From the Department of Medical Epidemiology and Biostatistics Karolinska Institutet, Stockholm, Sweden

Truncation and Missing Family Links in Population-Based Registers

Monica Leu

Stockholm 2008

Published by Karolinska Institutet. Printed by

ISBN 978-91-7357-552-2

LIST OF PUBLICATIONS

I. Leu M, Czene K, Reilly M. "Population lab": the creation of virtual populations for genetic epidemiology research. ***Epidemiology*** 2007; 18(4):433-40

II. Leu M, Czene K, Reilly M. The impact of truncation and missing family links in population-based registers on familial risk estimates. ***Am J Epidemiol*** 2007; 166(12):1461-7

III. Leu M, Czene K, Reilly M. Bias-correction of estimates of familial risk from population based registers. ***Submitted***

IV. Leu M, Reilly M, Czene K. Evaluation of bias in familial risk estimates: a study of common cancers using Swedish population-based registers. ***Submitted***

Contents

1 Introduction 2

2 Background 3

2.1 Family studies in genetic epidemiology 3

2.2 Various measures of familial aggregation 5

2.3 Familial risk in studies of cancer 7

2.4 Register-based studies 9

2.4.1 The Swedish MultiGeneration Register (MGR) 10

2.4.2 The Swedish Cancer Register (SCR) 12

2.4.3 Potential limitations in using register-based data 13

2.5 Forms of bias: selection bias, information bias and confounding 16

2.5.1 Misclassification of exposure 17

2.6 Simulation Methods 19

3 Aims 21

4 Methods 22

4.1 Simulating virtual populations: the Poplab package (Paper 1-4) . . . 22

4.1.1 Creation of the baseline population 24

4.1.2 The trimming process 26

4.1.3 The evolution of a population 27

4.2 Statistical analyses of familial risk (Paper 1-4) 28

4.3 Simulating the virtual Swedish population (Paper 1) 29

4.4 Family data with missing links and truncation (Paper 2) 31

4.4.1 Range of investigations 31

4.4.2 Mimicking the incompleteness 32
4.4.3 The apparent relative risk 33
4.5 Bias due to registration start-up, theoretical considerations (Paper 3,4) 33
4.5.1 Testing the assumption of non-differential misclassification . . 34
4.5.2 The apparent relative risk 34
4.5.3 Bias-corrected relative risk, R_{bc}: point estimate and variance . 35
4.5.4 Estimating the prevalence and sensitivity of exposure 36
4.5.5 The bias-correction 37
4.6 Correcting for the bias due to registration start-up (Paper 3) 37
4.7 Evaluating the bias in the real data (Paper 4) 38
4.7.1 The apparent relative risk 38
4.7.2 Sensitivity and true prevalence of exposure 40

5 Results 41
5.1 Paper 1 41
5.2 Paper 2 41
5.3 Paper 3 46
5.4 Paper 4 50

6 Discussion 51
6.1 General overview of Poplab (Paper 1) 51
6.2 How large is the bias, and what is causing it? (Paper 2) 53
6.3 From evaluating the bias to bias correction (Paper 3) 55
6.4 Bias in studies of cancer using the Swedish registers (Paper 4) 56
6.5 Simplifying assumptions and limitations 58

7 Conclusions 60

8 Future studies 62

List of abbreviations

MGR	MultiGeneration Register
SCR	Swedish Cancer Register
FRR	familial relative risk
IRR	incidence rate ratio
CI	confidence interval
Poplab	The software package Population Lab

Chapter 1

Introduction

To estimate the familial contribution to the risk of diseases that aggregate in families valuable information is provided by considering the number of affected relatives, their degree of relationship and age at diagnosis. If such information is recorded in population-based registers, then these offer very efficient means of immediate electronic follow-up of study cohorts and identifying cases. Regional or nation-wide registers have contributed vastly to the study of familial cancer [1],[2],[3]. However, such resources are incomplete due to the start-up date truncation, ascertainment bias from inclusion/exclusion criteria [4],[5],[6], length-time bias [7],[8] and broken family links due to unknown parents [9]. While these realities of register-based data induce bias in the estimates of familial aggregation, the statistical methodology offered to correct for such shortcomings [10],[11] is many times cumbersome to implement.

This thesis aims to investigate the potential bias from analyzing incomplete family data in population-based registers. Secondly, it will focus on developing simple methodology to correct for such biases and demonstrating the methods on follow-up cohorts from the Swedish MultiGeneration Register [12] and the Swedish Cancer Register [13].

Chapter 2

Background

2.1 Family studies in genetic epidemiology

Genetic epidemiology focuses on related individuals and their family histories. Family studies first aim to identify whether a certain disease or trait clusters in families. They are usually based on specific relationships, such as twin pairs or parent-offspring. The presence of familial clustering and excess familial risk can be due to many factors other than common genes, including shared family environment, cultural influences, enhanced awareness among family members, or chance [14], [15]. Even though evidence of familial aggregation is not a sufficient condition to claim a genetic background for a disease, the absence of such an evidence makes the genetic component much less likely, especially when environmental factors are considered in the analyses [16].

Once the familial aggregation is established, further research is conducted in suggesting the most plausible explanation of this finding (e.g., via segregation analysis), and to distinguish the relative contributions of environmental and genetic factors [17], [18], [19], [20], [21]. When disease-associated genes are identified, the next step is to determine the genetic model that underlies the disease (e.g, possibly several loci on different chromosomes might trigger the onset of disease), measure the increased risk for individuals with the putative genetic susceptibility (penetrance) and study interactions with other genetic and environmental risk factors [22], [23], [24].

Various designs provide the framework for quantifying the relative importance of environment and genes determining susceptibility. They include case-control fam-

ily studies, case-families with or without population controls, and population-based case-control-family design, and they can all be grouped into population-based family designs.

The case-control family design includes information about the disease status (or other characteristics) of the relatives of cases. Since the accuracy of the information may depend on the disease and on the degree of relatedness, these studies are often restricted to first-degree relatives.

Case-family studies recruit relatives of cases for comparisons between cases and unaffected family members. The most typical comparison is with siblings, and a special example is disease discordant twin pairs. In this design, genetic effects and effects of environmental exposures, separately as well as their interactions can be estimated. For rare genetic variants it may be the only feasible design.

The population-based case-control-family design recruits cases and their relatives as well as controls and their relatives. It has been used especially in cancer studies where complete population registries facilitate recruitment. Over-sampling of cases with earlier onset is often a feature, given that familial and hence most likely genetic factors are more pronounced in those case families.

One application of family studies has been to determine genetic models for susceptibility. The inclusion of different relative types permits the examination of the consistency of familial aggregation with the assumed model of inheritance. This approach is known as segregation analysis and asks what model best explains the pattern of familial aggregation of the disease. It involves the specification of the mode of inheritance, the population frequency of individuals at high genetic risks, and the associated genetic risks themselves. A particular example of family studies is the kin-cohort design, where relatives of a case with a particular genetic variant constitute a subpopulation with increased likelihood of carrying the same variant.

Twin pairs and twin families also constitute special examples of the family design, and are particulary efficient in teasing apart the effects of shared genes and shared environment. Twin models compare the similarity between monozygotic twins (MZ) with same-sex dizygotic twins (DZ) via a measure called pairwise concordance rate, which is the proportion of pairs with both twins affected of all twin pairs with at least one affected [15]. The classical assumption behind these models is that MZ

and DZ twins display a comparable degree of similarity due to shared environment, and the difference in concordance rates is only a reflection of genetic factors. For example, a higher risk to the co-twin of an affected twin in monozygotic rather than dizygotic pairs suggests that genetic effects may be explaining familial aggregation; a higher risk to siblings of cases than to their offspring suggests recessive or X-linked genetic components; a greater risk in the maternal aunts or uncles in the absence of increased risks between paternal aunts or uncles indicates X-linked effects; and a rapid decline in familial relative risk with the degree of relationship may indicate a polygenic model.

Adoption studies are very efficient family designs, and compare the biological relatives of affected with control adoptees [25]. They have been especially used in dissecting genetic from environmental contributions to human behavioral variations [26], [27]. Although they target an extended range of investigations, from the relative importance of nurture and nature on the cognitive development to the etiology of various psychiatric disorders [28], data for such studies are unfortunately rarely available.

In summary, the population-based family design to be employed is a function of the research question under investigation, how common/uncommon the genetic variant or phenotype of interest and recruitment possibilities (access to family members). Potential analytical limitations should also be considered, due to the non-independence and possible incompleteness of data within families and the way families have been chosen for study.

2.2 Various measures of familial aggregation

Several measures of familial aggregation have been proposed, but only few are designed to be implemented at the individual level, among which the most common is an indicator of whether one or more first degree relatives (parents, siblings, and offspring) have been diagnosed with the disease. Moreover, these measures will be influenced by the definition of "family history", which can consider various types of relatives, number of affected members or the age at incidence [29], [30]. A positive family history is a function of the number of relatives, the background incidence risk

and the correlation in risk among relatives [31]. Understanding the dependence of disease on the familial history is essential in distinguishing hereditary from non-familial forms, and thus identifying causal factors.

In a paper from 2001 [32], Boucher and Kerber propose to consider the complete risk experience of all observable biological relatives, adjusted for the age, sex, number and degree of the relatives. This total familial risk is summarized either as a familial standardized incidence ratio (FSIR) or as a familial rate (FR). FSIR is defined in terms of the kinship coefficient c(i,j)

$$c(i,j) = (1/2)\sum_{p=1}^{P_{i,j}} 2^{-l(p)} \tag{2.1}$$

where $P_{i,j}$ is the total number of distinct shortest paths through a common ancestor between individuals i and j, and l(p) is the length in number of reproductive events of the path p. As a simple example, if i and j represent indices for full siblings, this coefficient is 0.25.

FSIR for the ith individual is defined as

$$FSIR_i = \frac{\sum_{j\neq i} I_j c(i,j)}{\sum_{k=1}^{K}\sum_{j\neq i} t_{jk}\lambda_k c(i,j)} \tag{2.2}$$

where the summation over j runs over all related individuals in the pedigree, K is the number of strata in the population, λ_k is the incidence in the kth stratum and I_j is the indicator for the disease of individual j.

The authors point that FSIR is a very efficient measure for identifying the individuals at high risk, but such detailed family history data is often not available.

Another example of familial aggregation measure is the familial risk ratio (FRR), defined as the risk to a given type of relative of an affected individual divided by the population prevalence [15]. A more common terminology for this measure is familial relative risk (or recurrence risk ratio). While for the most common cancers, the familial relative risk for first-degree relatives ranges from 2 to 4, for less prevalent cancers such as thyroid and testis [33], [34], [35], and for many nonmalignant diseases, such as multiple sclerosis, schizophrenia and type I diabetes it can range from 5 to 20 [17].

2.3 Familial risk in studies of cancer

It is widely believed that most forms of cancers occur either sporadically, or from a hereditary background [15], [36]. The inherited predisposition to cancer has been documented as far back as from the sixteenth century, when familial clustering of some distinctive phenotypic features was noticed [37]. Long-established hereditary examples include the cancer of colon in familial adenomatous polyposis, and the breast-ovarian cancer syndrome with BRCA1/BRCA2 mutations. The familial component contributing to the development of cancer is widely illustrated in the modern era of scientific research [38], [39], [40], [41], [42], [43], [44]. Evaluating the "familiarity" of cancer constitutes the basis for identifying genes associated with familial cancer syndromes, and it is essential to genetic counseling targeting individuals at high risk of disease [36].

Typically, familial cancers are presumed to account for 5-10% of all cancers [165]. They are characterized by an early age of onset of cancer and the occurrence of cancer in multiple members of the same family [14], [45]. Table 2.1 presents evidence of the familial component of cancer from two studies: the Utah study, with data obtained from merging the Utah Genealogic Database to the Utah Cancer Register [1], [38], [46], and a population based study based on the Swedish MultiGeneration and Cancer Registers. In the first study, FRR was calcuated as the ratio of the observed number of cancer cases among the first degree relatives of the cases divided by the expected number among relatives of controls, with matching for the year of birth. Similarly, FRR for relatives of cases with a young age at incidence was calculated, where young age was defined as "before 50 years" for breast, melanoma and brain/central nervous system (CNS) and "before 60 years" for the other sites. Risch combined the SIRs as originally calculated in the second study to obtain a more comparable measure with FRR, and the SIRs presented in the table are assessed for offspring with an affected parent and with/without an affected sibling, and for offspring with an affected sibling and with/without an affected parent. These results show that the most prevalent cancers should not be expected to be the most familial, FRR spans a relatively narrow range for these cancer sites, and that the familial risk increases with an early age at diagnosis.

Cancer site	Utah study [35]		Sweden study [33]	
	FRR (total)	FRR (early onset)	FRR (offspring)	FRR (sibling)
Prostate	2.21	4.08	2.82	9.41
Breast	1.83	3.70	1.86	2.01
Colorectal	2.54	4.53	1.86	4.41
Lung	2.55	2.50	1.68	3.16
Melanoma	2.10	6.43	2.50	3.41
Bladder	1.53	5.00	1.53	3.30
Non-Hodgkins lymphoma	1.68	2.40	1.68	2.37
Brain/CNS	1.97	8.95	1.72	2.37
Cervix	1.73		1.93	2.39

Table 2.1: "Familiarity" of cancer for first degree relatives of cancer, in decreasing magnitude of prevalence in the Utah populations, assessed in 2 population-based studies. Adapted from Risch 2001, Cancer Epidemiol Biomarkers Prev [15]

The various settings that are available for assessing the familial risk of cancer, such as clinical studies, twin studies, population based studies or studies of informative families, have already been mentioned. Their validity, generality and interpretation rely on complete information being obtained on the considered family members. With interviews as the means to obtain family history for cases or controls, the accuracy of the data (a high and/or non-differential response independent of the disease status) may be questionable [17], [47], [48], [49]. Table 2.2, adapted from [17], shows the positive predictive value (the probability that a self-reported family history is true) by case/control status and cancer site. The accuracy of self-reported family history decreases with the rarity of the cancer, and is higher for cases than for controls.

	Positive predictive value	
	Case	Control
Breast	93%	74%
Prostate	85%	68%
Colon	81%	71%
Ovarian	69%	25%
Endometrial	37%	17%

Table 2.2:

Thus, when available, merged population and disease registers are invaluable sources in obtaining medically verified family data. Estimating familial risk of cancer from using registered data will be discussed below.

2.4 Register-based studies

The emerging potential of registries for studying the epidemiology of cancer has been acknowledged as early as the 1970s [50]. Present efforts include setting-up of national cancer registers linked to pedigree (genealogical) information [51] and the use of multinational registers [52], [53], [54].

Nation-wide population- and disease-registers have been available in the Nordic countries for the past 50 years [55], [56], [57], [58]. Not surprisingly, an impressive

number of scientific publications emerged from these excellent resources [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75]. In Sweden, the MultiGeneration Register [12] and the Cancer Register [13] have been extensively used in studying the epidemiology of cancer.

2.4.1 The Swedish MultiGeneration Register (MGR)

MGR (Swedish: Flergenerationsregistret) is a database of individuals, initiated in 1961 from written records maintained by church parishes and county registration offices. Updates of the MGR are carried out yearly. The recorded individuals, called index persons, are persons registered in Sweden at some time since 1961, including those born in Sweden since 1932 and individuals who immigrated to Sweden. Each index person has information on the personal identity number of biological (or adoptive) parents, the year of birth or immigration, year of death, sex and country of birth. In 2002, the registry contained 9 million index persons and a total of 13.5 million individuals. Figure 2.1 shows a simplified flowchart of how MGR is produced from various registered data sources. The target population in the MultiGeneration Register is illustrated in Figure 2.2.

