Int J Mol Epidemiol Genet 2011;2(2):95-113
www.ijmeg.org /ISSN:1948-1756/IJMEG1102002

Original Article

Design and validity of a clinic-based case-control study on the molecular epidemiology of lymphoma

James R Cerhan[1], Zachary S. Fredericksen[2], Alice H. Wang[2], Thomas M. Habermann[3], Neil E. Kay[3], William R. Macon[5], Julie M. Cunningham[6], Tait D. Shanafelt[3], Stephen M. Ansell[3], Timothy G. Call[3], Thomas E. Witzig[3], Susan L. Slager[2], Mark Liebow[4]

[1]Division of Epidemiology and [2]Division of Biostatistics, Department of Health Sciences Research; [3]Division of Hematology and [4]Division of General Internal Medicine, Department of Medicine; [5]Division of Hematopathology and [6]Division of Experimental Pathology, Department of Laboratory Medicine and Pathology; College of Medicine, Mayo Clinic, Rochester, Minnesota, USA.

Received February 17, 2011; accepted April 3, 2011; Epub April 5, 2011; published May 15, 2011

Abstract: We present the design features and implementation of a clinic-based case-control study on the molecular epidemiology of lymphoma conducted at the Mayo Clinic (Rochester, Minnesota, USA), and then assess the internal and external validity of the study. Cases were newly diagnosed lymphoma patients from Minnesota, Iowa and Wisconsin seen at Mayo and controls were patients from the same region without lymphoma who had a pre-scheduled general medical examination, frequency matched on age, sex and residence. Overall response rates were 67% for cases and 70% for controls; response rates were lower for cases and controls over age 70 years, cases with more aggressive disease, and controls from the local area, although absolute differences were modest. Cases and controls were well-balanced on age, sex, and residence characteristics. Demographic and disease characteristics of NHL cases were similar to population-based cancer registry data. Control distributions were similar to population-based data on lifestyle factors and minor allele frequencies of over 500 SNPs, although smoking rates were slightly lower. Associations with NHL in the Mayo study for smoking, alcohol use, family history of lymphoma, autoimmune disease, asthma, eczema, body mass index, and single nucleotide polymorphisms in *TNF* (rs1800629), *LTA* (rs909253), and *IL10* (rs1800896) were at a magnitude consistent with estimates from pooled studies in InterLymph, with history of any allergy the only directly discordant result in the Mayo study. These data suggest that this study should have strong internal and external validity. This framework may be useful to others who are designing a similar study.

Keywords: Case-control study, etiology, lymphoma, molecular epidemiology, validity

Introduction

Molecular epidemiology approaches have revolutionized the field of cancer epidemiology [1-4], and this has led to the need for efficient and timely collection of biologic specimens from participants, including blood (for serum, plasma, DNA, and cells), urine, and tumor tissue (including paraffin-embedded and fresh frozen tissue) on large numbers of cases and controls. The added complexity of collecting biologic samples on a population basis ranges from relatively easy (buccal samples) to difficult (pre-treatment peripheral blood sample) to extremely difficult in most settings (frozen tumor tissue). The need for high-quality biologic specimens has in part driven a renewed interest in the use of hospital- or clinic-based study designs for molecular epidemiology studies.

There are several advantages of clinic- or hospital-based designs. While all types of epidemiologic studies appear to be subject to declining participation rates, this is most acute for controls in population-based case-control studies [5]. Further, more aggressive cases, who may die quickly, can be more rapidly identified and enrolled into a clinic- or hospital-based study; for example, in recent population-based non-Hodgkin lymphoma (NHL) case-control studies, 10 to 20% of eligible cases were dead by the time of ascertainment, and therefore could not be enrolled into the study [6-10]. Specimen collection can be more standardized and under closer quality control in the clinical setting, and research samples can often be collected in

combination with routine clinical samples. While standardized collection is somewhat less critical for DNA, it is important for serum collection, and extremely important for the collection of diagnostic tumor tissue. Pre-treatment (but not pre-disease) serum can also be efficiently collected, which is extremely difficult to do on a population basis, where cases are often not identified until after therapy has been initiated.

The availability of serum and tumor tissue is likely to become even more critical as new methods of diagnosing and classifying tumors at the molecular level (e.g., through the use cDNA arrays, mRNA expression, microRNA; proteomics, methylation status, etc) are developed [11, 12] and can be adapted into molecular epidemiology studies. Central pathology review, which is becoming more important as tumor heterogeneity is integrated into molecular epidemiology studies, can be coordinated more efficiently in a clinical setting. All of these advantages in patient enrollment and biospecimens collection accrue at considerable cost efficiency over that of population-based studies.

Case-control studies, while being highly efficient for studying disease etiology, are subject to a greater potential for several key biases, most prominently selection and recall biases [13, 14]. This concern is further heightened for hospital- or clinic-based studies, particularly for bias related to the use of referral cases and in the selection of controls [15, 16]. While internal validity is the paramount concern in evaluating study findings, external validity also warrants assessment. For population-based case-control studies, the external validity (generalizability) is directly related to the underlying source population for cases and controls, and is related to the nature of the sampling frame from the source population, participation rates, and the characteristics of the non-participants [17]. For a clinic-based study, these considerations along with the features of the underlying study base need to be addressed.

We present the design features of a clinic-based molecular epidemiology study of lymphoma conducted at the Mayo Clinic in Rochester, Minnesota, with a focus on an assessment of both internal and external validity. We used the study base framework developed by Miettinen [18] and advocated by Wacholder et al. [14] to design and evaluate the internal validity of our study. To evaluate external validity, we compared characteristics of our cases and controls to population-based data. This framework may be useful for others who are considering designing a clinic- or hospital-based study.

Materials and methods

Human subjects review

This study was reviewed and approved by the Human Subjects Review Board at the Mayo Clinic, and all participants provided written, informed consent.

Study base principle

In its most basic conceptualization, a study base can be defined as a source population or hypothetical cohort that gives rise to cases during a specified period of time [14, 18]. There is complete flexibility in defining the hypothetical cohort, but a useful distinction is the primary versus secondary study base. A primary study base is defined *a priori*, generally based on a geographically defined area over a particular period of calendar time (e.g., classic population-based case-control study), and the main task is to find all (or a sub-sample) of the cases along with a sample of controls from the study base. In contrast, for a secondary study base (e.g., a classic hospital-based case-control study), cases are defined *a priori* (e.g., all cases of newly diagnosed NHL seen at the Mayo Clinic over a certain time frame) and the definition of the study base is secondary to the case selection. For a secondary base, the main task is to define (reconstruct) the study base in order to validly select controls.

Case selection

A first step of a case-control study using a secondary base is to define case eligibility. Starting on 9/1/02, we offered enrollment to all consecutive cases of histologically-confirmed Hodgkin (HL) and non-Hodgkin (NHL) lymphoma (ICD-O-3 codes 9590-9729) who met the following inclusion criteria: 1) aged 20 years and older; 2) a resident of Minnesota, Iowa or Wisconsin at the time of diagnosis; 3) within 9 months of their initial diagnosis at presentation to Mayo Clinic Rochester; 4) no history of HIV infection; 5) English-speaking; and 6) able to provide written informed consent. A Mayo hematopathologist reviewed all materials for each case to verify the diagnosis and to classify each case into

the World Health Organization (WHO) Classification of Neoplastic Diseases of the Hematopoietic and Lymphoid Tissues [19]. The WHO classification included chronic lymphocytic leukemia (CLL) as a NHL subtype that is classified with small lymphocytic leukemia (SLL). We grouped lymphoma subtypes using the approach developed by InterLymph [20].

Two practical considerations were whether the initial diagnosis needed to be made at Mayo and what was an allowable lag time between diagnosis and enrollment. We included diagnoses made at both Mayo and from outside the institution, since all outside diagnoses were clinically verified (including pathology review) at Mayo. This was a trade off to increase sample size balanced against the potential for referral bias and the increased difficulty of collecting specific types of biologic specimens (e.g., tumor tissue, pre-treatment serums) from patients diagnosed outside of Mayo. With respect to time from diagnosis, we used 9 months as an interval between the date of first clinical encounter leading to the NHL diagnosis and the date of consent. It was not practical to enroll all patients on the day of diagnosis, but any delay increased the potential for incidence-prevalence bias [13], recall bias (e.g., disease and treatment experience impacting recall of exposures prior to diagnosis), and biospecimen collection issues (e.g., collection of serum before the initiation of therapy). In the setting of lymphoma, a 9 month timeframe from diagnosis to consent seemed a reasonable compromise for enrollment eligibility.

