S, Potter M, Rabkin CS. Lymphoma- and leukemia-associated chromosomal translocations in healthy individuals. Genes, Chromosomes and Cancer 2003;36:211-23.
145. Vega F, Medeiros LJ. Chromosomal Translocations Involved in Non-Hodgkin Lymphomas. Archives of Pathology & Laboratory Medicine 2003;127:1148-60.
146. Chiu BC-H, Lan Q, Dave BJ, Blair A, Zahm SH, Weisenburger DD. The Utility of t(14;18) in Understanding Risk Factors for Non-Hodgkin Lymphoma. JNCI Monographs 2008;2008:69-73.
147. Roulland S, Lebailly P, Lecluse Y, Briand M, Pottier D, Gauduchon P. Characterization of the t(14;18) BCL2-IGH Translocation in Farmers Occupationally Exposed to Pesticides. Cancer Research 2004;64:2264-9.
148. Schüler F, Hirt C, Dolken G. Chromosomal translocation t(14;18) in healthy individuals. Semin Cancer Biol 2003;13:203-9.
149. Paltiel O, Zelenetz A, Sverdlin I, Gordon L, Ben-Yehuda D. Translocation t(14;18) in healthy individuals: preliminary study of its association with family history and agricultural exposure. Ann Oncol 2000;11 Suppl 1:75-80.
150. Figgs LW, Holland NT, Rothmann N, et al. Increased lymphocyte replicative index following 2,4-dichlorophenoxyacetic acid herbicide exposure. Cancer causes & control: 2000;11:373-80.
151. Burns CJ, Beard KK, Cartmill JB. Mortality in chemical workers potentially exposed to 2,4-dichlorophenoxyacetic acid (2,4-D) 1945-94: an update. Occup Environ Med 2001;58:24-30.
152. Zeljezic D, Vrdoljak AL, Radic B, et al. Comparative evaluation of acetylcholinesterase status and genome damage in blood cells of industrial workers exposed to carbofuran. Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association 2007;45:2488-98.
153. Lan Q, Zhang L, Tang X, et al. Occupational exposure to trichloroethylene is associated with a decline in lymphocyte subsets and soluble CD27 and CD30 markers. Carcinogenesis 2010;31:1592-6.
154. Purdue MP, Hofmann JN, Kemp TJ, et al. A prospective study of 67 serum immune and inflammation markers and risk of non-Hodgkin lymphoma. Blood 2013;122:951-7.
155. Vermeulen R, Hosnijeh FS, Portengen L, et al. Circulating Soluble CD30 and Future Risk of Lymphoma; Evidence from Two Prospective Studies in the General Population. Cancer Epidemiology Biomarkers & Prevention 2011;20:1925-7.
156. Purdue MP, Lan Q, Bagni R, et al. Prediagnostic Serum Levels of Cytokines and Other Immune Markers and Risk of Non-Hodgkin Lymphoma. Cancer Research 2011;71:4898-907.
157. Taylor KH, Briley A, Wang Z, Cheng J, Shi H, Caldwell CW. Aberrant epigenetic gene regulation in lymphoid malignancies. Semin Hematol 2013;50:38-47.
158. Zekri AR, Hassan ZK, Bahnassy AA, et al. Gene expression profiling of non-hodgkin lymphomas. Asian Pac J Cancer Prev 2013;14:4393-8.
159. Ottman R. Gene–Environment Interaction: Definitions and Study Design. Preventive Medicine 1996;25:764-70.
160. Robinette SL, Holmes E, Nicholson JK, Dumas ME. Genetic determinants of metabolism in health and disease: from biochemical genetics to genome-wide associations. Genome Med 2012;4:30.

161. Wheelock AM, Wheelock CE. Trials and tribulations of 'omics data analysis: assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine. Mol Biosyst 2013;9:2589-96.
162. Vineis P, Kriebel D. Causal models in epidemiology: past inheritance and genetic future. Environ Health 2006;5:21.
163. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics 2010;11:484-98.
164. Papathomas M, Molitor J, Hoggart C, Hastie D, Richardson S. Exploring Data From Genetic Association Studies Using Bayesian Variable Selection and the Dirichlet Process: Application to Searching for Gene × Gene Patterns. Genetic Epidemiology 2012;36:663-74.
165. Papathomas M, Molitor J, Richardson S, Riboli E, Vineis P. Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. Environ Health Perspect 2011;119:84-91.
166. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-11.
167. Grulich AE, Vajdic CM. The epidemiology of non-Hodgkin lymphoma. Pathology 2005;37:409-19.
168. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. Am J Epidemiol 2009;169:227-30; discussion 34-5.
169. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet 2005;6:287-98.
170. Rava M, Ahmed I, Demenais F, Sanchez M, Tubert-Bitter P, Nadif R. Selection of genes for gene-environment interaction studies: a candidate pathway-based strategy using asthma as an example. Environmental Health 2013;12:56.
171. Wang SS, Nieters A. Unraveling the interactions between environmental factors and genetic polymorphisms in non-Hodgkin lymphoma risk. Expert review of anticancer therapy 2010;10:403-13.
172. Garrod A. The Incidence of Alkaptonuria: A Study in Chemical Individuality. The Lancet 1902;160:1616-20.
173. Chatterjee N, Hartge P, Cerhan JR, et al. Risk of Non-Hodgkin's Lymphoma and Family History of Lymphatic, Hematologic, and Other Cancers. Cancer Epidemiology Biomarkers & Prevention 2004;13:1415-21.
174. Morton L, Holford T, Leaderer B, et al. Alcohol use and risk of non-Hodgkin's lymphoma among Connecticut women (United States). Cancer causes & control: 2003;14:687-94.
175. Smedby KE, Eloranta S, Duvefelt K, et al. Vitamin D Receptor Genotypes, Ultraviolet Radiation Exposure, and Risk of Non-Hodgkin Lymphoma. American Journal of Epidemiology 2011;173:48-54.
176. Fernberg P, Chang E, Duvefelt K, et al. Genetic variation in chromosomal translocation breakpoint and immune function genes and risk of non-Hodgkin lymphoma. Cancer causes & control: 2010;21:759-69.
177. Morton LM, Bernstein L, Wang SS, et al. Hair dye use, genetic variation in N-acetyltransferase 1 (NAT1) and 2 (NAT2), and risk of non-Hodgkin lymphoma. Carcinogenesis 2007;28:1759-64.
178. Koshiol J, Lam TK, Gridley G, Check D, Brown LM, Landgren O. Racial Differences in Chronic Immune Stimulatory Conditions and Risk of Non-Hodgkin's Lymphoma in Veterans From the United States. Journal of Clinical Oncology 2011;29:378-85.
179. Qiao Y, Zhou Y, Wu C, et al. Risk of genome-wide association study-identified genetic variants for non-Hodgkin lymphoma in a Chinese population. Carcinogenesis 2013;34:1516-9.
180. Hall MA, Ritchie MD, Peissig P, et al. Environment-wide association study (EWAS) for type 2 Diabetes in the Marshfield Personalised Medicine Research Project Biobank. In: Biocomputing 2014:200-11.
181. Armstrong BK, Kricker A. Sun Exposure and Non-Hodgkin Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2007;16:396-400.
182. Han X, Zheng T, Lan Q, et al. Genetic Polymorphisms in Nitric Oxide Synthase Genes Modify the Relationship between Vegetable and Fruit Intake and Risk of Non-Hodgkin Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2009;18:1429-38.

183. Li Y, Zheng T, Kilfoy BA, et al. Genetic polymorphisms in cytochrome P450s, GSTs, NATs, alcohol consumption and risk of non-Hodgkin lymphoma. Am J Hematol 2010;85:213-5.
184. Hoeft B, Becker N, Deeg E, Beckmann L, Nieters A. Joint effect between regular use of non-steroidal anti-inflammatory drugs, variants in inflammatory genes and risk of lymphoma. Cancer causes & control: 2008;19:163-73.
185. Holly EA, Bracci PM. Population-based Study of Non-Hodgkin Lymphoma, Histology, and Medical History among Human Immunodeficiency Virus-negative Participants in San Francisco. American Journal of Epidemiology 2003;158:316-27.
186. Kane EV, Painter D, Roman E, Allan J, Law G, Lightfoot T. Melanocortin 1 receptor (MC1R), pigmentary characteristics and sun exposure: Findings from a case–control study of diffuse large B-cell and follicular lymphoma. Cancer Epidemiology 2010;34:136-41.
187. Morton LM, Schenk M, Hein DW, et al. Genetic variation in N-acetyltransferase 1 (NAT1) and 2 (NAT2) and risk of non-Hodgkin lymphoma. Pharmacogenet Genomics 2006;16:537-45.
188. Gathany AH, Hartge P, Davis S, et al. Relationship between interferon regulatory factor 4 genetic polymorphisms, measures of sun sensitivity and risk for non-Hodgkin lymphoma. Cancer causes & control: 2009;20:1291-302.
189. Wang SS, Cozen W, Cerhan JR, et al. Immune mechanisms in non-Hodgkin lymphoma: joint effects of the TNF G308A and IL10 T3575A polymorphisms with non-Hodgkin lymphoma risk factors. Cancer Res 2007;67:5042-54.
190. Purdue M, Hartge P, Davis S, et al. Sun exposure, vitamin D receptor gene polymorphisms and risk of non-Hodgkin lymphoma. Cancer causes & control: 2007;18:989-99.
191. Chen Y, Zheng T, Lan Q, et al. Cytokine polymorphisms in Th1/Th2 pathway genes, body mass index, and risk of non-Hodgkin lymphoma. Blood 2011;117:585-90.
192. Colt JS, Rothman N, Severson RK, et al. Organochlorine exposure, immune gene variation, and risk of non-Hodgkin lymphoma. Blood 2009;113:1899-905.
193. Ng CH, Janoo-Gilani R, Sipahimalani P, et al. Interaction between organochlorines and the AHR gene, and risk of non-Hodgkin lymphoma. Cancer Causes Control 2010;21:11-22.
194. De Roos AJ, Gold LS, Wang S, et al. Metabolic Gene Variants and Risk of Non-Hodgkin's Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2006;15:1647-53.
195. Kelemen L, Wang S, Lim U, et al. Vegetables- and antioxidant-related nutrients, genetic susceptibility, and non-Hodgkin lymphoma risk. Cancer causes & control: 2008;19:491-503.
196. Kilfoy BA, Zheng T, Lan Q, et al. Genetic polymorphisms in glutathione S-transferases and cytochrome P450s, tobacco smoking, and risk of non-Hodgkin lymphoma. American Journal of Hematology 2009;84:279-82.
197. Kilfoy B, Zheng T, Lan Q, et al. Genetic variation in N-acetyltransferases 1 and 2, cigarette smoking, and risk of non-Hodgkin lymphoma. Cancer causes & control: 2010;21:127-33.
198. Zhang Y, Hughes KJ, Zahm SH, et al. Genetic variations in xenobiotic metabolic pathway genes, personal hair dye use, and risk of non-Hodgkin lymphoma. Am J Epidemiol 2009;170:1222-30.
199. Kerridge I, Lincz L, Scorgie F, Hickey D, Granter N, Spencer A. Association between xenobiotic gene polymorphisms and non-Hodgkin's lymphoma risk. Br J Haematol 2002;118:477-81.
200. Jiao J, Zheng T, Lan Q, et al. Occupational solvent exposure, genetic variation of DNA repair genes, and the risk of non-Hodgkin's lymphoma. European Journal of Cancer Prevention 2012;21:580-4
201. Skibola CF, Holly EA, Forrest MS, et al. Body Mass Index, Leptin and Leptin Receptor Polymorphisms, and Non-Hodgkin Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2004;13:779-86.
202. Willett EV, Skibola CF, Adamson P, et al. Non-Hodgkin's lymphoma, obesity and energy homeostasis polymorphisms. Br J Cancer 2005;93:811-6.
203. Barry KH, Zhang Y, Lan Q, et al. Genetic Variation in Metabolic Genes, Occupational Solvent Exposure, and Risk of Non-Hodgkin Lymphoma. American Journal of Epidemiology 2011;173:404-13.
204. Lim U, Wang SS, Hartge P, et al. Gene-nutrient interactions among determinants of folate and one-carbon metabolism on the risk of non-Hodgkin lymphoma: NCI-SEER Case-Control Study. Blood 2007;109:3050-9.

205. García-Closas M, Lubin JH. COMMENTARY: Power and Sample Size Calculations in Case-Control Studies of Gene-Environment Interactions: Comments on Different Approaches. American Journal of Epidemiology 1999;149:689-92.
206. Hein R, Beckmann L, Chang-Claude J. Sample size requirements for indirect association studies of gene–environment interactions (G × E). Genetic Epidemiology 2008;32:235-45.
207. Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. Am J Epidemiol 2012;175:177-90.
208. Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. The choice of a genetic model in the meta-analysis of molecular association studies. International Journal of Epidemiology 2005;34:1319-28.
209. Gilthorpe MC, D. Statistical Interactions and Gene-Environment Joint Effects. In: DCG Y-KT, ed. Modern Methods for Epidemiology. Netherlands: Springer; 2012:219-331.
210. García-Closas M, Malats N, Silverman D, et al. NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. The Lancet 2005;366:649-59.
211. Clayton DG. Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes. PLoS Genet 2009;5:e1000540.
212. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. International Journal of Epidemiology 1997;26:S6.
213. IARC. EPIC Project. 2010. (http://epic.iarc.fr/)
214. Riboli E, Hunt K, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutrition 2002;5:1113-24.
215. Hermann S, Rohrmann S, Linseisen J, et al. Level of education and the risk of lymphoma in the European prospective investigation into cancer and nutrition. J Cancer Res Clin Oncol 2010;136:71-7.
216. Kelly RS, Lundh T, Porta M, et al. Blood Erythrocyte Concentrations of Cadmium and Lead and the Risk of B-Cell Non-Hodgkin's Lymphoma and Multiple Myeloma: A Nested Case-Control Study. PLoS One 2013;8:e81892.
217. Palli D, Berrino F, Vineis P, et al. A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. Tumori 2003;89:586-93.
218. Hallmans G, Agren A, Johansson G, et al. Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. Scand J Public Health Suppl 2003;61:18-24.
219. Barany E, Bergdahl IA, Schutz A, Skerfving S, Oskarsson A. Inductively Coupled Plasma Mass Spectrometry for Direct Multi-element Analysis of Diluted Human Blood and Serum. Journal of Analytical Atomic Spectrometry 1997;12:1005-9.
220. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Analytical Chemistry 2006;78:779-87.
221. Neal R. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics 2000;9:249-65.
222. Dahl D. Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In: Do K-A MP, Vannucci M, ed. Bayesian Inference for Gene Expression and Proteomics: Cambridge University Press; 2006:201-18.
223. Liverani S HD, Papathomas M, Richardson S. PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. In. R package version 3.0.21 ed; 2013.
224. Chadeau-Hyam M, Athersuch TJ, Keun HC, et al. Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. Biomarkers 2011;16:83-8.
225. Wang P, Dai M, Xuan W, et al. SNP Function Portal: a web database for exploring the function implication of SNP alleles. Bioinformatics 2006;22:e523-e9.

226. Wang J, Ronaghi M, Chong SS, Lee CGL. pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. Human Mutation 2011;32:19-24.
227. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols 2008;4:44-57.
228. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumor MARKer prognostic studies (REMARK). Nat Clin Pract Urol 2005;2:416-22.
229. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309-14.
230. Xia J, Broadhurst D, Wilson M, Wishart D. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics 2013;9:280-99.
231. Hill AB. The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine 1965;58:295-300.
232. Ioannidis JP, Boffetta P, Little J, et al. Assessment of cumulative evidence on genetic associations: interim guidelines. International Journal of Epidemiology 2008;37:120-32.
233. Wilson JM, Jungner YG. Principles and practice of mass screening for disease. Geneva, Switzerland. World Health Organisation. 1968
234. Soucek P, Šarmanová J, Kristensen V, Apltauerová M, Gut I. Genetic polymorphisms of biotransformation enzymes in patients with Hodgkin's and non-Hodgkin's lymphomas. Int Arch Occup Environ Health 2002;75:86-92.
235. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
236. Furrow RE, Christiansen FB, Feldman MW. Environment-sensitive Epigenetics and the Heritability of Complex Diseases. Genetics 2011.
237. Goldstein DB. Common genetic variation and human traits. N Engl J Med 2009;360:1696-8.
238. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet 2010;86:6-22.
239. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 2012;8:e1002375.
240. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics;81:559-75.
241. Raval A, Tanner SM, Byrd JC, et al. Downregulation of Death-Associated Protein Kinase 1 (DAPK1) in Chronic Lymphocytic Leukemia. Cell;129:879-90.
242. Turner S, Armstrong LL, Bradford Y, et al. Quality Control Procedures for Genome-Wide Association Studies. In: Current Protocols in Human Genetics: John Wiley & Sons, Inc.; 2001.
243. Baranzini SE, Galwey NW, Wang J, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Human Molecular Genetics 2009;18:2078-90.
244. Braun R, Buetow K. Pathways of Distinction Analysis: A New Technique for Multi–SNP Analysis of GWAS Data. PLoS Genet 2011;7:e1002101.
245. Thomas D. Methods for Investigating Gene-Environment Interactions in Candidate Pathway and Genome-Wide Association Studies. Annual Review of Public Health 2010;31:21-36.
246. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. Bioinformatics 2003;19:376-82.
247. Kramer S, Hikel SM, Adams K, Hinds D, Moon K. Current status of the epidemiologic evidence linking polychlorinated biphenyls and non-hodgkin lymphoma, and the role of immune dysregulation. Environ Health Perspect 2012;120:1067-75.
248. Ramis R, Diggle P, Boldo E, Garcia-Perez J, Fernandez-Navarro P, Lopez-Abente G. Analysis of matched geographical areas to study potential links between environmental exposure to oil refineries and non-Hodgkin lymphoma mortality in Spain. International Journal of Health Geographics 2012;11:4.
249. Engel LS, Laden F, Andersen A, et al. Polychlorinated Biphenyl Levels in Peripheral Blood and Non-Hodgkin's Lymphoma: A Report from Three Cohorts. Cancer Research 2007;67:5545-52.

250. Golden R, Kimbrough R. Weight of Evidence Evaluation of Potential Human Cancer Risks from Exposure to Polychlorinated Biphenyls: An Update Based on Studies Published Since 2003. Critical Reviews in Toxicology 2009;39:299-331.
251. Kimbrough RD, Doemland ML, Mandel JS. A Mortality Update of Male and Female Capacitor Workers Exposed to Polychlorinated Biphenyls. Journal of Occupational and Environmental Medicine 2003;45:271-82.
252. Lauby-Secretan B, Loomis D, Grosse Y, et al. Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls. The Lancet Oncology 2013;14:287-8.
253. Faroon OM, Keith S, Jones D, De Rosa C. Carcinogenic effects of polychlorinated biphenyls. Toxicology and Industrial Health 2001;17:41-62.
254. Consonni D, Pesatori AC, Zocchetti C, et al. Mortality in a Population Exposed to Dioxin after the Seveso, Italy, Accident in 1976: 25 Years of Follow-Up. American Journal of Epidemiology 2008;167:847-58.
255. Strauss HS, Heiger-Bernays W. Methodological Limitations May Prevent the Observation of Non-Hodgkin's Lymphoma in Bioassays of Polychlorinated Biphenyls. Toxicologic Pathology 2012;40:995-1003.
256. Cocco P, Kazerouni N, Zahm SH. Cancer mortality and environmental exposure to DDE in the United States. Environ Health Perspect 2000;108:1-4.
257. Petreas M, She J, Brown FR, et al. High body burdens of 2,2',4,4'-tetrabromodiphenyl ether (BDE-47) in California women. Environ Health Perspect 2003;111:1175-9.
258. García-Lestón J, Méndez J, Pásaro E, Laffon B. Genotoxic effects of lead: An updated review. Environment International 2010;36:623-36.
259. García-Lestón J, Roma-Torres J, Vilares M, et al. Biomonitoring of a population of Portuguese workers exposed to lead. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 2011;721:81-8.
260. Iavicoli I, Fontana L, Bergamaschi A. The Effects of Metals as Endocrine Disruptors. Journal of Toxicology and Environmental Health, Part B 2009;12:206-23.
261. Jiao W, Chen W, Chang AC, Page AL. Environmental risks of trace elements associated with long-term phosphate fertilizers applications: A review. Environmental Pollution 2012;168:44-53.
262. Guzel S, Kiziler L, Aydemir B, et al. Association of Pb, Cd, and Se Concentrations and Oxidative Damage-Related Markers in Different Grades of Prostate Carcinoma. Biol Trace Elem Res 2012;145:23-32.
263. Jomova K, Valko M. Advances in metal-induced oxidative stress and human disease. Toxicology 2011;283:65-87.
264. Menke A, Muntner P, Silbergeld EK, Platz EA, Guallar E. Cadmium levels in urine and mortality among U.S. adults. Environ Health Perspect 2009;117:190-6.
265. Satarug S, Garrett SH, Sens MA, Sens DA. Cadmium, environmental exposure, and health outcomes. Environ Health Perspect 2010;118:182-90.
266. Fewtrell L KR, Prüss-Üstün A. Lead: assessing the environmental burden of disease at national and local level. Geneva; 2003.
267. Järup L, Åkesson A. Current status of cadmium as an environmental health problem. Toxicology and Applied Pharmacology 2009;238:201-8.
268. IARC. Beryllium, cadmium, mercury, and exposures in the glass manufacturing industry. . In: International Agency for Research on Cancer Monographs on the Evaluation of Carcinogenic Risks to Humans. Lyon: IARC Scientific Publications; 1993:119-237.
269. IARC. Inorganic and organic lead compounds. Lyon; IARC Scientific Publications; 2006.
270. Kayama F, Yoshida T, Elwell MR, Luster MI. Cadmium-Induced Renal Damage and Proinflammatory Cytokines: Possible Role of IL-6 in Tubular Epithelial Cell Regeneration. Toxicology and Applied Pharmacology 1995;134:26-34.
271. Mishra KP. Lead exposure and its impact on immune system: A review. Toxicology in Vitro 2009;23:969-72.
272. Demir C, Demir H, Esen R, Sehitogullari A, Atmaca M, Alay M. Altered serum levels of elements in acute leukemia cases in Turkey. Asian Pac J Cancer Prev 2011;12:3471-4.
273. Waalkes MP, Rehm S, Sass B, Ward JM. Induction of tumours of the haematopoietic system by cadmium in rats. IARC Sci Publ 1992:401-4.