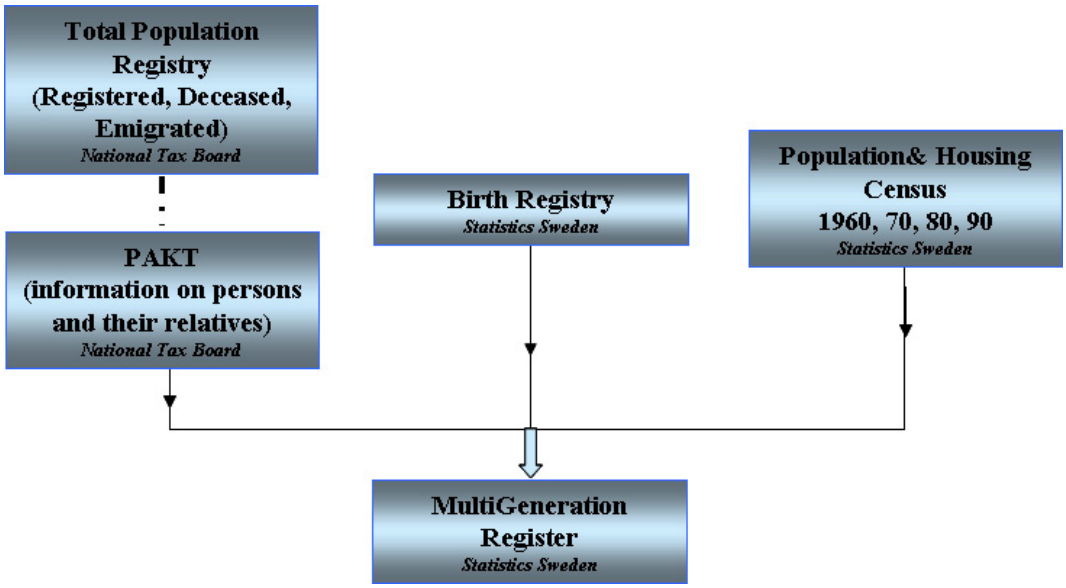


Figure 2.1: How the MultiGeneration Register is produced

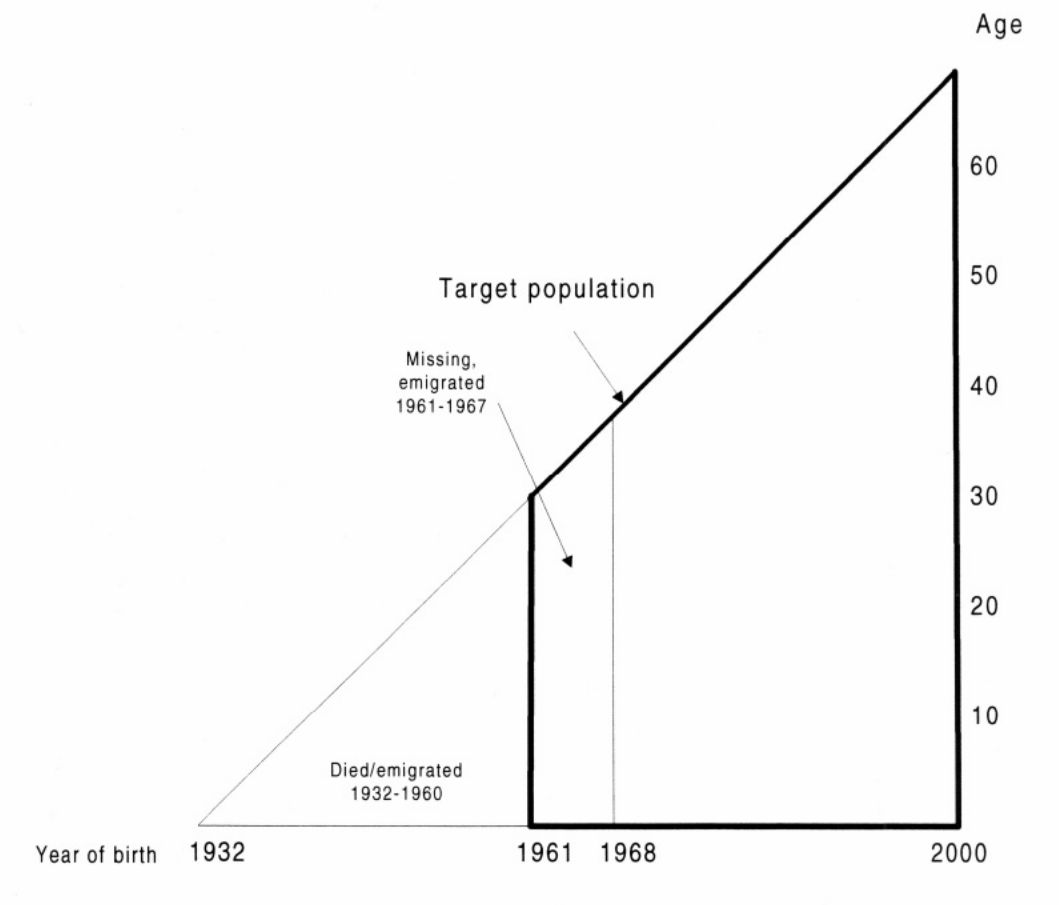


Figure 2.2: Population in the MultiGeneration Register

2.4.2 The Swedish Cancer Register (SCR)

SCR records primary cancers incident in Sweden since 1958, based on compulsory reports from all physicians in hospitals and other establishments for medical treatment under public or private administration in Sweden [13]. Furthermore, pathologists and cytologists report separately every cancer diagnosis. Only persons with an official residency in Sweden are included in the Cancer Register. If a person has more than one primary tumor, each tumor is registered separately. The register records the unique personal identification number, site of the tumor, ICD 7, stage and date of diagnosis. New cancers are not recorded based on death certificates, due to the uncertainty in the death certificates, especially for older individuals. The completeness of this register, and the accuracy of the reported cases of cancer are estimated to be close to 100% [76]. Based on the data published in the Swedish Cancer Register, Figure 2.3 illustrates the incidence rates (per 100000), by calendar year for all cancer sites. Figure 2.4 shows the incidence profile for five cancer sites that are going to be studied in the present thesis.

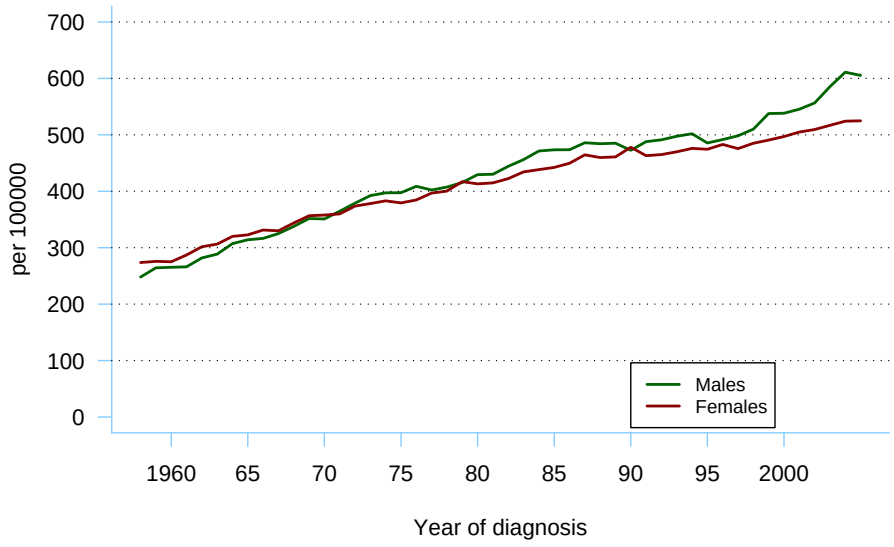


Figure 2.3: Cancer incidence in Sweden, all sites combined. Data source: the Swedish Cancer Register [13].

2.4.3 Potential limitations in using register-based data

The unquestionable potential of registers for research into familial aggregation of diseases has to be balanced by the awareness of the possible limitations of the data, such as incompleteness, causes of "missing-ness" [9], or lack of recording various exposures of interest [77]. Even with extensive coverage of the population members, the Scandinavian population registers were initially started for administrative, and not for research, purposes. In the Swedish MultiGeneration Register, for example, familial relationships are not complete due to inclusion/exclusion criteria for the index subjects at the start-up time of registration, non-identifiability of the parents of index persons who immigrated to Sweden as adults, and technical matters associated with the conversion of written archives to computerized information as well as the management of the register by different authorities.

Figure 2.5 illustrates the poor knowledge on parental identity in the initial years of registration, and how it differs substantially between Swedish-born individuals and immigrants. In addition, a large proportion of individuals who died before 1990s have the parental identity missing, and furthermore information on the father is usually

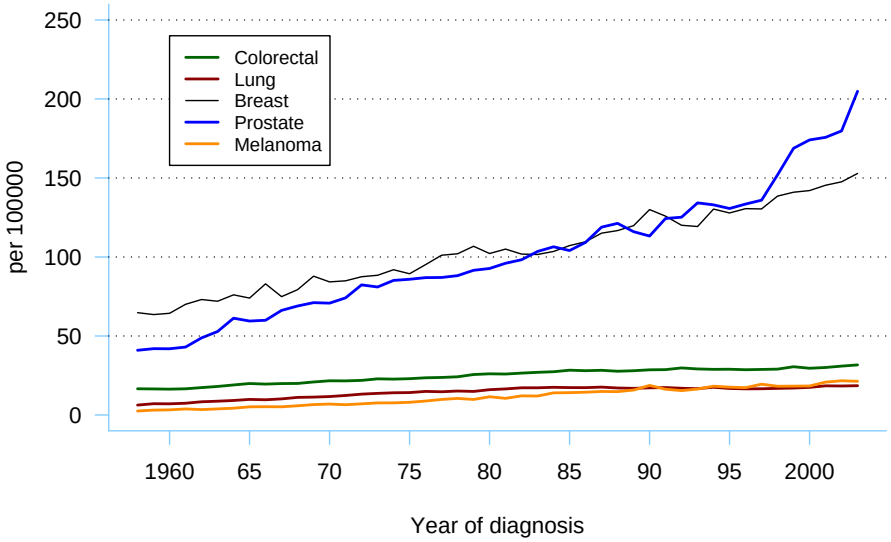


Figure 2.4: Cancer incidence in Sweden, five cancer sites. Data source: the Swedish Cancer Register [13].

less well recorded than information on the mother (Figure 2.6).

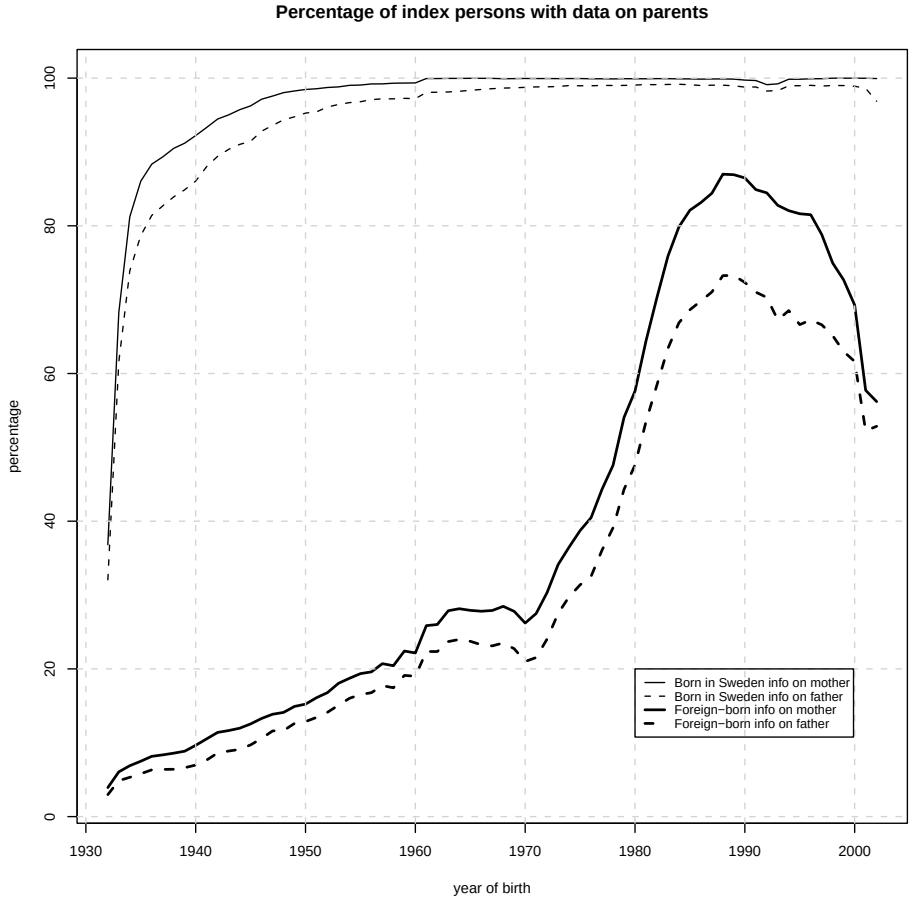


Figure 2.5: Percentage of index persons with known parents in the MultiGeneration Register, by year of birth, stratified by the place of birth (whether or not in Sweden)

The Swedish Cancer Register records incident cases after 1958, and cancers incident before this start-up date are not recorded. Thus, when merging this register with MGR to collect the family history of cancer, this exposure will be misclassified from failing to identify relatives as affected when disease occurs before the start-up date of registration. Such left-truncation of family history may cause dramatic biases in familial aggregation measures.

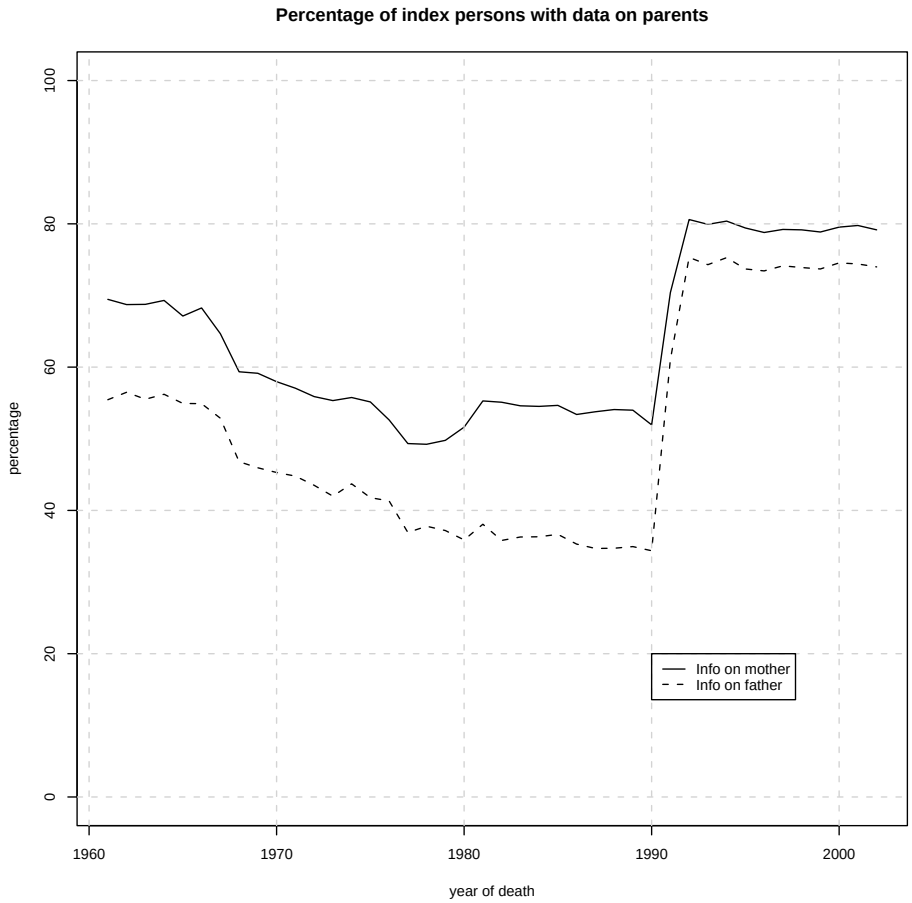


Figure 2.6: Percentage of index persons with known parents in the MultiGeneration Register, by year of death

2.5 Forms of bias: selection bias, information bias and confounding

Bias is a systematic error, and as opposed to random error, it cannot be reduced by increasing the population/sample size. Researchers should design epidemiologic studies in such a way as to avoid or minimize known or suspected biases. If they are unavoidable biases, a required step is to explain how they may affect the study results, in terms of direction and magnitude [78].

There are many and various forms of bias, and they have been listed over time [79], [80]. Among them, confounding, selection bias and information bias are the most common encountered in epidemiological studies. If the bias is not acknowledged and accounted for it can render illusory associations, and thus undermine both the internal and external validity of the findings [81], [83], [84].

Confounding bias is a distortion of the estimated effect of exposure on the outcome, as a result of an extraneous factor associated both with the exposure and outcome. Selection bias results from systematic differences between those participating in a study and those who do not, while information bias results from inaccurate measurement (or classification) of the study participants [82]. Selection biases can be further classified into self-selection bias, diagnostic suspicion bias, assembly bias, publication/reporting bias, and information biases into recall bias, surveillance (detection) bias and misclassification bias.

For example, when film photography was used in astronomy, observations typically found more blue galaxies than red ones. This was not because blue galaxies are actually more common, but rather because photographic film was more sensitive to blue light than red light. With the conversion of astronomy to digital cameras, which are more sensitive to red light than blue, the opposite bias is now the case.

http://en.wikipedia.org/wiki/Selection_bias

An example of detection bias that results in the inflation of the familial risk of a disease can be imagined when increased surveillance among family members shortly after diagnosis of the first familial tumor results in the earlier detection of asymptomatic familial cancers [85]. Such overestimation of the familial risk can obviously

impact on clinical and genetic counseling policies.

Confounding can be controlled for through randomisation (random allocation to make the experimental and control groups similar with respect to the suspected confounders), restriction and matching. Confounding can be also adjusted for in the analysis phase via stratification or multivariate analysis (modeling that estimates the effect of one variable while controlling for many other factors simultaneously). The later approach has a strong advantage over stratification in situations where sparse numbers would result for certain strata [82].

Misclassification of exposure, which is a form of information bias, will be discussed in the next section, as this is a central theme in the present thesis.

2.5.1 Misclassification of exposure

Two types of misclassification are usually considered: non-differential and differential. The existent statistical literature has dedicated ample attention [86] to distinguish between these two forms and to the methods to correct for these biases. Shortly, non-differential misclassification of exposure occurs when the probability of exposure misclassification is not related to disease status i.e. diseased and non-diseased individuals have the same probability to be misclassified according to exposure [87]. Similarly, probabilities of exposure misclassification which are dependent of the disease status render differential misclassification [88], [89], [90]. Misclassification of exposure is usually quantified in terms of sensitivity and specificity. Sensitivity represents the probability of correct classification among those who are truly exposed, while specificity is the probability of correct classification of those truly unexposed.

An example of differential misclassification is provided by a Norwegian study of respiratory disorders in relation to occupational exposures (quartz dust), where exposures where assessed both through self-reporting and interview [91]. The sensitivity of self-reported exposure is calculated with respect to the interview assessment, which is regarded as the gold standard. The sensitivity among those individuals presenting respiratory symptoms was in the range 50% to 65%, while for free-symptom participants the sensitivity was below 30%.

Assessing the form of misclassification that would be present in a study is essential

both for the interpretation of study results, in terms of direction and magnitude of bias, and for the bias-correction approach. For a binary exposure, researchers assuming non-differential misclassification will claim that the true disease-exposure association is even stronger than their estimate, while even slight deviations from this condition may result in bias away from the null. Where the exposure variable has several categories, the consequences are even more intractable [92]. When the non-differential hypothesis does not hold, simple correction methods are inappropriate [93], [94], [95], and may lead to "corrected" relative risks considerably higher than the truth [96].

The apparent (biased) relative risk in the presence of non-differential misclassification can be easily derived from basic principles [97]:

$$\widehat{R} = \frac{[SRP(E) + (1-V)P(\overline{E})] \cdot [(1-S)P(E) + VP(\overline{E})]}{[SP(E) + (1-V)P(\overline{E})] \cdot [(1-S)RP(E) + VP(\overline{E})]} \tag{2.3}$$

where R is the true relative risk, S and V are the sensitivity and, respectively, specificity of exposure, P(E) the true prevalence of exposure, and $P(\overline{E})$ the complement of P(E). Thus, the apparent relative risk is a function of the true relative risk, sensitivity and specificity. Flegal et al, 1986 [97] tabulate the effects of these parameters on the apparent relative risk (table 2.3).

As it will be presented later in this thesis (section 4.5), left-truncation due to disease registration start-up causes non-differential sensitivity of the observed exposure (i.e., family history of disease), and perfect specificity, as all unexposed individuals are classified as such. In this context, with $V = 1$, expression (2.3) becomes:

$$\widehat{R} = R \cdot (1 - SP(E)) \cdot \frac{1}{(1-S)RP(E) + P(\overline{E})} \tag{2.4}$$

When this last equation is re-arranged to provide an expression for R, R is referred to as the bias-corrected relative risk, and subsequently denoted as R_{bc}:

$$R_{bc} = \widehat{R} \cdot \frac{P(\overline{E})}{1 + P(E)((\widehat{R}-1)S - \widehat{R})} \tag{2.5}$$

It is obvious from the expression of R_{bc} that, to carry out a bias-correction of the relative risk, estimates of the true prevalence and sensitivity of exposure are needed. The challenge is to obtain these quantities since the extent of misclassification in

the real data is rarely known. Common approaches include the specification of a misclassification model [98] or use of validation samples [99], [100], [101], [102], [103], [104], [105]. The practical drawback of the later approach is that the validation samples are many times expensive or cumbersome to obtain.

2.6 Simulation Methods

In the present thesis, a simulation approach is developed in order to obtain estimates of the sensitivity of the observed (misclassified) exposure without the use of validation samples.