Another consideration was whether to restrict the cases to a certain distance or geographic region around Rochester, Minnesota (location of Mayo Clinic). This consideration was two-fold. First, cases ascertained from a farther distance increased the probability of introducing a referral bias, such that cases from greater distances became less representative of all cases of the disease in the target population based on disease characteristics, clinical complexity (requiring tertiary evaluation), or socioeconomic factors. This could have a particularly strong impact on the external validity of the study results. Second, we were concerned that it would become more difficult to reconstruct the study base for valid control selection if all cases seen at Mayo were utilized. It is critical to understand the fundamentals of the Mayo Rochester clinical practice. There is a major primary care practice for Olmsted County (location of Mayo Clinic) and the surrounding counties, a larger secondary care practice for the 120-mile region surrounding Rochester, and then a tertiary care practice that is based on referral patients from the region, the United States, and other countries. We decided to restrict case and control selection to the regional practice to reduce the potential impact of referral bias, maintain access to a sufficient number of cases (power considerations), and take advantage of the large regional general medicine practice in Rochester for controls (discussed below). We defined the regional practice as residents of Minnesota, Iowa, and Wisconsin, as these 3 states make up the majority of lymphoma patients and the general medicine practice. This restriction is valid if it is applied equally to cases and controls (simply refining of the scope of the study base).

Control selection

Having specified a secondary base to define our case population, the next step was to select controls from this study base. Most fundamentally, controls must represent the underlying exposure distribution (genetic and environmental) of the study base that generated the cases [14, 18]. Selection of population-based controls could introduce bias related to use of and access to medical care (i.e., utilization of Mayo Clinic). We therefore selected controls from the general medicine divisions of Mayo Clinic due to the large number of patients who are seen for pre-scheduled general medical examinations. The eligibility criteria were: 1) age 20 years and older; 2) a resident of Minnesota, Iowa or Wisconsin at the time of appointment at Mayo; 3) no history of lymphoma or leukemia; 4) no history of HIV infection; 5) English-speaking; and 6) able to provide written informed consent. Controls were frequency matched to the regional case distribution on 5-year age group, sex, and geographic area using a computer algorithm that randomly selects subjects from eligible patient appointments. The 3 state region was divided into 12 geographic areas: Minnesota (Olmsted County; southeast rural; southwest rural; north rural; all other rural; central urban; north urban), Iowa (within 120 miles of Rochester; >120 miles and rural; >120 miles and urban), and Wisconsin (within 120 miles of Rochester; >120 miles and rural; >120 miles and urban). Urban/rural status was based on 2000 census categorization.

We chose to use clinic-based controls since other patients seen at the same institution (Mayo) constitute a sample, albeit not a random sample, of our study base. Non-random selection is a reasonable alternative to use in this situation, as long as similar catchment areas are used and control selection is independent of the exposure under study [15]. With respect to catchment areas, we applied the 3 state restrictions to controls, and further applied a frequency matching system to balance controls on urban/rural status and distance from Rochester. This was a practical solution for defining the catchment, although there remains a potential for bias if the catchment differs for NHL cases and the general medicine controls. Furthermore, by including geographic region (a function of distance from Rochester and urban/rural status) as one of the frequency matching factors and controlling for this variable in the analysis, we are able to decrease confounding and increase the internal validity of the study with respect to correlates of any referral patterns to Mayo [15].

With respect to the independence of the exposure under study from selection of controls, most authors recommend selecting controls across many diagnoses, and excluding those controls whose active diagnosis was the reason for the current visit (and therefore selection) and was related to the exposure under study [13, 15]. Because our study was focused on assessing multiple hypotheses related to genetic susceptibility, lifestyle and environmental factors, and gene-environment interactions, it was not possible to pick diagnosis categories to include or exclude in general. We therefore elected to study patients presenting for prescheduled general medical examinations, as this avoided a specific active diagnosis leading to selection, as the appointments are prescheduled several weeks to months in advance and are general in nature (i.e., no specific active diagnosis *per se* leading to selection). In addition, there was a very large local and regional practice available from which to select controls, with an electronic list of appointments available that allowed pre-selection based on age, sex and residence. Of note, we did not exclude controls due to a history of certain diseases or conditions, as the exclusion only applies to the cause of the clinic visit used to select a control [13, 15, 21].

The general medicine practice at Mayo is dynamic (open), and so controls were selected from updated lists generated weekly throughout the study. Controls were eligible to become cases, and control exposure data (questionnaire, medical record and serum) were all from the point of time at selection, and no information after enrollment was included. This mirrors the incidence-density sampling approach used in studies of a primary study base (e.g., population-based case-control studies and case-control studies nested in defined cohorts), so that controls are validly selected such that "the exposure distribution among the controls is, apart from random error, the same as it is among the person-time in the population that is the source of the cases" [13].

Risk factor data collection

All participants were requested to complete a self-administered risk-factor questionnaire that included data on demographics, ethnicity, family cancer history, medical history, lifestyle and other putative NHL risk factors, much of which was adapted from previous studies of NHL [6, 22] or other cancer studies (http://dceg.cancer.gov/QMOD). Women were asked to complete a questionnaire focused on menstrual and reproductive health, which was based on questions derived from the Women's Health Initiative (www.whiscience.org/data/forms/F31v2_1.pdf). Participants who had lived or worked on a farm or with pesticides for more than one year were asked to complete a questionnaire on farming and pesticide exposures. A majority of the items on this questionnaire were derived from the Agricultural Health Study (http://aghealth.nci.nih.gov/), and focused on details of farming history (e.g., years farmed, size of farm, crops grown, livestock raised, etc.) and personal mixing and application of a variety of pesticides (e.g., herbicides, insecticides, fumigants, and fungicides).

A socioeconomic status score based on the occupational titles provided by participants for their longest held occupation was calculated according to a standardized score developed by Green [23] for use in research on health behavior. The scores among controls were summed to determine quintile cutpoints.

Biologic specimen collection

A peripheral blood sample was collected for serologic and genetic studies. Serum was proc-

essed and aliquoted into tubes and stored at -70°C. Serum samples were classified as pre-treatment (i.e., blood sample collected before initiation of therapy) or not (i.e., blood sample collected during or after treatment). DNA was extracted from samples using a Gentra Systems automated salting-out methodology (Gentra Inc., Minneapolis, MN). Tumor tissue, both paraffin-embedded and fresh frozen, was collected through the University of Iowa/Mayo Clinic Lymphoma Specialized Program of Research Excellence (SPORE) [24] according to a standardized protocol. For diagnostic tissue not at Mayo, we requested outside slides and blocks.

Genotyping data

Genotyping data: from a ParAllele (now Affymetrix) Immune and Inflammation SNP panel (9412 SNPs from 1253 genes) supplemented with a custom Illumina GoldenGate 384 SNP OPA (from 100 genes) was available on 441 cases and 475 controls enrolled through October 2005, as previously described [25, 26].

Other data sources

Mayo Cancer Registry: The Mayo Cancer Registry abstracts newly diagnosed cancer patients seen at Mayo Clinic Rochester for reporting to the Minnesota Cancer Surveillance System (Minnesota's population-based cancer registry) and the American College of Surgeons (ACoS) Commission on Cancer (www.facs.org/cancer), for which it is fully accredited. We linked our cases to the Mayo Cancer Registry to obtain case class (based on place of diagnosis – Mayo, outside Mayo; and place of treatment – Mayo, outside Mayo) and lymphoma subtype as abstracted by the registry. We also obtained the same data in anonymized format on subjects who did not enroll into the study; patients for this group who declined research authorization for the state of Minnesota (<5%) were excluded.
SEER Data: We obtained publically available incidence data from the Surveillance, Epidemiology and End-Results (SEER) program (SEER 2010 release) for the 17 registries limited use + Hurricane Katrina Impacted Louisiana cases November 2009 submission (released April 2010). We used SEER*Stat to obtain characteristics of NHL cases (including CLL/SLL) aged 20 to 79 years and diagnosed from 2002 to 2007 (inclusive) for both the Iowa SEER Registry and all 17 SEER Registries. We also obtained 12 month observed survival on these cases.

NCI-SEER Case-Control Study: We obtained data for controls from the Iowa component of the NCI-SEER Case-Control Study of NHL, which has been previously described [6, 27]. Briefly, population controls were identified by random digit dialing (under age 65 years) and from Medicare eligibility files (65 years and older), and were frequency matched to the case distribution on age, sex, and race. For the Iowa component of the study, of the 478 controls that were selected, 6 died before interview (1.3%), 36 were not locatable (7.5%), 15 were too ill or cognitively impaired (3.1%), 145 refused (30.3%), and 276 participated (58%). Demographic and anthropometric data were collected on all participants, while smoking was collected on a 50% random sample.

DNA was extracted from either a blood (N=242) or buccal (N=33) sample, which was collected on 275 (99%) of the controls. Data on 257 controls from the Iowa dataset were available from a custom-designed Infinium assay (Illumina, www.illumina.com) that included 6679 successfully genotyped SNPs; full details on the genotyping and quality control are available elsewhere [28]. Of these SNPs, 617 SNPs were genotyped in both studies, 514 with a MAF of 10% or higher.

Data analysis

For group comparisons, we compared percent distributions and mean and medians. Group differences were tested using t-test (for means) and chi-square (for contingency tables). Due to the relatively large sample size that make trivial differences statistically significant at $p<0.05$, we focused our interpretation on the effect size of group differences. To evaluate the association of genetic and lifestyle factors with risk of NHL, we calculated odds ratios (OR) and 95% confidence intervals (CI) using unconditional logistic regression. We adjusted for the design variables of age, sex and residence. All statistical tests were two-sided, and all analyses were carried out using SAS (SAS Institute, Inc., Cary, NC).