274. Warner J, Osuch JR, Karmaus W, et al. Common classification schemes for PCB congeners and the gene expression of CYP17, CYP19, ESR1 and ESR2. Science of The Total Environment 2012;414:81-9.
275. Engel LS, Lan Q, Rothman N. Polychlorinated Biphenyls and Non-Hodgkin Lymphoma. Cancer Epidemiology Biomarkers & Prevention 2007;16:373-6.
276. Hardell L, Lindström G, van Bavel B, et al. Adipose Tissue Concentrations of Dioxins and Dibenzofurans, Titers of Antibodies to Epstein–Barr Virus Early Antigen and the Risk for Non-Hodgkin Lymphoma. Environmental Research 2001;87:99-107.
277. Grover P, Rekhadevi PV, Danadevi K, Vuyyuri SB, Mahboob M, Rahman MF. Genotoxicity evaluation in workers occupationally exposed to lead. International Journal of Hygiene and Environmental Health 2010;213:99-106.
278. Palus J, Rydzynski K, Dziubaltowska E, Wyszynska K, Natarajan AT, Nilsson R. Genotoxic effects of occupational exposure to lead and cadmium. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 2003;540:19-28.
279. Navas-Acien A, Selvin E, Sharrett AR, Calderon-Aranda E, Silbergeld E, Guallar E. Lead, Cadmium, Smoking, and Increased Risk of Peripheral Arterial Disease. Circulation 2004;109:3196-201.
280. Fischbein A, Tsang P, Luo JC, Bekesi JG. The immune system as target for subclinical lead related toxicity. British journal of industrial medicine 1993;50:185-6.
281. Knox EG, Gilman EA. Hazard proximities of childhood cancers in Great Britain from 1953-80. Journal of epidemiology and community health 1997;51:151-9.
282. Adams SV, Passarelli MN, Newcomb PA. Cadmium exposure and cancer mortality in the Third National Health and Nutrition Examination Survey cohort. Occupational and Environmental Medicine 2012;69:153-6.
283. Fernandez MF, Kiviranta H, Molina-Molina JM, et al. Polychlorinated biphenyls (PCBs) and hydroxy-PCBs in adipose tissue of women in Southeast Spain. Chemosphere 2008;71:1196-205.
284. Weaver VM, Kim N-S, Jaar BG, et al. Associations of low-level urine cadmium with kidney function in lead workers. Occupational and Environmental Medicine 2011;68:250-6.
285. González-Estecha M, Trasobares EM, Tajima K, et al. Trace elements in bipolar disorder. Journal of Trace Elements in Medicine and Biology 2011;25, Supplement 1:S78-S83.
286. Akesson A, Bjellerup P, Lundh T, et al. Cadmium-induced effects on bone in a population-based study of women. Environ Health Perspect 2006;114:830-4.
287. Porta M, Pumarega J, Gasull M. Number of persistent organic pollutants detected at high concentrations in a general population. Environment International 2012;44:106-11.
288. Vahter M, Åkesson A, Lidén C, Ceccatelli S, Berglund M. Gender differences in the disposition and toxicity of metals. Environmental Research 2007;104:85-95.
289. Edwards TM, Myers JP. Environmental exposures and gene regulation in disease etiology. Environ Health Perspect 2007;115:1264-70.
290. Whitfield JB, Dy V, McQuilty R, et al. Genetic effects on toxic and essential elements in humans: arsenic, cadmium, copper, lead, mercury, selenium, and zinc in erythrocytes. Environ Health Perspect 2010;118:776-82.
291. Bergdahl IA, Grubb A, Schutz A, et al. Lead binding to delta-aminolevulinic acid dehydratase (ALAD) in human erythrocytes. Pharmacology & toxicology 1997;81:153-8.
292. Sarmanová J, Benesová K, Gut I, Nedelcheva-Kristensen V, Tynková L, Soucek P. Genetic polymorphisms of biotransformation enzymes in patients with Hodgkin's and non-Hodgkin's lymphomas. Human Molecular Genetics 2001;10:1265-73.
293. Kiyohara C, Ohno Y. Sex differences in lung cancer susceptibility: A review. Gender Medicine 2010;7:381-401.
294. Ramchandran K, Patel JD. Sex differences in susceptibility to carcinogens. Semin Oncol 2009;36:516-23.
295. Kirsch-Volders M, Bonassi S, Herceg Z, Hirvonen A, Moller L, Phillips DH. Gender-related differences in response to mutagens and carcinogens. Mutagenesis 2010;25:213-21.
296. Wolff MS, Camann D, Gammon M, Stellman SD. Proposed PCB congener groupings for epidemiological studies. Environ Health Perspect 1997;105:13-4.

297. Porta M. Role of Organochlorine Compounds in the Etiology of Pancreatic Cancer: A Proposal to Develop Methodological Standards. Epidemiology 2001;12:272-6.
298. Mendy A, Gasana J, Vieira ER. Urinary heavy metals and associated medical conditions in the US adult population. International Journal of Environmental Health Research 2011;22:105-18.
299. Gaskins AJ, Schisterman EF. The effect of lipid adjustment on the analysis of environmental contaminants and the outcome of human health risks. Methods in molecular biology 2009;580:371-81.
300. Rylander L, Björkdahl C-M, Axmon A, et al. Very high correlations between fresh weight and lipid-adjusted PCB-153 serum concentrations: Irrespective of fasting status, age, body mass index, gender, or exposure distributions. Chemosphere 2012;88:828-31.
301. Aronson KJ, Miller AB, Woolcott CG, et al. Breast Adipose Tissue Concentrations of Polychlorinated Biphenyls and Other Organochlorines and Breast Cancer Risk. Cancer Epidemiology Biomarkers & Prevention 2000;9:55-63.
302. Tuomisto J, Pekkanen J, Kiviranta H, et al. Dioxin cancer risk--example of hormesis? Dose-response : a publication of International Hormesis Society 2005;3:332-41.
303. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends in Biotechnology 2004;22:245-52.
304. Fiehn O. Metabolomics--the link between genotypes and phenotypes. Plant molecular biology 2002;48:155-71.
305. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 1999;29:1181-9.
306. Nicholson JK, Lindon JC. Systems biology: Metabonomics. Nature 2008;455:1054-6.
307. Liesenfeld DB, Habermann N, Owen RW, Scalbert A, Ulrich CM. Review of Mass Spectrometry–Based Metabolomics in Cancer Research. Cancer Epidemiology Biomarkers & Prevention 2013;22:2182-201.
308. Holmes E, Loo RL, Stamler J, et al. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 2008;453:396-400.
309. Denkert C, Bucher E, Hilvo M, et al. Metabolomics of human breast cancer: new approaches for tumor typing and biomarker discovery. Genome Medicine 2012;4:37.
310. Spratlin JL, Serkova NJ, Eckhardt SG. Clinical Applications of Metabolomics in Oncology: A Review. Clinical Cancer Research 2009;15:431-40.
311. Hilvo M, Denkert C, Lehtinen L, et al. Novel Theranostic Opportunities Offered by Characterization of Altered Membrane Lipid Metabolism in Breast Cancer Progression. Cancer Research 2011;71:3236-45.
312. Ritchie S, Ahiahonu P, Jayasinghe D, et al. Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. BMC Medicine 2010;8:13.
313. Davis V, Schiller D, Eurich D, Sawyer M. Urinary metabolomic signature of esophageal cancer and Barrett's esophagus. World Journal of Surgical Oncology 2012;10:271.
314. Zhang A, Sun H, Yan G, Wang P, Han Y, Wang X. Metabolomics in diagnosis and biomarker discovery of colorectal cancer. Cancer letters 2014;345:17-20.
315. Odunsi K, Wollman RM, Ambrosone CB, et al. Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. International Journal of Cancer 2005;113:782-8.
316. Nordström A, Lewensohn R. Metabolomics: Moving to the Clinic. J Neuroimmune Pharmacol 2010;5:4-17.
317. Ye G, Zhu B, Yao Z, et al. Analysis of Urinary Metabolic Signatures of Early Hepatocellular Carcinoma Recurrence after Surgical Removal Using Gas Chromatography–Mass Spectrometry. Journal of Proteome Research 2012;11:4361-72.
318. Asiago VM, Alvarado LZ, Shanaiah N, et al. Early Detection of Recurrent Breast Cancer Using Metabolite Profiling. Cancer Research 2010;70:8309-18.
319. Wishart DS, Jewison T, Guo AC, et al. HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Research 2013;41:D801-D7.
320. Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrometry Reviews 2007;26:51-78.

321. Nicholson G, Rantalainen M, Maher AD, et al. Human metabolic profiles are stably controlled by genetic and environmental variation. Molecular systems biology 2011;7.
322. Nicholson JK. Global systems biology, personalized medicine and molecular epidemiology. Molecular systems biology 2006;2:52.
323. Bictash M, Ebbels TM, Chan Q, et al. Opening up the "Black Box": Metabolic phenotyping and metabolome-wide association studies in epidemiology. Journal of clinical epidemiology 2010;63:970-9.
324. Hebels DG, Georgiadis P, Keun HC, et al. Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. Environ Health Perspect 2013;121:480-7.
325. Sangster TP, Wingate JE, Burton L, Teichert F, Wilson ID. Investigation of analytical variation in metabonomic analysis using liquid chromatography/mass spectrometry. Rapid Communications in Mass Spectrometry 2007;21:2965-70.
326. Gieger C, Geistlinger L, Altmaier E, et al. Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum. PLoS Genet 2008;4:e1000282.
327. Trygg J WS. Orthogonal projections to latent structures (O-PLS). . J Chemometrics 2002;16:119-28.
328. Yoo B, Kong S-Y, Jang S-G, et al. Identification of hypoxanthine as a urine marker for non-Hodgkin lymphoma by low-mass-ion profiling. BMC Cancer 2010;10:55.
329. MacIntyre DA, Jimenez B, Lewintre EJ, et al. Serum metabolome analysis by 1H-NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. Leukemia 2010;24:788-97.
330. Dettmer K, Vogl FC, Ritter AP, et al. Distinct metabolic differences between various human cancer and primary cells. Electrophoresis 2013;34:2836-47.
331. Calvo-Vidal MN, Cerchietti L. The metabolism of lymphomas. Current Opinion in Hematology 2013;20:345-54.
332. Roulland S, Kelly RS, Morgado E, et al. t(14;18) Translocation: A Predictive Blood Biomarker for Follicular Lymphoma. J Clin Oncol 2014;32:1347-55.
333. Ott G, Rosenwald A. Molecular pathogenesis of follicular lymphoma. Haematologica 2008;93:1773-6.
334. Schroeder JC, Olshan AF, Baric R, et al. Agricultural risk factors for t(14;18) subtypes of non-Hodgkin's lymphoma. Epidemiology 2001;12:701-9.
335. Dölken G, Illerhaus G, Hirt C, Mertelsmann R. BCL-2/JH rearrangements in circulating B cells of healthy blood donors and patients with nonmalignant diseases. Journal of Clinical Oncology 1996;14:1333-44.
336. Limpens J, Stad R, Vos C, et al. Lymphoma-associated translocation t(14;18) in blood B cells of normal individuals. Blood 1995;85:2528-36.
337. Schüler F, Dölken L, Hirt C, et al. Prevalence and frequency of circulating t(14;18)-MBR translocation carrying cells in healthy individuals. International Journal of Cancer 2009;124:958-63.
338. Roulland S, Navarro JM, Grenot P, et al. Follicular lymphoma-like B cells in healthy individuals: a novel intermediate step in early lymphomagenesis. J Exp Med 2006;203:2425-31.
339. Hirt C, Dölken G, Janz S, Rabkin CS. Distribution of t(14;18)-positive, putative lymphoma precursor cells among B-cell subsets in healthy individuals1. British Journal of Haematology 2007;138:349-53.
340. Freedman A. Follicular lymphoma: 2014 update on diagnosis and management. Am J Hematol 2014;89:429-36.
341. Bretherick KL, Bu R, Gascoyne RD, Connors JM, Spinelli JJ, Brooks-Wilson AR. Elevated circulating t(14;18) translocation levels prior to diagnosis of follicular lymphoma. Blood 2010;116:6146-7.
342. Hart J, Robert Turner A, Larratt L, et al. Transmission of a follicular lymphoma by allogeneic bone marrow transplantation – evidence to support the existence of lymphoma progenitor cells. British Journal of Haematology 2007;136:166-7.
343. Weigert O, Kopp N, Lane AA, et al. Molecular ontogeny of donor-derived follicular lymphomas occurring after hematopoietic cell transplantation. Cancer discovery 2012;2:47-55.

344. Pepe MS, Etzioni R, Feng Z, et al. Phases of Biomarker Development for Early Detection of Cancer. Journal of the National Cancer Institute 2001;93:1054-61.
345. Mandigers CMPW, Meijerink JPP, Mensink EJBM, et al. Lack of correlation between numbers of circulating t(14;18)-positive cells and response to first-line treatment in follicular lymphoma. Blood 2001;98:940-4.
346. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 2008;100:1432-8.
347. Montoto S, Fitzgibbon J. Transformation of Indolent B-Cell Lymphomas. Journal of Clinical Oncology 2011;29:1827-34.
348. Landgren O, Albitar M, Ma W, et al. B-Cell Clones as Early Markers for Chronic Lymphocytic Leukemia. New England Journal of Medicine 2009;360:659-67.
349. Landgren O, Kyle RA, Pfeiffer RM, et al. Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. Blood 2009;113:5412-7.
350. Leich E, Salaverria I, Bea S, et al. Follicular lymphomas with and without translocation t(14;18) differ in gene expression profiles and genetic alterations. Blood 2009;114:826-34.
351. Buchonnet G, Jardin F, Jean N, et al. Distribution of BCL2 breakpoints in follicular lymphoma and correlation with clinical features: specific subtypes or same disease? Leukemia 2002;16:1852-6.
352. Roulland S, Faroudi M, Mamessier E, Sungalee S, Salles G, Nadel B. Chapter 1 - Early Steps of Follicular Lymphoma Pathogenesis. In: Frederick WA, ed. Advances in Immunology: Academic Press; 2011:1-46.
353. Carlotti E, Wrench D, Matthews J, et al. Transformation of follicular lymphoma to diffuse large B-cell lymphoma may occur by divergent evolution from a common progenitor cell or by direct evolution from the follicular lymphoma clone. Blood 2009;113:3553-7.
354. Ruminy P, Jardin F, Picquenot J-M, et al. Sμ mutation patterns suggest different progression pathways in follicular lymphoma: early direct or late from FL progenitor cells. Blood 2008;112:1951-9.
355. Zhu CS, Pinsky PF, Cramer DW, et al. A Framework for Evaluating Biomarkers for Early Detection: Validation of Biomarker Panels for Ovarian Cancer. Cancer Prevention Research 2011;4:375-83.
356. Bentham G, Wolfreys AM, Liu Y, et al. Frequencies of hprt– mutations and bcl-2 translocations in circulating human lymphocytes are correlated with United Kingdom sunlight records. Mutagenesis 1999;14:527-32.
357. Agopian J, Navarro J-M, Gac A-C, et al. Agricultural pesticide exposure and the molecular connection to lymphomagenesis. The Journal of Experimental Medicine 2009;206:1473-83.
358. Vodicka P, Koskinen M, Stetina R, et al. The role of various biomarkers in the evaluation of styrene genotoxicity. Cancer detection and prevention 2003;27:275-84.
359. Veglia F, Matullo G, Vineis P. Bulky DNA Adducts and Risk of Cancer: A Meta-Analysis. Cancer Epidemiology Biomarkers & Prevention 2003;12:157-60.
360. O'Kane AA, Chevallier OP, Graham SF, Elliott CT, Mooney MH. Metabolomic Profiling of In Vivo Plasma Responses to Dioxin-Associated Dietary Contaminant Exposure in Rats: Implications for Identification of Sources of Animal and Human Exposure. Environmental Science & Technology 2013;47:5409-18.
361. Shi X, Wahlang B, Wei X, et al. Metabolomic Analysis of the Effects of Polychlorinated Biphenyls in Nonalcoholic Fatty Liver Disease. Journal of Proteome Research 2012;11:3805-15.
362. Ellis J, Athersuch T, Thomas L, et al. Metabolic profiling detects early effects of environmental and lifestyle exposure to cadmium in a human population. BMC Medicine 2012;10:61.
363. Koedrith P, Kim H, Weon J-I, Seo YR. Toxicogenomic approaches for understanding molecular mechanisms of heavy metal mutagenicity and carcinogenicity. International Journal of Hygiene and Environmental Health 2013;216:587-98.
364. Caro P, Kishan Amar U, Norberg E, et al. Metabolic Signatures Uncover Distinct Targets in Molecular Subsets of Diffuse Large B Cell Lymphoma. Cancer Cell;22:547-60.

365. Cano KE, Li L, Bhatia S, Bhatia R, Forman SJ, Chen Y. NMR-Based Metabolomic Analysis of the Molecular Pathogenesis of Therapy-Related Myelodysplasia/Acute Myeloid Leukemia. Journal of Proteome Research 2011;10:2873-81.
366. Lim U, Gayles T, Katki HA, et al. Serum High-Density Lipoprotein Cholesterol and Risk of Non-Hodgkin Lymphoma. Cancer Research 2007;67:5569-74.
367. Kyrtopoulos SA. Making sense of OMICS data in population-based environmental health studies. Environmental and Molecular Mutagenesis 2013;54:468-79.
368. Rawstron AC, Bennett FL, O'Connor SJM, et al. Monoclonal B-Cell Lymphocytosis and Chronic Lymphocytic Leukemia. New England Journal of Medicine 2008;359:575-83.
369. Shanafelt TD, Ghia P, Lanasa MC, Landgren O, Rawstron AC. Monoclonal B-cell lymphocytosis (MBL): biology, natural history and clinical management. Leukemia 2010;24:512-20.
370. Korde N, Kristinsson SY, Landgren O. Monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM): novel biological insights and development of early treatment strategies. Blood 2011;117:5573-81.
371. Patnala R, Clements J, Batra J. Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC Genetics 2013;14:39.
372. Chadeau-Hyam M, Vermeulen RCH, Hebels DGAJ, et al. Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. Annals of Oncology 2014;25:1065-72.
373. Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. Journal of clinical epidemiology 2007;60:1205-19.
374. Wolf SM, Crock BN, Van Ness B, et al. Managing incidental findings and research results in genomic research involving biobanks and archived data sets. Genet Med 2012;14:361-84.
375. Casey G, Conti D, Haile R, Duggan D. Next generation sequencing and a new era of medicine. Gut 2013;62:920-32.
376. Robson ME, Storm CD, Weitzel J, Wollins DS, Offit K. American Society of Clinical Oncology policy statement update: genetic and genomic testing for cancer susceptibility. J Clin Oncol 2010;28:893-901.