The idea of simulating populations is not new. Even from the early 1980s, important demographic work focused on creating electronic populations with the purpose of predicting future population structures. These methods require the construction of a realistic version of the present population with its relationships, and estimates of future vital rates. Examples include studies on factors affecting the household formation in England [106], forecasting kin counts in the US population [107], depicting the kinship networks of elderly for the United States of the year 2030 [108], investigating

True risk	**Sum S+V**	**Apparent risk**	**Direction of bias**
$R > 1.0$	$S + V > 1.0$	$1.0 < \widehat{R} < R$	Underestimation
	$S + V = 1.0$	$\widehat{R} = 1.0$	Underestimation
	$S + V < 1.0$	$\widehat{R} < 1.0$	Reversal of direction
R=1.0		$\widehat{R} = 1.0$	No bias
$R < 1.0$	$S + V > 1.0$	$R < \widehat{R} < 1.0$	Underestimation
	$S + V = 1.0$	$\widehat{R} = 1.0$	Reversal of direction
	$S + V < 1.0$	$1.0 < \widehat{R}$	Reversal of direction

Adapted from *Flegal et al, 1986* [97]

Table 2.3: Dependency of the apparent relative risk ($\widehat{R}$) on the true relative risk (R), sensitivity (S) and specificity (V).

the effects of serial monogamy on the 2035 predicted Netherlands population [109] and predicting the Chinese kinship of the year 2060 [110]. These mechanisms of demographic simulation use as the unit of study either the individual, or a group defined by certain characteristics (such as sex, age or residence location). The former method is referred to as microsimulation and the later as macrosimulation. A mixture of these two approaches can also be employed [109]. Numerous computer programs have been developed in the field of population demography, including CAMSIM [111], POPSIM [112], and SOCSIM [113], [114], [115], [116]. While SOCSIM creates virtual populations of a structure suitable for genetic epidemiology research (family registers where all relationships are known), the emphasis is on the processes of cohabitation, marriage and divorce, with the requirement of monthly follow-up of the events affecting the simulated individuals and the implementation of competing risk schemes. As opposed to demographers, we are interested to obtain a correct version of the past (i.e., family history of disease), in order to study disease etiology, which can be approached by investigating familial clustering, familial risk or predisposed sub-populations. For example, to appreciate the magnitude of cancer risk posed by having a positive family history, it is essential to have accurate information on whether the mother, siblings or other identifiable relatives, were previously diagnosed with cancer. The simulation package, Poplab, was designed to use simple and easily-available vital statistics, such as age profile, fertility and cause-specific or general mortality, and disease incidence to create virtual population registers.

Chapter 3

Aims

This thesis aims to contribute to the understanding of potential biases arising from analyzing incomplete family data from population-based registers. Dedicated focus will be given to the left-truncation of family history of disease due to registration start-up date and the missing family links due to the death of an index person, and the effects of such incompleteness on estimates of familial association. Specifically, the following objectives are to be addressed:

1) Creating a software package, Poplab for simulating virtual populations of related individuals using real vital rates (such as birth and death), with user-controlled parameters of familial aggregation of diseases (background incidence rates, value of familial association, familial model of disease) (Paper 1)

2) Studying the impact of truncation and missing family links in population-based registers on familial risk estimates (Paper 2)

3) Deriving an easy to use correction for the bias in estimates of familial risk due to the left-truncation of familial exposure (Paper 3)

4) Applying the methodology developed in 3) to correct for the bias in overall and age-specific familial risks estimated from follow-up cohorts extracted from the Swedish MultiGeneration Register and the Swedish Cancer Register (Paper 4)

Chapter 4

Methods

In this thesis, we have used virtual populations simulated through the Poplab package, that will be described in the following section, to mimic the Swedish population with complete families. The size of the simulated populations is specific to each Paper.

Each virtual population experiences cancer incidence, and cancer is assumed to aggregate in families. We are interested in the impact of incomplete family links and/or incomplete exposure on estimates of familial risk of cancer. Our exposure of interest is defined as a family history of cancer (i.e., affected first-degree relatives). We have mainly used female breast cancer throughout Paper 1 to Paper 3, and investigated other four most common cancers (colorectal, lung, prostate and melanoma) in Paper 4.

4.1 Simulating virtual populations: the *Poplab* package (Paper 1-4)

The simulation package Poplab, described in Paper 1, was developed in order to create virtual populations of complete families. This package was written for the R environment [117] (see [118] for free download). The first simulated year (baseline year) is the earliest time point for which data are available, and the populations can be constructed for as long as there are vital statistics. The technique allows the simulated population to evolve dynamically over calendar years, not only to arrive at a correct representation of the present, but to approximate the evolving population

in all its intermediary states.

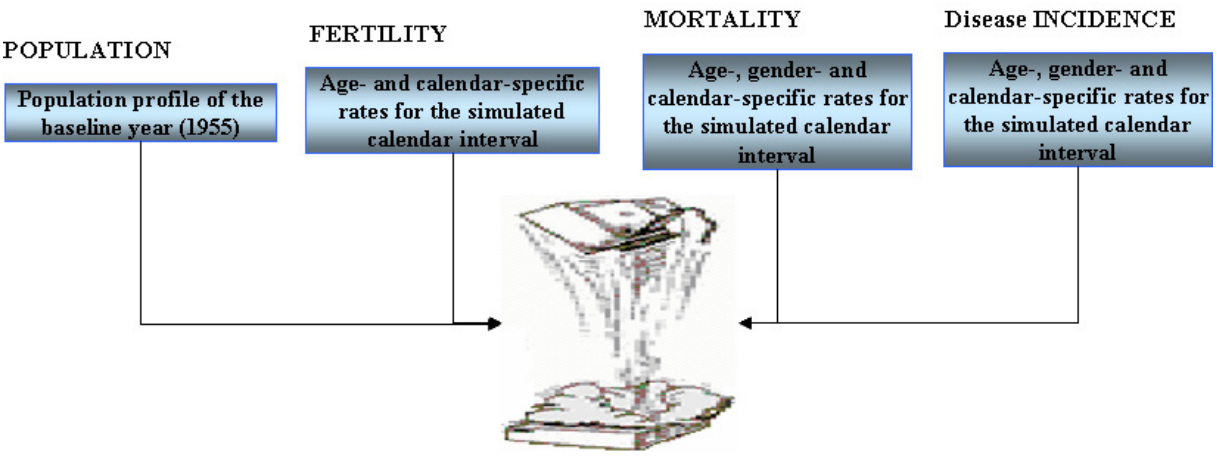


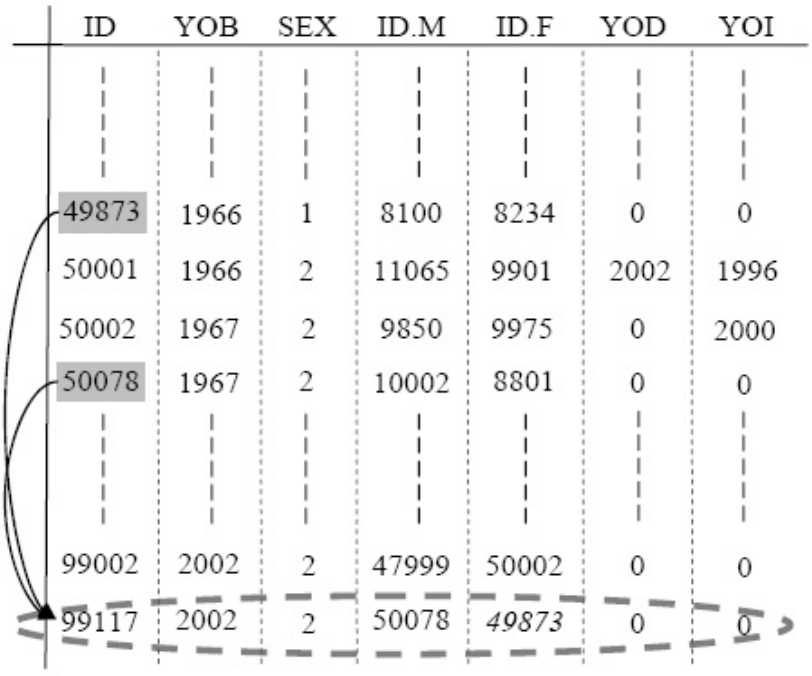
Figure 4.1: Input files required by *Poplab*

The required input files are, as illustrated in figure 4.1, a) the population profile for the baseline year, separately for males and females; and b) for the each year that is simulated, the female fertility rates, the mortality rates separately for males and females, and disease incidence rates, all age-specific. The simulation creates a pedigree structure (figure 4.2), which increases with every calendar year, as it contains all individuals who ever belonged to the population at any time from the base year to the latest simulated year. Each row in this data matrix represents an individual with his or her vital information: ID number (ID), year of birth (YOB), sex (SEX = 1 for males, 2 for females), ID number of mother (ID.M), ID number of father (ID.F), year of death (YOD), and year of incidence (YOI). These individuals are called index persons. If death or disease have not yet occurred, YOD and YOI, respectively, are set to missing (coded as 0 in the data).

For each new birth, the baby is added to the pedigree file with the parents correctly identified. Figure 4.2 highlights that the baby girl (ID = 99117) is born in the year 2002, and her parents' identity is recorded. "Death" of a person is simply recorded in the pedigree, by updating the year of death (YOD) for this person (see ID=50001). The pedigree structure enables dynamic storage of all individuals and straightforward identification of kinship at any time point. The simulation starts with the creation of the baseline population of related individuals for the base year. The birth and death processes are applied each year and the population is updated accordingly.

4.1.1 Creation of the baseline population

The baseline population is a pedigree structure containing a scaled version of the real population for the base year, and it constitutes the input population for the calendar period to be simulated. The only feature regarding the baseline population that is known before the simulation is the gender-specific age-profile, but a correct representation of the base year should include related individuals with this age profile. Starting with an assumed number of females and approximately the same number of males with the age distributions of baseline year, we assign unique ID numbers and an indicator for sex. The number of males is calculated based on the number of females so that the sex ratio for the baseline year is preserved. We assume these individuals were alive 100 years back in time, and regard them as the "founder" population, thus setting their parents ID numbers to missing. Their year of birth is calculated as the difference between the year they were sent back to and their age. Their year of death is set to missing (since we started with the living population). A "run-in" process of 100 years is simulated, where these founders and their descendants give



ID	YOB	SEX	ID.M	ID.F	YOD	YOI
49873	1966	1	8100	8234	0	0
50001	1966	2	11065	9901	2002	1996
50002	1967	2	9850	9975	0	2000
50078	1967	2	10002	8801	0	0
99002	2002	2	47999	50002	0	0
99117	2002	2	50078	*49873*	0	0

Figure 4.2: Pedigree structure generated by *Poplab*

birth and die with the base year fertility and mortality rates, respectively. The "run-in" time is by default 100 years, but the value of this parameter can be changed by the user. The details of the fertility and mortality processes are presented in Paper 1, Appendix 1. Also, figure 4.3 illustrates schematically the technical steps in procedure Give Birth(). In the population alive that results from this process, even the oldest individuals (100-years-old) are linked to their parents, as they were born during the run-in simulation. Thus, we construct a population in which no one is left unrelated.

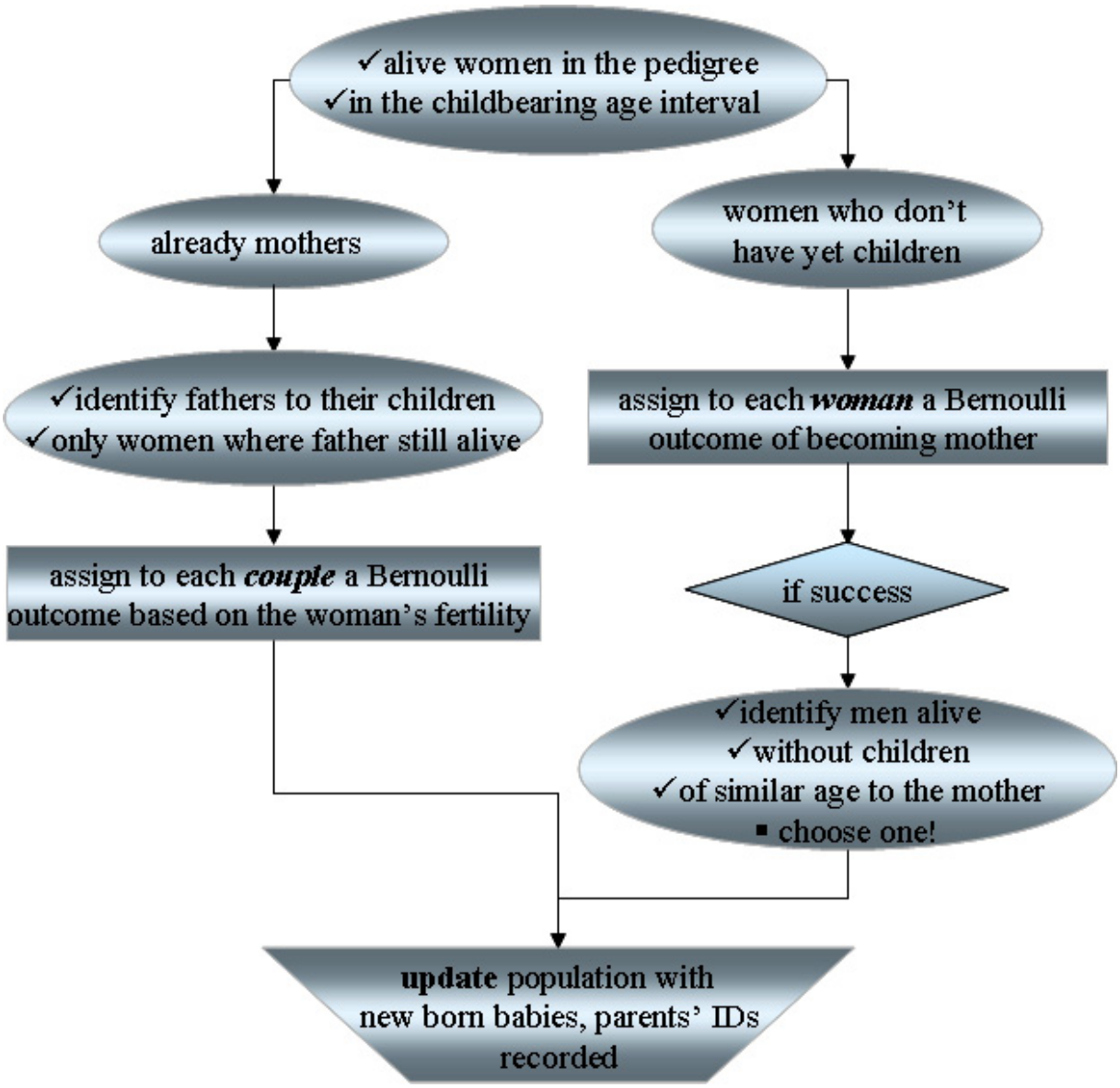


Figure 4.3: Schematic representation of the GiveBirth procedure in *Poplab*: successive selections of women (and men) eligible to become parents (spherical representation) and random identification of parents (rectangular).

Finally, we trim the created population to yield the real age profile for the base year, using the trimming algorithm that is explained in the next section. Because our goal is to obtain a baseline population of complete families, we add to the trimmed population any parents that were removed by the trimming.

4.1.2 The trimming process

The basic idea behind the trimming process is to randomly delete a proportion of the simulated individuals in each age group, in such a way as to yield the real age profile. The proportion of individuals trimmed in each age group depends on two descriptive features of the real population: the modal (i.e. most frequently occurring) age category, which we call the reference age, and the ratios between each age-specific count and this reference age count. The trimming process 'tailors' the simulated population so that it reflects these same two features, by using the reference age group to "tie down" the distribution. To do this, we calculate the ratio of age-specific counts to the reference age count for the simulated population using the same reference age as in the real population, and these latter ratios are referred to as simulated ratios (Table 4.1). Since the method aims to impose the real ratios on the simulated population, then any age groups which have a simulated ratio greater than the real ratio, will have enough individuals from which to eliminate the surplus proportion to match the real age profile. However, if this is not the case and there are age categories in which the real ratio exceeds the simulated ratio (e.g, ages 8 and 34 in the illustration), we must first diminish the reference age count in the simulated population to ensure that there are enough individuals in all age groups to impose the real ratios. The algorithm for adjusting the reference age count is presented in Paper 1, Appendix 2, together with a proof that all age categories can then have the real ratios imposed.

By multiplying the adjusted reference age count in the simulated population by the age-specific real ratios, we obtain the number of individuals in each age group such that the real population age profile is imposed on our simulated population; these age-specific counts are referred to as corrected counts. The last step, based on these corrected counts, is to eliminate individuals that do not belong in the final population.

age	*real* count	*real* ratio	*simulated* count	*simulated* ratio	ratio	*corrected* count
0	50251	0.797	1164	1.089	1.366	827
1	52342	0.830	1192	1.115	1.343	861
2	52298	0.829	1144	1.070	1.291	860
3	52202	0.828	1058	0.990	1.196	859
.....						
8	63010	0.999	1044	0.977	0.978	1036
9	*63077*	*1*	*1069*	*1*	*1*	*1037*
.....						
34	**61599**	**0.977**	**1013**	**0.948**	**0.970**	**1013**
35	50829	0.806	946	0.885	1.098	836
36	51217	0.812	981	0.918	1.131	842
.....						

Table 4.1: Illustration of the "trimming" technique, based on 50.000 female- and 50.000 male-founders. The *real* and *simulated* ratios are calculated by dividing the age-specific real and simulated counts, respectively, to the count for the modal age (here 9 years). Column 6, *ratio* represents the relative magnitude of the simulated and real ratio, and is used to identify the adjustment factor for the simulated modal count (see Paper 1, Appendix 2 for more details).

This is achieved by an age-specific Bernoulli process, applied to every individual, with the probability of deletion equal to the proportion of surplus individuals.

4.1.3 The evolution of a population

The evolution of a simulated population over a calendar-period is achieved through yearly birth and death events, as well as disease incidence, as this is a common feature of interest for genetic epidemiology. Incident cancers are assigned both during the run-in and evolution of the population, using a Bernoulli process that operates on individuals who are alive and cancer free through procedure AssignCancer() (see Paper 1, Appendix 1). Since cancer incidence is included in the run-in period, the

baseline population has prevalent cases with the year of incidence recorded. The AssignCancer() procedure requires the specification of the familial aggregation model, with choice between the following: the model where the age-specific risk of disease incidence for an individual is increased by a constant factor (the incidence rate ratio) if their relative is a case, the odds-ratio model, where the odds of disease increases by a constant factor in relatives of cases, and the age-dependent versions of these models, where the risk (and odds, respectively) increases by a constant factor as a function of the age at incidence in an affected relative. The user has also to choose for "affected relative" between a parent and a sibling. In simulating death events, the age-specific mortality for diseased individuals, referred to as the case mortality ratio, can be increased by a constant value, which is also specified by the user.