Results

Case enrollment

Participation: This analysis included all Mayo lymphoma (NHL and HL) patients enrolled into the study from 9/1/02 through 2/28/07; par-

Table 1. Comparison of participants versus non-participants, case patients

	Regional Cases (3 State)				
	Participants		Non-Participants		
Characteristic	N	%	N	%	p-value
Sex					
Male	560	58.9%	264	56.4%	0.37
Female	391	41.1%	205	43.7%	
Age Distribution					
≤40	97	10.2%	52	11.1%	<0.01
41-50	148	15.6%	71	15.1%	
51-60	197	20.7%	72	15.4%	
61-70	281	29.5%	118	25.1%	
71+	228	24.0%	156	33.4%	
Case class (American College of Surgeons)†					
Diagnosis Mayo, Treatment Mayo	412	48.8%	197	47.5%	<0.01
Diagnosis Mayo, Treatment Outside	32	3.8%	29	7.0%	
Diagnosis Outside, Treatment Mayo	179	21.2%	57	13.7%	
Diagnosis Outside, Treatment Outside	221	26.2%	131	31.6%	
Not classified	107		55		
Lymphoma Subtype†					
CLL/SLL	228	27.0%	47	12.8%	<0.01
Follicular	196	23.2%	59	16.1%	
DLBCL	173	20.5%	123	33.5%	
Marginal Zone	48	5.7%	33	9.0%	
Mantle Cell	34	4.0%	12	3.3%	
T-Cell	38	4.5%	26	7.1%	
Other/NOS	58	6.9%	22	6.0%	
Hodgkin lymphoma	70	8.3%	45	12.3%	
Not available	106		102		

†As abstracted by the Mayo Cancer Registry

ticipation was defined as consenting within 9 months of diagnosis and either completing a risk factor questionnaire or providing a blood sample within 12 months of diagnosis. Of the 1420 eligible patients identified during this time frame, 951 (67%) participated, 148 (10%) refused, 23 (2%) could not be contacted, and 298 (21%) had their eligibility expire (i.e., after identification they did not consent within 9 months of diagnosis or after consent they did not complete data collection within 12 months of diagnosis). The median time from diagnosis to consent was 32 days, the 10th percentile was 5 days, and the 90th percentile was 153 days.

Impact of non-participation: There were statistically significant differences ($p<0.05$) between participants and non-participants for age group, case class and lymphoma subtype, but not sex distribution (**Table 1**). While statistically significant, the absolute differences were relatively modest. For example, the age distribution for non-participants was shifted towards the older age groups, particularly for the oldest age group (age 71 years and older). The distribution of case class was roughly similar between participants and non-participants, although participants where more likely to receive their initial treatment at Mayo (70%) compared to non-participants (60%). Differences were more pronounced for several NHL subtypes as defined by the Mayo Cancer Registry, with participants more likely to have CLL/SLL (27% versus 13%) or follicular lymphoma (23% versus 16%) and less likely to have DLBCL (21% versus 34%) compared to non-participants.

Impact of early deaths: Higher survival at one-year was observed for participants (95%) compared to non-participants (85%) (**Table 2**) . This difference in one year survival rates between participants and non-participants was slightly higher for men (12%) compared to women (7.1%). For age, there was little difference in survival between participants and non-participants for the age groups 41-50 (0.1%)

Table 2. One-year survival for participants versus non-participants, case patients

	Participants		Non-Participants	
Characteristic	N	Survival	N	Survival
All	951	94.8%	470	85.3%
Sex				
Male	560	94.1%	265	82.6%
Female	391	95.9%	205	88.8%
Age Distribution				
≤40	97	99.0%	52	92.3%
41-50	148	95.9%	71	95.8%
51-60	197	95.4%	72	94.4%
61-70	281	94.3%	118	81.4%
71+	228	92.5%	157	77.1%
Case Class†				
Diagnosis Mayo, Treatment Mayo	412	94.9%	197	82.2%
Diagnosis Mayo, Treatment Outside	32	93.8%	29	82.8%
Diagnosis Outside, Treatment Mayo	179	92.7%	57	80.7%
Diagnosis Outside, Treatment Outside	221	94.1%	131	89.3%
Lymphoma Subtype†				
CLL/SLL	228	99.1%	47	85.1%
Follicular	196	95.9%	59	94.9%
DLBCL	173	86.7%	123	82.1%
Marginal Zone	48	97.9%	33	87.9%
Mantle Cell	34	94.1%	12	66.7%
T-Cell	38	86.8%	26	53.8%
Other/NOS	58	94.8%	22	81.8%
Hodgkin lymphoma	70	92.9%	45	91.1%

†As abstracted by the Mayo Cancer Registry

and 51-60 years (1%), but the difference then increased with age group, and the largest difference was observed for the oldest age group, 71+ years (15%). For case class, differences in one-year survival rates were approximately the same (11-13%) for all groupings except cases diagnosed and treated outside of Mayo, where the difference between participants and non-participants was smaller (5%). Finally, for lymphoma subtypes, the smallest differences in one year survival rates between participants and non-participants were observed for follicular (1%), DLBCL (5%) and HL (2%), while the largest differences were observed for CLL/SLL (14%), mantle cell (27%), and T-cell (33%) lymphomas, although the latter two subtypes each had a small number of non-participants (≤26). These data suggest that non-participants appear to have slightly more aggressive disease (impacting survival) than participants, although the absolute impact was relatively modest (10% range).

Control enrollment

Participation: This analysis included all controls enrolled into the study from 9/1/02 through 2/28/07; participation was defined as enrolling within 9 months of selection and either completing a risk factor questionnaire or providing a blood sample within 12 months of selection. For the case-control study, we restricted cases to the 3-state region, and then frequency matched the control group based on age, sex, and geographic area (county groupings based on distance from Rochester, MN and urban/rural status; see methods section). Of the 1737 eligible controls identified, 1209 (70%) participated, 500 (29%) refused and 27 (1%) had their eligibility expire (i.e., did not complete data collection within 12 months of selection). Control participation (70%) was similar to case participation (67%), but the refusal rate for controls (30%) was much higher than cases (10%), with the remaining difference due to a small percent-

Table 3. Comparison of participants and non-participants, controls

Variable	Participants (N=1209)		Non-participants (N=527)		p-value
	N	%	N	%	
Age at selection, years					0.023
≤40	100	8.3%	38	7.2%	
41-50	166	13.7%	69	13.1%	
51-60	248	20.5%	108	20.5%	
61-70	366	30.3%	130	24.7%	
71+	329	27.2%	182	34.5%	
Mean Age (± SD)	60.8 ± 13.7		62.5 ± 15.0		0.018
Sex					0.009
Male	658	54.4%	251	47.6%	
Female	551	45.6%	276	52.4%	
Residence					0.53
Minnesota	243	20.1%	100	19.0%	
Iowa	803	66.4%	364	69.1%	
Wisconsin	163	13.5%	63	12.0%	
Distance from Rochester, MN					0.012
Olmsted County	141	11.7%	80	15.2%	
Outside Olmsted, <50 miles	249	20.6%	129	24.5%	
50-119 miles	471	39.0%	203	38.5%	
120-249 miles	311	25.7%	104	19.7%	
250+ miles	37	3.1%	11	2.1%	
Mean distance ± SD (miles)	91.0 ± 73.4		78.6 ± 70.4		0.001
Density					0.61
Urban	482	39.9%	217	41.2%	
Rural	727	60.1%	310	58.8%	
Mayo Characteristics					
Time in system ± SD (years)	22.2 ± 17.4		24.0 ± 17.9		0.047
Total visits ± SD	32.8 ± 30.2		35.8 ± 34.4		0.063
Year of first visit ± SD	1984.2 ± 16.0		1982.3 ± 16.7		0.027

age of controls with 'eligibility expired' status (1%) versus cases (21%). Three controls have subsequently developed lymphoma, but after 2/28/07.

Impact of control non-participation: With exception of state of residence and urban/rural density, all other characteristics showed statistically significant differences ($p<0.05$) between participants and non-participants (**Table 3**). However, the absolute differences were relatively modest. Compared to participants, non-participants were on average 1.7 years older, lived 12.4 miles closer to Mayo, were Mayo patients 1.8 years longer, and had 3 more total visits to Mayo prior to clinical contact that led to enrollment. Non-participants were also slightly more likely to be female (52%) compared to participants (46%).

Data and specimen collection

The main questionnaire was completed by 82% of the cases and 87% of the controls (**Table 4**). The female reproductive questionnaire was completed by 86% of female cases and 88% of female controls. Participants who worked on a farm or with pesticides for more than one year were eligible to complete a Farming and Pesticide questionnaire, and completion rates were somewhat somewhat higher for controls (81%) than cases (74%). Blood samples were obtained for >95% of cases and controls, and for cases approximately 50% of the serum was obtained prior to the initiation of any therapy. Formalin-

Table 4. Questionnaire data and biologic specimen collection by case-control status

	Cases (N=951)		Controls (N=1209)	
	N	%	N	%
Questionnaires				
Main Questionnaire	780	82.0%	1053	87.1%
Female Reproductive†	335	85.7%	485	88.0%
Farming & Pesticide‡	215	74.1%	272	80.5%
DNA and serum sample				
No	37	3.9%	10	0.8%
Yes	914	96.1%	1199	99.2%
Tissue Blocks at Mayo				
No	357	38.1%		
Yes	579	61.9%		
Missing	15			

†Calculation of percentage is based on using a denominator for the number of female cases (N=391) and controls (N=551). ‡Calculation of percentage is based using a denominator for the number of participants who worked on a farm or with pesticides for more than one year for cases (N=290) and controls (N=338).

fixed, paraffin-embedded tissue blocks were housed at Mayo for 62% of the cases.