APPENDIX A

Supporting Tables, Figures and Information by Chapter

CONTENTS

3. METHODS

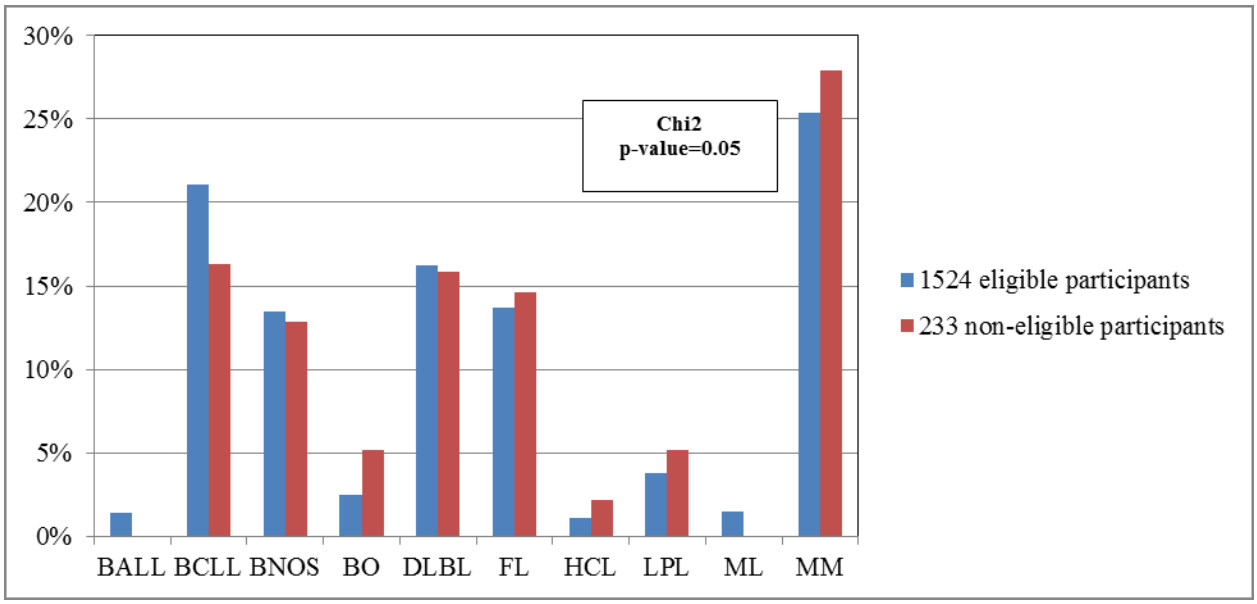
Table 3.1.1; Baseline characteristics of the 1757 incident B-cell NHL cases identified in the EPIC cohort during follow-up stratified by eligibility for the case-control analyses

Eligibility was determined by the availability of a baseline blood sample and two suitable matched controls who were cancer free at the time of diagnosis in the corresponding case

		Eligible Participants (n=1524)		Non-eligible Participants (n=233)		Difference (*p-value*)
Sex	Males (%)	759	(49.8)	59	(25.3)	
	Females (%)	765	(50.2)	174	(74.7)	*<0.001*
Age	Mean yrs (range)	57.01	(22 -79)	55.70	(24 - 94)	*0.028*
Country	France (%)	0	(0)	5	(2.1)	
	Italy (%)	173	(11.4)	9	(3.9)	
	Spain (%)	128	(8.4)	6	(2.6)	
	UK (%)	216	(14.2)	126	(5.4)	
	Netherlands (%)	119	(7.8)	8	(3.4)	
	Greece (%)	49	(3.2)	0	(0.0)	
	Germany (%)	156	(0.2)	10	(4.3)	
	Sweden (%)	309	(20.3)	1	(4.3)	
	Denmark (%)	358	(23.5)	2	(8.6)	
	Norway (%)	16	(1.0)	66	(28.3)	*<0.001*

Figure 3.1.1 Subtype Classification of Incident Cases in Eligible and Non-eligible Participants

Percentage of the total cases classified into each subtype. Chi-squared p-value refers to the differences across the subtype categories between the two groups



BALL- B-cell Acute Lymphatic Leukemia , BCLL- B-cell Chronic Lymphocytic Leukaemia, BNOS - B-cell NHL Not Otherwise Specified, BO – BCLL NHL, other, DLBCL - Diffuse Large B-cell Lymphoma, FL – Follicular Lymphoma, HCL - Hairy Cell Leukaemia, LPL - Lymphoplasmacytic Lymphoma, ML - Marginal Zone B-cell Lymphoma, MM- Multiple Myeloma

R script to run Profile Regression for chapter 4 (4.3.4) and in chapter 8 (8.3.1.1.1)

The sections in italics represent variable indices that were altered to run the sensitivity analysis, and the GxE analysis. Code is developed from the R package PReMiuM [223].

```
library(PReMiuM)
gxgmatrix<-read.table(file.choose(),header=T,sep="\t")
inputs_yModel<-"Bernoulli"
inputs_xModel<-"Discrete"
inputs_covnames<-names(gxgmatrix[c(2:1001)])
inputs_fixedeffectsnames<-names(gxgmatrix[c(1002:1014)])
inputs<-list(yModel=inputs_yModel, xModel=inputs_xModel, inputData=gxgmatrix,
covNames=inputs_covnames, fixedEffectsNames=inputs_fixedeffectsnames)

runInfoObj<-profRegr(covNames=inputs$covNames, fixedEffectsNames=inputs$fixedEffectsNames,
outcome="outcome", data=inputs$inputData, output="output", nSweeps=285000, nBurn=285000,
nClusInit=5, yModel=inputs$yModel, xModel=inputs$xModel, nProgress=1, nFilter=100,
varSelectType="BinaryCluster", seed=1)

dissimObj<-calcDissimilarityMatrix(runInfoObj)
clusObj<-calcOptimalClustering(dissimObj,maxNClusters=20)
riskProfileObj<-calcAvgRiskAndProfile(clusObj, includeFixedEffects=T)

rho<-summariseVarSelectRho(runInfoObj)
clusObj
rho$rhoMedian
whichcovariates<-which(rho$rhoMedian>0.45)
clusterOrderObj<-plotRiskProfile(riskProfileObj, "summary.png", showRelativeRisk=F,
orderBy=NULL, whichCovariates=whichcovariates,  useProfileStar=T)
```

Box 3.5: Criteria for a population-based screening test

(modified from Wilson & Junger, 1968 [233])

1. The condition should represent an important public health problem
2. The natural history of the condition should be well understood
3. There should be a recognizable latent or early sympotomatic stage
4. There should be a test that is easy to perform and interpret and which in addition is acceptable, accurate, reliable, sensitive and specific
5. There should be an accepted treatment for the condition, which is more effective if started early
6. There should be an accepted policy on who should be treated
7. Diagnosis and treatment should be cost effective
8. Case-finding should be a continuous process

4.PROFILE REGRESSION AS A TOOL FOR IDENTIFYING SUSCEPTIBILTY BIOMARKERS FROM GENETIC PROFILES

4.2.4 R script to thin the list of top 5000 SNPs found to be significantly associated with NHL

When the top 5000 SNPs are ordered by chromosome and then base-pair position, this code picks up the top SNPs for each region, according to p-value, by assuming that SNPs within 100kb of each other are in the same region and those more than 100kb apart are independent.

```
minp<-1
      data<-read.table("preGC-top1k4s.txt",header=F)
      for(i in 1:dim(data)[1]){
 if(i==1){
            if(data[i,4]<minp){
                  minp<-data[i,4]
                  minpRow<- i
            }else{
                  print("!!!")
            }
            }else{
            if(data[i,4]<minp && data[i,2]<data[i-1,2]+100000 && data[i,3]==data[i-1,3]){
                  minp<-data[i,4]
                  minpRow<- i
            }
                  if(data[i,2]>data[i-1,2]+100000 || data[i,3]!=data[i-1,3]){
       write.table(data[minpRow,],file="selectedSNP",quote=F,row.names=F,col.names=F,append=T)
                  minp<-data[i,4]
                   minpRow<- i
            }
         }
      }
```

Table 4.3.4: Sensitivity analyses exploring the number of iterations (nSweep) of the sampler required to obtain consistent clustering in the profile regression analysis

nBurn, nSweep, and nFilter refer to the script in appendix 3.4.4. nClusInit=*5 and* maxNClusters=*20* were kept consistent throughout. which(rho$rhoMedian>xx) varied dependent on the preceding results

				Cluster				No.	Significant Covariates
nBurn	**nSweep**	**nFilter**	**Outcome**	**1**	**2**	**3**	**4**	**clusters**	**(β)**
1000	***1000***	***40***	***Cluster sizes***	*408*	*30*			***2***	Spain; 1.40,
			Empirical risk	0.355	1				Greece; 1.43,
			Profile regression risk	0.358	0.941				Age; -0.02
1500	***1500***	***40***	***Cluster sizes***	*407*	*31*			***2***	
			Empirical risk	0.350	1				Italy; -2.97,
			Profile regression risk	0.360	0.930				Age; -0.04
3000	***3000***	***40***	***Cluster sizes***	*415*	*23*			***2***	
			Empirical risk	0.366	1				Italy; -2.52,
			Profile regression risk	0.369	0.915				UK; 0.66
5000	***5000***	***40***	***Cluster sizes***	*407*	*17*	*10*	*4*	***4***	Italy; -2.28,
			Empirical risk	0.354	1	1	1		UK; 1.78,
			Profile regression risk	0.369	0.601	0.632	0.919		Age; -0.02
10000	***10000***	***40***	***Cluster sizes***	*418*	*20*			***2***	
			Empirical risk	0.371	1				Italy; -2.54,
			Profile regression risk	0.374	0.917				Age; -0.03
50000	***50000***	***40***	***Cluster sizes***	*414*	*24*				
			Empirical risk	0.365	1				Italy; -2.16,
			Profile regression risk	0.370	0.861				UK; 1.72
250000	***250000***	***100***	***Cluster sizes***	*413*	*25*			***2***	
			Empirical risk	0.363	1				
			Profile regression risk	0.370	0.870				Italy; -2.31
275000	***275000***	***100***	***Cluster sizes***	*412*	*26*			***2***	
			Empirical risk	0.362	1				
			Profile regression risk	0.371	0.816				Italy; 2.05
285000	***285000***	***100***	***Cluster sizes***	*413*	*25*			***2***	
			Empirical risk	0.363	1				Italy; -2.09,
			Profile regression risk	0.371	0.850				UK; 1.71

5. BIOMARKERS OF INTERNAL DOSE: EXPOSURE TO ENVIRONMENTAL POLLUTANTS AND FUTURE RISK OF NHL

Table 5.3i; Mean, median and range of six PCB congeners, HCB, DDE, Cadmium and Lead concentrations in cases and controls from the EnviroGenoMarkers Study stratified by cohort and by gender

Exposure	Population	Cases			Controls			Difference
		Median	Mean	Range	Median	Mean	Range	p-value[a]
PCB118	**Sweden**	121.1	161.2	(8.5,901.6)	129.0	153.7	(11.6,832.1)	0.890
	Italy	180.1	228.1	(50,882.9)	201.9	224.8	(52.4,578.1)	0.885
	Men	109.9	140.7	(8.5,708)	133.7	166.4	(11.6,832.1)	0.0244*
	Women	171.5	222.1	(12.6,901.6)	162.8	184.7	(16.5,753.3)	0.0267*
PCB138	**Sweden**	530.0	608.1	(11,1770.9)	562.6	652.1	(51.2,2675.4)	0.279
	Italy	519.0	600.1	(186.8,1810.1)	562.0	601.4	(220.5,1768.6)	0.647
	Men	499.5	580.1	(11,1770.9)	576.3	669.0	(51.2,2675.4)	0.036*
	Women	562.1	630.5	(15.1,1810.1)	555.1	604.8	(66.7,1603.9)	0.581
PCB153	**Sweden**	1036.0	1142.8	(32.5,2698.4)	1083.5	1210.4	(120.6,4334.1)	0.364
	Italy	981.2	1169.0	(368.2,3308)	1100.1	1194.5	(472.4,3010)	0.448
	Men	974.9	1110.0	(32.5,2698.4)	1116.1	1256.9	(120.6,4334.1)	0.0333*
	Women	1032.0	1190.7	(69.1,3308)	1013.2	1155.6	(139,2860.4)	0.670
PCB156	**Sweden**	92.2	104.0	(19.8,307.8)	101.3	109.0	(15.5,394.9)	0.309
	Italy	83.9	97.5	(22.6,243.9)	90.8	101.7	(32.6,326.7)	0.391
	Men	89.8	100.9	(31.4,307.8)	102.2	114.2	(24.5,394.9)	0.0168*
	Women	91.8	103.0	(19.8,243.9)	90.8	99.4	(15.5,326.7)	0.614
PCB170	**Sweden**	343.3	384.1	(62.7,1082.3)	375.7	399.1	(50.4,1211.8)	0.384
	Italy	303.3	355.4	(98.6,934.3)	306.1	367.7	(127.5,1291.2)	0.465
	Men	341.3	384.6	(62.7,1082.3)	383.5	421.1	(84,1211.8)	0.088
	Women	309.0	366.0	(89.5,934.3)	311.5	358.6	(50.4,1291.2)	0.842
PCB180	**Sweden**	677.9	753.1	(139.7,2111.6)	733.2	784.0	(100.1,2159.5)	0.367
	Italy	712.3	868.0	(279.9,2708.8)	720.0	861.2	(349.2,2431.5)	0.525
	Men	718.2	800.0	(139.7,2708.8)	782.3	860.2	(180.3,2159.5)	0.134
	Women	625.4	778.0	(177.4,2430.6)	678.9	757.1	(100.1,2431.5)	0.984
HCB	**Sweden**	201.8	243.0	(62.5,659.7)	218.1	248.9	(64.5,1098.8)	0.696
	Italy	568.7	741.5	(115.8,3604)	628.5	841.7	(106.9,3882.2)	0.274
	Men	199.7	298.2	(62.5,2011.4)	257.9	378.2	(64.5,3661.2)	0.0113*
	Women	332.8	495.1	(82.9,3604)	309.4	486.0	(83.3,3882.2)	0.597
DDE	**Sweden**	1595.3	2223.6	(16.4,13776.1)	1992.8	2706.3	(76.4,18041.6)	0.084
	Italy	5946.1	8056.7	(1052.1,30992.8)	5579.2	7547.6	(1098.7,23858.3)	0.490
	Men	1919.7	3299.0	(16.4,30992.8)	2374.1	3617.0	(76.4,22536.8)	0.230
	Women	3305.3	4756.1	(78.6,30124.3)	3408.7	4770.4	(321,23858.3)	0.854
Cadmium	**Sweden**	0.41	0.72	(0.10, 4.11)	0.41	0.78	(0.10, 5.22)	0.998
	Italy	0.63	0.86	(0.22, 3.59)	0.59	0.71	(0.18, 2.76)	0.112
	Men	0.36	0.61	(0.13, 3.78)	0.38	0.73	(0.10, 5.22)	0.315
	Women	0.64	0.91	(0.10, 4.11)	0.54	0.79	(0.18, 4.32)	0.046*
Lead	**Sweden**	44.5	55.2	(15.4, 342.8)	45.5	54.7	(11.2, 672.5)	0.633
	Italy	92.8	104.6	(48.5, 400.8)	89.6	110.5	(39.3, 378.9)	0.287
	Men	64.1	78.5	(15.4, 342.8)	60.9	85.2	(24.3, 672.5)	0.794
	Women	51.9	62.9	(16.8, 400.8)	53.4	59.6	(11.2, 220.94)	0.823

[a]Differences between median concentrations in cases and controls according to the Wilcoxon signed-rank test for paired samples
One participant with exposure levels above the 99th percentile was excluded from the PCB analysis

Table 5.3ii Association between NHL and quartiles of exposure to persistent organic pollutants and metals stratified by cohort

POP by quartile		NSHDS								EPIC-Italy							
		Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*	Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*
PCB118	1	68/54	**1**				**1**				29/20	**1**				**1**			
	2	53/47	**0.84**	(0.48,1.48)	0.551		**1.14**	(0.58,2.23)	0.701		16/22	**0.48**	(0.19,1.23)	0.126		**0.61**	(0.21,1.8)	0.372	
	3	32/47	**0.45**	(0.23,0.88)	0.019*		**0.36**	(0.16,0.82)	0.014*		14/21	**0.46**	(0.17,1.27)	0.133		**0.5**	(0.17,1.5)	0.216	
	4	33/37	**0.61**	(0.32,1.17)	0.14	*0.047**	**0.4**	(0.16,0.99)	0.047*	*0.012**	25/20	**1.08**	(0.37,3.16)	0.895	*0.672*	**1.12**	(0.31,4.01)	0.864	*0.754*
PCB138	1	53/46	**1**				**1**				34/21	**1**				**1**			
	2	48/47	**0.85**	(0.48,1.52)	0.579		**0.67**	(0.33,1.36)	0.268		11/20	**0.28**	(0.09,0.81)	0.020*		**0.17**	(0.05,0.63)	0.008*	
	3	41/46	**0.68**	(0.35,1.33)	0.262		**0.48**	(0.21,1.1)	0.082		18/22	**0.38**	(0.13,1.14)	0.084		**0.27**	(0.07,0.98)	0.046*	
	4	44/46	**0.75**	(0.38,1.5)	0.418	*0.344*	**0.47**	(0.2,1.15)	0.097	*0.064*	21/20	**0.44**	(0.13,1.47)	0.181	*0.198*	**0.32**	(0.07,1.47)	0.143	*0.136*
PCB153	1	54/46	**1**				**1**				29/21	**1**				**1**			
	2	46/47	**0.75**	(0.39,1.42)	0.376		**0.76**	(0.36,1.61)	0.48		19/21	**0.57**	(0.22,1.48)	0.245		**0.46**	(0.15,1.38)	0.166	
	3	44/47	**0.66**	(0.32,1.35)	0.251		**0.62**	(0.25,1.51)	0.294		13/20	**0.31**	(0.09,1.08)	0.066		**0.22**	(0.05,0.95)	0.043*	
	4	42/45	**0.66**	(0.31,1.39)	0.276	*0.282*	**0.39**	(0.15,1)	0.05	*0.048**	23/21	**0.5**	(0.15,1.71)	0.272	*0.33*	**0.46**	(0.11,1.93)	0.285	*0.361*
PCB156	1	46/46	**1**				**1**				28/21	**1**				**1**			
	2	59/47	**1.16**	(0.59,2.28)	0.666		**1.46**	(0.67,3.17)	0.343		21/20	**0.67**	(0.23,1.9)	0.448		**0.46**	(0.13,1.61)	0.227	
	3	32/47	**0.62**	(0.28,1.35)	0.23		**0.56**	(0.21,1.47)	0.242		11/22	**0.3**	(0.09,0.99)	0.049*		**0.16**	(0.03,0.72)	0.017*	
	4	49/45	**1**	(0.47,2.14)	0.995	*0.638*	**0.75**	(0.3,1.89)	0.544	*0.241*	24/20	**0.61**	(0.16,2.26)	0.458	*0.377*	**0.42**	(0.08,2.11)	0.291	*0.294*
PCB170	1	45/47	**1**				**1**				27/21	**1**				**1**			
	2	61/46	**1.31**	(0.7,2.44)	0.403		**1.22**	(0.59,2.51)	0.595		19/20	**0.6**	(0.21,1.7)	0.336		**0.4**	(0.12,1.36)	0.141	
	3	34/46	**0.73**	(0.36,1.51)	0.403		**0.5**	(0.2,1.3)	0.155		17/21	**0.45**	(0.13,1.49)	0.19		**0.34**	(0.08,1.38)	0.13	
	4	46/46	**0.96**	(0.47,1.96)	0.92	*0.529*	**0.64**	(0.25,1.6)	0.339	*0.16*	21/21	**0.51**	(0.14,1.82)	0.299	*0.368*	**0.41**	(0.09,1.83)	0.241	*0.382*
PCB180	1	45/47	**1**				**1**				26/21	**1**				**1**			
	2	59/46	**1.34**	(0.71,2.56)	0.367		**1.41**	(0.65,3.05)	0.382		19/20	**0.74**	(0.3,1.81)	0.504		**0.48**	(0.16,1.43)	0.187	
	3	35/46	**0.76**	(0.37,1.55)	0.444		**0.48**	(0.19,1.26)	0.139		20/22	**0.6**	(0.19,1.87)	0.378		**0.44**	(0.11,1.68)	0.23	
	4	47/46	**1**	(0.5,2.01)	0.996	*0.651*	**0.63**	(0.24,1.64)	0.342	*0.168*	19/20	**0.62**	(0.2,1.96)	0.414	*0.447*	**0.45**	(0.11,1.89)	0.274	*0.336*