4.2 Statistical analyses of familial risk (Paper 1-4)

Familial risk of cancer was estimated from three standard analyses (Poisson analysis, conditional logistic regression and Cox analysis), which are described below. The study cohort consists of all individuals who were alive and cancer-free at the beginning of follow-up and those who were born before the end of follow-up. Familial exposure enters analyses as a binary variable for the parental-risk, sibling-risk and parental-odds models, and as a categorical variable with the reference group consisting of unexposed daughters for the model where familial risk changes with maternal age at incidence.

Familial risk was assessed from all three regression models in Paper 1, while in Paper 2-4 we use just the Poisson analysis.

For the Poisson analysis, data are first summarized for each calendar year to yield the total number of persons at risk and the total number of cases in each stratum defined by the familial exposure and age group. In illustrations where we have used age-specific incidence rates constant over the entire simulated calendar period (Paper 1 and 2), the analysis will be adjusted for age groups, but not for the calendar period. Thus, in these instances data are further collapsed to obtain the total number of cases and total number of individuals at risk in each age stratum over the entire period. In Paper 3 and 4, the familial risk was analyzed both within each stratum defined by

age group and calendar period and as overall estimates adjusted for these strata (see Methods section in these Papers for more details).

We also used a nested case-control design, where three controls are selected at random for each incident case. The controls are chosen from individuals who are alive and cancer-free in the year of incidence of the case, and of the same age with the case. The data are analyzed by means of conditional logistic regression.

When the data are analyzed by means of Cox regression, the entry time is age at the beginning of follow-up or zero for those born later. The exit time is age at incidence, death, or end of follow-up, whichever is smallest. If, during the follow-up of any individual, their relative (mother or sister, depending on the familial simulated model) becomes a case, family history enters the analysis as a time-varying covariate.

4.3 Simulating the virtual Swedish population (Paper 1)

We have used Poplab (see section 4.1) to create the virtual Swedish population for the calendar period 1955-2002 from 50,000 female- and 50,000 male-founders, and considered female breast cancer as the disease with familial aggregation. Figure 4.4 presents a simplified overview of running Poplab to create this population. To preserve simplicity in this illustrative instance, we used age-specific constant incidence rates (1980's rates) over calendar time throughout the simulated calendar period. The value of the case mortality ratio is 2. We simulate several models of familial aggregation: (i) the maternal relative-risk model of disease aggregation, where a woman's age-specific risk of disease incidence is increased by a constant factor if her mother is a case, (ii) the maternal odds-ratio model, and (iii) a model where the relative risk is modified by maternal age at incidence. For the first two models, we simulate separately a "null hypothesis" population of no familial aggregation of disease, and an "alternative hypothesis" population where the risk and the odds, respectively, are doubled in daughters of affected mothers. For the third model, the risk is increased by a factor of 4 for women whose mothers were incident before the age of 50 years, compared with daughters of unaffected mothers, and by a factor of 2 for daughters

of women diagnosed after the age of 50. We also considered the sibling relative-risk model where a woman's age-specific disease risk is doubled after a diagnosis in any of her sisters.

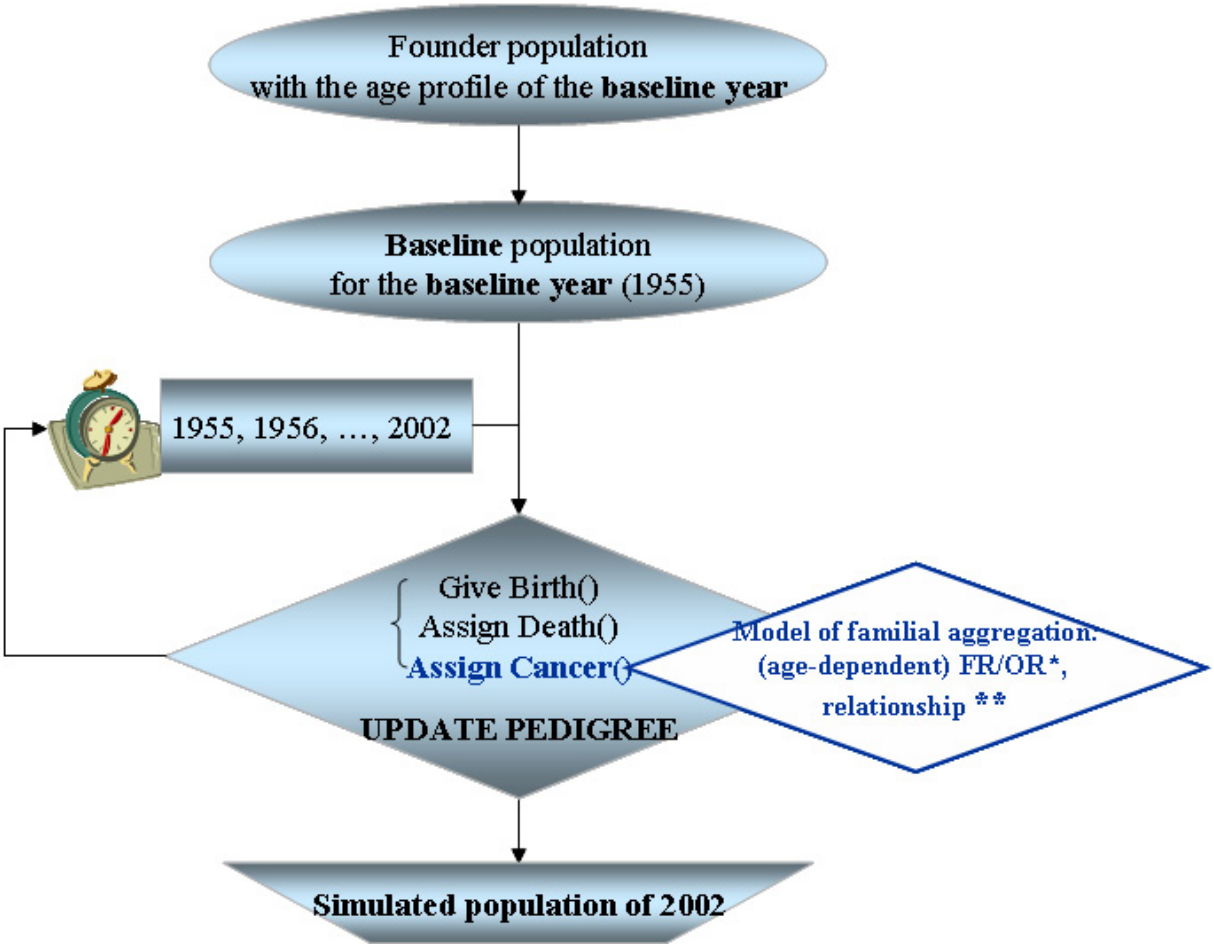


Figure 4.4: Simplified overview of running *Poplab* to create a virtual population for the calendar-period 1955-2002. *FR: familial risk, OR: familial odds ratio; **relationship: parent/sibling.

In order to verify that the values of familial association parameter of incident breast cancer that were employed in the simulation can be recovered accurately from the virtual populations, we analyzed the cohort at risk for the calendar period 1955-2002 using the three models introduced in Section 4.2. The study population consisted of all individuals who were alive and cancer-free at the beginning of follow-up (1955) and those who were born before 2002.

4.4 Family data with missing links and truncation (Paper 2)

Population and disease registers are subjected to various sources of missingness (see Section 2.4.3), and, when merged, the effect on familial aggregation estimates from using data that combines several patterns of incompleteness should be evaluated. Virtual population registers with complete family history of disease and complete family links, as created by Poplab, offer a golden standard for examining such effects. We first mimic the left-truncation of family history of disease due to the start-up date of the Swedish Cancer Register [13] and the missingness patterns seen in the Swedish MultiGeneration Register [12]. Next we evaluate the bias in familial risk estimates in relation with various background rates, age-pattern of incidence and population structure.

4.4.1 Range of investigations

We simulated virtual populations for the calendar time 1955-2002, starting from 500,000 female- and 500,000 male-founders. We used the real Swedish age- and calendar-year-specific mortality and fertility rates and the female breast cancer age-specific incidence rates of the year 1980. Several features of this cancer, such as the age-specific incidence profile, familial risk, and case mortality, indicate it as representative for the common cancers [119]. Familial aggregation of disease is simulated based on the maternal relative risk model where age-specific rates are multiplied by a constant value for daughters of affected mothers from the time that a mother becomes a case. Also cases experience a higher mortality compared to the general population.

Our investigations range over several levels of familial risk (2, 5, and 10), case mortality ratio (2, 5, and 10) and background incidence rates (1980 Swedish breast cancer rates scaled by a factor of 0.5, 1, and 2). The contribution of the disease age pattern is examined by simulating populations with a "constructed" incidence profile where the actual 1980 age-specific breast cancer rates were shifted to younger ages by 20 years (figure 4.5), so that in these populations a woman aged 30, for example, experiences the 50 year breast cancer rate. For each combination of these parameters,

an independent population was created.

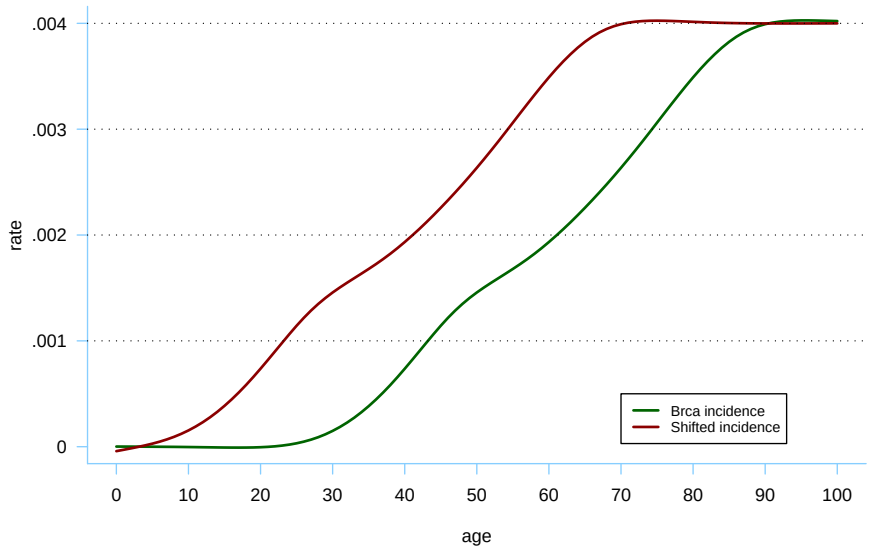


Figure 4.5: Breast cancer incidence and shifted incidence profile (smoothed curves).

4.4.2 Mimicking the incompleteness

Starting with the simulated populations, with complete family information, two types of incompleteness are considered: the left-truncation of family history of disease due to the start-up date of the Swedish Cancer Register and missingness patterns seen in the Swedish MultiGeneration Register. While the accuracy and completeness of incident cancer registration after the start-up date are prime features in the Cancer Register [76], cancer events occurred before this date are not identifiable. To reflect this, maternal cancer incident before the first year of registration is "hidden" in the simulated populations (i.e. maternal YOI becomes 0).

The completeness of family relationships in the MultiGeneration Register, such as the identification of a biologic parent, has been shown to depend on the date of birth and date of death (see Figures 2.5 and 2.6). Approximately 50 percent of those persons who died prior to 1991 have unidentified parents, and, similarly, about 10% of those who died between 1990 and 2002. We mimicked this by assigning to each individual who died before 1990 a Bernoulli event that his/her mother is unknown

with probability 0.5, and with probability 0.1 to those who died between 1990 and 2002.

4.4.3 The apparent relative risk

After imposing the two missingness patterns described above, we analyzed the cohort at risk for the calendar time 1955-2002 in a fashion that ignored the incompleteness and we obtained what will subsequently be called the apparent relative risk: for the left-truncated data, daughters of mothers incident before 1955 were treated as unexposed; in the analysis of data with missing family links, only those persons with a known mother were included. Populations affected by these two forms of incompleteness were created, first separately and later simultaneously, and analyzed in Poisson regression models (see section 4.2).

4.5 Bias due to registration start-up, theoretical considerations (Paper 3,4)

Familial risk estimates resulting from merged population and disease registers [119], [120] are potentially biased, especially due to the effect of start-up registration of disease events (see Paper 2). Left-truncation results in misclassifying the family history of disease for those individuals whose relatives became affected before the disease registration was initiated. In the following, the definition of family history is restricted to affected parent. Since truncation depends on the registration start-up, we expect that misclassification will be non-differential for diseased and non-diseased individuals. In attempting to correct for such bias, common practice would employ validation studies to collect, for a subsample of subjects, additional error-free covariate data, that may offer information on the true prevalence and sensitivity of the observed exposure. In the context of family studies, such proxy data are often not available.

We present in the following sections a bias-correction method that uses the Poplab package to estimate the sensitivity of exposure. We illustrate this approach in a framework mimicking the Swedish reality. First, we simulate virtual population registers of related individuals where the family history of disease is complete. We will refer to

these registers as the complete populations. Next, we mimic the lack of family history knowledge due to the start-up of cancer registration. In addition, demonstrating the extent of bias and the efficacy of this bias-correction approach in the real data will follow in subsequent sections.

4.5.1 Testing the assumption of non-differential misclassification

The non-differential misclassification of exposure assumption claims that there is no relationship between misclassification and the disease status, or in our framework, diseased and healthy individuals have an equal probability of "loosing" their family history of disease. We tested (Paper 3) a possible statistical association between disease status and knowledge of the family history of cancer (i.e. sensitivity of exposure) [121]. The logistic regression model analyzed all exposed daughters from the cohort at risk in the complete populations, where the dichotomous outcome is an indicator for truncated maternal cancer (i.e. mother incident before registration start-up), and the predictor variable is the disease incidence of the daughter. The model adjusted for age, categorized in 5-year age-groups. The odds ratios, close to the null value, suggested that the sensitivity of the observed exposure is non-differential. This form if misclassification is assumed throughout the following sections.

4.5.2 The apparent relative risk

Let R denote the true relative risk and P(E) denote the true prevalence of exposure i.e. the proportion of exposed individuals in the population at risk. The observed prevalence of exposure (in the truncated population) is denoted as $P(\hat{E})$. The expression developed by Flegal et al [97] for the apparent relative-risk $\widehat{R}$ was already introduced in Section 2.5.1 of the present thesis. Basically, $\widehat{R}$ is a function of the sensitivity, specificity and the true prevalence of exposure. The truncation of disease events due to registration start-up affects only the sensitivity, and not the specificity, of exposure. That is because all subjects truly unexposed are recorded as such, i.e.

the specificity of exposure is 1. The sensitivity of exposure, S can be written as:

$$S = P[\hat{E}|E] = \frac{P(\hat{E})}{P(E)} \tag{4.1}$$

Thus, sensitivity is simply the proportion of exposures that are recorded by the register.

The apparent relative risk can be written as:

$$\widehat{R} = R \cdot (1 - SP(E)) \cdot \frac{1}{(1-S)RP(E) + P(\overline{E})} \tag{4.2}$$

where $P(\overline{E})$ is the complement of P(E). Appendix 1 in Paper 3 provides the derivation of this expression at length, as well as the form of the apparent relative-risk in the context of differential misclassification of exposure.

From equation (4.2), the following algebraic properties of $\widehat{R}$ can be derived:

a) For $R > 1$, $\widehat{R} <= R$ i.e. the bias is towards the null

b) Keeping R and S constant, $\widehat{R}$ is a decreasing function of P(E). This means that for a given true value of R, with higher P(E) (as would it be the case with more incident diseases), the apparent relative risk becomes increasingly biased.

c) Assuming P(E) and R constant, $\widehat{R}$ is an increasing function of S; thus with a longer life-span of a register, which improves sensitivity, $\widehat{R}$ becomes less biased.

Rearranging equation (4.2), the bias-corrected relative risk R_{bc} can be written as

$$R_{bc} = \widehat{R} \cdot \frac{P(\overline{E})}{1 + P(E)((\widehat{R} - 1)S - \widehat{R})} \tag{4.3}$$

4.5.3 Bias-corrected relative risk, R_{bc}: point estimate and variance

In any real data context, the prevalence and sensitivity of the observed exposure may depend on several factors. Among these, the age-group of the individuals at risk (the sensitivity for older age-groups will be lower than that of younger groups, at least in the beginning of registration, since the former category has relatives whose disease is more likely to be overlooked), and the calendar-period (sensitivity of exposure increases with time after registration start-up as more familial exposures have

the chance to be captured by the register). If this is the case, stratum-specific corrections should first be calculated (for example, in the strata defined by age-groups and calendar periods). Denoting the variance of the bias-corrected log relative-risk, $\beta_i = log(R_{bc,i})$, in stratum i as $var(\beta_i)$, the overall bias-corrected estimate of the log-relative-risk for a specific subpopulation, such as a certain age-group, can be calculated as the weighted average of the appropriate stratum-specific parameter estimates:

$$\beta_W = \frac{\sum w_i \beta_i}{\sum w_i} \tag{4.4}$$

where the weight w_i is the inverse of $var(\beta_i)$, and thus the variance of β_W simplifies to:

$$var(\beta_W) = \frac{1}{\sum w_i} \tag{4.5}$$

From expressions (4.4) and (4.5), $R_{bc,W}$ and its variance can be written:

$$R_{bc,W} = exp(\frac{\sum w_i \beta_i}{\sum w_i}) \tag{4.6}$$

$$var(R_{bc,W}) = var(exp(\beta_W)) = (exp(\beta_W))^2 var(\beta_W) \tag{4.7}$$

4.5.4 Estimating the prevalence and sensitivity of exposure

We mimicked the lack of family history knowledge before the start-up date of registration by adding a new variable to the pedigree structure that stores the complete populations (see Figure 4.2), called the apparent family history. Technically, it records the year of incidence of the parent if this was affected after the start-up date and is 0 otherwise (i.e. parents incident before registration are assumed to be cancer free). This variable is similar to family history as it would be observed in the real Swedish data.

The sensitivity of exposure is calculated in the population at risk as the ratio between the proportion of those with an apparent family history and the proportion of those who truly have a positive family history. Since both these quantities use the

number of individuals at risk as the denominator, the sensitivity reduces to a simple ratio between the number of offspring in the study cohort with a parent incident after the start-up date of registration and all offspring with an affected parent. As the prevalence of exposure depends on the familial risk [129], the true prevalence of exposure should be estimated from the apparent prevalence of exposure and the sensitivity, according to expression (4.1).

4.5.5 The bias-correction

Using the estimates of $\widehat{R}$, S and P(E), we corrected the apparent relative risks (on the log scale) in each of the strata defined by age-group and calendar-period. The standard errors of these log-bias-corrected risks were estimated from 100 bootstrap samples of the population at risk. Weighted averages (equation (4.4)) of the stratum-specific estimates were then computed to obtain the calendar-period-specific and age-specific log-relative risks. These were further exponentiated according to equations (4.6) and (4.7)) to obtain the bias-corrected relative risks and their standard errors.