Additional considerations for internal validity

A key assumption of using non-random selection of controls from a secondary base is that the controls come from the same catchment area as the cases. As shown in **Table 5**, the case and control groups were well balanced on the design variables of age and residence characteristics, including state of residence, distance from Rochester, MN, and urban/rural status; there was a slight sex imbalance, with a higher percentage of female controls compared to cases (46% compared to 41%, $p<0.05$). The geographic distribution based on the county of residence is shown in **Figure 1**, and visually highlights the balance of cases and controls and provides support that the cases and controls likely represent the same geographic catchment area. While not matching factors, cases and controls were also well balanced on race, marital status, education, and SES. To assess the impact of health services factors, we compared cases and controls on health care utilization at Mayo (**Table 5**). Controls had more time in the Mayo system (defined as time from date of first Mayo visit to date of diagnosis/selection), 22 years versus 14 years, and also had more total visits 32.7 versus 17.3 (all comparisons $p<0.05$), which is consistent with using general medicine patients as a source for the control group.

We next compared the associations in the Mayo case-control study to published associations from seven pooled InterLymph analyses [29-35], of which five included Mayo data from Phase 1 (enrollment from 2002-2005, N=626 cases and N=572 controls). Given the much smaller sample size of this study relative to the pooled analyses, the direction and magnitude of the ORs, rather than statistical significance *per se,* was most relevant to compare. As shown in **Table 6**, the ORs identified in the Mayo study for the association of NHL with smoking, alcohol use, family history of lymphoma, history of autoimmune disease, BMI, and history of asthma or eczema were all very similar to the InterLymph pooled estimates, while the Mayo study was directly discordant only for history of any allergy (i.e., showed opposite associations). There was significant heterogeneity for the BMI pooled analysis by region of study, and for the North American studies (5 studies including Mayo Phase 1; total N=3545 cases and N=4752 controls) there was a weak positive association with BMI. Specifically, compared to a BMI of 18.5-24.99 kg/m^2 (reference group), risks were increased for <18.5 kg/m^2 (OR=1.03, 95% CI 0.68-1.56), 25.0-29.99 kg/m^2 (OR=1.17; 95% CI 1.05-1.29), and 30-39.99 kg/m^2 (OR=1.34; 95% 1.08-1.65), and then declined to the null for 40+ kg/m^2 (OR=1.00; 95% CI 0.69-1.46). The risk estimates in the Mayo study fell between the overall pooled analysis and the subset from North America. The Mayo study also had genotype data available, and the associa-

Table 5. Descriptive characteristics of matching and related demographic factors by case-control status

Variable	Cases		Controls		
	N	%	N	%	p-value
Age at diagnosis/enrollment (years)					0.23
≤40	97	10.2%	100	8.3%	
41-50	148	15.6%	166	13.7%	
51-60	197	20.7%	248	20.5%	
61-70	281	29.5%	366	30.3%	
71+	228	24.0%	329	27.2%	
Mean Age (± SD)	59.6 ± 14.1		60.8 ± 13.7		0.041
Sex					0.038
Male	560	58.9%	658	54.4%	
Female	391	41.1%	551	45.6%	
Residence					0.79
Minnesota	642	67.5%	803	66.4%	
Iowa	180	18.9%	243	20.1%	
Wisconsin	129	13.6%	163	13.5%	
Distance from Rochester, MN					0.075
Olmsted County	90	9.5%	141	11.7%	
Outside Olmsted, <50 miles	177	18.6%	249	20.6%	
50-119 miles	363	38.2%	471	39.0%	
120-249 miles	293	30.8%	311	25.7%	
250+ miles	28	2.9%	37	3.1%	
Mean distance ± SD (miles)	100.7 ± 75.8		91.0 ± 73.4		0.003
Density					0.22
Urban	404	42.5%	482	39.9%	
Rural	547	57.5%	727	60.1%	
Race					<0.01
White	890	99.1%	1177	98.3%	
All Other	8	0.8%	20	1.7%	
Don't know/refused	53		12		
Marital status					0.61
Married	632	81.2%	871	82.9%	
Widowed	52	6.7%	61	5.8%	
Divorced/Separated	52	6.7%	58	5.5%	
Never Married	42	5.4%	61	5.8%	
Missing	173		158		
Education					0.27
Less than high school graduate	43	5.6%	40	3.8%	
High School graduate/GED	182	23.5%	237	22.5%	
Some college/vocational school	216	27.9%	296	28.2%	
College graduate	154	19.9%	201	19.1%	
Graduate school+	179	23.1%	277	26.4%	
Missing	177		158		
Socioeconomic Status					0.49
Group 1 (lower)	85	10.9%	92	8.8%	
Group 2	253	32.4%	370	35.2%	
Group 3	163	20.9%	218	20.7%	
Group 4	200	25.6%	259	24.6%	
Group 5 (higher)	79	10.1%	112	10.7%	
Missing	171		158		
Mayo Characteristics					
Time in Mayo system	14.2 ± 17.5		22.2 ± 17.4		<0.01
Total visits	17.3 ± 30.3		32.8 ± 30.2		<0.01
Year of first visit (include selection)	1992.0 ± 16.6		1984.2 ± 16.0		<0.01

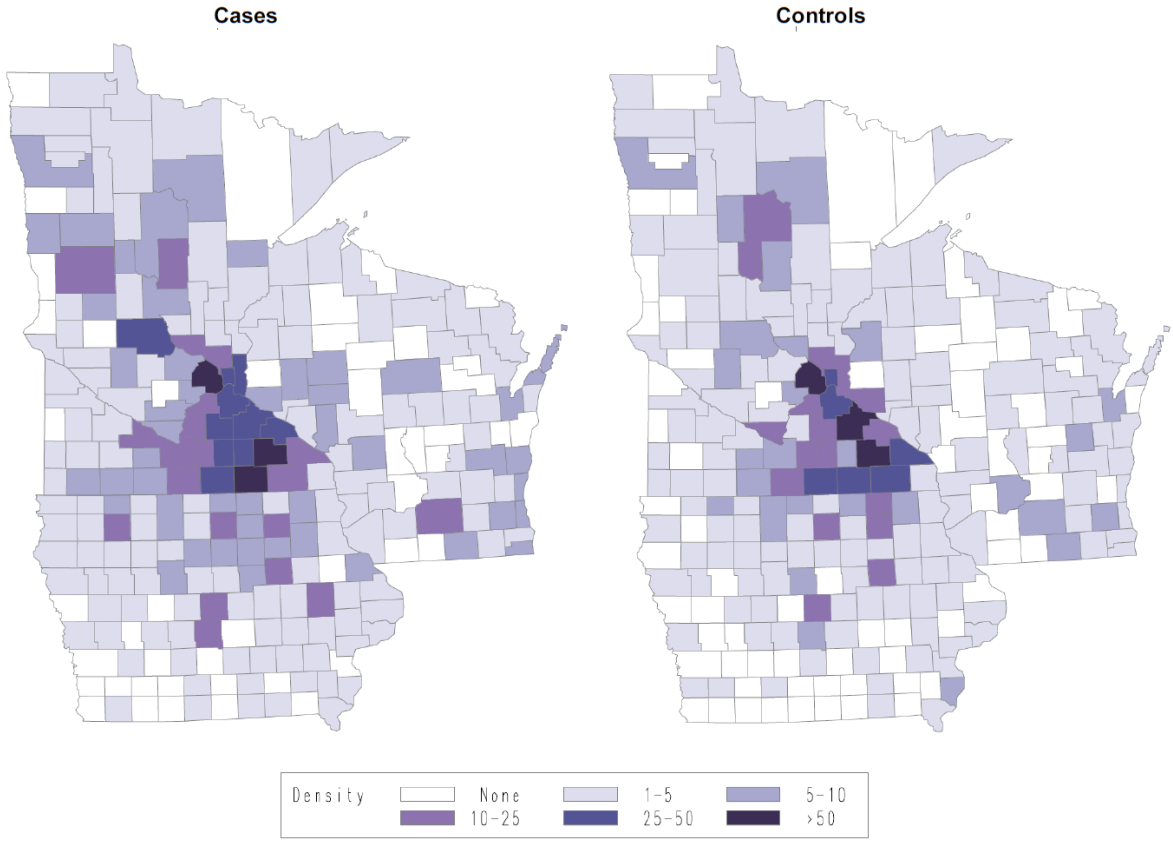

Figure 1. Spot map of the number of cases and controls from each county in Minnesota, Iowa and Wisconsin, Mayo Clinic Case-Control Study of NHL, 2002-2007.

tions for *TNF* rs1800629, *LTA* rs909253, and *IL10* rs1800896 were all consistent with the InterLymph estimates (**Table 6**).