Table 5.3ii *continued*

POP by quartile		NSHDS								EPIC-Italy							
		Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*	Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*
Dioxin-like PCBs (118, 156)	1	51/47	**1**				**1**				25/21	**1**				**1**			
	2	48/46	**0.87**	(0.45,1.68)	0.674		**0.94**	(0.44,2.03)	0.878		20/21	**0.83**	(0.35,1.94)	0.666		**1.02**	(0.36,2.85)	0.974	
	3	36/46	**0.64**	(0.3,1.36)	0.249		**0.52**	(0.22,1.26)	0.149		16/21	**0.67**	(0.25,1.79)	0.422		**0.66**	(0.22,1.96)	0.451	
	4	51/46	**0.9**	(0.42,1.95)	0.794	*0.711*	**0.65**	(0.25,1.67)	0.367	*0.234*	23/20	**1.07**	(0.32,3.54)	0.918	*0.846*	**1.12**	(0.28,4.55)	0.871	*0.871*
Non-Dioxin like PCBs (138, 153, 170, 180)	1	56/47	**1**				**1**				26/21	**1**				**1**			
	2	43/46	**0.72**	(0.39,1.34)	0.301		**0.66**	(0.33,1.34)	0.254		21/21	**0.73**	(0.29,1.86)	0.511		**0.6**	(0.21,1.69)	0.33	
	3	44/46	**0.68**	(0.34,1.37)	0.28		**0.61**	(0.25,1.49)	0.281		15/21	**0.43**	(0.13,1.45)	0.174		**0.28**	(0.07,1.17)	0.081	
	4	43/46	**0.67**	(0.32,1.37)	0.272	*0.294*	**0.41**	(0.16,1.04)	0.061	*0.072*	22/20	**0.66**	(0.2,2.2)	0.498	*0.553*	**0.57**	(0.14,2.33)	0.433	*0.497*
Immunotoxic PCBs (118, 138, 156, 170)	1	51/47	**1**				**1**				25/21	**1**				**1**			
	2	48/46	**0.92**	(0.5,1.72)	0.797		**0.76**	(0.36,1.6)	0.47		21/21	**0.81**	(0.3,2.18)	0.682		**0.59**	(0.2,1.78)	0.353	
	3	41/46	**0.75**	(0.38,1.47)	0.397		**0.56**	(0.25,1.28)	0.173		14/21	**0.54**	(0.18,1.63)	0.273		**0.39**	(0.11,1.4)	0.151	
	4	46/46	**0.86**	(0.43,1.72)	0.663	*0.572*	**0.51**	(0.21,1.25)	0.139	*0.107*	24/20	**0.94**	(0.28,3.18)	0.922	*0.843*	**0.83**	(0.2,3.46)	0.801	*0.753*
ΣPCBs	1	54/47	**1**				**1**				27/21	**1**				**1**			
	2	48/46	**0.82**	(0.44,1.54)	0.54		**0.76**	(0.37,1.56)	0.45		22/21	**0.69**	(0.25,1.84)	0.454		**0.51**	(0.17,1.54)	0.233	
	3	40/46	**0.62**	(0.3,1.28)	0.195		**0.52**	(0.21,1.27)	0.152		12/21	**0.32**	(0.1,1.1)	0.07		**0.2**	(0.04,0.85)	0.030*	
	4	44/46	**0.7**	(0.34,1.46)	0.343	*0.289*	**0.41**	(0.16,1.05)	0.064	*0.048**	23/20	**0.62**	(0.18,2.1)	0.441	*0.435*	**0.46**	(0.11,2.05)	0.31	*0.386*
HCB	1	53/46	**1**				**1**				27/21	**1**				**1**			
	2	49/46	**0.80**	(0.42,1.53)	0.496		**0.90**	(0.42,1.9)	0.774		18/20	**0.64**	(0.27,1.51)	0.305		**0.68**	(0.27,1.73)	0.419	
	3	36/46	**0.53**	(0.24,1.2)	0.129		**0.51**	(0.18,1.4)	0.19		20/22	**0.61**	(0.23,1.64)	0.329		**0.69**	(0.24,2)	0.489	
	4	48/46	**0.71**	(0.31,1.6)	0.408	*0.424*	**0.5**	(0.18,1.4)	0.189	*0.156*	19/20	**0.6**	(0.2,1.85)	0.377	*0.368*	**0.59**	(0.17,2.07)	0.414	*0.411*
DDE	1	68/47	**1**				**1**				20/21	**1**				**1**			
	2	41/46	**0.50**	(0.26,0.95)	0.035*		**0.39**	(0.18,0.86)	0.020*		19/20	**1.01**	(0.38,2.66)	0.984		**1.18**	(0.42,3.34)	0.753	
	3	39/46	**0.45**	(0.23,0.89)	0.022*		**0.33**	(0.14,0.78)	0.011*		22/22	**1.05**	(0.41,2.7)	0.915		**1.25**	(0.44,3.53)	0.672	
	4	38/46	**0.46**	(0.24,0.89)	0.022*	*0.026**	**0.19**	(0.07,0.48)	0.001*	*<0.001**	23/20	**1.34**	(0.46,3.86)	0.589	*0.607*	**1.44**	(0.43,4.79)	0.551	*0.555*

Table 5.3ii *continued*

POP by quartile		NSHDS									EPIC-Italy								
		Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*	Ca/Co	OR	95% CI	P-value	*p for trend*	$OR^{(adj)}$	95% CI	P-value	*p for trend*
Cadmium	1	46/46	**1**				**1**				16/21	**1**				**1**			
	2	48/47	**1.04**	(0.58,1.88)	0.886		**1.12**	(0.57,2.21)	0.741		17/21	**1.00**	(0.43,2.33)	0.994		**1.06**	(0.42,2.67)	0.896	
	3	47/46	**1.05**	(0.57,1.91)	0.882		**1.09**	(0.54,2.23)	0.802		24/21	**1.49**	(0.64,3.5)	0.357		**1.68**	(0.68,4.14)	0.263	
	4	45/46	**1.00**	(0.54,1.86)	0.992	*0.999*	**1.10**	(0.51,2.39)	0.806	*0.837*	27/21	**1.91**	(0.75,4.86)	0.177	*0.168*	**2.11**	(0.75,5.91)	0.157	*0.106*
Lead	1	59/47	**1**				**1**				23/21	**1**				**1**			
	2	36/45	**0.57**	(0.29,1.11)	0.097		**0.61**	(0.28,1.32)	0.211		16/21	**0.71**	(0.3,1.71)	0.449		**0.77**	(0.31,1.91)	0.573	
	3	37/46	**0.59**	(0.3,1.14)	0.115		**0.53**	(0.23,1.2)	0.126		26/21	**1.16**	(0.49,2.7)	0.738		**1.17**	(0.48,2.87)	0.724	
	4	54/47	**0.83**	(0.44,1.58)	0.574	*0.755*	**0.80**	(0.37,1.75)	0.58	*0.648*	19/21	**0.79**	(0.31,2.05)	0.636	*0.945*	**0.88**	(0.3,2.53)	0.809	*0.921*

OR – conditional logistic regression accounting for matching factors

$OR^{(adj)}$ – conditional logistic regression additionally adjusting for BMI, height, educational level, vegetables, dairy, protein, total fat, alcohol

**Significant at the 95% confidence level*

Table 5.3.1i: Association between log-transformed exposure levels and NHL risk by Subtype; stratified by cohort

Exposure	CLL			DLBCL			FL			MM		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Sweden	***n=31***			***n=34***			***n=19***			***n=55***		
PCB118	**0.64**	(0.26,1.56)	0.324	**0.66**	(0.23,1.93)	0.451	**0.68**	(0.23,1.98)	0.48	**1.64**	(0.75,3.6)	0.215
PCB138	**0.61**	(0.23,1.6)	0.313	**0.47**	(0.12,1.88)	0.286	**0.63**	(0.16,2.56)	0.519	**0.87**	(0.36,2.1)	0.764
PCB153	**0.56**	(0.17,1.82)	0.336	**0.46**	(0.1,2.01)	0.299	**0.61**	(0.12,3.03)	0.548	**0.88**	(0.32,2.38)	0.795
PCB156	**0.61**	(0.14,2.74)	0.52	**0.34**	(0.08,1.39)	0.133	**0.64**	(0.15,2.67)	0.541	**0.99**	(0.37,2.64)	0.978
PCB170	**0.62**	(0.17,2.32)	0.481	**0.37**	(0.08,1.6)	0.182	**0.82**	(0.16,4.27)	0.811	**0.92**	(0.32,2.62)	0.873
PCB180	**0.62**	(0.15,2.49)	0.496	**0.37**	(0.08,1.59)	0.18	**0.89**	(0.15,5.19)	0.892	**0.93**	(0.32,2.68)	0.893
HCB	**0.20**	(0.03,1.29)	0.091	**0.60**	(0.12,2.94)	0.526	**0.37**	(0.06,2.28)	0.286	**1.21**	(0.44,3.34)	0.712
DDE	**0.64**	(0.29,1.4)	0.265	**0.77**	(0.35,1.71)	0.523	**0.73**	(0.32,1.67)	0.46	**0.69**	(0.39,1.24)	0.216
Dioxin-like PCBs	**0.74**	(0.39,1.39)	0.344	**0.62**	(0.3,1.29)	0.201	**0.76**	(0.38,1.53)	0.446	**1.22**	(0.74,1.99)	0.435
Non-Dioxin like PCBs	**0.87**	(0.63,1.18)	0.367	**0.78**	(0.53,1.15)	0.207	**0.91**	(0.61,1.37)	0.661	**0.97**	(0.75,1.26)	0.818
Immunotoxic PCBs	**0.85**	(0.62,1.18)	0.332	**0.76**	(0.51,1.13)	0.179	**0.88**	(0.6,1.29)	0.513	**1.04**	(0.8,1.34)	0.772
ΣPCBs	**0.9**	(0.72,1.12)	0.344	**0.83**	(0.64,1.09)	0.183	**0.92**	(0.71,1.21)	0.561	**1.01**	(0.85,1.21)	0.897
Cadmium	**0.94**	(0.42,2.11)	0.884	**1.14**	(0.56,2.29)	0.722	**1.1**	(0.35,3.47)	0.872	**0.96**	(0.61,1.52)	0.87
Lead	**1.61**	(0.47,5.55)	0.45	**0.62**	(0.21,1.83)	0.39	**1.46**	(0.28,7.76)	0.655	**1.32**	(0.66,2.63)	0.428
Italy	***n=11***			***n=11***			***n=20***			***n=21***		
PCB118	**1.27**	(0.15,11.01)	0.827	**2.57**	(0.07,96.51)	0.609	**1.94**	(0.4,9.33)	0.408	**0.22**	(0.04,1.21)	0.082
PCB138	**0.82**	(0.06,10.79)	0.878	**0.11**	(0,3.7)	0.22	**1.64**	(0.28,9.75)	0.585	**0.18**	(0.01,2.68)	0.214
PCB153	**0.28**	(0.02,4.17)	0.357	**0.04**	(0,2.74)	0.135	**1.63**	(0.26,10.07)	0.599	**0.27**	(0.02,3.32)	0.304
PCB156	**0.48**	(0.04,6.15)	0.574	**0.02**	(0,2.53)	0.111	**2.22**	(0.3,16.39)	0.433	**0.21**	(0.02,2.32)	0.203
PCB170	**0.28**	(0.02,3.64)	0.331	**0**	(0,1.92)	0.08	**2.48**	(0.41,14.82)	0.321	**0.32**	(0.03,3.62)	0.361
PCB180	**0.21**	(0.01,4.2)	0.31	**0.01**	(0,1.31)	0.062	**2.11**	(0.41,10.89)	0.374	**0.82**	(0.11,6.24)	0.845
Dioxin-like PCBs	**0.88**	(0.22,3.5)	0.861	**0.30**	(0.02,3.76)	0.353	**1.63**	(0.57,4.65)	0.36	**0.38**	(0.12,1.18)	0.093
Non-Dioxin like PCBs	**0.72**	(0.35,1.5)	0.384	**0.32**	(0.08,1.28)	0.107	**1.20**	(0.76,1.91)	0.432	**0.76**	(0.41,1.41)	0.386
Immunotoxic PCBs	**0.84**	(0.41,1.74)	0.636	**0.38**	(0.1,1.46)	0.16	**1.27**	(0.75,2.16)	0.365	**0.62**	(0.33,1.17)	0.14
ΣPCBs	**0.84**	(0.52,1.37)	0.496	**0.47**	(0.18,1.23)	0.124	**1.16**	(0.83,1.62)	0.39	**0.77**	(0.51,1.17)	0.216
HCB	**0.69**	(0.19,2.5)	0.571	**1.17**	(0.33,4.08)	0.809	**1.05**	(0.41,2.69)	0.92	**0.17**	(0.03,1.16)	0.07
DDE	**2.50**	(0.44,14.14)	0.299	**2.52**	(0.24,26.58)	0.441	**0.85**	(0.29,2.46)	0.763	**0.39**	(0.06,2.49)	0.321
Cadmium	**0.68**	(0.13,3.67)	0.652	**1.21**	(0.24,5.98)	0.817	**2.12**	(0.71,6.29)	0.178	**1.79**	(0.56,5.71)	0.322
Lead	**0.13**	(0.01,1.76)	0.126	**5.08**	(0.16,159.28)	0.355	**1.31**	(0.34,5.08)	0.692	**0.25**	(0.03,2.16)	0.21

OR – conditional logistic regression accounting for matching factors

**Significant at the 95% confidence level*

Table 5.3.1ii: Association between log-transformed exposure levels and NHL risk by Subtype; stratified by gender

Exposure	CLL			DLBCL			FL			MM		
	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Males	*n=25*			*n=24*			*n=19*			*n=34*		
PCB118	**0.58**	(0.21,1.64)	0.305	**0.51**	(0.13,2.01)	0.335	**0.31**	(0.07,1.28)	0.105	**0.74**	(0.25,2.23)	0.596
PCB138	**0.29**	(0.06,1.46)	0.135	**0.69**	(0.14,3.37)	0.645	**0.38**	(0.09,1.6)	0.186	**0.48**	(0.12,1.9)	0.296
PCB153	**0.20**	(0.03,1.48)	0.115	**0.56**	(0.1,3.12)	0.508	**0.38**	(0.08,1.89)	0.237	**0.54**	(0.13,2.35)	0.415
PCB156	**0.16**	(0.02,1.53)	0.11	**0.38**	(0.07,2.02)	0.258	**0.38**	(0.07,1.99)	0.253	**0.57**	(0.14,2.3)	0.426
PCB170	**0.14**	(0.01,1.41)	0.097	**0.51**	(0.09,2.85)	0.44	**0.56**	(0.1,3.11)	0.51	**0.59**	(0.15,2.35)	0.457
PCB180	**0.15**	(0.01,1.63)	0.119	**0.51**	(0.09,2.96)	0.453	**0.61**	(0.11,3.47)	0.575	**0.79**	(0.21,2.99)	0.732
Dioxin-like PCBs	**0.58**	(0.26,1.28)	0.178	**0.56**	(0.23,1.41)	0.22	**0.51**	(0.22,1.18)	0.116	**0.76**	(0.38,1.55)	0.458
Non-Dioxin like PCBs	**0.62**	(0.35,1.11)	0.106	**0.85**	(0.55,1.34)	0.488	**0.81**	(0.53,1.23)	0.316	**0.86**	(0.6,1.25)	0.433
Immunotoxic PCBs	**0.68**	(0.42,1.11)	0.127	**0.79**	(0.49,1.26)	0.324	**0.74**	(0.48,1.13)	0.168	**0.84**	(0.58,1.23)	0.377
ΣPCBs	**0.75**	(0.53,1.07)	0.118	**0.86**	(0.63,1.18)	0.362	**0.83**	(0.63,1.11)	0.209	**0.90**	(0.7,1.16)	0.424
HCB	**0.27**	(0.06,1.22)	0.089	**0.42**	(0.09,2.02)	0.28	**0.70**	(0.21,2.34)	0.565	**0.52**	(0.17,1.57)	0.248
DDE	**0.73**	(0.31,1.76)	0.487	**0.9**	(0.36,2.29)	0.832	**0.59**	(0.25,1.37)	0.22	**0.67**	(0.28,1.59)	0.36
Cadmium	**0.43**	(0.14,1.27)	0.125	**0.84**	(0.38,1.89)	0.681	**1.51**	(0.57,3.98)	0.404	**0.77**	(0.36,1.62)	0.489
Lead	**0.66**	(0.23,1.92)	0.448	**0.65**	(0.22,1.97)	0.45	**1.13**	(0.23,5.62)	0.885	**1.11**	(0.45,2.7)	0.822
Females	*n=17*			*n=21*			*n=20*			*n=42*		
PCB118	**1.00**	(0.26,3.88)	0.999	**1.23**	(0.26,5.86)	0.795	**11.12**	(0.88,140.92)	0.063	**1.32**	(0.56,3.07)	0.523
PCB138	**1.00**	(0.35,2.79)	0.993	**0.14**	(0.01,1.29)	0.083	**6.54**	(0.69,61.6)	0.101	**0.97**	(0.33,2.8)	0.951
PCB153	**0.87**	(0.25,3.11)	0.836	**0.12**	(0.01,1.37)	0.089	**5.61**	(0.55,57.19)	0.145	**0.90**	(0.27,2.99)	0.866
PCB156	**1.87**	(0.27,12.8)	0.525	**0.12**	(0.01,1.22)	0.073	**3.85**	(0.51,29)	0.191	**0.96**	(0.29,3.15)	0.945
PCB170	**0.97**	(0.24,3.9)	0.971	**0.05**	(0,0.94)	0.045*	**3.81**	(0.55,26.44)	0.176	**0.99**	(0.26,3.8)	0.988
PCB180	**0.94**	(0.2,4.33)	0.936	**0.06**	(0,0.86)	0.038*	**3.29**	(0.53,20.56)	0.202	**1.03**	(0.27,3.92)	0.961
HCB	**1.16**	(0.46,2.96)	0.753	**0.62**	(0.21,1.88)	0.399	**4.15**	(1,17.26)	0.05	**1.11**	(0.64,1.92)	0.702
DDE	**0.99**	(0.7,1.38)	0.932	**0.5**	(0.25,0.98)	0.045	**1.51**	(0.87,2.6)	0.14	**0.99**	(0.72,1.37)	0.956
Dioxin-like PCBs	**1.02**	(0.69,1.51)	0.905	**0.6**	(0.31,1.14)	0.118	**1.86**	(0.98,3.54)	0.059	**1.03**	(0.76,1.39)	0.848
Non-Dioxin like PCBs	**1.00**	(0.78,1.29)	0.982	**0.67**	(0.43,1.04)	0.077	**1.44**	(0.96,2.15)	0.08	**1.01**	(0.82,1.25)	0.912
Immunotoxic PCBs	**1.16**	(0.15,9.07)	0.885	**2.11**	(0.44,9.99)	0.348	**0.94**	(0.3,2.92)	0.918	**1.12**	(0.28,4.44)	0.875
ΣPCBs	**0.99**	(0.37,2.65)	0.985	**0.85**	(0.26,2.81)	0.795	**1.36**	(0.42,4.36)	0.609	**0.65**	(0.31,1.36)	0.251
Cadmium	**4.10**	(0.69,24.28)	0.12	**2.06**	(0.62,6.81)	0.235	**1.71**	(0.5,5.81)	0.389	**1.23**	(0.73,2.08)	0.443
Lead	**1.57**	(0.15,15.83)	0.704	**1.69**	(0.13,21.7)	0.686	**1.59**	(0.39,6.55)	0.52	**1.13**	(0.45,2.82)	0.793

OR – conditional logistic regression accounting for matching factors

**Significant at the 95% confidence level*

Table 5.3.2i; Association between log-transformed exposure levels and risk stratified by age group in NSHDS and EPIC-Italy