4.6 Correcting for the bias due to registration start-up (Paper 3)

In Paper 3 we illustrate the bias-correction methodology on simulated populations. We mimic the Swedish population for the time period 1955-2002. Female breast cancer was chosen as an example of disease with familial aggregation, and for presenting methodological "advantages" such as the age-specific incidence profile, magnitude of the familial risk and case mortality ratio, suggesting this disease representative of many cancers [119]. For simplicity, we used constant age-specific incidence rates (of the year 1980).

Our follow-up period was 1955-2002, and the study cohort is defined as in section 4.2. In the real Swedish population, approximately 3.5 million female individuals match the follow-up definition. To create a virtual population of similar size, the simulation started with 4 million founders (2 million males and 2 million females), resulting in a baseline population of approximately 5 million related individuals for

the year 1955. This population evolves until 2002, with an assumed relative risk model, where for daughters of affected women, we multiplied age-specific rates of cancer by a constant factor (the incidence rate ratio, IRR). Three familial risks are considered, IRR = 2, 5 and 10, and an independent population is created for each of these values. The case mortality ratio was 5.

The apparent relative risks in strata defined by age-groups (5-year) and calendar period (decades) were estimated from Poisson regression models. Estimates of S and P(E) were calculated in each strata, and the corrections were performed as explained in section 4.5.5.

4.7 Evaluating the bias in the real data (Paper 4)

Using the bias-correction methodology described in Section 4.5, we examined the bias in familial risk estimates in study cohorts obtained by merging the Swedish Multi-Generation Register [12] and the Swedish Cancer Register [13]. From the real data, we estimate the observed prevalence of exposure and the apparent relative risk (RR). From the simulated data, we estimate the sensitivity of the observed exposure. These quantities are used in expression (4.3) to calculate the bias-corrected RR, which is then compared to the apparent RR to evaluate the bias. Figure 4.6 displays this process for five of the most common cancer sites (colorectal, lung, female breast, prostate and melanoma). We investigated how age-group specific and overall estimates are affected by the left-truncation in the Cancer Register.

4.7.1 The apparent relative risk

Follow-up for each individual was started at birth or January 1, 1961, whichever occurred latest, and was terminated on diagnosis of first cancer, death, emigration, or the closing date of the study, December 31, 2002, whichever occurred first. Individuals who appear in different generations in the MultiGeneration Register, first as offspring and later as parents, were considered independently. Multiple affected offspring in the same family were treated as independent events. Apparent relative risks were calculated from Poisson regression models.

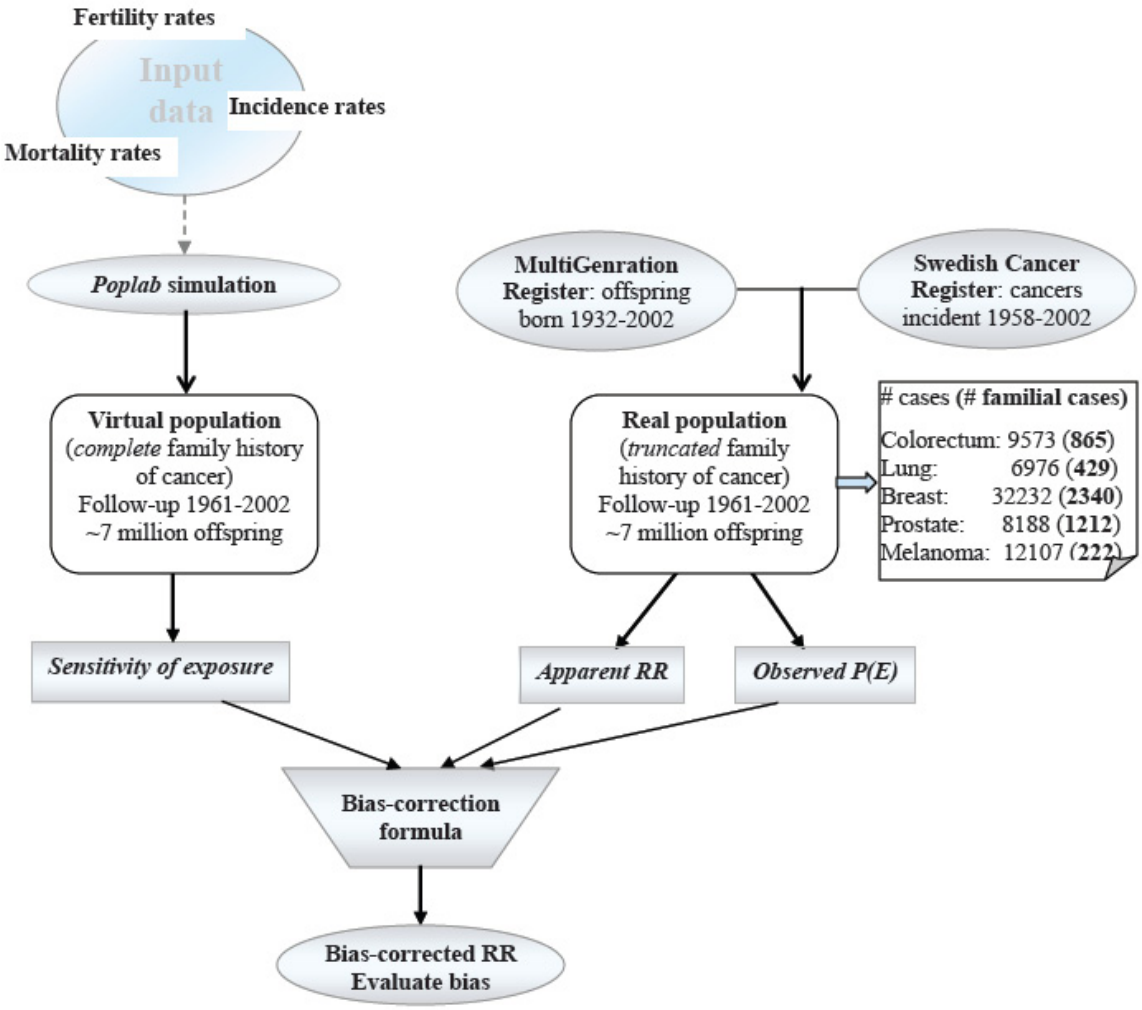


Figure 4.6: Schematic representation of the steps involved in the evaluation of the bias in the apparent relative risk for five common cancers.

4.7.2 Sensitivity and true prevalence of exposure

We use Poplab to simulate, for each investigated cancer site, an independent virtual population register with complete individual family history, for the calendar period 1961-2002. Each baseline population is created from 6 million unrelated founders (3000000 males and 3000000 females), and evolves to the year 2002. Cases have an age-specific mortality that is five times that of the general population. Positive family history of cancer is defined as an affected mother for breast cancer, an affected father for prostate cancer, and an affected mother or father for colorectal cancer, lung cancer or melanoma. We used a relative-risk model, where population age-specific rates of disease are multiplied by a constant factor for exposed individuals from the year of incidence of the parent. As we did not find an impact of the familial risk used in the simulation on the sensitivity of exposure (see Figure 5.3), we used in the simulation the apparent relative risk of that specific cancer.

We impose the effect of start-up of cancer registration on the complete virtual populations, and estimate the sensitivity and true prevalence of exposure as described in section 4.5.4.

Chapter 5

Results

5.1 Paper 1

We illustrated the use of Poplab with the creation of the virtual 1955 - 2002 Swedish population with related individuals, but of a smaller magnitude than the real population. The age profile of the 2002 simulated population was very similar to the real profile, and the sibship size distribution (the number of offspring in nuclear families) displayed a reasonable agreement. All these investigations were stable across repeated simulations. The average age at first birth for mothers of offspring born after 1932 and alive in 2002 resembled closely the real data in the MultiGeneration Register (Paper 1, Figure 4). When female breast cancer incidence was included in the simulation, we defined familial exposure as having an affected mother, and consequently increased the age-specific risk of incidence for exposed individuals with a constant factor. For all assumed models of familial aggregation, the parameters used in the simulations were faithfully estimated (table 5.1). The age-specific incidence rate ratio (IRR) extracted from the simulated populations were similar to those employed in the simulation (figure 5.1).

5.2 Paper 2

Table 5.2 shows the results of Poisson analyses of three populations simulated with different values of familial risk (2, 5 and 10), after imposing the left-truncation due

Familial aggregation	**True value**	**Poisson regression IRR**	**(95% CI)**
RR	1	0.94	(0.81, 1.09)
	2	1.97	(1.77, 2.19)
OR	1	1.02	(0.88, 1.18)
	2	2.05	(1.85, 2.28)
RR changing with MAI†			
$^{<}50$	4	3.71	(3.15, 4.36)
$^{\geq}$ 50	2	1.96	(1.75, 2.20)

CI confidence interval; IRR incidence rate ratio

RR relative risk; OR odds ratios

† MAI = maternal age at incidence

Table 5.1: Statistical analyses of familial aggregation of cancer for several simulation scenarios

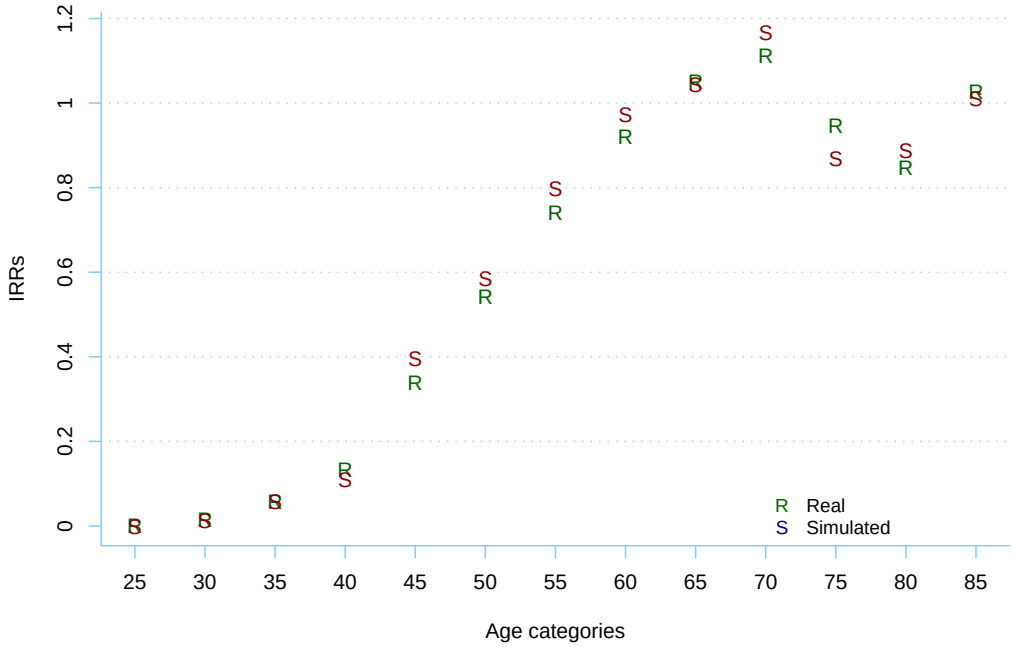


Figure 5.1: Comparison of incident rate ratios (IRRs) used to simulate disease and the IRRs estimated from the simulated population. Data is shown for the model with FR = 2. Age is categorized in 5-year groups (represented here by their upper limits), with the 85+ group as reference.

to registration start-up, missing maternal identity based on the year of death of the index person and the combined effects of these two sources of missingness. The bias due to left-truncation increased with the value of familial risk, and it had a dramatic impact when the familial risk was 10. With missing maternal identity, which results in less individuals being analyzed as only those with maternal information are included in the analyses, the risk estimates were very close to the true values, but the associated 95%CI were somewhat broader than when analyzing the complete populations (see Paper 2). Nevertheless, when the populations were simulated with differential mortality for familial and non-familial cases, the estimates were biased for all 3 values of familial risk, as it can be seen in table 5.3. When imposing both left-truncation and missing mothers simultaneously, the patterns of bias were similar to those caused by left-truncation only (i.e. increasing with familial risk), but of a smaller magnitude. All these investigations were performed for 3 different values of case mortality ratio (2, 5 and 10), but we found no impact of this parameter. The results in Table 5.2 are shown for a mortality ratio of 5.

Truncated*		
FRR = 2	**FRR = 5**	**FRR = 10**
1.88(1.79, 1.98)	4.33(4.20, 4.47)	7.47(7.31, 7.64)
Missing mothers†		
FRR = 2	**FRR = 5**	**FRR = 10**
2.01(1.94, 2.09)	5.09(4.96, 5.23)	10.14(9.93, 10.35)
Combined‡		
FRR = 2	**FRR = 5**	**FRR = 10**
1.98(1.87, 2.09)	4.64(4.49, 4.80)	8.16(7.96, 8.36)

* Maternal cancer truncated before 1955.

† Missing mothers based on daughters' YOD.

‡ Combining the truncation and missing mothers.

Table 5.2: Overall effects on familial risk

We continue to investigate the impact of left-truncation for background incidence rates of different magnitude, by simulating populations with age-specific incidence rates that were half and double, respectively, those of breast cancer. Figure 5.2 shows the results from analyzing the populations simulated with a familial risk of 5. The bias was more pronounced as the disease was more common, and this effect was even more visible for higher values of risk (Paper 2, Table 3).

We also investigated the contribution of the age distribution of disease and age structure of the population to the bias from left-truncation. We "constructed" a theoretical disease incidence, in which the rate applied to each age group was the breast cancer incidence rate for women 20 years older (figure 4.5), which results in substantial incidence rates at young ages. We compared the populations simulated with this incidence to those simulated with breast cancer incidence, and found a much larger bias especially for a high familial risk (for FR = 10, the estimates were 5.87 95%CI (5.79, 5.95) and 7.47 95%CI (7.31, 7.64), respectively). We also mimicked the real Swedish family data context, by restricting the study cohort to persons born after 1932. The bias did increase with the true value of familial risk but was of a lesser magnitude than the bias resulting from analyzing the unrestricted cohorts.

Based on these investigations, we concluded that the left-truncation of family history due to registration start-up would be the cause of most dramatic bias in familial

Non-differential MR†		
FRR = 2	**FRR = 5**	**FRR = 10**
2.01(1.94, 2.09)	5.09(4.96, 5.23)	10.14(9.93, 10.35)
Differential MR‡		
FRR = 2	**FRR = 5**	**FRR = 10**
1.72(1.66, 1.79)	4.61(4.49, 4.74)	9.23(9.04, 9.43)

† Case mortality ratio (MR) is 10.

‡ Case mortality ratio is 2 for nonfamilial cases and 10 for familial cases.

Table 5.3 Effects of missing family links on familial risk estimates

risk estimates, and bias-correction methodology in the framework of population-based registers should be developed.

5.3 Paper 3

In Paper 3 we focused on finding a bias-correction formulation of familial risk estimates, that uses quantities which can be obtained from the simulated context. For a range of familial risks, the truncation of family history is demonstrated to result in non-differential misclassification of exposure, and sensitivity that has little or no dependence on the familial risk or the incidence rates (figure 5.3). Figure 5.4 illustrates the dependence of the prevalence of exposure on the value of familial risk.

As already concluded, the bias is most pronounced for high familial risks. We found that estimates could be dramatically biased especially when data are extracted from registers with a short life-span (Figure 5.5, panels (B:I) and (C:I)), and for older study cohorts (panels (B:II) and (C:II)). In all the situations studied, the bias-corrected estimates are in excellent agreement with the true values.

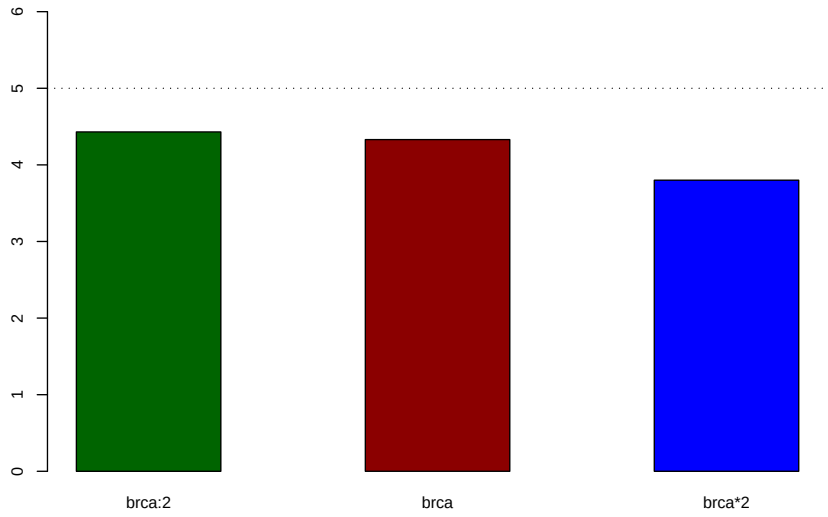


Figure 5.2: The impact of left-truncation on familial risk estimates changes with the background incidence rates. The dotted line represents the true value of risk.

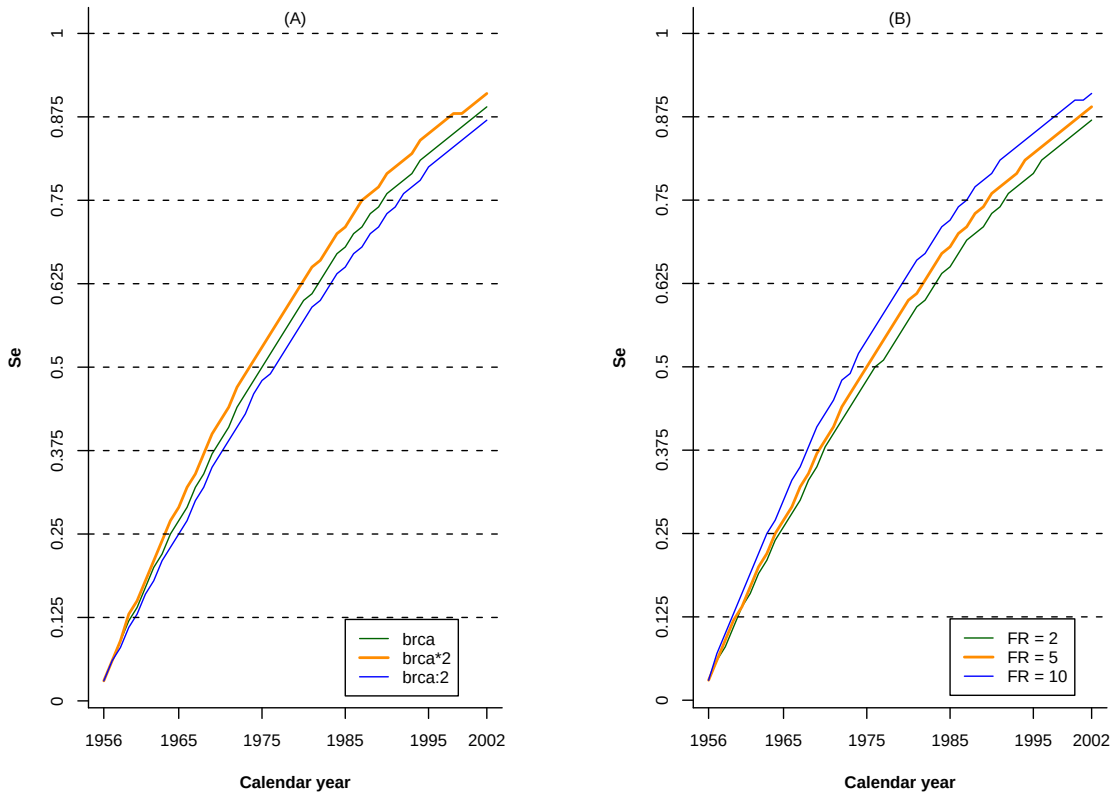


Figure 5.3: Sensitivity for a cohort at risk of breast cancer, for the calendar period 1956-2002, for various background incidence rates (panel (A)) and 3 values of familial risk (panel (B)).