External validity

Cases: To assess the external validity of the cases who participated in the Mayo study, we compared the sex, age and NHL subtype distribution of our NHL cases with SEER data for the upper Midwestern United States (Iowa SEER Registry) and for the U.S. (17 SEER Registries) (**Table 7**). Since we had few cases above age 79 years (4%), we restricted our comparisons to ages 20-79 years. There were slightly more males enrolled as cases (60%) compared to Iowa SEER data (56%) and the age distributions were shifted towards older age groups for the Iowa SEER data, particularly for the age group 70-79 years for Iowa SEER data (35%) compared to the Mayo study (23%). Mayo cases were slightly over-represented with CLL/SLL and follicular lymphoma and under-represented with DLBCL compared to Iowa SEER data, although absolute differences were modest (≤10%). The one-year survival of the Mayo cases (95%) was higher than Iowa SEER data (88%), and this difference in one-year survival between Mayo cases and Iowa SEER data was generally consistent across subgroups defined by sex, age, and NHL subtype. Comparison to national SEER data, which included all race/ethnicities, showed even lower one-year survival overall (84%) and this was also consistent across subgroups defined by sex, age, and NHL subtype.

Controls: We compared the prevalence of exposure for several key exposures with a population -based control group recruited from 1998-2000 in Iowa as part of the NCI-SEER case-control study [6]. As shown in **Table 8**, while Mayo controls were slightly younger and had a slightly

Table 6. InterLymph pooled estimates of associations for selected risk factors, and the same estimates from the Mayo Case-Control Study (Phases 1-2)

Exposure and Reference	Mayo in Published Analysis?*	N studies	Cases/ controls	Exposure Category	InterLymph Pooled Estimates OR	(95% CI)	Mayo Case-Control Study (Phases 1-2)† OR	(95% CI)
Smoking [30]	No	9	6594/ 8892	Never smoker	1.00	(reference)	1.00	(reference)
				Ever	1.07	(1.00-1.15)	1.10	(0.91-1.33)
				Current	1.10	(1.00-1.20)	1.03	(0.68-1.56)
				Pack-years				
				1 to 10	0.99	(0.90-1.08)	1.12	(0.84-1.49)
				11 to 20	1.04	(0.93-1.17)	0.95	(0.66-1.35)
				21 to 35	1.14	(1.02-1.27)	1.02	(0.71-1.45)
				36+	1.21	(1.09-1.34)	1.25	(0.91-1.71)
				Duration (years)				
				1 to 10	0.94	(0.84-1.06)	1.30	(0.93-1.81)
				11 to 20	1.02	(0.91-1.13)	0.95	(0.68-1.33)
				21 to 35	1.12	(1.02-1.23)	0.97	(0.73-1.29)
				36+	1.16	(1.05-1.28)	1.33	(0.97-1.83)
Alcohol [29]	No	9	6492/ 8683	Never drinker	1.00	(reference)	1.00	(reference)
				Ever drinker	0.83	(0.76-0.89)	0.84	(0.61-1.14)
				Current drinker	0.73	(0.64-0.84)	0.81	(0.59-1.11)
				Former drinker	0.95	(0.80-1.14)	0.99	(0.68-1.45)
				Frequency for ever (servings/week)				
				1 to 6	0.81	(0.74-0.88)	0.88§	(0.68-1.13)
				7 to 13	0.83	(0.74-0.92)		
				14 to 27	0.85	(0.76-0.95)	0.82	(0.63-1.06)
				28+	0.87	(0.76-0.99)	0.65	(0.46-0.90)
Family History [31]	Yes	11	10211/ 11905	No Family Hx	1.0	(reference)	1.00	(reference)
				NHL	1.5	(1.2-1.9)	1.92	(1.23-3.01)
				Any Heme	1.5	(1.3-1.6)	2.20	(1.62-2.98)
Autoimmune Diseases [32]	No	12	12982/ 16441	SLE	2.69	(1.68-4.30)	1.79	(0.39-8.14)
				Sjogren	6.56	(3.10-13.9)	5.70	(1.18-27.5)
				RA	1.06	(0.87-1.29)	1.26	(0.85-1.87)
Obesity [33]	Yes	18	10453/ 16507	<18.5	0.88	(0.71-1.08)	1.05	0.42-2.62
				18.5-24.99	1.00	(reference)	1.00	(reference)
				25-29.99	0.95	(0.85-1.07)	0.97	0.77-1.22
				30-39.99	0.97	(0.81-1.15)	1.11	0.86-1.44
				40+	0.99	(0.70-1.41)	0.64	0.33-1.27
Atopy [34]	Yes	13	13535/ 16388	Asthma	0.97	(0.84-1.11)	0.95	(0.69-1.29)
				Eczema	1.04	(0.87-1.25)	1.12	(0.94-1.32)
				Any allergy	0.80	(0.68-0.94)	1.20	(0.97-1.48)
SNPs‡ [35]	Yes	14	7999/ 8452	*TNF* -308G>A (rs1800629)				
				GG	1.00	(reference)	1.00	(reference)
				AG	1.12	(1.04-1.21)	1.00	(0.74-1.34)
				AA	1.34	(1.10-1.62)	2.14	(0.94-4.85)
				AG/AA	1.14	(1.06-1.23)	1.08	(0.81-1.43)
				LTA 252A>G (rs909253)				
				AA	1.00	(reference)	1.00	(reference)
				AG	1.01	(0.94-1.09)	1.11	(0.84-1.47)
				GG	1.12	(1.00-1.27)	1.20	(0.77-1.85)
				AG/GG	1.04	(0.96-1.11)	1.13	(0.87-1.47)
				IL10 1082A>G (rs1800896)				
				AA	1.00	(reference)	1.00	(reference)
				AG	1.06	(0.97-1.14)	1.06	(0.77-1.46)
				GG	1.08	(0.98-1.19)	0.87	(0.59-1.26)
				AG/GG	1.06	(0.98-1.15)	0.99	(0.73-1.35)

*Phase 1 Mayo case-control data included in the pooled analysis. †Adjusted for design variables (age, sex and geographic region). §Estimate is for 1-13 servings/week. ‡Phase 1 Mayo data only.

Table 7. Comparison of Mayo NHL cases ages 20-79 years to Iowa SEER and U.S. SEER data

	Mayo Cases			Iowa SEER Data			U.S. SEER Data		
	N	% Distribution	One-year Survival	N	% Distribution	One-year Survival	N	% Distribution	One-year Survival
All	831	n/a	95.1%	3,311	n/a	87.9%	67,900	n/a	83.9%
Sex									
Male	502	60.4%	94.6%	1,861	56.2%	88.0%	38,324	56.4%	82.5%
Female	329	39.6%	95.7%	1,450	43.8%	87.7%	29,576	43.6%	85.9%
Age Distribution									
20-39	45	5.4%	100.0%	203	6.1%	92.0%	5,555	8.2%	85.8%
40-49	128	15.4%	96.9%	357	10.8%	94.3%	8,965	13.2%	86.8%
50-59	185	22.3%	94.6%	679	20.5%	89.0%	15,323	22.6%	88.3%
60-69	283	34.1%	95.1%	923	27.9%	89.6%	18,393	27.1%	85.8%
70-79	190	22.9%	93.2%	1149	34.7%	83.0%	19,664	29.0%	76.9%
NHL Subtype*									
CLL/SLL	282	34.6%	99.3%	898	27.1%	95.3%	14,557	21.4%	93.1%
Follicular	207	25.4%	96.1%	689	20.8%	93.8%	12,167	17.9%	93.5%
DLBCL	147	18.0%	87.8%	926	28.0%	80.6%	20,143	29.7%	74.8%
Marginal Zone	51	6.3%	98.0%	249	7.5%	97.4%	5,264	7.8%	94.9%
Mantle Cell	38	4.7%	92.1%	111	3.4%	84.3%	1,989	2.9%	83.3%
T-Cell, NOS	39	4.8%	84.6%	225	6.8%	73.6%	5,953	8.8%	76.8%

*NOS/other subtypes not shown

higher percentage of men, the race distribution, prevalence of family history of lymphoma, and anthropometric characteristics were very similar. In contrast, there were more current smokers (16% versus 7%) and a higher percentage of long-term smokers (26% with 30+ years of smoking versus 14%) in the NCI-SEER Iowa controls. We also compared the difference in minor allele frequency (MAF) in the control group on 514 SNPs with a MAF >10% that were genotyped in both studies (**Figure 2**). No significant differences were observed (p-value = 0.31); the mean difference in minor allele frequency was 0.001 (SD 0.03).

Discussion

The clinic-based case-control study (or the more common hospital-based case-control study) is often considered to be highly susceptible to bias and to have lower external validity, although these concerns are not absolute and need to be evaluated for a specific study in the context of epidemiologic design principles and goals of the study [36]. While the Mayo study *per se* would not directly replicate to other clinical centers, the framework used for designing the study and evaluating both internal and external validity should be generalizable.