Exposure	30-44 years			45-59 years			60-75 years		
	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Sweden	*n cases=30, n controls=30*			*n cases=119, n controls=109*			*n cases=37, n controls=47*		
PCB118	0.34	(0.08,1.44)	0.145	0.90	(0.56,1.47)	0.685	2.08	(0.77,5.58)	0.146
PCB138	0.66	(0.22,1.98)	0.461	0.77	(0.46,1.3)	0.334	1.04	(0.34,3.16)	0.941
PCB153	0.62	(0.17,2.27)	0.471	0.76	(0.42,1.39)	0.378	0.88	(0.25,3.12)	0.841
PCB156	0.36	(0.06,2.31)	0.282	0.85	(0.4,1.79)	0.666	1.01	(0.27,3.68)	0.994
PCB170	0.58	(0.14,2.48)	0.464	0.86	(0.41,1.77)	0.677	0.83	(0.19,3.66)	0.802
PCB180	0.66	(0.15,2.93)	0.582	0.87	(0.41,1.84)	0.711	0.72	(0.15,3.44)	0.678
Dioxin-like PCBs	0.46	(0.17,1.27)	0.134	0.93	(0.67,1.28)	0.642	1.39	(0.74,2.64)	0.309
Non-Dioxin like PCBs	0.88	(0.63,1.24)	0.474	0.94	(0.8,1.11)	0.460	0.97	(0.69,1.37)	0.855
Immunotoxic PCBs	0.78	(0.5,1.21)	0.267	0.95	(0.8,1.12)	0.513	1.10	(0.78,1.53)	0.593
ΣPCBs	0.87	(0.66,1.15)	0.334	0.96	(0.86,1.08)	0.503	1.03	(0.82,1.3)	0.781
HCB	0.19	(0.02,1.94)	0.160	0.65	(0.3,1.42)	0.280	1.38	(0.27,7.08)	0.701
DDE	0.96	(0.32,2.91)	0.943	0.76	(0.55,1.06)	0.111	0.78	(0.39,1.56)	0.482
Cadmium	0.93	(0.48,1.79)	0.824	1.19	(0.83,1.7)	0.346	0.81	(0.41,1.58)	0.536
Lead	0.64	(0.26,1.53)	0.314	1.65	(0.81,3.35)	0.169	0.65	(0.17,2.41)	0.516
Italy	*n cases=10, n controls=10*			*n cases=50, n controls=49*			*n cases=24, n controls=25*		
PCB118	Δ			0.83	(0.31,2.21)	0.709	0.72	(0.18,2.89)	0.643
PCB138				0.93	(0.29,2.95)	0.904	0.85	(0.16,4.6)	0.854
PCB153				0.76	(0.23,2.52)	0.655	0.76	(0.16,3.62)	0.732
PCB156				0.73	(0.23,2.31)	0.599	0.65	(0.13,3.32)	0.602
PCB170				0.84	(0.3,2.39)	0.742	0.75	(0.13,4.27)	0.742
PCB180				1.05	(0.38,2.92)	0.919	0.89	(0.17,4.74)	0.887
Dioxin-like PCBs				0.86	(0.47,1.56)	0.617	0.79	(0.33,1.84)	0.579
Non-Dioxin like PCBs				0.97	(0.72,1.3)	0.839	0.95	(0.62,1.44)	0.799
Immunotoxic PCBs				0.94	(0.69,1.28)	0.698	0.91	(0.58,1.42)	0.669
ΣPCBs				0.97	(0.79,1.19)	0.755	0.95	(0.71,1.27)	0.713
HCB				0.84	(0.45,1.57)	0.582	0.33	(0.07,1.54)	0.158
DDE				1.42	(0.69,2.95)	0.344	1.01	(0.41,2.5)	0.978
Cadmium				2.14	(1.03,4.42)	0.041*	0.74	(0.19,2.93)	0.667
Lead				0.84	(0.27,2.6)	0.767	0.66	(0.15,2.9)	0.580

Δ *model would not converge*

OR – conditional logistic regression accounting for matching factors

**Significant at the 95% confidence level*

Table 5.3.2ii; Association between log-transformed levels and risk stratified by age group in males and females

Exposure	30-44 years			45-59 years			60-75 years		
	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Males	*n cases=25, n controls=25*			*n cases=89, n controls=78*			*n cases=19, n controls=30*		
PCB118	0.16	(0.02,1.19)	0.073	0.59	(0.33,1.07)	0.082	0.94	(0.28,3.08)	0.913
PCB138	0.16	(0.02,1.45)	0.103	0.59	(0.3,1.18)	0.135	0.75	(0.19,2.92)	0.677
PCB153	0.21	(0.02,1.99)	0.175	0.52	(0.23,1.19)	0.123	0.64	(0.15,2.8)	0.558
PCB156	0.14	(0.01,1.61)	0.113	0.45	(0.17,1.14)	0.092	0.74	(0.18,3.08)	0.679
PCB170	0.29	(0.04,2.19)	0.230	0.56	(0.23,1.35)	0.196	0.59	(0.11,3.14)	0.539
PCB180	0.46	(0.07,3.25)	0.436	0.63	(0.26,1.49)	0.292	0.56	(0.1,3.21)	0.516
Dioxin-like PCBs	0.21	(0.05,0.91)	0.037	0.67	(0.45,1.01)	0.057	0.90	(0.43,1.87)	0.779
Non-Dioxin like PCBs	0.68	(0.39,1.21)	0.191	0.86	(0.69,1.06)	0.150	0.89	(0.6,1.33)	0.566
Immunotoxic PCBs	0.49	(0.23,1.02)	0.056	0.83	(0.67,1.02)	0.081	0.92	(0.63,1.36)	0.683
ΣPCBs	0.68	(0.43,1.07)	0.095	0.89	(0.77,1.02)	0.100	0.94	(0.72,1.22)	0.630
HCB	0.05	(0,0.79)	0.033	0.57	(0.29,1.12)	0.103	0.48	(0.1,2.36)	0.363
DDE	0.57	(0.12,2.66)	0.472	0.83	(0.57,1.2)	0.315	0.59	(0.23,1.5)	0.268
Cadmium	0.82	(0.33,2.06)	0.674	1.13	(0.76,1.68)	0.550	0.23	(0.04,1.23)	0.086
Lead	0.53	(0.2,1.38)	0.194	2.00	(0.89,4.52)	0.094	0.19	(0.03,1.45)	0.110
Females	*n cases=15, n controls=15*			*n cases=80, n controls=80*			*n cases=42, n controls=42*		
PCB118	0.91	(0.18,4.55)	0.907	1.75	(0.83,3.71)	0.142	2.03	(0.73,5.63)	0.176
PCB138	0.65	(0.18,2.34)	0.513	1.22	(0.57,2.59)	0.613	1.27	(0.34,4.71)	0.718
PCB153	0.47	(0.08,2.79)	0.410	1.22	(0.53,2.83)	0.644	1.04	(0.27,4.1)	0.951
PCB156	0.17	(0.01,3.05)	0.227	1.42	(0.59,3.45)	0.435	0.97	(0.23,4.17)	0.972
PCB170	0.34	(0.03,3.26)	0.347	1.31	(0.55,3.11)	0.544	1.03	(0.21,5)	0.966
PCB180	0.25	(0.02,3.44)	0.303	1.41	(0.58,3.41)	0.444	1.05	(0.22,4.92)	0.953
Dioxin-like PCBs	0.68	(0.21,2.17)	0.512	1.37	(0.86,2.18)	0.188	1.39	(0.7,2.74)	0.342
Non-Dioxin like PCBs	0.81	(0.5,1.31)	0.400	1.07	(0.86,1.34)	0.539	1.03	(0.71,1.49)	0.885
Immunotoxic PCBs	0.82	(0.49,1.38)	0.456	1.12	(0.89,1.41)	0.328	1.13	(0.78,1.65)	0.512
ΣPCBs	0.86	(0.6,1.23)	0.417	1.07	(0.92,1.25)	0.384	1.06	(0.82,1.37)	0.639
HCB	2.18	(0.31,15.15)	0.430	1.09	(0.52,2.27)	0.824	0.80	(0.18,3.55)	0.771
DDE	1.09	(0.26,4.58)	0.903	0.89	(0.55,1.43)	0.619	1.15	(0.54,2.43)	0.722
Cadmium	1.09	(0.46,2.58)	0.851	1.87	(1.06,3.27)	0.030	1.39	(0.61,3.14)	0.432
Lead	1.37	(0.21,8.86)	0.742	0.81	(0.32,2.04)	0.660	1.19	(0.35,4.07)	0.785

OR – conditional logistic regression accounting for matching factors

**Significant at the 95% confidence level*

Table 5.3.2iii Association between log transformed exposure levels and NHL risk stratified by time to diagnosis in NSHDS and EPIC-Italy

Exposure (log pg/ml)	<5 years			≥ 5 years		
	OR	95% CI	p-value	OR	95% CI	p-value
Sweden	*n cases=70, n controls=186*			*n cases=116, n controls=186*		
PCB118	1.08	(0.65,1.78)	0.766	0.91	(0.6,1.39)	0.662
PCB138	0.66	(0.39,1.12)	0.123	0.74	(0.45,1.19)	0.214
PCB153	0.63	(0.33,1.17)	0.144	0.71	(0.4,1.27)	0.253
PCB156	0.77	(0.35,1.69)	0.51	0.79	(0.41,1.5)	0.467
PCB170	0.62	(0.28,1.36)	0.23	0.84	(0.43,1.67)	0.627
PCB180	0.56	(0.25,1.27)	0.167	0.84	(0.42,1.69)	0.623
Dioxin-like PCBs	0.99	(0.7,1.39)	0.933	0.91	(0.69,1.22)	0.536
Non-Dioxin like PCBs	0.88	(0.73,1.04)	0.137	0.93	(0.79,1.09)	0.346
Immunotoxic PCBs	0.93	(0.78,1.1)	0.4	0.94	(0.8,1.09)	0.404
ΣPCBs	0.94	(0.83,1.06)	0.287	0.95	(0.86,1.06)	0.389
HCB	1.26	(0.56,2.84)	0.579	0.56	(0.28,1.12)	0.103
DDE	0.74	(0.52,1.05)	0.089	0.66	(0.48,0.91)	0.012*
Cadmium	0.92	(0.63, 1.33)	0.652	0.99	(0.74, 1.33)	0.939
Lead	0.68	(0.36, 1.28)	0.23	0.98	(0.58, 1.65)	0.941
Italy	*n cases=38, n controls=84*			*n cases=33, n controls=84*		
PCB118	0.35	(0.12,1)	0.050*	1.38	(0.46,4.19)	0.566
PCB138	0.53	(0.16,1.77)	0.3	0.56	(0.15,2.11)	0.39
PCB153	0.50	(0.14,1.77)	0.285	0.38	(0.1,1.48)	0.162
PCB156	0.58	(0.18,1.89)	0.368	0.32	(0.08,1.24)	0.099
PCB170	0.65	(0.2,2.09)	0.468	0.28	(0.07,1.18)	0.084
PCB180	0.81	(0.26,2.57)	0.723	0.50	(0.14,1.79)	0.286
Dioxin-like PCBs	0.59	(0.32,1.11)	0.1	0.83	(0.42,1.63)	0.589
Non-Dioxin like PCBs	0.88	(0.64,1.2)	0.407	0.79	(0.55,1.12)	0.182
Immunotoxic PCBs	0.80	(0.58,1.11)	0.179	0.83	(0.58,1.2)	0.32
ΣPCBs	0.88	(0.71,1.09)	0.253	0.87	(0.69,1.11)	0.269
HCB	0.41	(0.19,0.9)	0.025*	0.77	(0.36,1.64)	0.495
DDE	0.78	(0.39,1.55)	0.474	1.72	(0.85,3.47)	0.13
Cadmium	0.40	(0.12, 1.31)	0.132	1.56	(0.68, 3.58)	0.292
Lead	1.56	(0.79, 3.08)	0.204	0.59	(0.17, 2.07)	0.41

OR – conditional logistic regression accounting for matching factors
**Significant at the 95% confidence level*

Table 5.3.2iv Association between log transformed exposure levels and NHL risk stratified by time to diagnosis in Males and females

Exposure (log pg/ml)		<5 years			≥ 5 years		
		OR	95% CI	p-value	OR	95% CI	p-value
Males		*n cases=60, n controls=133*			*n cases=71, n controls=133*		
	PCB118	0.59	(0.32,1.07)	0.083	0.54	(0.3,0.95)	0.034*
	PCB138	0.57	(0.3,1.07)	0.082	0.47	(0.23,0.95)	0.036*
	PCB153	0.46	(0.21,1.01)	0.054*	0.4	(0.17,0.92)	0.031*
	PCB156	0.50	(0.2,1.28)	0.148	0.32	(0.13,0.8)	0.015*
	PCB170	0.42	(0.17,1.06)	0.066	0.42	(0.16,1.1)	0.078
	PCB180	0.44	(0.17,1.13)	0.088	0.49	(0.18,1.29)	0.149
	Dioxin-like PCBs	0.69	(0.46,1.04)	0.075	0.61	(0.41,0.9)	0.013*
	Non-Dioxin like PCBs	0.82	(0.66,1.01)	0.059	0.79	(0.63,1)	0.047*
	Immunotoxic PCBs	0.82	(0.66,1.01)	0.057	0.77	(0.62,0.96)	0.018*
	ΣPCBs	0.87	(0.75,1)	0.055	0.84	(0.72,0.98)	0.024*
	HCB	0.46	(0.22,0.96)	0.039*	0.48	(0.23,0.98)	0.045*
	DDE	0.75	(0.5,1.13)	0.174	0.71	(0.46,1.08)	0.111
	Cadmium	0.78	(0.50, 1.20)	0.26	1.63	(0.99, 2.68)	0.055
	Lead	0.38	(0.17, 0.83)	0.016*	1.02	(0.54, 1.92)	0.947
Females		*n cases=48, n controls=137*			*n cases=78, n controls=137*		
	PCB118	1.38	(0.68,2.8)	0.369	1.61	(0.91,2.85)	0.099
	PCB138	0.83	(0.39,1.81)	0.647	0.94	(0.51,1.72)	0.829
	PCB153	0.96	(0.39,2.34)	0.925	0.87	(0.43,1.78)	0.71
	PCB156	1.10	(0.43,2.82)	0.85	1.11	(0.51,2.41)	0.796
	PCB170	1.12	(0.44,2.89)	0.808	0.92	(0.42,2.01)	0.833
	PCB180	1.12	(0.42,2.94)	0.825	0.94	(0.43,2.08)	0.884
	Dioxin-like PCBs	1.17	(0.74,1.86)	0.496	1.26	(0.87,1.83)	0.222
	Non-Dioxin like PCBs	0.99	(0.78,1.26)	0.961	0.98	(0.81,1.18)	0.8
	Immunotoxic PCBs	1.03	(0.81,1.31)	0.781	1.05	(0.87,1.28)	0.602
	ΣPCBs	1.02	(0.86,1.2)	0.835	1.02	(0.89,1.16)	0.794
	HCB	0.99	(0.43,2.24)	0.973	1.04	(0.49,2.21)	0.917
	DDE	0.73	(0.46,1.16)	0.186	0.86	(0.57,1.28)	0.452
	Cadmium	1.66	(1.01, 2.73)	0.046*	1.20	(0.80, 1.81)	0.376
	Lead	0.89	(0.39, 2.06)	0.79	0.75	(0.35, 1.60)	0.456

OR – conditional logistic regression accounting for matching factors
**Significant at the 95% confidence level*

Table 5.3.2v Association between log-transformed levels and risk stratified by BMI category in NSHDS and EPIC-Italy

Exposure (log pg/ml)	Normal			Overweight			Obese		
	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Sweden	*n cases=73, n controls=79*			*n cases=83, n controls=76*			*n cases=26, n controls=27*		
PCB118	0.96	(0.41,2.22)	0.919	0.77	(0.35,1.72)	0.526	0.69	(0.09,5.47)	0.723
PCB138	0.71	(0.32,1.59)	0.406	0.24	(0.06,0.96)	0.044*	0.31	(0.02,5.24)	0.415
PCB153	0.64	(0.25,1.67)	0.363	0.26	(0.06,1.17)	0.079	0.10	(0,5.30)	0.252
PCB156	0.55	(0.19,1.64)	0.285	0.53	(0.13,2.13)	0.374	0.09	(0,5.17)	0.242
PCB170	0.61	(0.2,1.87)	0.390	0.51	(0.12,2.13)	0.359	0.00	(0,11.1)	0.171
PCB180	0.59	(0.18,1.9)	0.373	0.39	(0.09,1.78)	0.226	0.00	(0,13.48)	0.174
Dioxin-like PCBs	0.85	(0.5,1.44)	0.542	0.78	(0.44,1.4)	0.407	0.61	(0.15,2.47)	0.488
Non-Dioxin like PCBs	0.89	(0.69,1.15)	0.371	0.73	(0.5,1.06)	0.096	0.45	(0.13,1.52)	0.198
Immunotoxic PCBs	0.90	(0.69,1.17)	0.438	0.80	(0.57,1.11)	0.185	0.65	(0.28,1.53)	0.328
ΣPCBs	0.93	(0.78,1.11)	0.406	0.84	(0.66,1.07)	0.151	0.69	(0.36,1.31)	0.258
HCB	1.02	(0.26,3.94)	0.982	0.47	(0.11,2.01)	0.307	0.00	(0,6.08)	0.132
DDE	0.82	(0.43,1.55)	0.546	0.28	(0.1,0.76)	0.013*	0.74	(0.06,9.35)	0.818
Cadmium	0.94	(0.53,1.68)	0.846	2.24	(1.03,4.88)	0.043*	0.33	(0.02,4.84)	0.421
Lead	3.66	(0.93,14.4)	0.063	0.96	(0.34,2.67)	0.932	1.31	(0.19,9.14)	0.785
Italy	*n cases=32, n controls=30*			*n cases=38, n controls=42*			*n cases=14, n controls=12*		
PCB118	0.90	(0.06,14.73)	0.942	0.17	(0.02,1.2)	0.075	Δ		
PCB138	2.23	(0.11,44.39)	0.599	0.03	(0,1.51)	0.079			
PCB153	4.34	(0.14,135.04)	0.403	0.03	(0,1.61)	0.085			
PCB156	39.77	(0.37,4280.14)	0.123	0.09	(0,1.85)	0.119			
PCB170	8.10	(0.26,250.69)	0.232	0.12	(0.01,1.84)	0.126			
PCB180	6.62	(0.26,166.65)	0.251	0.19	(0.02,1.79)	0.147			
Dioxin-like PCBs	12.96	(0.19,887.66)	0.235	0.32	(0.09,1.12)	0.074			
Non-Dioxin like PCBs	1.60	(0.65,3.91)	0.303	0.50	(0.22,1.13)	0.097			
Immunotoxic PCBs	2.56	(0.58,11.3)	0.214	0.49	(0.22,1.09)	0.082			
ΣPCBs	1.65	(0.73,3.72)	0.231	0.61	(0.35,1.08)	0.089			
HCB	1.55	(0.34,7.19)	0.573	0.26	(0.05,1.4)	0.117			
DDE	0.85	(0.14,5.01)	0.856	0.30	(0.06,1.36)	0.118			
Cadmium	6.09	(0.7,52.63)	0.101	0.26	(0.05,1.39)	0.115			
Lead	1.02	(0.03,31.01)	0.991	0.38	(0.08,1.83)	0.228			

Δ *model would not converge*

OR – conditional logistic regression accounting for matching factors

**Significant at the 95% confidence level*

Table 5.3.2vi Association between log-transformed exposure levels and risk stratified by BMI category in males and females