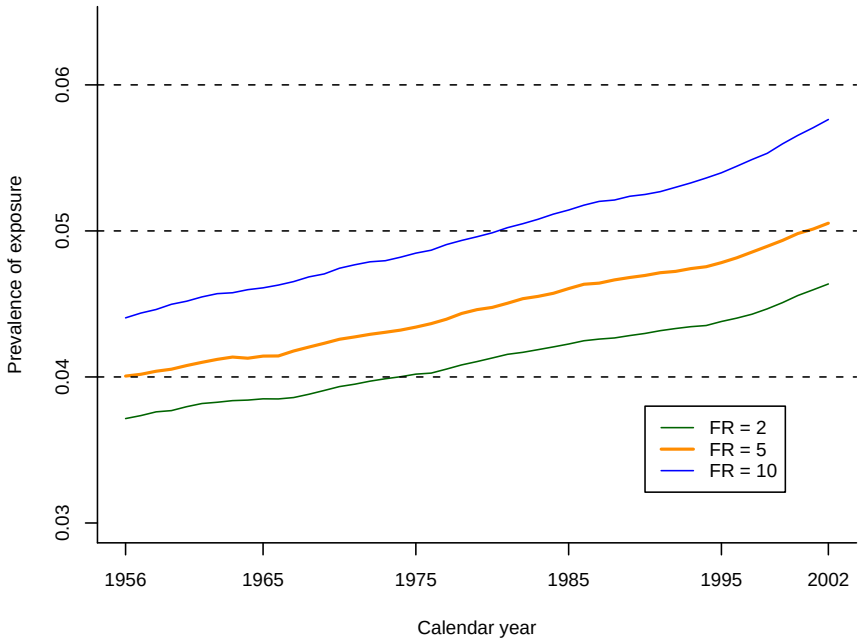


Figure 5.4: Prevalence of exposure (defined as "affected mother") for a cohort at risk of breast cancer, for the calendar period 1956-2002, for 3 values of familial risk .

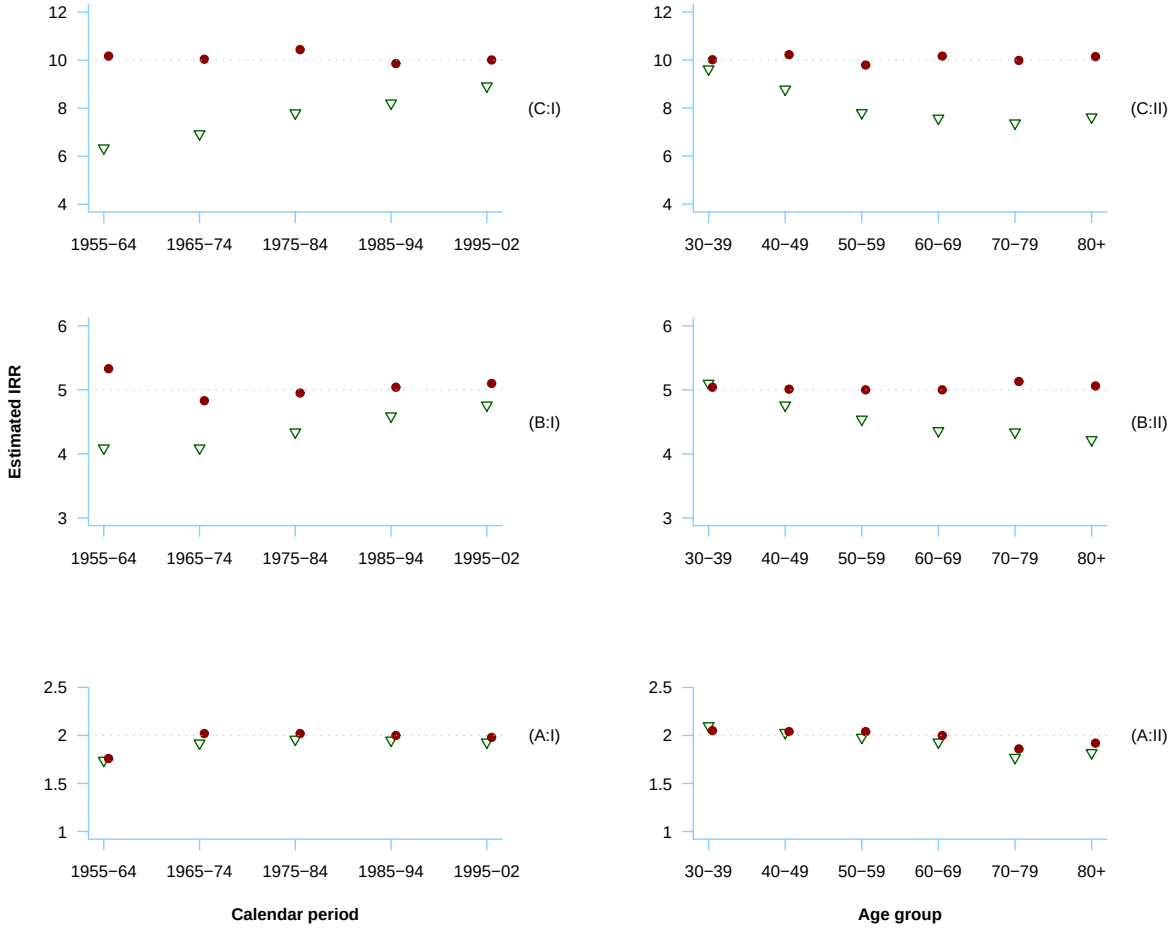


Figure 5.5: Apparent relative risk (IRR) estimates recovered from Poisson analyses of truncated populations (triangles), and corrected estimates computed by the bias-correction formula (solid circles). The panels show, for three values of familial risk, calendar-specific (on the left) and age-specific estimates (on the right): IRR = 2 (A:I and A:II), IRR = 5 (B:I and B:II) and IRR = 10 (C:I and C:II). The dotted lines represent the true values of familial risk.

5.4 Paper 4

In the last study we evaluated the bias in familial risk estimates from using the Swedish MultiGeneration and Cancer Registers, for five cancers: colorectum, lung, breast, prostate cancer and melanoma. Corrected age-group specific and overall estimates were close to the apparent relative risks for the first four cancers, with overall values of 1.99 95%CI (1.85, 2.14), 2.05 (1.86, 2.26), 1.84 (1.76, 1.92) and 2.33 (2.19, 2.48), respectively. For melanoma, the apparent estimate, 2.68 (2.35, 3.07) was somewhat smaller than the corrected estimate, 3.18 (2.73, 3.64), and the associated apparent 95% CI did not include the corrected value. When the exposure of interest is a parent affected at a younger age, the bias is more pronounced (Figure 5.6), with the naive estimate for melanoma changing from 4.07 (3.21, 5.16) to 5.67 (4.51, 6.83) after correction.

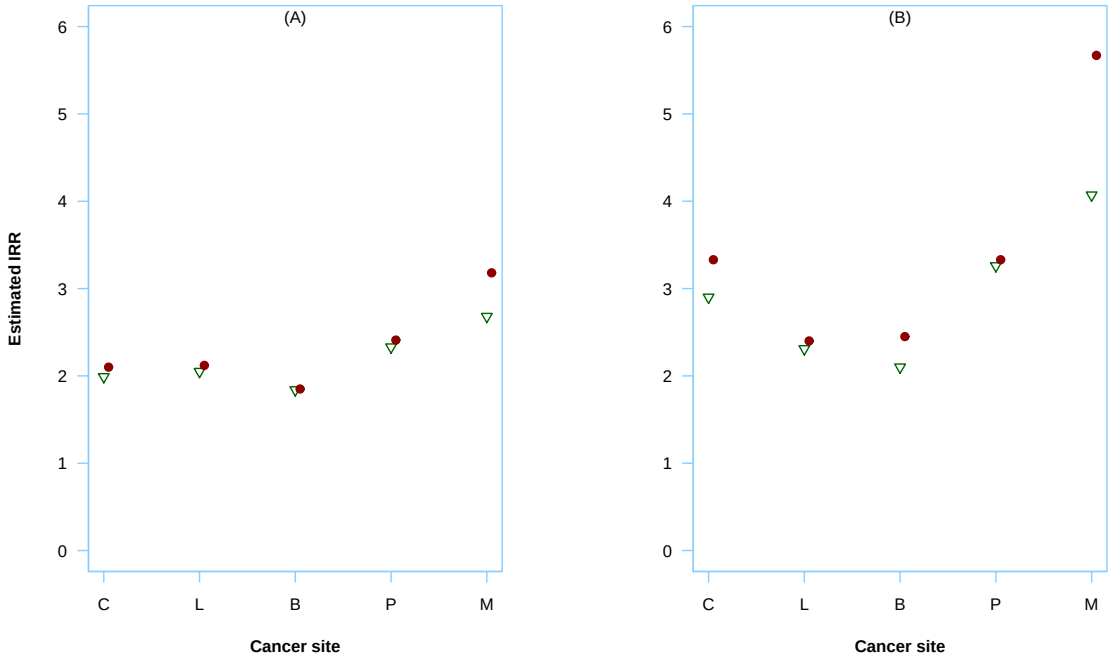


Figure 5.6: Overall apparent relative risk (triangles) and bias-corrected estimates (solid circles) in offspring with parental history of concordant cancer for 5 cancers: **C**olorectal, **L**ung, **B**reast, **P**rostate, **M**elanoma. Exposure is defined as an affected parent (panel (A)), and as a parent incident before 60 years of age for colorectal, lung and prostate cancer, and before 50 years for breast cancer and melanoma (panel (B)).

Chapter 6

Discussion

6.1 General overview of *Poplab* (Paper 1)

The objective of the present thesis was to assess the reliability of using population-based registers in estimating measures of familial aggregation of diseases. The impact of potential limitations of these data sources can be evaluated against a golden standard. We have developed a simulation tool, Poplab that creates virtual populations of related individuals evolving over calendar time, with complete family history of disease, by use of simple vital statistics, such as population fertility, mortality and incidence rates. This tool generates first the web of family relations at a given point in time, the baseline year, which by itself may be of interest. At the completion of the evolved population, an ideal population register was mimicked, where all relationships are fully known, and demographic and disease characteristics can be easily extracted for any given year.

Poplab allows the choice between several family models, such as the relative-risk and odds-ratio model, which could be constant across age-groups or vary with the age at onset of the affected relative, as this is a known feature of many heritable diseases [41], [122], [123], [124]. In addition, the familial aggregation could operate through parents or siblings. A simulation tool equipped with several disease models enables the investigation of the usual analytical strategies for their robustness to the assumed underlying disease model.

We have applied this methodology to mimic the Swedish population, and found

that the simulated population mirrors the real age-profile and the resulted average age-gap between mother and first born was in excellent agreement with the real population [12], as expected when using the correct age-specific fertility rates for each year. These features were reproduced in repeated simulations. The baseline and evolved populations both exhibit a reasonable sibship-size distribution, but they differ somewhat from the real population. Some of this discrepancy may be due to the incomplete family links in the MultiGeneration Register, especially for earlier birth cohorts. However, the surplus of families of size one and the deficit of families of size two are to be expected from simulating each birth as an independent event. In reality, a woman's childbearing is influenced by many factors, including desired family size, economic status and societal norms [125], [126], [127], [128]. As an initial illustration of how this environment can be used in studies of familial diseases, female breast cancer incidence was simulated under several familial models which increased cancer risk in offspring of affected mothers, by a known factor. We chose this disease due to a well acknowledged contribution of family history to this malignancy and a middle-range magnitude of familial risk (approximately 2.0) among those cancers that aggregate in families [119], [15]. Through standard epidemiological analyses, the value of the familial component of cancer used in the simulations was extracted from the virtual populations. This provided the basis for using the created populations to investigate how known values of familial risk would be modified by various sources of incompleteness. We mimicked and examined mainly two such sources encountered in the real Swedish registers: the lack of family history information and missing/broken family links between population members. The former problem originates from the absence of knowledge on affected individuals before the Cancer Register [13] was started, while the later one arises from the various inclusion/exclusion criteria and other administrative considerations of individuals present in the MultiGeneration Register.

6.2 How large is the bias, and what is causing it? (Paper 2)

Our illustrations included levels of familial risk (2, 5 and 10) and case mortality ratio (2, 5 and 10) that cover a spectrum of diseases with familial aggregation [33], [132], [133], [134], [135]. Due to already mentioned methodological advantages, we mainly relied on female breast cancer. To understand the contribution of each register to the magnitude of bias, in a first step only cancer history is truncated and family information is preserved, while in a second step family links are broken and the history of cancer is kept intact. Naive statistical analyses were then performed by using just the available complete data.

Left-truncation resulted in a downward bias for all models studied, thus yielding conservative estimates of risk, as expected when the misclassification of exposure is nondifferential [97], [136], [103]. The bias was of increased magnitude at high levels of familial risk and for large background incidence rates. For example, the incidence rate ratio estimates corresponding to the three levels of familial risk were 1.88 (resulting in 6 percent bias), 4.33 (13 percent bias), and 7.47 (25 percent bias), respectively, for the model with 1980 breast cancer incidence rates and a mortality ratio of 5. It is clear that, in the early years of registration, the information from the cancer register cannot give valid estimates of familial risk. However, as time passes, family history is more faithfully recorded; thus, study designs that use, for example, only recent years from a longer registry accrual time can produce valid estimates [137], [138]. We noted (Paper 2, Figure 2) that, for cancers with a low familial risk, reasonable estimates of familial risk can be achieved in a relatively short time (approximately 20 years). However, when this risk is higher, there is a more long-lasting bias.

The age pattern of disease also has an impact on the magnitude of bias due to left-truncation. In our illustrations, follow-up began in 1955; thus, cohort members affected by loss of family history were in their twenties (the Multi-Generation Register records only those persons born after 1932), and many had parents younger than 50 years of age at start-up. Thus, when the studied disease affects mainly older persons (for breast cancer, substantial incidence occurs after the age of 50 years),

left-truncation of maternal cancer will have only a modest impact on familial risk estimates since these mothers are unlikely to have developed the disease before registration start-up. However, when the disease affects younger individuals, as reflected by the population simulated with an incidence profile that was shifted with 20 years towards young ages, the loss of family history due to truncation can be substantial, resulting in serious biases (Paper 2, Table 3). Thus diseases like melanoma, thyroid and testicular cancer which are characterized by younger incidence profiles [70], [131], [139], [134], [140] are expected to be the most impacted by this problem, and we have demonstrated in Study 4 that truncating the family history of melanoma results in the most serious bias. Left-truncation of cancer diagnoses in siblings biased the estimates of familial risk to a lesser extent compared with the loss of maternal cancer diagnoses, as the overall age difference between sisters was considerably smaller than the gap for mother-daughter pairs.

By using patterns of missingness similar to the Swedish MultiGeneration Register, maternal identity was hidden in realistic proportions for individuals deceased within certain time frames. For all studied levels of familial risk and mortality ratio, there was little or no bias when mortality for familial and non-familial cases was the same. However, for some diseases, familial cases may experience a differential mortality. There are indications that carriers of the BRCA1 mutation for breast cancer might have a poorer prognosis than sporadic cases [130]. When we simulated populations with a higher mortality for familial cancer cases, we noted a bias with magnitude depending on the value for familial risk (Table 5.3). This higher mortality leads to preferential exclusion of exposed cases from the study cohort, when they do not survive to the time point when good-quality parental information is available (1991 in the case of MGR). Thus, the differential mortality would be expected to lead to a downward bias in the familial risk estimates, as we observed.

After separately studying left-truncation of family history and missing maternal identity, we constructed study cohorts that would result from using real register data subject to both of these sources of incompleteness. The bias in the familial risk estimates followed patterns similar to the bias resulting from left truncation only but was of a somewhat lesser magnitude (Table 5.2). This is a predictable consequence of some age groups (mainly older persons) being subject to exposure truncation and

exclusion from the study cohort because of an unidentified mother. Consequently, persons whose family history would otherwise be truncated are excluded from the study cohort, so that the bias of familial risk estimates is mitigated.

6.3 From evaluating the bias to bias correction (Paper 3)

Failure to account for left-truncation of exposure (family history of disease) due to registration start-up is likely to result in biased relative risks [10], [141], and the magnitude of such biases, as illustrated in Study 2, is influenced by several factors, such as the value of familial association, background incidence rates, the mortality mechanism for cases, disease age pattern and the biological relationship between case and relative. Considering the dramatic impact on the validity of estimates that truncation may cause, we have developed a bias-correction method, for which the required sensitivity of the observed exposure is estimated based on the simulation context Poplab. The expression of the bias-corrected relative risk is adapted from the published literature on non-differential misclassification of exposure [97].

The same values of familial risk were assumed as in Paper 2 (2, 5 and 10), and we continued exemplifying with female breast cancer incident according to age-specific 1980's Swedish rates. As the previous study suggested no impact of different values of case mortality ratio on the apparent relative risk values, we assumed a value of 5 for this parameter in all our investigations. We simulated populations of larger sizes (approximately 5 million founders as opposed to 1 million in the previous study) so that the bias-correction can be performed both for overall estimates and for age-group specific.

The correction performed well for all values of familial risk, and we obtained valid estimates even for registers with a short life-span (10-20 years) and for sparse age-groups. Our approach is a feasible alternative to the use of validation samples in correcting for left-truncation bias in family studies.