Internal validity

We have presented evidence that this case-control study has robust internal validity. Response rates were reasonably high for both cases (67%) and controls (70%). In many lymphoma studies, particularly population-based studies, the response rates are often much lower for controls compared to cases [6-10, 37], increasing the potential for bias, although similar response rates do not necessarily provide protection against bias [38]. Participation has been declining in epidemiologic studies, and the decline has been most sharp for population-based studies, particularly among controls, while declines in participation have not been as steep for hospital/clinic-based studies [5]. This study had very high collection rates for blood samples (96% for cases and 99% for controls). Biospecimen collection rates are often not reported; for example, a recent review [5] found that only 27% of studies reported the participation rate for the biologic component of their study.

We were further able to evaluate characteristics of non-participants compared to participants, and found differences on age, lymphoma subtype, and one-year survival for cases, and age,

Table 8. Comparison of Mayo and Iowa controls aged 20-74 years of age

Variable	Mayo Controls N	Mayo Controls %	Iowa Controls N	Iowa Controls %	P-value
Age, mean ± SD	57.4 ± 12.2		60.7 ± 11.4		<0.01
<40	87	8.6%	17	6.2%	0.0002
40-64	561	55.3%	121	43.8%	
65-74	367	36.2%	138	50.0%	
Sex					0.28
Male	552	54.4%	140	50.7%	
Female	463	45.6%	136	49.3%	
Race					0.19
Asian	5	0.5%	0	0.0%	
Any Black	5	0.5%	3	1.1%	
Any White	987	96.9%	271	98.2%	
Other/Unknown	21	2.1%	2	0.7%	
Family history of NHL					0.53
No	838	96.4%	264	95.7%	
Yes	31	3.6%	12	4.3%	
Anthropometrics					
Weight (pounds), mean ± SD	183.5 ± 42.3		178 ± 38.6		0.051
Height (inches), mean ± SD	68 ± 3.9		67.3 ± 4.0		0.009
BMI (kg/m^2), mean ± SD	27.7 ± 5.4		27.5 ± 4.9		0.58
BMI (kg/m^2) distribution					0.94
<20	31	3.7%	9	3.4%	
20-24.9	236	28.0%	75	28.5%	
25-29.9	346	41.0%	114	43.3%	
30-34.9	154	18.3%	43	16.3%	
35+	76	9.0%	22	8.4%	
Smoking history					0.0005
Never	472	54.3%	64	45.4%	
Former	338	38.9%	54	38.3%	
Current	59	6.8%	23	16.3%	
Smoking duration (years)					0.0024
0	472	55.7%	64	45.4%	
<10	74	8.7%	10	7.1%	
10-19	100	11.8%	13	9.2%	
20-29	86	10.2%	17	12.1%	
30+	115	13.6%	37	26.2%	

sex, distance of residence from Rochester, and characteristics of Mayo utilization for controls. This appears to be mainly due to lower participation by the oldest cases and controls (>70 years); cases with more aggressive disease (as evidence by lower one-year survival) and more aggressive lymphoma subtypes (e.g., DLBCL); and controls from the local area (residence <50 miles). However, these differences were overall fairly modest, suggesting that non-participation was unlikely to introduce major biases.

The choice of control group is critical for clinic-based designs. While hospital and clinic-based studies have used controls from blood donor clinics, other hospital patients, the local community, spouses, friends or companions of cases and other sources, the specific choice generally depends on scientific and logistic issues [13, 15, 36]. Based on the specifics of the Mayo regional practice, we elected to use controls recruited from the general medical practice. Since this was a non-random selection of

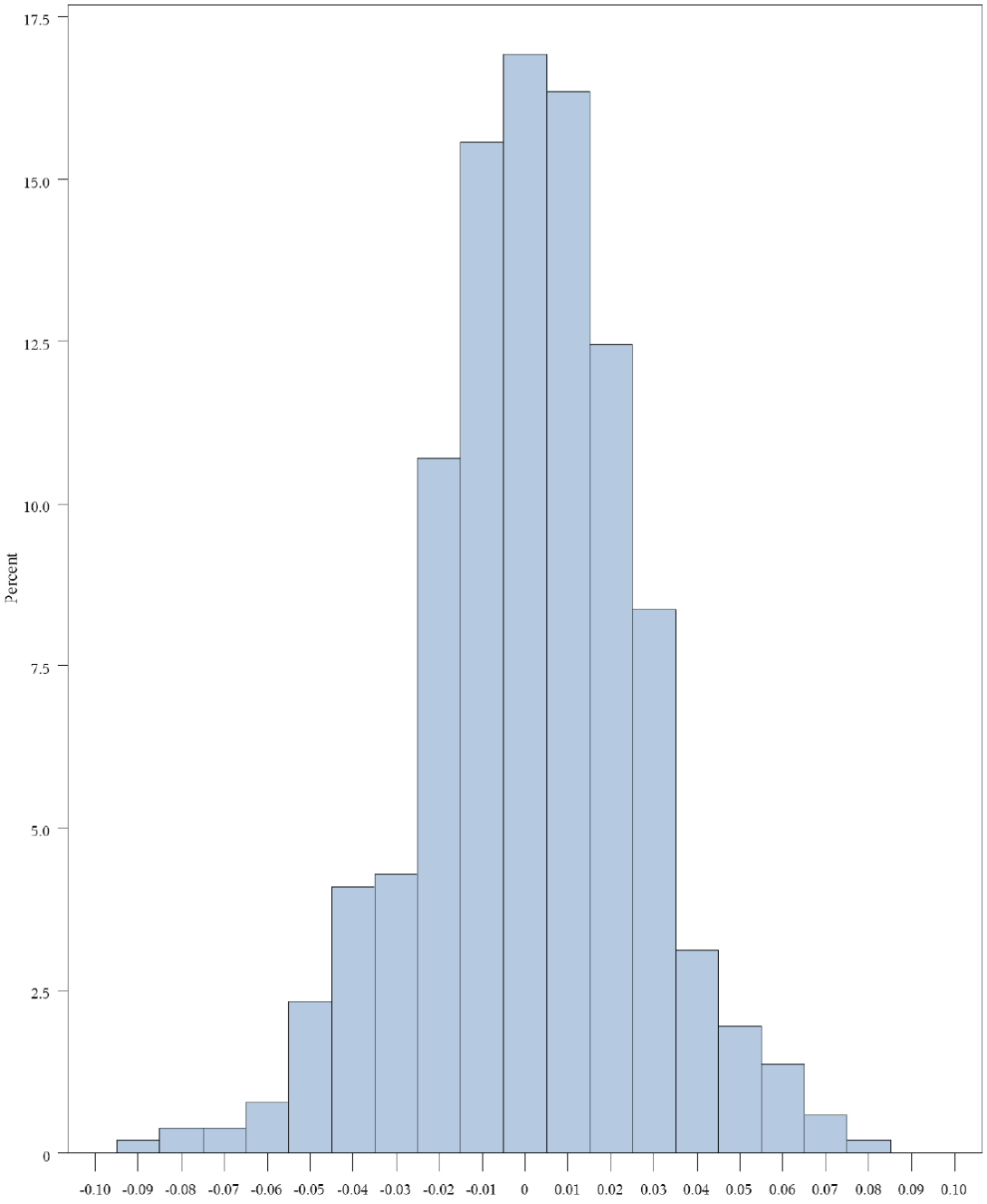

Figure 2. Difference in the minor allele frequency (MAF) of 514 SNPs with a MAF >10% that were genotyped in controls from the Mayo Clinic NHL Case-Control Study compared to the Iowa controls from the NCI-SEER NHL Case-Control Study.

controls from a secondary base, it was important to assess whether the controls came from the same catchment as the cases and whether the controls were selected independent of the exposure under study. In this study, controls and cases were well balanced on race, marital status, education, and occupational socioeconomic status, but cases, compared to controls, were slightly more likely to be from a farther distance from Rochester (9 miles) and to have

had a shorter time in the Mayo system (8 years), and total number of visits (15). While these differences were modest, it will be important to consider adjusting for time in the Mayo system and total visits when evaluating risk factors that are correlated with these types of variables in order to minimize any potential confounding by health care utilization at Mayo [15]. With respect to selection of controls independent of exposure, there was no single approach to assess this since we designed the study to address a variety of genetic, serologic and questionnaire-based risk factors. However, by selecting controls with pre-scheduled appointments for general medical examinations, as opposed to targeting patients being seen for a specific medical condition, the potential for this source of bias should be greatly decreased.

Finally, we found that the Mayo case-control study was able to replicate (in terms of the direction and magnitude of association) nearly all of the published InterLymph associations across a range of lifestyle [29, 30], medical history [31-33] and genetic [35] risk factors, with only the atopy association [34] not consistent with the InterLymph pooled results. While only an indirect assessment of internal validity, it does suggest that if there was a selection bias in case or control recruitment that was sufficiently large to jeopardize internal validity, then this study would be unlikely to observe most of the pooled associations, acknowledging that other factors including chance and true population differences could also account for discrepant findings.