Exposure (log pg/ml)	Normal			Overweight			Obese		
	OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Males	*n cases=41, n controls=39*			*n cases=73, n controls=73*			*n cases=18, n controls=19*		
PCB118	0.74	(0.16,3.4)	0.695	0.48	(0.18,1.23)	0.127	Δ		
PCB138	1.61	(0.16,16.34)	0.689	0.15	(0.03,0.73)	0.019*			
PCB153	2.13	(0.15,30.8)	0.580	0.14	(0.03,0.78)	0.024*			
PCB156	1.08	(0.2,5.87)	0.925	0.28	(0.06,1.22)	0.090			
PCB170	2.22	(0.22,22.17)	0.497	0.28	(0.07,1.17)	0.081			
PCB180	3.18	(0.25,41.12)	0.375	0.30	(0.07,1.23)	0.095			
Dioxin-like PCBs	0.90	(0.33,2.46)	0.838	0.56	(0.29,1.06)	0.076			
Non-Dioxin like PCBs	1.26	(0.65,2.45)	0.495	0.65	(0.43,0.97)	0.037*			
Immunotoxic PCBs	1.05	(0.58,1.92)	0.866	0.67	(0.46,0.99)	0.043*			
ΣPCBs	1.09	(0.71,1.67)	0.708	0.76	(0.58,0.99)	0.040*			
HCB	0.46	(0.05,3.84)	0.472	0.24	(0.06,0.92)	0.037*			
DDE	1.13	(0.14,9)	0.908	0.39	(0.16,0.94)	0.037*			
Cadmium	1.97	(0.55,7.12)	0.299	0.94	(0.44,2.01)	0.873			
Lead	3.46	(0.43,27.64)	0.242	0.81	(0.3,2.2)	0.674			
Females	*n cases=64, n controls=70*			*n cases=48, n controls=45*			*n cases=22, n controls=20*		
PCB118	1.05	(0.41,2.74)	0.913	0.82	(0.25,2.75)	0.753	0.76	(0.1,5.68)	0.793
PCB138	0.70	(0.31,1.61)	0.402	0.20	(0.02,1.78)	0.150	0.31	(0.02,4.66)	0.394
PCB153	0.65	(0.24,1.71)	0.380	0.25	(0.03,2.35)	0.223	0.14	(0,5.93)	0.307
PCB156	0.65	(0.19,2.25)	0.501	0.60	(0.06,5.71)	0.653	0.17	(0.01,5.82)	0.326
PCB170	0.64	(0.2,2.06)	0.457	0.60	(0.06,5.92)	0.659	0.01	(0,37.71)	0.270
PCB180	0.60	(0.18,1.99)	0.400	0.33	(0.03,3.84)	0.376	0.01	(0,28.84)	0.235
Dioxin-like PCBs	0.92	(0.51,1.68)	0.790	0.83	(0.34,2.02)	0.682	0.66	(0.19,2.31)	0.520
Non-Dioxin like PCBs	0.89	(0.69,1.16)	0.395	0.70	(0.39,1.25)	0.226	0.53	(0.17,1.64)	0.268
Immunotoxic PCBs	0.92	(0.7,1.21)	0.549	0.80	(0.5,1.31)	0.380	0.71	(0.33,1.53)	0.385
ΣPCBs	0.94	(0.78,1.13)	0.484	0.84	(0.59,1.19)	0.324	0.75	(0.42,1.35)	0.332
HCB	1.73	(0.52,5.78)	0.371	0.97	(0.15,6.35)	0.971	0.01	(0,23.08)	0.240
DDE	0.80	(0.43,1.5)	0.491	0.05	(0,1.14)	0.061	0.26	(0.02,3.42)	0.304
Cadmium	1.04	(0.57,1.88)	0.895	2.54	(0.81,7.95)	0.109	1.09	(0.01,92.54)	0.969
Lead	2.91	(0.61,13.91)	0.181	0.52	(0.11,2.56)	0.422	0.50	(0.01,42.85)	0.761

Δ *model would not converge*

OR – conditional logistic regression accounting for matching factors

**Significant at the 95% confidence level*

6. METABOLIC PROFILING OF AN NHL CASE CONTROL STUDY TO IDENTIFY PREDICTIVE METABOLIC BIOMARKERS

Table 6.3.1: Baseline characteristics of cases and controls in the EnviroGenoMarkers study

Baseline variable		Case (n=210) [no. (%)]	Control (n=210) [no. (%)]	Difference (p-value)
Cohort	**Epic-Italy**	27 (12.9)	27 (12.9)	
	NSHDS	183 (87.1)	183 (87.1)	
Sex	**Male**	108 (51.4)	108 (51.4)	
	Female	102 (48.6)	102 (48.6)	
Mean Age	**(yrs)**	52.75	52.75	0.997
Mean Height	**(cm)**	171.10	169.30	0.046*
Mean Weight	**(kg)**	77.78	76.25	0.283
BMI	**Underweight**	1 (0.5)	0 (0.0)	
	Normal	82 (39.0)	84 (40.0)	
	Overweight	94 (44.8)	91 (43.3)	
	Obese	29 (13.8)	31 (14.8)	
	nk	4 (1.9)	4 (1.9)	0.768
Smoking status	**Never**	87 (41.4)	101 (48.1)	
	Former	73 (34.8)	55 (26.2)	
	Current	48 (22.9)	44 (21.0)	
	nk	2 (1.0)	10 (4.8)	0.166
Highest educational level	**None**	2 (1.0)	1 (0.5)	
	Primary	65 (31.0)	64 (30.5)	
	Technical/professional	56 (26.7)	49 (23.3)	
	Secondary	42 (20.0)	51 (24.3)	
	University/college	40 (19.1)	34 (16.2)	
	nk	5 (2.4)	11 (5.2)	0.722
Cambridge physical activity index	**Inactive**	64 (30.5)	58 (27.6)	
	Moderately inactive	82 (39.1)	76 (36.2)	
	Moderately active	53 (25.2)	60 (28.6)	
	Active	10 (4.8)	15 (7.1)	
	nk	1 (0.5)	1 (0.5)	0.581

Figure 6.3.1i: Percentage Coefficient of Variance (CV%) of feature intensity for the 750 included metabolite features in the discovery cohort and validation cohorts of the EnviroGenoMarkers Study

(A) Correlation between CV% in NSHDS (NS) and EPIC-ITALY (EPIC), (B) Histogram of CV% in EPIC-Italy, (C) Histogram of CV% in NSHDS

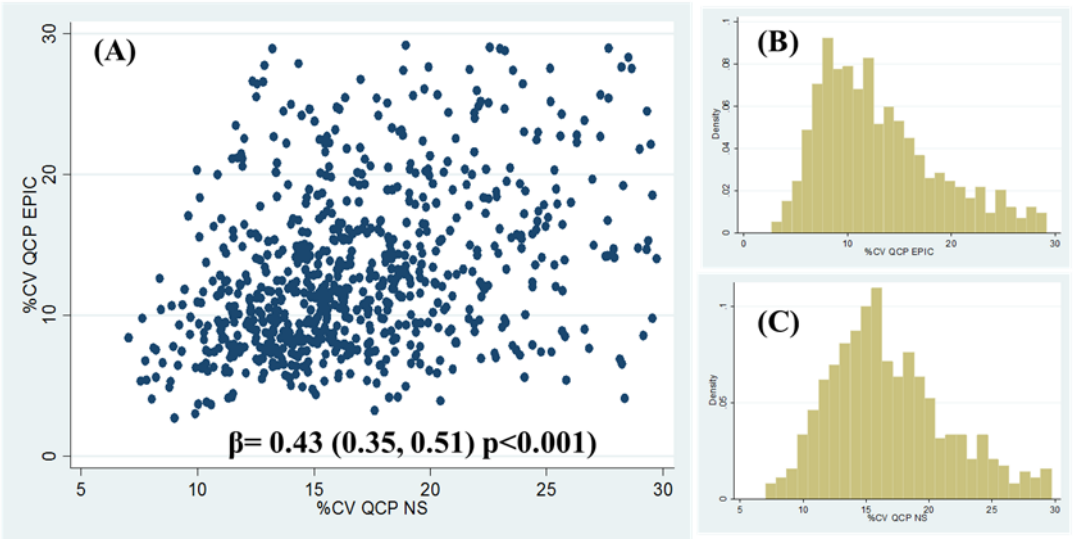


Figure 6.3.1ii: Correlation matrix describing the correlation between feature intensity for the 750 metabolite features in the discovery cohort and validation cohorts of the EnviroGenoMarkers Study

Computed using Spearman's rank correlation coefficient and raw intensity values

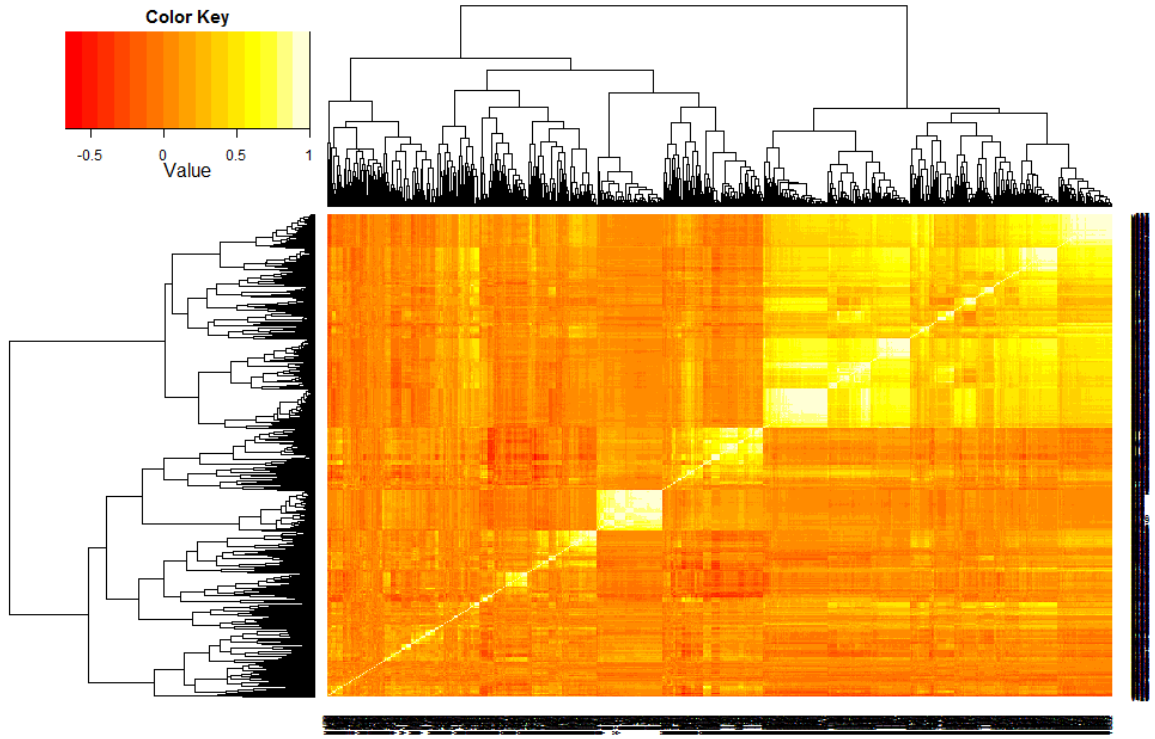


Figure 6.3.1iii: Score plot of the first two components following principal components analysis of 750 metabolite features in the discovery cohort and validation cohorts of the EnviroGenoMarkers Study

Coloured by sex, cohort, study phase and metabolomics processing batch

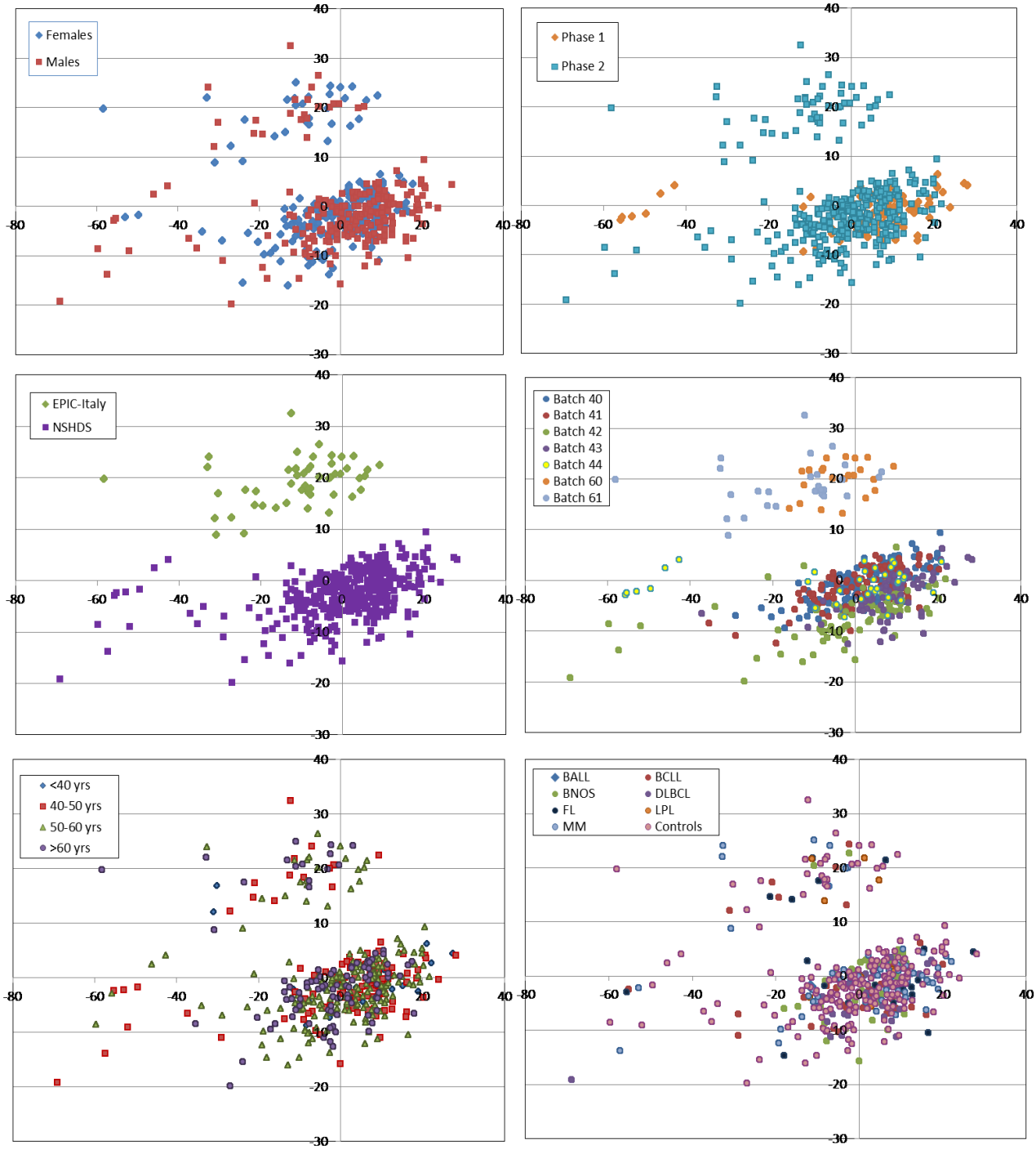


Table 6.3.2i: Number of metabolite features associated with lifestyle, anthropometric and lifestyle covariates at differing levels of statistical significance

Covariate	*n* Significant Associations (p)		
	<0.05	**<0.01**	**<6.67x10^{-4} (*bonferroni*)**
BMI (kg/m2)[2,3]	196	138	59
Height (cm)	214	149	39
Weight (kg)	59	23	2
Educational level[a,2,3]	99	49	4
Smoking status[b]	37	13	0
Cambridge physical activity index[c]	24	1	0
Alcohol (g/day)[2,3]	67	26	6
Vegetable intake (g/day)[2]	46	11	0
Dairy products (g/day)[2]	50	10	0
Protein (g/day)[2]	94	17	0
Total fats (g/day)[2,3]	192	89	4
Energy (kcal/day)[2,3]	147	36	2
Fish (g/day)[2]	65	12	1
Fruit (g/day)[2]	42	12	0

Computed using linear regression where log transformed metabolite feature intensity levels were considered as the outcome variable, and the covariate was considered as a continuous explanatory variable. [a]Educational level; 0-None, 1-primary, 2-Technical/professional, 3-secondary, 4-university/college. [b]Smoking status 0-never, 1-former, 2-current. [c]Physical activity 0-inactive, 1-moderately inactive, 2-moderately active, 3-active.

[2] variables included in model 2

[3]variables included in model 3

Table 6.3.2ii: Association between NHL risk and metabolite feature intensity quartiles for 19 selected features

Metabolite Feature	Conditional				Model1 (conditional + 10 covars)				Model2 (conditional + 5 covars)			
	OR	95% CI	p-value	*p-trend*	OR	95% CI	p-value	*p-trend*	OR	95% CI	p-value	*p-trend*
EPIC_NSHDS_15												
1	1				1				1			
2	1.66	(0.86,3.21)	0.134		3.93	(1.53,10.1)	0.004*		2.68	(1.19,6.03)	0.018*	
3	2.29	(1.08,4.84)	0.031*		5.72	(1.98,16.53)	0.001*		3.77	(1.47,9.65)	0.006*	
4	2.21	(0.99,4.95)	0.054*	*0.058*	4.85	(1.58,14.94)	0.006*	*0.013**	3.18	(1.17,8.6)	0.023*	*0.037**
EPIC_NSHDS_172												
1	1				1				1			
2	1.35	(0.7,2.61)	0.366		1.19	(0.55,2.59)	0.662		1.12	(0.54,2.32)	0.754	
3	1.50	(0.76,2.97)	0.246		1.80	(0.82,3.97)	0.145		1.80	(0.84,3.88)	0.13	
4	1.93	(1,3.72)	0.051*	*0.050**	1.87	(0.86,4.07)	0.112	*0.07*	1.72	(0.83,3.59)	0.146	*0.089*
EPIC_NSHDS_231												
1	1				1				1			
2	0.95	(0.52,1.75)	0.867		0.81	(0.39,1.71)	0.581		0.71	(0.35,1.44)	0.349	
3	0.76	(0.39,1.48)	0.419		0.59	(0.26,1.34)	0.209		0.60	(0.27,1.31)	0.196	
4	1.57	(0.69,3.53)	0.281	*0.612*	1.37	(0.5,3.77)	0.544	*0.936*	1.41	(0.54,3.67)	0.484	*0.988*
EPIC_NSHDS_286												
1	1				1				1			
2	1.17	(0.54,2.53)	0.686		1.24	(0.52,2.96)	0.626		1.37	(0.59,3.19)	0.46	
3	1.29	(0.54,3.12)	0.567		1.55	(0.55,4.4)	0.408		1.39	(0.52,3.71)	0.511	
4	2.93	(0.98,8.73)	0.054*	*0.079*	4.13	(1.07,15.98)	0.040*	*0.063*	3.97	(1.07,14.78)	0.040*	*0.078*
EPIC_NSHDS_375												
1	1				1				1			
2	1.52	(0.82,2.84)	0.186		1.79	(0.87,3.7)	0.115		1.64	(0.82,3.28)	0.166	
3	1.47	(0.76,2.84)	0.255		2.17	(0.98,4.81)	0.055*		1.88	(0.88,4)	0.102	
4	1.42	(0.73,2.75)	0.304	*0.437*	1.15	(0.51,2.58)	0.735	*0.749*	1.16	(0.54,2.49)	0.707	*0.765*
EPIC_NSHDS_482[#]												
1	1				1				1			
2	1.82	(0.94,3.53)	0.074		1.31	(0.58,2.95)	0.514		1.42	(0.65,3.11)	0.379	
3	1.97	(1.08,3.61)	0.027*		1.45	(0.7,2.99)	0.32		1.59	(0.79,3.19)	0.197	
4	2.57	(1.32,5.01)	0.005*	*0.007**	2.13	(0.96,4.72)	0.064	*0.053**	2.20	(1.03,4.69)	0.042*	*0.035**
EPIC_NSHDS_513												
1	1				1				1			
2	1.17	(0.65,2.13)	0.603		1.20	(0.59,2.45)	0.609		1.12	(0.57,2.21)	0.742	
3	1.49	(0.79,2.78)	0.215		1.39	(0.65,2.96)	0.395		1.63	(0.79,3.37)	0.182	
4	1.54	(0.77,3.08)	0.224	*0.166*	1.38	(0.6,3.14)	0.446	*0.388*	1.46	(0.66,3.26)	0.353	*0.225*