The effects of making the assumption of non-differential misclassification of exposure on both the direction and magnitude of bias and on the performance of bias-

correction methods have received a lot of attention [93], [94]. If this assumption does not hold, erroneous conclusions are drawn regarding the direction of bias, and "corrected" relative risks considerably higher than the truth are calculated [96]. In our case, we tested for an association between an indicator for family history knowledge (or equivalently, a mother becoming incident after registration start-up) and the disease status of her daughter. We did not expect the misclassification of exposure to be differential, and the odds ratios close to the null value supported this hypothesis. This finding simplifies the bias-correction methodology, as differential misclassification would require the sensitivity among diseased individuals to be estimated separately, which can prove difficult with sparse number of cases registered as exposed.

As the prevalence of exposure depends on the level of familial risk [142], the true prevalence of exposure should be calculated as the ratio between the observed prevalence of exposure and the sensitivity, two quantities that are estimable without knowledge of the true value of familial risk. Thus the bias-correction formula becomes a simple function of factors readily provided by the available data (i.e. the apparent relative-risk and observed prevalence of exposure) and an estimate of the sensitivity of the observed exposure.

6.4 Bias in studies of cancer using the Swedish registers (Paper 4)

We studied the familial risk for common cancers using cohorts from the MultiGeneration Register merged to the Swedish Cancer Register. We observed an age-dependent familial risk for colorectal, breast, prostate cancer, and also for melanoma, with relatively high risks at younger ages. For breast and prostate cancer, these observations are in agreement with numerous previous studies [143], [42], [144]. For colorectal cancer and melanoma, the literature has focused on the modification of familial risk by age of the index case, although some studies of colorectal cancer provide indirect evidence that younger relatives of cases are at higher risk [145], [146]. We found no differences in familial risks by age at onset for lung cancer, indicating a low impor-

tance of genetic factors (as compared to the effect of smoking habits) in this cancer [147], [148].

We observed little or no bias in the overall estimates of familial risk of cancer, with the exception of melanoma, a cancer with relatively young age at onset. The lack of bias for most of these cancers is due to the relatively low familial risk (RR approximately 2) and/or relatively low incidence in the population. Since melanoma has both a poor sensitivity and a relatively large value of familial risk, it would be expected to show the most biased RR (see Paper 2), as we observed. This bias was worst for exposure defined as a young age at onset in a parent, where the apparent relative risk in offspring of parents diagnosed before the age of 50 substantially underestimated the true (bias-corrected) risk. Such dependence on age is to be expected since more parental cancers will be truncated at registration start-up. A similar effect was observed for breast cancer.

We have focused on biases due to truncation of disease events, which is only one of the potential sources of incompleteness in any real population. For family studies, missing parental links will result in subjects whose exposure information (family history) is missing and these will usually be dropped from analysis and only complete nuclear families analyzed [149], [150]. Where the missing links depend on calendar time (for example, improvement in completeness of family registration over time) as is the case in the Swedish MultiGeneration Register [12], we have shown that significant bias only occurs where there is a very strong differential mortality between familial and non familial cases (Paper 2). We are aware of no differential mortality of such magnitude; even BRCA1/2 breast cancer has been shown to have a somewhat similar prognosis to sporadic breast cancer [151], [152], [153], [154].

In this study we have explored biases in the context of a positive family history being defined as an affected parent. Since a sibling is also a first-degree relative, an affected sibling provides important information about genetic susceptibility. Although some studies investigate the risk due to an affected sibling [155] or any affected first-degree relative [156], studies of parental relative risk are predominant in the literature. This is understandable, as cancer is generally a disease of older people, so that the parental generation provides more complete information about the disease profile in a family. On the other hand, due to their younger age relative to parents, siblings will

be less subject to left-truncation, so we would expect minimal bias in the estimates of sibling relative risks based on the findings presented here.

6.5 Simplifying assumptions and limitations

In all the simulations, the following simplifying assumptions were introduced:

(i) For each mother, we chose a spouse close in age (from 1 year younger to 4 years older), which is realistic in our Swedish data. Although this could be extended to a stochastic model, the age gap between spouses is not a primary factor of interest for our investigations.

(ii) In applying fertility rates we assigned each new birth as an independent event. Future extensions could accommodate more realistic fertility patterns (for example, influenced by parity and gap between offspring) and other family structures such as half-siblings and adoptions.

(iii) We simulated closed populations without immigration or emigration. Thus our method is suitable for studies of homogeneous populations; it could be extended to include immigration/emigration, provided the data are available.

When some of these simplifications interfere with the research questions under consideration, additional programming effort could transform the software as needed, provided also that appropriate population data are available.

One potential limitation, in Study 3 and 4, is that in creating the baseline population we use the first available incidence rates of cancer in the run-in simulation. This assumption has been used previously [10] in calculating the probability of a parent being disease-free before the start-up of disease registration. Since most cancers have a rising incidence with calendar time, these will overestimate the true incidence rates experienced by the population prior to the baseline year. We investigated the impact of the baseline incidence rates on the bias-corrected estimates, by simulating the breast cancer baseline population with age-specific incidence rates that were half the true rates. Both the age-group specific and the overall bias-corrected estimates were very similar to those obtained for the population where the baseline was simulated with the baseline rates.

Another potential limitation of our method is the assumption of a constant factor

increasing the age-specific mortality rates from the general population for all cancers. This is likely to overestimate the mortality in all but the oldest age groups or most fatal cancers [157]. However, we have shown that the bias in the familial risk is essentially independent of the mortality rate ratio.

Finally, we did not simulate with age-specific familial risks, but instead we used an average value of familial risk across all ages for every simulated cancer. We do not expect this assumption to influence our results, as we have found in Paper 3 that the sensitivity does not depend on the value of familial risk used in simulations.

Chapter 7

Conclusions

The software package Poplab is available for free download. With a running version of R [117], the user can investigate the performance of various analytical approaches to family data. Realistic settings for epidemiological exploration of diseases can be created by specifying appropriate input data. The simulation can also be run to a specified time point in the future (using projected vital rates or the latest rates available) to extrapolate the estimates of population disease burden or other features of interest.

Extensions of the package to incorporate additional features specific to other research questions can create valuable tools for experimentation and investigation.

We studied the impact on familial risk estimates of left-truncation of family history of disease and of missing parental identity, and we showed that truncation induces considerable bias. The magnitude of such bias and the time needed for the register to recover are specific to each study because they depend on the value of familial association, background incidence rates, and the mortality mechanism for cases. They also depend indirectly on disease age pattern and family relationship through loss of family history at start-up.

Disease registration start-up was found to induce non-differential misclassification of exposure. The sensitivity of the observed exposure has little or no dependence on the value of familial risk or magnitude of background incidence rates, but only on the start-up date of registration (i.e. whether the family relative was incident after the beginning of registration). This implies that the estimates of sensitivity required

to correct for bias will be specific to each study, and that they can be obtained from a population simulated with any reasonable assumed value for the familial risk, and feasible incidence rates that display a realistic age-profile.

The strength of our bias-correction methodology resides in the use of estimates of sensitivity obtained from simulated populations, without the need for validation samples. With simple population vital statistics and disease incidence rates, familial risks can thus be corrected for the unavoidable bias due to registration start-up.

Our final study of biases in familial risk of cancer based on register data is reassuring of the validity of the large body of literature that has used the Swedish and other Scandinavian registers to estimate overall familial risks for common cancers. However, where the exposure of interest is early age of onset in a parent, commonly considered to be an indication of genetically determined cancer, estimates may be biased, especially where familial risk is high.

Conclusions, in brief

- We offer Poplab for free download
- Poplab can be used to simulate populations with desired demographic and disease features
- Left-truncation due to start-up of disease registration induced considerable bias
- Missing links in the MultiGenation Register can potentially induce bias for diseases with differential mortality for familial and non-familial cases
- We propose a bias-correction method that performs well and does not necessitate validation samples
- Familial studies of cancer based on data from the Swedish Registers are expected to be generally unbiased, except for those instances where large familial risks and young age at onset combine

Chapter 8

Future studies

As illustrated in the present thesis, both, the life-span of a register and the age of the index persons at the beginning of registration, impact on the proportion of lost family history. Since the Swedish incidence rates of cancer are not dramatically different from the Scandinavian rates, comparisons of biases from population and cancer registries in the Nordic countries could offer added insight into the relative weight of these contributing factors. Our methodology can be used for such comparisons.

Future disease models that may be implemented in Poplab can specify an increased risk for individuals having an affected parent from the time they reach that parent's age at incidence, a genetic "doom" (from birth) for the descendants of a subpopulation of cases, or offspring risk modified by sex of parent (i.e., affected mother versus affected father).

Features of reproductive history known to be associated with disease, such as reduced fertility, delayed childbearing, age at first birth and parity, can be included in future risk models to study their role in familial aggregation [158]. For example, in breast cancer studies one could model the input risk parameter as a function of age at first birth, parity and age at menopause [159].

As studies of disease aggregation are complicated by the definition of familial "exposure", such as the biological relatives considered, the family size or age-gap between relatives, the influence of various definitions of exposure on estimates of familial risk can also be explored. These considerations impact on the power of a research study, and are especially relevant when needing to incorporate expensive

genetic data (e.g., biomarker or molecular information [160], [161], [162], [163], [164]) and for finding more efficient strategies for predictive genetic testing [165], [166].

Immigrant sub-populations may well have different background incidence rates and a different age structure than their host population [167], [168], and left-truncation is no longer a fixed point in calendar time but is specific to the person's immigration date. We have shown that the bias in familial risk estimates depends on all of these factors, but understanding the nature of the dependence for migrant populations requires further dedicated research.

While this thesis addresses left-truncation biases and simple corrections, it would be of interest to investigate these issues for diseases that aggregate under family patterns that are not specific to cancers, and with high incidence rates. Future research could also address biases in disease-related survival.

Acknowledgements

My time as a PhD student gave me the opportunity to meet wonderful colleagues, and it has been an interesting journey (an inner one most of all). I would like to acknowledge the following people:

Marie Reilly, my main supervisor. Thank you for introducing me in this extraordinarily interesting field of research and for your non compromising scientific enthusiasm.

Kamila Czene, my co-supervisor. For sharing your profound knowledge and expertise always with generosity and patience, for your constructive criticism, and for your unwavering support and guidance.

Juni Palmgren. For your continuous support, for all your wonderful advice and for your inspiring attitude.

All my friends and colleagues in the Biostatics group. For making it fun to go to work and for creating such a warm atmosphere. You have become my big extended family.

Bibliography

[1] Kerber RA, O'Brien E. A cohort study of cancer risk in relation to family histories of cancer in the Utah population database. Cancer, 2005;103(9):1906-15

[2] Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer, 2002; 99(2):260-6

[3] Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med, 2000;343(2):78-85

[4] Smith DG, Sing CF. Sampling biases in longitudinal genetic-epidemiologic surveys. Hum Biol, 1976;48(3):529-39

[5] Burton PR, Palmer LJ, Jacobs K, et al. Ascertainment adjustment: where does it take us? Am J Hum Genet, 2000. 67(6):1505-14

[6] Pfeiffer RM, Gail MH, Pee D. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. Biometrika, 2001;88:933-48

[7] Davidov O, Zelen M. Referent sampling, family history and relative risk: the role of length-biased sampling. Biostatistics, 2001;2(2):173-81

[8] Lindstrm L, Pawitan Y, Reilly M, et al. Estimation of genetic and environmental factors for age-of-onset of disease from population-based family data. Stat Med, 2006;25(18):3110-23

[9] Paltiel O, Schmit T, Adler B, et al. The incidence of lymphoma in first-degree relatives of patients with Hodgkin disease and non-Hodgkin lymphoma: results and limitations of a registry-linked study. Cancer, 2000;88(10):2357-66

[10] Andersen EW, Andersen PK. Adjustment for misclassification in studies of familial aggregation of disease using routine register data. Stat Med, 2002;21(23):3595-607

[11] Pfeiffer RM, Goldin LR, Chatterjee N, et al. Methods for testing familial aggregation of diseases in population-based samples: application to Hodgkin lymphoma in Swedish registry data. Ann Hum Genet, 2004;68:498-508

[12] Statistics Sweden. Multi-Generation Register 2005, A description of contents and quality (Serie:BE96 Bakgrundsfakta. Befolknings- och Vlfrdsstatistik);2006:6. Available at: http://www.scb.se

[13] Information regarding the Swedish Cancer Register, available at http://www.socialstyrelsen.se/en/Statistics/statsbysubject/Cancer+Registry.htm

[14] Eldon BJ, Jonsson E, Tomasson J, et al. Familial risk of prostate cancer in Iceland. BJU Int, 2003;92(9):915-9

[15] Risch N. The Genetic Epidemiology of Cancer: Interpreting Family and Twin Studies and Their Implications for Molecular Genetic Approaches. Cancer epidemiol Biomarkers Prev,2001;10:733-41

[16] Laird NM, Cuenco KT. Regression methods for assessing familial aggregation of disease. Stat Med, 2003;22(9):1447-55

[17] Hopper JL, Bishop T, Easton DF. Population-based family studies in genetic epidemiology. Lancet,2005;366:1397-406

[18] Antoniou A, Pharoah PD, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet, 2003;72(5):1117-30

[19] Baffoe-Bonnie AB, Kiemeney LA, Beaty TH, et al. Segregation analysis of 389 Icelandic pedigrees with Breast and prostate cancer. Genet Epidemiol, 2002;23(4):349-63

[20] Walss-Bass C, Escamilla MA, Raventos H, et al. Evidence of genetic overlap of schizophrenia and bipolar disorder: linkage disequilibrium analysis of chromosome 18 in the Costa Rican population. Am J Med Genet B Neuropsychiatr Genet, 2005;139(1):54-60

[21] Gauderman WJ, Morrison JL. Evidence for age-specific genetic relative risks in lung cancer. Am J Epidemiol, 2000;151(1):41-9

[22] Klein AP, Beaty TH, Bailey-Wilson JE, et al. Evidence for a major gene influencing risk of pancreatic cancer. Genet Epidemiol, 2002;23(2):133-49

[23] Ciske DJ, Rich SS, King RA, et al. Segregation analysis of breast cancer: a comparison of type-dependent age-at-onset versus type-dependent susceptibility models. Genet Epidemiol, 1996;13(4):317-28

[24] Gauderman WJ, Morrison JL, Carpenter CL, Thomas DC. Analysis of gene-smoking interaction in lung cancer. Genet Epidemiol, 1997;14(2):199-214

[25] Murray RM, Clifford CA, Gurling HM. Twin and adoption studies. How good is the evidence for a genetic role? Recent Dev Alcohol, 1983;1:25-48

[26] Saudino KJ. Behavioral genetics and child temperament. J Dev Behav Pediatr, 2005;26(3):214-23

[27] Maes HH, Neale MC, Kendler KS, et al. Genetic and cultural transmission of smoking initiation: an extended twin kinship model, Behav Genet, 2006;36(6):795-808

[28] McGuffin P. Nature and nurture interplay: schizophrenia. Psychiatr Prax, 2004;31 Suppl 2:S189-93

[29] Baglietto L, Jenkins MA, Severi G, et al. Measures of familial aggregation depend on definition of family history: meta-analysis for colorectal cancer. J Clin Epidemiol, 2006;59(2):114-24

[30] Zelen M. Risks of cancer and families (editorial). J Natl Cancer Inst. 2005;97:1556 1557

[31] Khoury MJ, Flanders WD. Bias in using family history as a risk factor in case-control studies of disease. Epidemiology, 1995;6(5):511-9

[32] Boucher KM, Kerber RA. Measures of familial aggregation as predictors of breast-cancer risk. J Epidemiol Biostat, 2001;6(5):377-85

[33] Dong C, Hemminki K. Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. Int J Cancer, 2001;92:14450

[34] Hemminki K, Dong C. Population-based study of familial medullary thyroid cancer. Fam Cancer, 2001;1(1):45-9

[35] Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst,1994;86:1600-8

[36] Marsh D, Zori R. Genetic insights into familial cancers– update and recent discoveries. Cancer Lett, 2002;181(2):125-64

[37] Lynch HT, Shaw TG, Lynch JF. Inherited predisposition to cancer: a historical overview. Am J Med Genet C Semin Med Genet, 2004;129(1):5-22

[38] Kerber RA, Slattery ML. The impact of family history on ovarian cancer risk. The Utah Population Database. Arch Intern Med, 1995;155(9):905-12

[39] Hemminki K, Vaittinen P. Effect of paternal and maternal cancer on cancer in the offspring: a population-based study. Cancer Epidemiol Biomarkers Prev, 1997;6(12):993-7

[40] Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol, 2001;96(10):2992-3003

[41] Johns LE, Houlston RS. A systematic review and meta-analysis of familial prostate cancer risk. BJU Int, 2003;91(9):789-94

[42] Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without disease. Lancet, 2001;358:138999

[43] Oldenburg RA, Meijers-Heijboer H, Cornelisse CJ, Devilee P. Genetic susceptibility for breast cancer: how many more genes to be found? Crit Rev Oncol Hematol, 2007;63(2):125-49

[44] Bromen K, Pohlabeln H, Jahn I, et al. Aggregation of lung cancer in families: results from a population-based case-control study in Germany. Am J Epidemiol, 2000;152(6):497-505

[45] Carter BS, Beaty TH, Steinberg GD, et al. Mendelian inheritance of familial prostate cancer. Proc Natl Acad Sci USA, 1992;89(8):3367-71

[46] Slattery ML, Kerber RA. A comprehensive evaluation of family history and breast cancer risk. The Utah Population Database. JAMA, 1993;270(13):1563-8

[47] Soegaard M, Jensen A, Frederiksen K, et al. Accuracy of self-reported family history of cancer in a large case-control study of ovarian cancer. Cancer Causes Control, 2008;[Epub ahead of print]

[48] Chang ET, Smedby KE, Hjalgrim H, et al. Reliability of self-reported family history of cancer in a large case-control study of lymphoma. J Natl Cancer Inst, 2006;98(1):61-8

[49] Kerber RA, Slattery ML. Comparison of self-reported and database-linked family history of cancer data in a case-control study. Am J Epidemiol, 1997;146(3):244-8

[50] Goldberg J, Gelfand HM, Levy PS. Registry evaluation methods: a review and case study. Epidemiologic Reviews, 1980;2:210-20

[51] Brewster DH, Fordyce A, Black RJ; Scottish Clinical Geneticists. Impact of a cancer registry-based genealogy service to support clinical genetics services. Fam Cancer, 2004;3(2):139-41

[52] John EM, Hopper JL, Beck JC, et al; Breast Cancer Family Registry. The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. Breast Cancer Res, 2004;6(4):R375-89

[53] Sandeep TC, Strachan MW, Reynolds RM, et al. Second primary cancers in thyroid cancer patients: a multinational record linkage study. J Clin Endocrinol Metab, 2006;91(5):1819-25

[54] Sclo G, Boffetta P, Autier P, et al. Associations between ocular melanoma and other primary cancers: an international population-based study. Int J Cancer, 2007;120(1):152-9