External validity

Compared to the population-based Iowa SEER Registry data for NHL patients aged 20-79 years, the Mayo NHL cases age 20-79 years had a similar sex distribution, while the Mayo study under-represented patients aged 70-79 years (23% versus 35%), which highlights the difficulty enrolling older patients. Indeed, many recent NHL studies only recruit to age 70 [6, 9, 39] or 75 years [6, 7, 22, 37, 40]. The Mayo study also over-represented less aggressive histologies such as CLL/SLL and follicular lymphoma, although absolute differences were 10% or less. Similarly, the one-year survival of Mayo cases was higher than Iowa SEER data, although again the absolute difference was modest (7%) and was consistent across subgroups defined by age, sex and NHL subtype. Few clinic-based studies have addressed referral or survival bias, but those that have found similar or more extreme differences in case characteristics, depending on the specifics of the disease studied and the referral practice [16, 41]. For this study, our results suggest that it is reasonable to conclude that the cases enrolled have strong generalizability to the target population, which consists of mainly Caucasians from the upper Midwest. The Mayo cases were fewer representatives of the national SEER data, although the one-year survival differences in the national SEER data were only modestly less than those seen for the Iowa SEER data and the Mayo study.

Other study designs are not immune to concerns about external validity. Cases in population-based case-control studies are also underrepresented for older patients and more aggressive NHL subtypes due to early deaths and nonresponse [6, 22]. While cohort studies overcome the concerns about early deaths due to aggressive cases (since all cases should have been identified), few are population-based, and even those that are population-based, the initial response rate is often not particularly high (e.g., the participation rate to the Iowa Women's Health Study was 43%) [42].

Our control group was also similar to data on anthropometrics, lifestyle factors, dietary intake, and minor allele frequencies on a variety of SNPs, obtained from a population-based case-control study conducted in Iowa from 1998-2000, although lifetime smoking rates were somewhat lower. The latter control group had a response rate of 58% among the controls, so there may be bias in these estimates, although they would appear to affect both control groups in a similar manner. While there is relatively little work on the external validity of clinic-based designs, it has been reported that control groups from hospital-based and population-based studies have similar allele and genotype frequencies for a variety of metabolic genes [43].

Given that the cases and controls are reasonably similar to population-based data; this would suggest that the secondary study base for the Mayo regional practice is not dissimilar to a population-based sample from the upper Midwest, which greatly enhances the generalizabil-

ity of the study.

Strengths and limitations

There are several strengths of this study, including case selection that was designed to capture all cases as rapidly as possible to decrease survival bias and increase the amount of pretreatment serum; conducting a thorough pathology review for the diagnosis and determination of lymphoma subtypes; collection of pathology tissue for assessment of tumor heterogeneity and other molecular studies; collection of extensive risk factor data; and collection of blood samples for serum, plasma and DNA for molecular epidemiology studies. Controls were selected from the underlying source population, and were well matched on design variables. Response rates were reasonable, and there were no large demographic, health care or clinical differences between respondents and nonrespondents for both cases and controls. Both cases and controls compared well to population -based data.

There were also limitations. While our goal was to enroll cases rapidly, we missed some of the most aggressive cases who either did not participate or never came to Mayo, which was apparent when comparing the Mayo study to population-based data. Due to our need to increase samples size, we accepted cases who were diagnosed outside of Mayo and who were enrolled up to 9 months after diagnosis. However, this led to a lower percentage of cases that had their blood drawn before the initiation of treatment and that had pathology tissue outside of the institution, requiring outside collection and limiting the amount and type of tissue (e.g., frozen tissue) that was available for study. Not all participants completed risk factors questionnaires, which is also common in epidemiologic studies. There were also some modest imbalances between cases and controls on utilization of health care at Mayo; however, these factors can be evaluated in the analysis.

All epidemiologic study designs have strengths and limitations, and no study is completely free of bias. Because clinic or hospital-based studies are particularly susceptible to biases that can greatly impact both internal and external validity, it was particularly important to design our study using well-established epidemiologic principles and conduct a variety of empirical checks of the study design and implementation. While biases that impact internal and external validity are unlikely to be completely removed, they can be minimized and quantified as has been demonstrated for this study.

Acknowledgments

We thank Sondra Buehler for her editorial assistance. This work was supported by a grant from the National Institutes of Health (R01 CA92153). The NCI-SEER study was supported by the Intramural Research Program of the National Cancer Institute and by Public Health Service (PHS) contracts N01-PC-65064, N01-PC-67008, N01-PC-67009, N01-PC-67010, and N02-PC-71105.

Please correspondence to: Dr. James R. Cerhan, Department of Health Sciences Research, Mayo Clinic College of Medicine, 200 1st Street SW, Rochester, MN 55905, USA. E-mail: cerhan.james@mayo.edu

References

[1] Furberg AH, Ambrosone CB. Molecular epidemiology, biomarkers and cancer prevention. Trends Mol Med 2001 Nov; 7(11):517-521.

[2] Chen YC, Hunter DJ. Molecular epidemiology of cancer. CA Cancer J Clin 2005 Jan-Feb; 55 (1):45-54; quiz 57.

[3] Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev 2007 Oct; 16(10):1954-1965.

[4] Potter JD. Logistics and design issues in the use of biological samples in observational epidemiology. IARC Sci Publ 1997(142):31-37.

[5] Morton LM, Cahill J, Hartge P. Reporting participation in epidemiologic studies: a survey of practice. Am J Epidemiol 2006 Feb 1; 163 (3):197-203.

[6] Chatterjee N, Hartge P, Cerhan JR, Cozen W, Davis S, Ishibe N, Colt J, Goldin L, Severson RK. Risk of non-Hodgkin's lymphoma and family history of lymphatic, hematologic, and other cancers. Cancer Epidemiol Biomarkers Prev 2004 Sep; 13(9):1415-1421.

[7] Hughes AM, Armstrong BK, Vajdic CM, Turner J, Grulich A, Fritschi L, Milliken S, Kaldor J, Benke G, Kricker A. Pigmentary characteristics, sun sensitivity and non-Hodgkin lymphoma. Int J Cancer 2004 Jun 20; 110(3):429-434.

[8] Morton LM, Holford TR, Leaderer B, Zhang Y, Zahm SH, Boyle P, Flynn S, Tallini G, Owens PH, Zhang B, Zheng T. Alcohol use and risk of non-Hodgkin's lymphoma among Connecticut women (United States). Cancer Causes Control 2003 Sep; 14(7):687-694.