Table 6.3.2ii *continued*

Metabolite Feature	Conditional				Model1 (conditional + 10 covars)				Model2 (conditional + 5 covars)			
	OR	**95% CI**	**p-value**	**p-trend**	**OR**	**95% CI**	**p-value**	**p-trend**	**OR**	**95% CI**	**p-value**	**p-trend**
EPIC_NSHDS_514[#]												
1	1				1				1			
2	0.88	(0.46,1.68)	0.693		0.77	(0.35,1.67)	0.508		0.85	(0.41,1.79)	0.673	
3	1.52	(0.81,2.86)	0.191		1.11	(0.5,2.48)	0.799		1.34	(0.63,2.87)	0.447	
4	1.99	(1.05,3.79)	0.035*	*0.013**	2.55	(1.16,5.58)	0.019*	*0.011**	2.93	(1.37,6.27)	0.005*	*0.003**
EPIC_NSHDS_516												
1	1				1				1			
2	1.33	(0.69,2.54)	0.392		1.35	(0.59,3.06)	0.474		1.20	(0.54,2.64)	0.656	
3	1.13	(0.59,2.15)	0.708		1.16	(0.53,2.56)	0.712		1.17	(0.55,2.49)	0.685	
4	1.33	(0.68,2.62)	0.407	0.554	1.38	(0.6,3.18)	0.455	0.603	1.32	(0.59,2.95)	0.503	0.557
EPIC_NSHDS_523												
1	1				1				1			
2	1.45	(0.76,2.77)	0.257		1.42	(0.65,3.1)	0.385		1.51	(0.72,3.2)	0.277	
3	1.38	(0.73,2.6)	0.322		1.29	(0.6,2.8)	0.516		1.50	(0.71,3.17)	0.291	
4	2.78	(1.31,5.91)	0.008*	0.016*	2.56	(1.02,6.42)	0.045*	0.081	2.6	(1.08,6.29)	0.034*	0.051*
EPIC_NSHDS_558												
1	1				1				1			
2	1.40	(0.49,4.03)	0.533		1.71	(0.45,6.46)	0.429		1.43	(0.4,5.15)	0.583	
3	1.56	(0.51,4.77)	0.436		2.11	(0.51,8.81)	0.305		1.65	(0.42,6.49)	0.474	
4	2.18	(0.73,6.51)	0.162	0.105	2.89	(0.72,11.65)	0.136	0.105	2.48	(0.65,9.43)	0.183	0.102
EPIC_NSHDS_587[#]												
1	1				1				1			
2	0.93	(0.49,1.77)	0.833		1.01	(0.48,2.14)	0.975		0.96	(0.47,1.99)	0.923	
3	1.32	(0.72,2.44)	0.371		1.39	(0.68,2.85)	0.365		1.31	(0.66,2.62)	0.442	
4	1.86	(1,3.45)	0.048*	0.019*	2.10	(1.01,4.37)	0.048*	0.034*	1.94	(0.97,3.87)	0.061	0.042*
EPIC_NSHDS_61												
1	1				1				1			
2	4.40	(2.08,9.29)	<0.001*		4.09	(1.76,9.52)	0.001*		4.21	(1.86,9.53)	0.001*	
3	2.92	(1.36,6.29)	0.006*		2.47	(1.03,5.93)	0.043*		2.27	(0.98,5.27)	0.056	
4	3.37	(1.6,7.11)	0.001*	0.038*	2.89	(1.16,7.22)	0.023*	0.144	2.38	(1.01,5.59)	0.047*	0.299
EPIC_NSHDS_781												
1	1				1				1			
2	0.86	(0.44,1.68)	0.661		0.93	(0.45,1.92)	0.836		0.97	(0.48,1.99)	0.942	
3	0.41	(0.15,1.15)	0.092		0.36	(0.11,1.17)	0.09		0.39	(0.13,1.22)	0.107	
4	0.36	(0.1,1.35)	0.13	0.119	0.36	(0.07,1.81)	0.216	0.172	0.39	(0.09,1.78)	0.225	0.182
EPIC_NSHDS_796												
1	1				1				1			
2	0.92	(0.52,1.65)	0.791		0.82	(0.39,1.69)	0.585		0.84	(0.42,1.66)	0.612	
3	1.08	(0.6,1.96)	0.788		1.08	(0.53,2.2)	0.84		1.12	(0.57,2.2)	0.743	
4	0.60	(0.31,1.17)	0.135	0.264	0.47	(0.21,1.06)	0.069	0.173	0.51	(0.24,1.09)	0.085	0.219

Table 6.3.2ii *continued*

Metabolite Feature	Conditional				Model1 (conditional + 10 covars)				Model2 (conditional + 5 covars)			
	OR	95% CI	p-value	p-trend	OR	95% CI	p-value	p-trend	OR	95% CI	p-value	p-trend
EPIC_NSHDS_803												
1	1				1				1			
2	1.04	(0.62,1.77)	0.874		0.86	(0.47,1.6)	0.644		0.98	(0.55,1.78)	0.959	
3	0.52	(0.28,0.98)	0.041*		0.45	(0.21,0.97)	0.040*		0.53	(0.26,1.09)	0.084	
4	0.59	(0.3,1.16)	0.127	0.038*	0.55	(0.24,1.28)	0.167	0.064	0.61	(0.28,1.36)	0.231	0.103
EPIC_NSHDS_861												
1	1				1				1			
2	0.71	(0.39,1.27)	0.248		0.85	(0.42,1.73)	0.654		0.82	(0.42,1.64)	0.581	
3	0.74	(0.38,1.45)	0.388		0.91	(0.41,2.02)	0.819		0.87	(0.4,1.87)	0.713	
4	0.73	(0.38,1.42)	0.357	0.35	0.90	(0.41,1.99)	0.8	0.802	0.91	(0.42,1.95)	0.807	0.788
EPIC_NSHDS_871												
1	1				1				1			
2	1.00	(0.55,1.82)	0.99		0.69	(0.32,1.51)	0.355		0.63	(0.3,1.33)	0.228	
3	0.95	(0.5,1.81)	0.872		0.54	(0.23,1.26)	0.157		0.52	(0.24,1.14)	0.102	
4	1.29	(0.65,2.53)	0.467	0.505	1.10	(0.46,2.66)	0.83	0.787	0.89	(0.4,2.01)	0.787	0.837
EPIC_NSHDS_888#												
1	1				1				1			
2	2.10	(0.98,4.51)	0.057*		2.48	(0.99,6.23)	0.053		2.68	(1.09,6.56)	0.032*	
3	2.30	(1.01,5.24)	0.046*		2.56	(0.95,6.85)	0.062		2.77	(1.07,7.16)	0.035*	
4	2.79	(1.17,6.66)	0.021*	0.035*	4.56	(1.55,13.38)	0.006*	0.010*	3.77	(1.37,10.36)	0.010*	0.022*

Ordered quartile intensities were computed based on the distribution in the controls
The Odds ratio represents the increase in odds of incident NHL in each quartile of intensity
**Significant at the 95% confidence level*
Model 1 included BMI, educational level, alcohol consumption, vegetable consumption, dairy product consumption, protein consumption, total fat consumption, energy intake, fruit consumption and fish consumption
Model 2 included BMI, educational level, alcohol consumption, total fat and energy consumption

Table 6.3.2iii Pearson's correlation coefficient between the 19 top metabolite features identified in the discovery set

selected on the basis of a significant ($p<0.05$) association with NHL under at least one of the logistic regression models in the discovery set

	EPIC_NSHDS_15	**EPIC_NSHDS_61**	**EPIC_NSHDS_172**	**EPIC_NSHDS_231**	**EPIC_NSHDS_286**	**EPIC_NSHDS_375**	**EPIC_NSHDS_482**	**EPIC_NSHDS_513**	**EPIC_NSHDS_514**	**EPIC_NSHDS_516**	**EPIC_NSHDS_523**	**EPIC_NSHDS_558**	**EPIC_NSHDS_587**	**EPIC_NSHDS_781**	**EPIC_NSHDS_796**	**EPIC_NSHDS_803**	**EPIC_NSHDS_861**	**EPIC_NSHDS_871**	**EPIC_NSHDS_888**
EPIC_NSHDS_15	1.000																		
EPIC_NSHDS_61	0.207*	1.000																	
EPIC_NSHDS_172	0.141*	0.188*	1.000																
EPIC_NSHDS_231	0.101	-0.009	0.137*	1.000															
EPIC_NSHDS_286	-0.004	0.051	0.316*	0.254*	1.000														
EPIC_NSHDS_375	0.093	0.037	0.045	0.091	-0.054	1.000													
EPIC_NSHDS_482	0.072	0.076	0.150*	0.142*	0.036	0.208*	1.000												
EPIC_NSHDS_513	0.088	0.003	0.095	0.172*	-0.016	0.149*	0.426*	1.000											
EPIC_NSHDS_514	0.009	-0.003	0.157*	0.172*	0.158*	0.091	0.335*	0.347*	1.000										
EPIC_NSHDS_516	0.039	0.118*	0.155*	-0.005	0.078	-0.010	0.137*	0.233*	0.324*	1.000									
EPIC_NSHDS_523	0.160*	0.094	0.087	0.172*	-0.203*	0.214*	0.437*	0.462*	0.170*	0.182*	1.000								
EPIC_NSHDS_558	-0.062	-0.051	0.125*	0.157*	**0.508***	0.024	0.059	-0.093	0.115*	-0.040	-0.271*	1.000							
EPIC_NSHDS_587	0.000	0.105*	0.092	0.167*	0.105*	0.228*	**0.563***	0.269*	0.334*	0.030*	0.219*	0.361*	1.000						
EPIC_NSHDS_781	-0.016	-0.010	0.077	0.350*	**0.535***	-0.077	-0.217*	-0.143*	0.074	0.013	-0.354*	0.463*	0.067	1.000					
EPIC_NSHDS_796	0.169*	0.134*	0.036	0.140*	0.059	0.047	0.320*	0.201*	0.119*	0.029	0.225*	0.011	0.216*	0.016	1.000				
EPIC_NSHDS_803	0.215*	0.057	0.192*	0.135*	-0.046	0.028	0.031	0.127*	0.032	0.031	0.203*	-0.105*	0.086	0.063	0.238*	1.000			
EPIC_NSHDS_861	-0.014	0.006	-0.080	0.067	-0.099	0.037	0.039	0.079	0.058	0.043	0.054	0.010	0.104*	0.011	0.017	0.028	1.000		
EPIC_NSHDS_871	0.023	0.080	0.014	-0.015	-0.166*	0.105*	0.064	0.138*	0.017	0.023	0.131*	-0.137*	0.048	-0.183*	0.076	-0.035	0.060	1.000	
EPIC_NSHDS_888	0.045	0.012	-0.003	0.143*	0.393*	-0.071	-0.107*	-0.119*	0.075	0.093	-0.271*	0.404*	0.149*	**0.527***	0.199*	0.079	-0.007	-0.047	1.000

**Significant at the 95% confidence level*

metabolite features pairs with a rho median>0.5 are highlighted in bold

Figure 6.3.2i: Significant features in NSHDS (discovery set, blue) and EPIC-Italy (validation set, red) plotted according to *rt* (x-axis) and *m/z* (y-axis)

Where two features (one from EPIC-Italy and one from NSHDS) appear to have similar *m/z* and *rt* (highlighted by a dashed red box) the OR and p-value associated with each feature is shown in order to explore whether the metabolite features are related

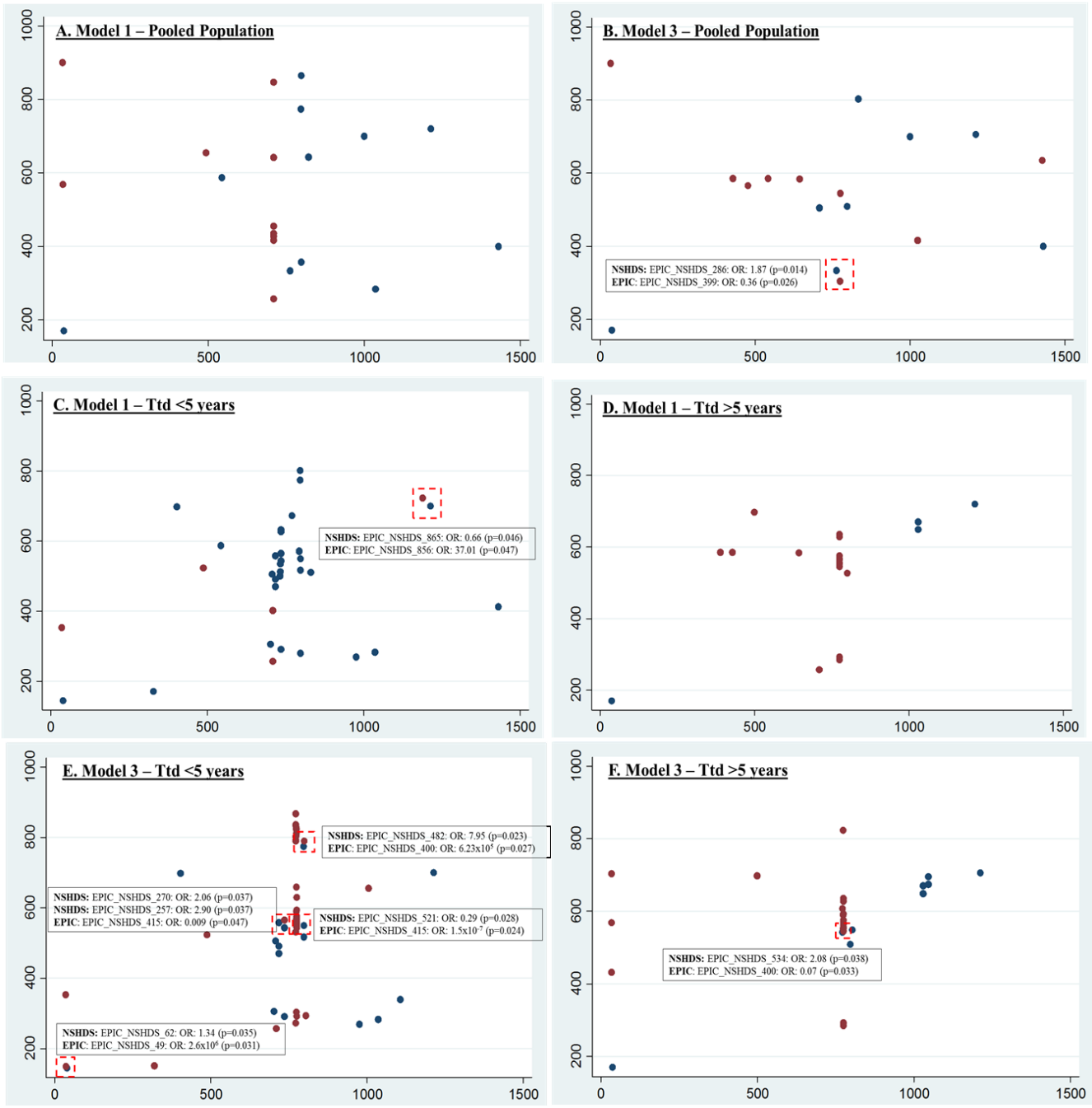


Figure 6.3.2ii Score Plot of the PCA Model in the discovery cohort

Coloured according to case-control status

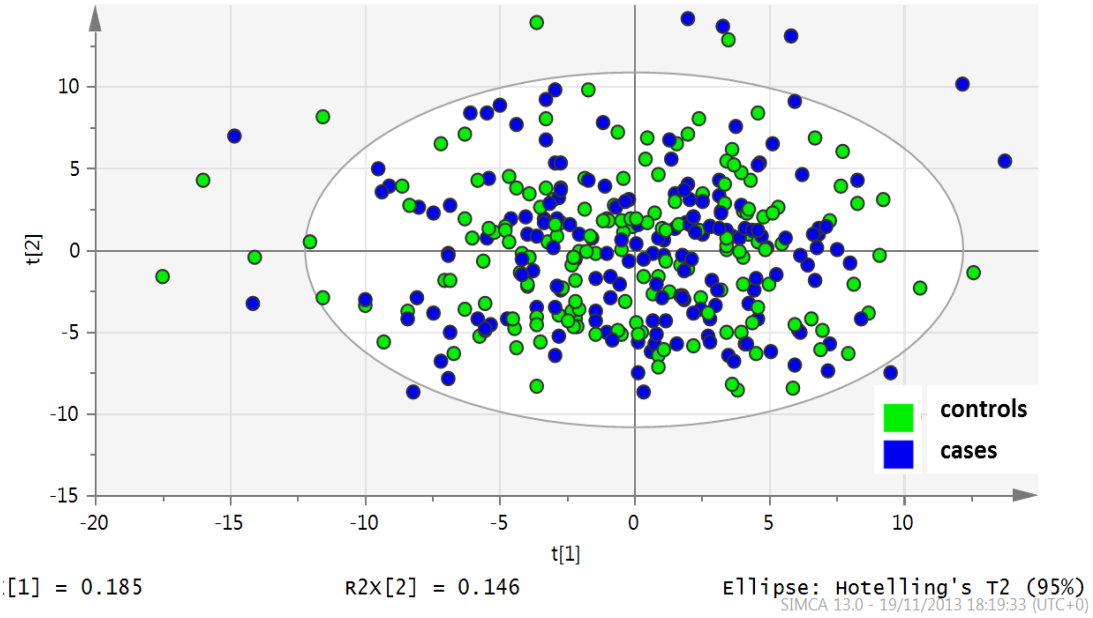
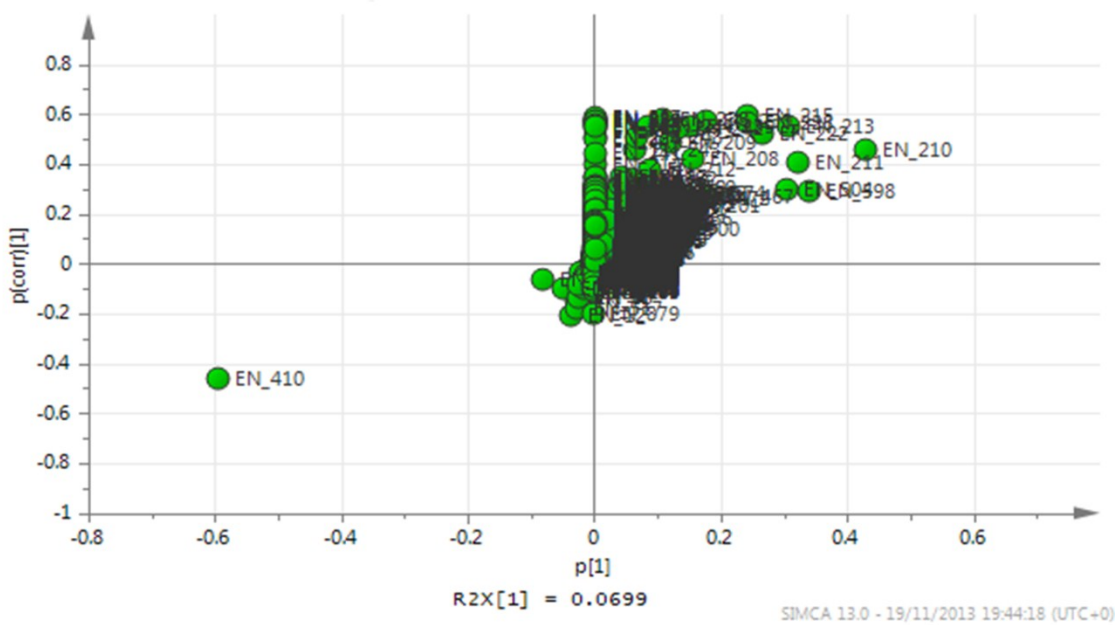


Figure 6.3.2iii: S-Plot of the OPLS-DA Model in the Discovery cohort



The S-plot shows the covariance (p) versus the correlation p(corr) of the metabolite features for the discriminating component of the OPLS-DA model. In theory features in the bottom left and top right of the plot represent reliable discriminators between cases and controls

Table 6.3.2iv: Predictive ability of the subtype specific OPLS-DA models based on the discovery cohort

The negative Q^2 for every subtype suggests none of these models have any predictive ability

Subtype	*N.*	*N.* excluded outliers	R^2	P1	O1	Q^2
Follicular lymphoma	36	2	0.299	0.065	0.234	-0.436
BCLL	62	3	0.329	0.046	0.280	-1.180
DLBCL	68	2	0.251	0.045	0.132	-1.040
MM	108	5	0.259	0.063	0.196	-0.469

R^2-Goodness of model fit; measure of the variance explained

P1-Proprtion of the explained variance due to disease risk

O1 –Proportion of the explained variance not due to disease risk (orthogonal variation)

Q^2 – goodness of prediction of the model

Table 6.3.3: Description of the metabolite feature models considered in the ROC analyses