[55] Information regarding the Finnish Cancer Register, available at

http://www.cancerregistry.fi/eng/general/

[56] Storm HH. The Danish Cancer Registry, a self-reporting national cancer registration system with elements of active data collection. Lyons: International Agency for Research on Cancer, 1991; IARC Scientific Publication No 95;22036

[57] Information regarding the Cancer Registry of Norway, available at

http://www.kreftregisteret.no/frame.htm?english.htm

[58] Information regarding the Icelandic Cancer Registry, available at

http://www.krabbameinsskra.is/indexen.jsp?id=g

[59] Brasso K, Ingimarsdottir IJ, Thomassen L, et al. Prostate cancer in Denmark 1943-2002 [Article in Danish]. Ugeskr Laeger, 2007;169(2):129-32

[60] Brasso K, Friis S, Kjaer SK, et al. Prostate cancer in Denmark: a 50-year population-based study. Urology, 1998;51(4):590-4

[61] Cantor-Graae E, Pedersen CB. Risk for schizophrenia in intercountry adoptees: a Danish population-based cohort study. J Child Psychol Psychiatry, 2007;48(11):1053-60

[62] Eising S, Svensson J, Skogstrand K, et al. Type 1 diabetes risk analysis on dried blood spot samples from population-based newborns: design and feasibility of an unselected case-control study. Paediatr Perinat Epidemiol, 2007;21(6):507-17

[63] Ahlgren M, Wohlfahrt J, Olsen LW, Srensen TI, Melbye M. Birth weight and risk of cancer. Cancer, 2007;110(2):412-9

[64] Winqvist S, Jokelainen J, Luukinen H, Hillbom M. Adolescents' drinking habits predict later occurrence of traumatic brain injury: 35-year follow-up of the northern Finland 1966 birth cohort. J Adolesc Health, 2006;39(2):275.e1-7

[65] Seregard S, Lundell G, Svedberg H, Kivel T. Incidence of retinoblastoma from 1958 to 1998 in Northern Europe: advantages of birth cohort analysis. Ophthalmology, 2004;111(6):1228-32

[66] Hinkula M, Pukkala E, Kyyrnen P, et al. A population-based study on the risk of cervical cancer and cervical intraepithelial neoplasia among grand multiparous women in Finland. Br J Cancer, 2004;90(5):1025-9

[67] Sankila R, Olsen JH, Anderson H, et al. Risk of cancer among offspring of childhood-cancer survivors. Association of the Nordic Cancer Registries and the Nordic Society of Paediatric Haematology and Oncology. N Engl J Med, 1998;338(19):1339-44

[68] Hakulinen T, Andersen A, Malker B, et al. Trends in cancer incidence in the Nordic countries. A collaborative study of the five Nordic Cancer Registries. Acta Pathol Microbiol Immunol Scand Suppl. 1986;288:1-151

[69] Stang A, Pukkala E, Sankila R, et al. Time trend analysis of the skin melanoma incidence of Finland from 1953 through 2003 including 16,414 cases. Int J Cancer, 2006;119(2):380-4

[70] Richiardi L, Bellocco R, Adami HO, et al. Testicular cancer incidence in eight northern European countries: secular and recent trends. Cancer Epidemiol Biomarkers Prev, 2004;13(12):2157-66

[71] Wilcox AJ, Skaerven R, Lie RT. Familial patterns of preterm delivery: maternal and fetal contributions. Am J Epidemiol, 2008;167(4):474-9

[72] Nagenthiraja K, Ewertz M, Engholm G, Storm HH. Incidence and mortality of pancreatic cancer in the Nordic countries 1971-2000. Acta Oncol, 2007;[Epub ahead of print]

[73] Mller TR, Garwicz S, Barlow L, et al; Association of the Nordic Cancer Registries; Nordic Society for Pediatric Hematology and Oncology. Decreasing late mortality among five-year survivors of cancer in childhood and adolescence: a population-based study in the Nordic countries. J Clin Oncol, 2001;19(13):3173-81

[74] Franco-Lie I, Iversen T, Robsahm TE, Abdelnoor M. Birth weight and melanoma risk: a population-based case-control study. Br J Cancer, 2008;98(1):179-82

[75] Seppanen J, Heinavaara S, Hakulinen T. Influence of alternative mammographic screening scenarios on breast cancer incidence predictions (Finland). Cancer Causes Control, 2006;17(9):1135-44

[76] Cancer incidence in Sweden 1998. Center for Epidemiology, The National Board for Health and Welfare, Stockholm, Sweden, 2000.

(http://www.socialstyrelsen.se/NR/rdonlyres/24BF1378-6978-45A2-A290-BEA734B550A3/1810/0042004.pdf)

[77] Linet MS, McLaughlin JK, Malker HS, et al. Occupation and hematopoietic and lymphoproliferative malignancies among women: a linked registry study. J Occup Med, 1994;36(11):1187-98

[78] Westhoff CL. Epidemiologic studies: pitfalls in interpretation. Dialogues Contracept, 1995;4(5):5-6,8

[79] Sackett DL. Bias in analytic research. J Chronic Dis, 1979;32:51-63

[80] Delgado-Rodrguez M, Llorca J. Bias. J Epidemiol Community Health, 2004;58(8):635-41

[81] Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet, 2002;359(9302):248-52

[82] Zaccai JH. How to assess epidemiological studies. Postgrad Med J, 2004;80(941):140-7

[83] Tripepi G, Jager KJ, Dekker FW, et al. Bias in clinical research. Kidney Int, 2008;73(2):148-53

[84] Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ, 2001;323:1015

[85] Bermejo JL, Hemminki K. Familial risk of cancer shortly after diagnosis of the first familial tumor. J Natl Cancer Inst, 2005;97(21):1575-9

[86] Armstrong B, White E, Saracci R. Principles of exposure measurement in epidemiology. New York: Oxford University Press, 1992

[87] Pearce N, Checkoway H, Kriebel D. Bias in occupational epidemiology studies. Occup Environ Med, 2007;64(8):562-8

[88] Mezei G, Kheifets L. Is there any evidence for differential misclassification or for bias away from the null in the Swedish childhood cancer study? Epidemiology, 2001;12(6):750-2

[89] Levois M, Switzer P. Differential exposure misclassification in case-control studies of environmental tobacco smoke and lung cancer. J Clin Epidemiol, 1998;51(1):37-54

[90] Garca-Closas M, Thompson WD, Robins JM. Differential misclassification and the assessment of gene-environment interactions in case-control studies. Am J Epidemiol, 1998;147(5):426-33

[91] Bakke PS, Hanoa R, Gulsvik A. Relation of occupational exposure to respiratory symptoms and asthma in a general population sample: self-reported versus interview-based exposure data. Am J Epidemiol, 2001;154(5):477-83

[92] Birkett NJ. Effects of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure. Am J Epidemiol, 1992;136(3):356-62

[93] Greenland S, Gustafson P. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. Am J Epidemiol, 2006; 164(1):63-8.

[94] Brenner H. Inferences on the potential effects of presumed nondifferential exposure misclassification. Ann Epidemiol, 1993;3(3):289-94.

[95] Brenner H. Correcting for exposure misclassification using an alloyed gold standard. Epidemiology, 1996;7(4):406-10

[96] Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol, 1991;134(10):1233-44.

[97] Flegal KM, Brownie C, Haas JD. The effects of exposure misclassification on estimates of relative risk. Am J Epidemiol, 1986;123(4):736-51

[98] Prescott GJ and Garthwaite PH. A Bayesian approach to prospective binary outcome studies with misclassification in a binary risk factor. Stat Med, 2005;24:3463-77

[99] Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. Biometrics, 1999;55:1193-1201

[100] White E. Design and interpretation of studies of differential exposure measurement error. Am J Epidemiol, 2003;157(5):380-7

[101] Veierod MB and Laake P. Exposure misclassification: bias in category specific Poisson regression coefficients. Stat Med, 2001;20:771-84

[102] Buonaccorsi JP, Laake P, Veierod MB. On the Effect of Misclassification on Bias of Perfectly Measured Covariates in Regression. Biometrics, 2005;61:831-6

[103] Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of releative risk. Am J Epidemiol, 1977;105(5):488-95

[104] Szatmari P, Jones MB. Effects of misclassification on estimates of relative risk in family history studies. Genet Epidemiol, 1999;16:368-81

[105] Jurek AM, Greenland S, Maldonado G, Church TR. Proper interpretation of non-differential misclassification effects: expectations vs observations. Int J Epidemiol, 2005;34:680-7

[106] Wachter KW, Hammel EA and Laslett P. Statistical studies of historical social structure. New York: Academic Press:1978

[107] Hammel EA, Wachter KW, McDaniel CK. The Kin of the Aged in A.D. 2000. In: Keisler S, Morgan J, Oppenheimer V, eds. Aging. New York: Academic Press;1981:11-40

[108] Wachter KW. Kinship resources for the elderly. Philos Trans R Soc Lond B Biol Sci, 1997;352(1363):1811-7

[109] Bartlema J. Modelling Step-families: first results. Netherlands Interdisciplinary Demographic Institute: The Hague, Netherlands: 1989

[110] Hammel EA, Mason C, Wachter KW, et al. Rapid population change and kinship: the effects of unstable demographic changes on Chinese kinship networks, 1750-2250. In: Tapinos G, Blanchet D, Horlacher D, eds. Consequences of Rapid Population Growth in Developing Countries. New York: Taylor and Francis;1991:243-71

[111] Smith JE. The Computer Simulation of Kin Sets and Kin Counts. In: Bongaarts J, Burch T and Wachter KW, eds. Family Demography: Methods and their Applications. Oxford: The Clarendon Press;1987:249-66

[112] Hampe J, Wienker T, Schreiber S, Nurnberg P. POPSIM: a general population simulation program. Bioinformatics, 1998;14(5):458-64

[113] Hammel EA, Hutchinson D, Wachter KW, et al. The SOCSIM Demographic-Sociological Microsimulation Program Operating Manual. Institute of International Studies Monograph no. 27. California: University of California, Berkeley:1976

[114] Hammel EA, Mason C, Wachter KW. SOCSIM II, A Sociodemographic Microsimulation Program, Rev. 1.0, Operating Manual. Program in Population Research Working Paper no. 29. California: Institute of International Studies, University of California, Berkeley;1990

[115] Wachter KW. SOCSIM: Description of the Program. Available at: http://www.demog.berkeley.edu/ wachter/socstory.html

[116] SOCSIM Documentation http://www.demog.berkeley.edu/ marcia/c_doc.html

[117] The R Project for Statistical Computing home page http://www.r-project.org

[118] Leu M, Czene K, Reilly M. Population Lab website

http://www.mep.ki.se/ marrei/software/poplab/ or

http://cran.at.r-project.org/

[119] Hemminki K, Li X, Plna K, et al. The nation-wide Swedish family-cancer database - updated structure and familial rates. Acta Oncologica, 2001;40(6):772-7

[120] Nielsen NM, Westergaard T, Frisch M, et al. Type 1 diabetes and multiple sclerosis: A Danish population-based cohort study. Arch Neurol. 2006;63(7):1001-4

[121] Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet, 2006;7:781-91

[122] Burt RW. Familial risk and colorectal cancer. Gastroenterol Clin North Am, 1996;25(4):793-803

[123] Pharoah PD, Day NE, Duffy S, et al. Family history and the risk of breast cancer: a systematic review and meta-analysis. Int J Cancer, 1997;71(5):800-9

[124] Boardman LA, Morlan BW, Rabe KG, et al. Colorectal cancer risks in relatives of young-onset cases: is risk the same across all first-degree relatives? Clin Gastroenterol Hepatol, 2007;5(10):1195-8.

[125] Johansson L, Finnas F. Fertility of Swedish women born 1927-1960. rebro: Statistics Sweden, 1983

[126] Berinde D. Pathways to a third child in Sweden. Eur J Popul, 1999;15(4):349-78

[127] Therborn G. Families in today's world-and tomorrow's. Int J Health Serv. 2006;36(3):593-603

[128] Kulu H, Vikat A, Andersson G. Settlement size and fertility in the Nordic countries. Popul Stud (Camb), 2007;61(3):265-85

[129] Hemminki K, Sundquist J, Bermejo J. How common is familial cancer? Ann Oncol 2008;19(1):163-7

[130] Chappuis PO, Rosenblatt J, Foulkes WD. The influence of familial and hereditary factors on the prognosis of breast cancer. Ann Oncol 1999;10:116370

[131] Ekbom A, Akre O. Increasing incidence of testicular cancer–birth cohort effects. APMIS, 1998;106(1):225-9; discussion 229-31

[132] Briollais L, Chompret A, Guilloud-Bataille M, et al. Patterns of familial aggregation of three melanoma risk factors: great number of naevi, light phototype and high degree of sun exposure. Int J Epidemiol. 2000 Jun;29(3):408-15

[133] Frich L, Glattre E, Akslen LA. Familial occurrence of nonmedullary thyroid cancer: a population-based study of 5673 first-degree relatives of thyroid cancer patients from Norway. Cancer Epidemiol Biomarkers Prev, 2001;10(2):113-7

[134] Hemminki K, Eng C, Chen B. Familial risks for nonmedullary thyroid cancer. J Clin Endocrinol Metab, 2005;90(10):5747-53

[135] Pisani P, Parkin DM, Bray F, Ferlay J. Estimates of the worldwide mortality from 25 cancers in 1990. Int. J. Cancer, 1999;83(1):1829

[136] Hfler M. The effect of misclassification on the estimation of association: a review. Int J Methods Psychiatr Res, 2005;14(2):92-101

[137] Hemminki K, Li X, Czene K. Familial risk of cancer: data for clinical counseling and cancer genetics. Int J Cancer 2004; 108:10914

[138] Hemminki K, Chen B. Familial risks for colorectal cancer show evidence on recessive inheritance. Int J Cancer 2005; 115:8358

[139] van der Horst M, Winther JF, Olsen JH. Cancer incidence in the age range 0-34 years: historical and actual status in Denmark. Int J Cancer, 2006;118(11):2816-26

[140] Karlsson PM, Fredrikson M. Cutaneous malignant melanoma in children and adolescents in Sweden, 1993-2002: the increasing trend is broken. Int J Cancer, 2007;121(2):323-8

[141] Bergfeldt K, Rydh B, Granath F, et al. Risk of ovarian cancer in breast-cancer patients with a family history of breast or ovarian cancer: a population-based cohort study. Lancet, 2002;360(9337):891-4

[142] Hemminki K, Sundquist J, Bermejo J. How common is familial cancer? Ann Oncol. 2007 [Epub ahead of print]

[143] Hemminki K, Vaittinen P. Familial breast cancer in the family-cancer database. Int J Cancer, 1999;77(3):386-91

[144] Tulinius H, Sigvaldason H, Olafsdttir G, et al. Breast cancer incidence and familiality in Iceland during 75 years from 1921 to 1995. J Med Genet, 1999;36(2):103-7

[145] Planck M, Anderson H, Bladstrom A, et al. Increased cancer risk in offspring of women with colorectal carcinoma: a Swedish registerbased cohort study. Cancer, 2000;89(4):741-9

[146] Stefansson T, Moller PH, Sigurdsson F, et al. Familial risk of colon and rectal cancer in Iceland: evidence for different etiologic factors? Int J Cancer, 2006;119(2):304-8

[147] Jonsson S, Thorsteinsdottir U, Gudbjartsson DF, et al. Familial risk of lung carcinoma in the Icelandic population. JAMA, 2004;292(24):2977-83

[148] Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers–a different disease. Nat Rev Cancer, 2007;7(10):778-90

[149] Hemminki K, Vaittinen P. National database of familial cancer in Sweden. Genet Epidemiol, 1998;15(3):225-36

[150] Chang ET, Smedby KE, Hjalgrim H, et al. Family history of hematopoietic malignancy and risk of lymphoma. J Natl Cancer Inst, 2005;97(19):1466-74

[151] Verhoog LC, Brekelmans CT, Seynaeve C, et al. Survival in hereditary breast cancer associated with germline mutations of BRCA2. J Clin Oncol, 1999;17(11):3396-402

[152] Robson M, Gilewski T, Haas B, et al. BRCA-associated breast cancer in young women. J Clin Oncol, 1998;16(5):1642-9.

[153] Tommiska J, Eerola H, Heinonen M, et al. Breast cancer patients with p53 Pro72 homozygous genotype have a poorer survival. Clin Cancer Res, 2005;11(14):5098-103

[154] Verhoog LC, Brekelmans CT, Seynaeve C, et al. Survival and tumour characteristics of breast-cancer patients with germline mutations of BRCA1. Lancet, 1998;351(9099):316-21.

[155] Hemminki K, Chen B. Familial risks for cervical tumors in full and half siblings: etiologic apportioning. Cancer Epidemiol Biomarkers Prev 2006;15(7):1413-4

[156] Czene K, Adami HO, Chang ET. Sex- and kindred-specific familial risk of non-Hodgkin's lymphoma. Cancer Epidemiol Biomarkers Prev, 2007;16(11):2496-9.

[157] Damber L, Gronberg H, Damber JE. Familial prostate cancer and possible associated malignancies: nation-wide register cohort study in Sweden. Int J Cancer, 1998;78(3):293-7

[158] Susser E, Susser M. Familial aggregation studies. A note on their epidemiologic properties. Am J Epidemiol, 1989;129(1):23-30

[159] Lambe M, Hsieh CC, Tsaih SW, et al. Parity, age at first birth and the risk of carcinoma in situ of the breast. Int J Cancer, 1998;77:3302

[160] Ivansson EL, Gustavsson IM, Magnusson JJ, et al. Variants of chemokine receptor 2 and interleukin 4 receptor, but not interleukin 10 or Fas ligand, increase risk of cervical cancer. Int J Cancer, 2007;121(11):2451-7

[161] Yilmaz M, Bukan N, Ersoy R, et al. Glucose intolerance, insulin resistance and cardiovascular risk factors in first degree relatives of women with polycystic ovary syndrome. Hum Reprod, 2005;20(9):2414-20

[162] Locker GY, Lynch HT. Genetic factors and colorectal cancer in Ashkenazi Jews. Fam Cancer, 2004;3(3-4):215-21

[163] Cunningham JM, McDonnell SK, Marks A, et al; Mayo Clinic, Rochester, Minnesota. Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. Prostate, 2003;57(4):335-46

[164] Sanders J, Gill M. Unravelling the genome: a review of molecular genetic research in schizophrenia. Ir J Med Sci, 2007;176(1):5-9

[165] Amor D. Familial cancers. An overview. Aust Fam Physician, 2001;30:937-45

[166] Lakhani SR, Manek S, Penault-Llorca F, et al. Pathology of ovarian cancers in BRCA1 and BRCA2 carriers. Clin Cancer Res. 2004;10(7):2473-81

[167] Hemminki K, Li X, Czene K. Cancer risks in first-generation immigrants to Sweden. Int J Cancer, 2002;99:21828

[168] Hemminki K, Li X. Cancer risks in second-generation immigrants to Sweden. Int J Cancer, 2002;99:22937