[9] Willett EV, Smith AG, Dovey GJ, Morgan GJ, Parker J, Roman E. Tobacco and alcohol consumption and the risk of non-Hodgkin lymphoma. Cancer Causes Control 2004 Oct;15 (8):771-780.
[10] Spinelli JJ, Ng CH, Weber JP, Connors JM, Gascoyne RD, Lai AS, Brooks-Wilson AR, Le ND, Berry BR, Gallagher RP. Organochlorines and risk of non-Hodgkin lymphoma. Int J Cancer 2007 Dec 15; 121(12):2767-2775.
[11] Iqbal J, Liu Z, Deffenbacher K, Chan WC. Gene expression profiling in lymphoma diagnosis and management. Best Pract Res Clin Haematol 2009 Jun; 22(2):191-210.
[12] Gascoyne RD, Rosenwald A, Poppema S, Lenz G. Prognostic biomarkers in malignant lymphomas. Leuk Lymphoma 2010 Aug; 51 Suppl 1:11-19.
[13] Rothman KJ, Greenland S. Modern epidemiology, second edition. Philadelphia, PA: Lippincott -Raven Publishers 1998.
[14] Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. Am J Epidemiol 1992; 135(9):1019-1028.
[15] Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. Am J Epidemiol 1992; 135(9):1029-1041.
[16] Nelson LM, Franklin GM, Hamman RF, Boteler DL, Baum HM, Burks JS. Referral bias in multiple sclerosis research. J Clin Epidemiol 1988; 41(2):187-192.
[17] Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing Clinical Research, Edition 3. Philadelphia, PA: Lippincott Williams & Wilkins 2007.
[18] Miettinen OS. The "case-control" study: valid selection of subjects. J Chronic Dis 1985; 38 (7):543-548.
[19] Jaffe ES, Harris NL, Stein H, Vardiman JW. World Health Organization Classification of Tumours: Pathology and Genetics, Tumours of Hematopoietic and Lymphoid Tissues. Lyon: IARC Press 2001.
[20] Morton LM, Turner JJ, Cerhan JR, Linet MS, Treseler PA, Clarke CA, Jack A, Cozen W, Maynadie M, Spinelli JJ, Costantini AS, Rudiger T, Scarpa A, Zheng T, Weisenburger DD. Proposed classification of lymphoid neoplasms for epidemiologic research from the Pathology Working Group of the International Lymphoma Epidemiology Consortium (InterLymph). Blood 2007 Jul 15; 110(2):695-708.
[21] Lubin JH, Hartge P. Excluding controls: misapplications in case-control studies. Am J Epidemiol 1984 Nov; 120(5):791-793.
[22] Holly EA, Lele C, Bracci PM, McGrath MS. Case-control study of non-Hodgkin's lymphoma among women and heterosexual men in the San Francisco Bay Area, California. Am J Epidemiol 1999; 150(4):375-389.
[23] Green LW. Manual for scoring socioeconomic status for research on health behavior. Public Health Rep 1970 Sep; 85(9):815-827.
[24] Weiner GJ, Witzig TE, Link BK. The University of Iowa/Mayo Clinic Lymphoma SPORE - SPORE Update. Clin Adv Hematol Oncol 2004; 2:57-59.
[25] Cerhan JR, Ansell SM, Fredericksen ZS, Kay NE, Liebow M, Call TG, Dogan A, Cunningham JM, Wang AH, Liu-Mares W, Macon WR, Jelinek D, Witzig TE, Habermann TM, Slager SL. Genetic variation in 1253 immune and inflammation genes and risk of non-Hodgkin lymphoma. Blood 2007 Dec 15; 110(13):4455-4463.
[26] Cerhan JR, Liu-Mares W, Fredericksen ZS, Novak AJ, Cunningham JM, Kay NE, Dogan A, Liebow M, Wang AH, Call TG, Habermann TM, Ansell SM, Slager SL. Genetic variation in tumor necrosis factor and the nuclear factor-{kappa}B canonical pathway and risk of non-Hodgkin's Lymphoma. Cancer Epidemiol Biomarkers Prev 2008 Nov; 17(11):3161-3169.
[27] Shen M, Cozen W, Huang L, Colt J, De Roos AJ, Severson RK, Cerhan JR, Bernstein L, Morton LM, Pickle L, Ward MH. Census and geographic differences between respondents and nonrespondents in a case-control study of non-Hodgkin lymphoma. Am J Epidemiol 2008 Feb 1; 167(3):350-361.
[28] Wang SS, Menashe I, Cerhan JR, Cozen W, Severson RK, Davis S, Hutchinson A, Rothman N, Chanock SJ, Bernstein L, Hartge P, Morton LM. Variations in chromosomes 9 and 6p21.3 with risk of non-Hodgkin lymphoma. Cancer Epidemiol Biomarkers Prev 2011 Jan; 20(1):42 -49.
[29] Morton LM, Zheng T, Holford TR, Holly EA, Chiu BC, Costantini AS, Stagnaro E, Willett EV, Dal Maso L, Serraino D, Chang ET, Cozen W, Davis S, Severson RK, Bernstein L, Mayne ST, Dee FR, Cerhan JR, Hartge P. Alcohol consumption and risk of non-Hodgkin lymphoma: a pooled analysis. Lancet Oncol 2005 Jul; 6(7):469-476.
[30] Morton LM, Hartge P, Holford TR, Holly EA, Chiu BC, Vineis P, Stagnaro E, Willett EV, Franceschi S, La Vecchia C, Hughes AM, Cozen W, Davis S, Severson RK, Bernstein L, Mayne ST, Dee FR, Cerhan JR, Zheng T. Cigarette smoking and risk of non-Hodgkin lymphoma: a pooled analysis from the International Lymphoma Epidemiology Consortium (interlymph). Cancer Epidemiol Biomarkers Prev 2005 Apr; 14(4):925-933.
[31] Wang SS, Slager SL, Brennan P, Holly EA, De Sanjose S, Bernstein L, Boffetta P, Cerhan JR, Maynadie M, Spinelli JJ, Chiu BC, Cocco P, Mensah F, Zhang Y, Nieters A, Dal Maso L, Bracci PM, Costantini AS, Vineis P, Severson RK, Roman E, Cozen W, Weisenburger D, Davis S, Franceschi S, La Vecchia C, Foretova L, Becker N, Staines A, Vornanen M, Zheng T, Hartge P. 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 Apr 15; 109(8):3479-3488.

[32] Ekstrom Smedby K, Vajdic CM, Falster M, Engels EA, Martinez-Maza O, Turner J, Hjalgrim H, Vineis P, Seniori Costantini A, Bracci PM, Holly EA, Willett E, Spinelli JJ, La Vecchia C, Zheng T, Becker N, De Sanjose S, Chiu BC, Dal Maso L, Cocco P, Maynadie M, Foretova L, Staines A, Brennan P, Davis S, Severson R, Cerhan JR, Breen EC, Birmann B, Grulich AE, Cozen W. Autoimmune disorders and risk of non-Hodgkin lymphoma subtypes: a pooled analysis within the InterLymph Consortium. Blood 2008 Apr 15; 111(8):4029-4038.

[33] Willett EV, Morton LM, Hartge P, Becker N, Bernstein L, Boffetta P, Bracci P, Cerhan J, Chiu BC, Cocco P, Dal Maso L, Davis S, De Sanjose S, Smedby KE, Ennas MG, Foretova L, Holly EA, La Vecchia C, Matsuo K, Maynadie M, Melbye M, Negri E, Nieters A, Severson R, Slager SL, Spinelli JJ, Staines A, Talamini R, Vornanen M, Weisenburger DD, Roman E. Non-Hodgkin lymphoma and obesity: a pooled analysis from the InterLymph Consortium. Int J Cancer 2008 May 1; 122(9):2062-2070.

[34] Vajdic CM, Falster MO, de Sanjose S, Martinez-Maza O, Becker N, Bracci PM, Melbye M, Smedby KE, Engels EA, Turner J, Vineis P, Costantini AS, Holly EA, Kane E, Spinelli JJ, La Vecchia C, Zheng T, Chiu BC, Dal Maso L, Cocco P, Maynadie M, Foretova L, Staines A, Brennan P, Davis S, Severson R, Cerhan JR, Breen EC, Birmann B, Cozen W, Grulich AE. Atopic disease and risk of non-Hodgkin lymphoma: an InterLymph pooled analysis. Cancer Res 2009 Aug 15;69(16):6482-6489.

[35] Skibola CF, Bracci PM, Nieters A, Brooks-Wilson A, de Sanjose S, Hughes AM, Cerhan JR, Skibola DR, Purdue M, Kane E, Lan Q, Foretova L, Schenk M, Spinelli JJ, Slager SL, De Roos AJ, Smith MT, Roman E, Cozen W, Boffetta P, Kricker A, Zheng T, Lightfoot T, Cocco P, Benavente Y, Zhang Y, Hartge P, Linet MS, Becker N, Brennan P, Zhang L, Armstrong B, Smith A, Shiao R, Novak AJ, Maynadie M, Chanock SJ, Staines A, Holford TR, Holly EA, Rothman N, Wang SS. Tumor necrosis factor (TNF) and lymphotoxin-alpha (LTA) polymorphisms and risk of non-Hodgkin lymphoma in the InterLymph Consortium. Am J Epidemiol 2010 Feb 1; 171(3):267-276.

[36] Miller AB. Hospital or population controls? It depends on the question. Prev Med 1994 May; 23(3):263-266.

[37] Chang ET, Smedby KE, Hjalgrim H, Porwit-MacDonald A, Roos G, Glimelius B, Adami HO. Family history of hematopoietic malignancy and risk of lymphoma. J Natl Cancer Inst 2005 Oct 5; 97(19):1466-1474.

[38] Hartge P. Participation in population studies. Epidemiology 2006 May; 17(3):252-254.

[39] Nelson RA, Levine AM, Marks G, Bernstein L. Alcohol, tobacco and recreational drug use and the risk of non-Hodgkin's lymphoma. Br J Cancer 1997; 76(11):1532-1537.

[40] Chiu BC, Soni L, Gapstur SM, Fought AJ, Evens AM, Weisenburger DD. Obesity and risk of non-Hodgkin lymphoma (United States). Cancer Causes Control 2007 Aug; 18(6):677-685.

[41] Rocca WA, Maraganore DM, McDonnell SK, Schaid DJ. Validation of a telephone questionnaire for Parkinson's disease. J Clin Epidemiol 1998 Jun; 51(6):517-523.

[42] Cerhan JR, Wallace RB, Dick F, Kemp J, Parker AS, Zheng W, Sellers TA, Folsom AR. Blood transfusion and risk of non-Hodgkin lymphoma subtypes and chronic lymphocytic leukemia. Cancer Epidemiol Biomarkers Prev 2001; 10:361-368.

[43] Garte S, Gaspari L, Alexandrie AK, Ambrosone C, Autrup H, Autrup JL, Baranova H, Bathum L, Benhamou S, Boffetta P, Bouchardy C, Breskvar K, Brockmoller J, Cascorbi I, Clapper ML, Coutelle C, Daly A, Dell'Omo M, Dolzan V, Dresler CM, Fryer A, Haugen A, Hein DW, Hildesheim A, Hirvonen A, Hsieh LL, Ingelman-Sundberg M, Kalina I, Kang D, Kihara M, Kiyohara C, Kremers P, Lazarus P, Le Marchand L, Lechner MC, van Lieshout EM, London S, Manni JJ, Maugard CM, Morita S, Nazar-Stewart V, Noda K, Oda Y, Parl FF, Pastorelli R, Persson I, Peters WH, Rannug A, Rebbeck T, Risch A, Roelandt L, Romkes M, Ryberg D, Salagovic J, Schoket B, Seidegard J, Shields PG, Sim E, Sinnet D, Strange RC, Stucker I, Sugimura H, To-Figueras J, Vineis P, Yu MC, Taioli E. Metabolic gene polymorphism frequencies in control populations. Cancer Epidemiol Biomarkers Prev 2001 Dec; 10(12):1239-1248.