Model	Description	*n* features	Percentage of variation explained by the first component	Included features
1	Significant ($p<0.05$) features from the conditional logistic regression model1 in the discovery cohort	13	20.4%	EPIC_NSHDS_61 EPIC_NSHDS_172 EPIC_NSHDS_286 EPIC_NSHDS_375 EPIC_NSHDS_482 EPIC_NSHDS_513 EPIC_NSHDS_516 EPIC_NSHDS_523 EPIC_NSHDS_587 EPIC_NSHDS_781 EPIC_NSHDS_803 EPIC_NSHDS_861 EPIC_NSHDS_888
2	Significant ($p<0.05$) features from conditional logistic regression model2 in the discovery cohort	9	24.5%	EPIC_NSHDS_15 EPIC_NSHDS_61 EPIC_NSHDS_231 EPIC_NSHDS_286 EPIC_NSHDS_375 EPIC_NSHDS_781 EPIC_NSHDS_796 EPIC_NSHDS_803 EPIC_NSHDS_888
3	Significant ($p<0.05$) features from conditional logistic regression model3 in the discovery cohort	9	29.3%	EPIC_NSHDS_61 EPIC_NSHDS_231 EPIC_NSHDS_286 EPIC_NSHDS_375 EPIC_NSHDS_514 EPIC_NSHDS_558 EPIC_NSHDS_781 EPIC_NSHDS_871 EPIC_NSHDS_888
4	Features with a VIP value>1 and jack-knifed confidence intervals that do not include one from the OPLS-DA model in the discovery cohort	54	59.4%	EPIC_NSHDS_6 EPIC_NSHDS_61 EPIC_NSHDS_213 EPIC_NSHDS_215 EPIC_NSHDS_244 EPIC_NSHDS_248 EPIC_NSHDS_256 EPIC_NSHDS_266 EPIC_NSHDS_316 EPIC_NSHDS_326 EPIC_NSHDS_347 EPIC_NSHDS_357 EPIC_NSHDS_417 EPIC_NSHDS_421 EPIC_NSHDS_422 EPIC_NSHDS_431 EPIC_NSHDS_445 EPIC_NSHDS_450 EPIC_NSHDS_461 EPIC_NSHDS_463 EPIC_NSHDS_464 EPIC_NSHDS_465 EPIC_NSHDS_466 EPIC_NSHDS_467 EPIC_NSHDS_468 EPIC_NSHDS_470 EPIC_NSHDS_472 EPIC_NSHDS_473 EPIC_NSHDS_475 EPIC_NSHDS_476 EPIC_NSHDS_482 EPIC_NSHDS_483 EPIC_NSHDS_484 EPIC_NSHDS_485 EPIC_NSHDS_486 EPIC_NSHDS_490 EPIC_NSHDS_493 EPIC_NSHDS_499 EPIC_NSHDS_500 EPIC_NSHDS_502 EPIC_NSHDS_505 EPIC_NSHDS_514 EPIC_NSHDS_515 EPIC_NSHDS_516 EPIC_NSHDS_518 EPIC_NSHDS_557 EPIC_NSHDS_560 EPIC_NSHDS_579 EPIC_NSHDS_582 EPIC_NSHDS_583 EPIC_NSHDS_585 EPIC_NSHDS_587 EPIC_NSHDS_589 EPIC_NSHDS_598
5	Features identified in models 1,2 or 3 AND in model 4	5	38.6%	EPIC_NSHDS_61 EPIC_NSHDS_482 EPIC_NSHDS_514 EPIC_NSHDS_516 EPIC_NSHDS_587
6	Features identified in models 1,2 or 3 AND in model 4 when these models were run only in those participants who were diagnosed within five years of blood draw	30	36.1%	EPIC_NSHDS_62 EPIC_NSHDS_87 EPIC_NSHDS_120 EPIC_NSHDS_172 EPIC_NSHDS_202 EPIC_NSHDS_231 EPIC_NSHDS_252 EPIC_NSHDS_256 EPIC_NSHDS_257 EPIC_NSHDS_265 EPIC_NSHDS_266 EPIC_NSHDS_267 EPIC_NSHDS_269 EPIC_NSHDS_270 EPIC_NSHDS_271 EPIC_NSHDS_272 EPIC_NSHDS_273 EPIC_NSHDS_274 EPIC_NSHDS_290 EPIC_NSHDS_325 EPIC_NSHDS_450 EPIC_NSHDS_457 EPIC_NSHDS_482 EPIC_NSHDS_495 EPIC_NSHDS_518 EPIC_NSHDS_521 EPIC_NSHDS_525 EPIC_NSHDS_767 EPIC_NSHDS_803 EPIC_NSHDS_887

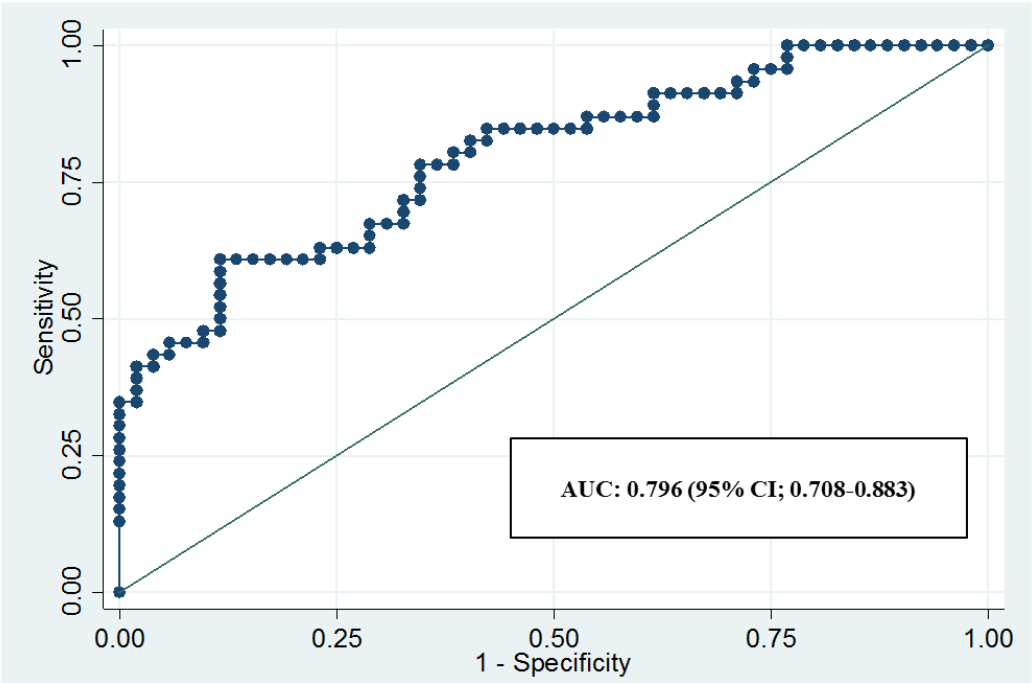
7. t(14;18) TRANSLOCATION: A PREDICTIVE BLOOD BIOMARKER FOR FOLLICULAR LYMPHOMA

Table 7.3.3: Ability of varying thresholds of t(14;18) frequency to predict risk of FL among t(14;18)$^{+}$ individuals in the validation cohort

t(14;18) frequency threshold	Sensitivity	Specificity	PPV	NPV	OR≠ (95% CI)
5x10^{-6}	69.57%	63.46%	62.75%	70.21%	4.43 (1.79, 10.94)
1x10^{-5}	60.87%	73.08%	66.67%	67.86%	5.93 (2.25, 15.60)
5x10^{-5}	36.96%	98.08%	94.44%	63.75%	41.82 (4.25, 411.75)
1x10^{-4}	26.09%	100.00%	100.00%	60.47%	ł
5x10^{-4}	17.39%	100.00%	100.00%	57.78%	ł
1x10^{-3}	13.04%	100.00%	100.00%	56.52%	ł
Continuous variable					1.04 (1.01, 1.07)

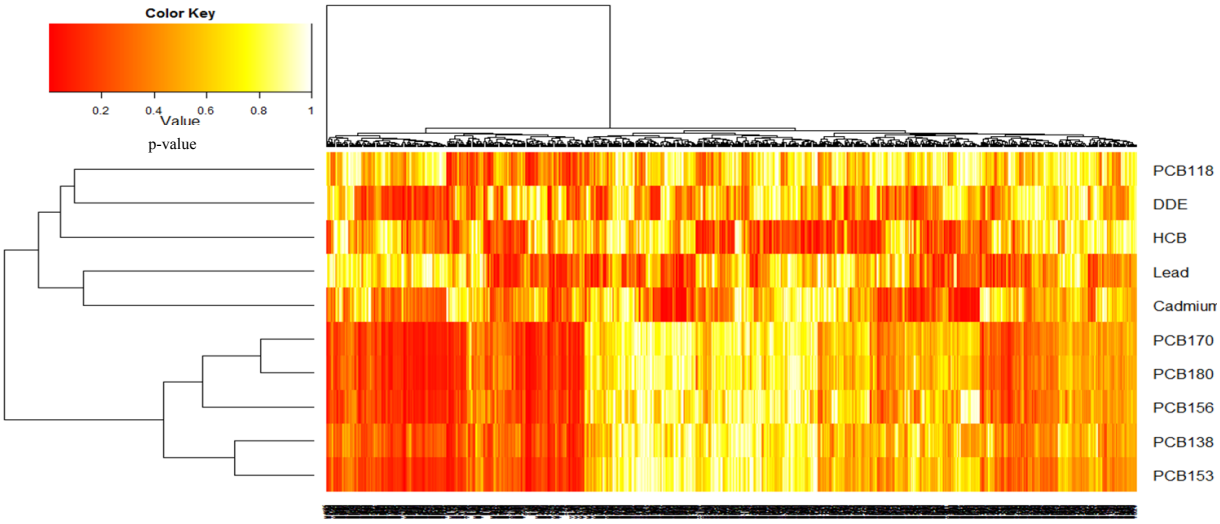
≠ OR adjusting for age at screening, sex, country and days between blood draw and frequency screening
ł No controls were above this threshold so an odds ratio could not be computed

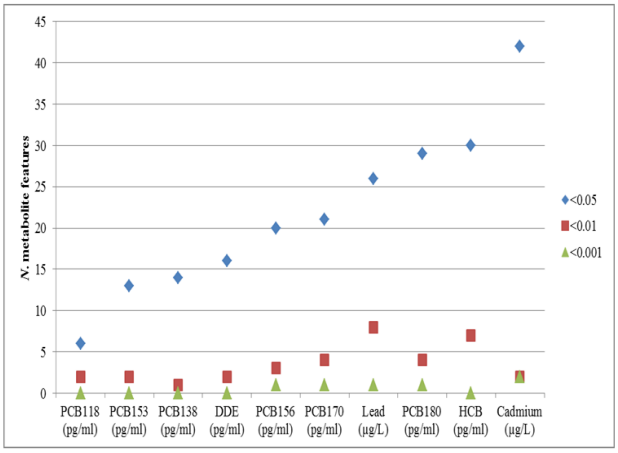
Figure 7.3.3: The receiver-operating-characteristic (ROC) curves for t(14;18) frequency in the discovery cohort (t(14;18)$^{+}$ samples)



8. INTEGRATING BIOMARKERS: THE "*MEET-IN-THE-MIDDLE*"APPROACH

Figure 8.3.3: Associations between ten environmental polluants and 750 metabolite features in the validation cohort (EPIC-Italy) according to significance level





Exposure	*n* Significant Associations (p)			
	<0.05	<0.01	<0.001	<6.67x10^{-4} (*bonferroni*)
PCB118 (pg/ml)	6	2	0	0
PCB153 (pg/ml)	13	2	0	0
PCB138 (pg/ml)	14	1	0	0
DDE (pg/ml)	16	2	0	0
PCB156 (pg/ml)	20	3	1	0
PCB170 (pg/ml)	21	4	1	0
Lead (µg/L)	26	8	1	0
PCB180 (pg/ml)	29	4	1	0
HCB (pg/ml)	30	7	0	0
Cadmium (µg/L)	42	2	2	0

APPENDIX B

Bibliography and Conference proceedings

BIBLIOGRAPHY:

Peer-reviewed publications:

1. **Kelly RS** and Vineis P. Biomarkers of Susceptibility to Chemical Carcinogens: The Example of Non-Hodgkin Lymphomas. *British Medical Bulletin* (accepted)

2. Roulland S, **Kelly RS** *et al.* t(14;18) Translocation: A predictive blood biomarker for Follicular Lymphoma. *Journal of Clinical Oncology.* 2014. 52.8190

3. Chadeau-Hyam M, Vermeulen RC, Hebels DG, Castagné R, Campanella G, Portengen L, **Kelly RS**, Bergdahl IA, Melin B, Hallmans G, Palli D, Krogh V,Tumino R, Sacerdote C, Panico S, de Kok TM, Smith MT, Kleinjans JC, Vineis P, Kyrtopoulos SA; on behalf of the EnviroGenoMarkers project consortium.
Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. *Annals of Oncology* 2014. 10.1093

4. Saberi Hosnijeh F, Peeters P, Romieu I, **Kelly R**, Riboli E, Olsen A, Tjønneland A, Fagherazzi G, Clavel-Chapelon F, Dossus L, Nieters A, Teucher B, Trichopoulou A, Naska A, Valanou E, Mattiello A, Sieri S, Parr C, Engest D, Skeie G, Dorronsoro M, Berricarte A, Sánchez, Ericson U, Sonestedt, Beuno-de-Mesquita H, Ros M, Travis R, Key T, Vineis P, Vermeulen R. Dietary Intakes and Risk of Lymphoid and Myeloid Leukemia in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Nutrition and Cancer*.2013. 66;1, 14-28

5. **Kelly RS,** Lundh T, Porta M, Bergdahl I, Palli D, Johansson A, Botsivali M, Vineis P, Vermeulen R, Kyrtopoulos S, Chadeau-Hyam M. Blood Erythrocyte Concentrations of Cadmium and Lead and the Risk of B-cell Non-Hodgkin's Lymphoma and Multiple Myeloma: A Nested Case-Control Study. *PLoS ONE* 8(11): e81892. doi:10.1371/journal.pone.0081892

6. Berndt SI, Skibola CF, Joseph V, Camp NJ, Nieters A, Wang Z, Cozen W, Monnereau A, Wang SS, **Kelly RS**, *et al*. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nature Genetics 2013; **45**, 868-876

7. Legood R; Sadique Z; Patnick J; Kitchener H; **Kelly R**; Moss S. Cost and logistics of alternative roll-out options for implementing human papillomavirus testing as a triage in cervical screening: results of the sentinel sites study. British Journal of Cancer 2012; 107, 1574–1579

8. **Kelly RS,** Patnick J, Kitchener HC and Moss SM on behalf of the NHSCSP HPV Special Interest Group. Human Papillomavirus Testing as a Triage for high grade cytological abnormalities: Evidence from the Sentinel Sites Study. British Journal of Cancer 2011; 105(7): 983–988.

9. **Kelly RS**, Walker P, Kitchener H, Moss S. Incidence of cervical intraepithelial neoplasia grade 2 or worse in colposcopy-negative/human papillomavirus-positive women with low-grade cytological abnormalities. The British Journal of Obstetrics and Gynecology 2011; 119: 20–25

10. Blanks RG and **Kelly RS.** Comparison of cytology and histology results in English cervical screening laboratories before and after liquid-based cytology conversion: do the data provide evidence for a single category of high-grade dyskaryosis? Cytopathology 2010; 21: 368-373

Under submission/revision:

11. Nieters A, Luczynska A, Becker S, Becker N, Vermeulen R,....,**Kelly RS**, Riboli E, Vineis P, Kaaks R. Association between prediagnostic total and specific Immunoglobulin E levels and risk of lymphoma – results of the EPIC study. *Cancer Research*

12. Sungalee S, Mamessier E, Morgado E…. **Kelly RS**, Vineis P, Ruminy P, Chetaille B , Jaffe E, Schiff C, Hardwigsen J, Tarte K, Roulland S, Bertrand Nadel. Iterative germinal center re-entries of memory B-cells with t(14;18) translocation drive early steps of follicular lymphoma progression. *Journal of Clinical Investigation*

13. Stringhini S, Polidoro S, Sacerdote C, **Kelly RS,** van Veldhoven K, Agnoli C, Grioni S, Tumino R, Giurdanella M, Panico S, Mattiello A, Palli D, Gallo V, Castagne R, Paccaud F, Campanella G, Chadeau-Hyam M, Vineis P. Association of life-course socioeconomic status with DNA methylation of genes regulating inflammation. *PLOS Medicine*

14. Cerhan J, Berndt S, Vijai J….**Kelly RS**…..Chanock S. Genome-wide association study identifies multiple susceptibility loci for diffuse large B-cell lymphoma. *Nature Genetics*

15. Kaaks R, Sookthai D, Luczynska A, Oakes C, Becker S, Johnson T, Johansson A, Melin B, Sjöberg K, Trichopoulou A, Lagiou P, Trichopoulos D, Mattiello A, Tumino R, Masala G, Agnoli C, Boeing H, Aleksandrova K, Brennan P, Franceschi S, Roulland S, Casabonne D, de Sanjose S, Sánchez M, Huerta J, Ardanaz E, Overvad K, Tjønneland A, Halkjær J, Weiderpass E, Bueno-de-Mesquita H, Vermeulen R, Peeters PH, **Kelly RS**, Vineis P, Khaw K, Travis R, Key TJ, Riboli E and Nieters A. Plasma soluble CD23 (sCD23) as pre-diagnostic marker for chronic lymphocytic leukemia (CLL) – a prospective analysis within the European EPIC cohort. *Clinical Cancer Research*

16. Vijai J, Wang Z, Berndt SI,…..,**Kelly RS**,…., Nieters A. A Genome-wide association study of marginal zone lymphoma shows association to the HLA region. *Nature Genetics*

17. Skibola CI, Berndt SI, Vijai J,…., **Kelly RS**,…., Rothman N. Genome-wide association study identifies new follicular lymphoma susceptibility loci. *Nature Genetics*

Agency Reports:

1. Moss SM, **Kelly RS**, Legood R, Sadique Z, Canfell K, Lew JB, Smith M, Walker R. Evaluation of HPV triage and test of cure: Report to the NHS Cancer Screening Programmes. 2011 (http://www.cancerscreening.nhs.uk/cervical/sentinelfinalreport.pdf)

Magazine Articles:

1. **Kelly RS**. From individual to collective cancer cell invasion. The Interactive Education Unit's Learning and Teaching newsletter. Institute of Cancer Research. 2009.

CONFERENCE PROCEEDINGS:

1. 2013 ASH (American Society of Haematology) Annual Meeting and Exposition, New Orleans 7-10 December 2013. Circulating t(14;18) Cells as Predictive Markers of Follicular

Lymphoma Development. Roulland S, Morgado E, **Kelly R**, Sungalee S, Colombat P, Solal-Celigny, Salles A, Vineis P, Nadal B

2. 9th Annual Conference of the Metabolomics Society, Glasgow 1-4 July 2013. Metabolic profiling of prospective human cohorts to identify biomarkers of exposure and disease: signatures of Cd exposure. Siskos AP, **Kelly RS**, Kamleh MA, Sood D, Vineis P, Lundh T, Keun H, and Kyrtopoulos S on behalf of the EnviroGenoMarkers Consortium

3. Inaugural UK and Ireland Exposure Science Meeting. Edinburgh 14 March 2013. Metabolic profiling of a Non-Hodgkin's Lymphoma case-control study to identify biomarkers of exposure and disease: Results from the EnviroGenoMarkers Study. **Kelly RS**, Chadeau-Hyam M, Kamleh MA, Keun H, Vineis P and Kyrtopoulos S on behalf of the EnviroGenoMarkers Consortium

4. 6th UK and Ireland Environmental and Occupational Epidemiology Meeting, LSHTM, UK 27 March 2012 Blood measurements of heavy metals, PCBs and DDT/DDE as predictive biomarkers of breast cancer and Lymphomas: The EnviroGenoMarkers Project. **Kelly RS**, Chadeau-Hyam M, Vineis P.

5. 21st Asia Pacific Cancer Conference: The science of strategy and cancer control; Kuala Lumpur. Malaysia 10-12 November 2011. Atopy, allergy and autoimmune conditions, polymorphisms in inflammatory pathway genes and non-hodgkin's lymphoma. Lim W, **Kelly RS**, Chen Y, Seow A and collaborators of the Singapore Lymphoma Study

6. From Exposure to Health Effects: Novel Approaches to Find the Linkage; University of Birmingham, UK 17th May 2011, The EnviroGenoMarkers Project: Design and Methodology **Kelly RS** et al. on behalf of the EGM Consortium

7. 26th International Papilloma Conference, Montréal, Canada 3-8 July 2010; Is switching from repeat cytology to HPV triage a good idea? Sadique Z, Moss S, **Kelly RS**, Kitchener H, Legood R

8. The British Society for Colposcopy and Cervical Pathology Annual Conference, Dublin 7-8 May 2009 Incidence of cervical intraepithelial neoplasia grade 2 or worse in colposcopy-negative/human papillomavirus-positive women with low-grade cytological abnormalities. **Kelly RS**, Walker P, Kitchener H, Moss S.

Invited Presentations:

2013 'Novel Strategies for the Identification of Predictive Biomarkers: Lessons from Non-Hodgkin Lymphoma' Harvard School of Public Health, Boston, USA

2010 'Epidemiology; The most important science of all?' Café Scientifique, Tampere, Finland

PATENTS:

Nadal B, Roulland S, Vineis P, **Kelly RS.** METHODS FOR PREDICTING WHETHER A SUBJECT IS AT RISK OF DEVELOPING A FOLLICULAR LYMPHOMA France: 12118, filed December 2013. Patent Pending