An Analysis of Using Pedigrees in Family Based Studies and an Exploration of Cancer Risk and Cancer Resistance UsingTwin Studies

An Analysis of Using Pedigrees in Family Based Studies and an Exploration of Cancer Risk and Cancer Resistance UsingTwin Studies

A dissertation presented

by

Christina Magdaline McIntosh

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

January 2017

Dissertation Advisor: Professor Giovanni Parmigiani Christina McIntosh

An Analysis of Using Pedigrees in Family Based Studies and an Exploration of Cancer Risk and Cancer Resistance UsingTwin Studies

Abstract

In the first section of this thesis, we explore the use of family pedigrees in association analysis. Family pedigrees were successfully used in linkage analysis to discover many genes for Mendelian traits, but less successful for identifying genes for complex diseases. The family-based association test (FBAT) can be used to test for association in pedigrees by two methods, conditioning on the parents to get the null distribution of the offsprings' genotypes separately for each nuclear family and then combining over families, or conditioning on the founders in the pedigree to get the offsprings' genotypes. In this study, we use simulations to compare the power of conditioning on the founders or parents when using the FBAT statistic to test for association in the family pedigree.We consider two scenarios where the disease outcome is represented as a simple Mendelian trait and the disease outcome is modeled as more complex as multiple factors influence the disease.

Two new results were found in our simulation study. Under the first assumption of a Mendelian disease outcome, conditioning on the founders is slightly more powerful than conditioning on the parents for detecting association. For complex diseases, the power of all of the ascertainment methods were reduced considerably, but using multiplex pedigrees were still more powerful than trios and sib pairs when the recurrence risk ratio was between family members was relatively low.

In the second part of this thesis we explore the use of twin studies in assessing cancer risk and cancer resistance. Twin studies provide unique information about familial risk of cancer. Typically, twin studies are used to quantify heritability. However, clinical

application of the insight gained from twin studies needs to rely on estimates of absolute risk, as these are directly relevant for both individual and population-level decision making. We provide estimates of risk using the Nordic Twin Studies of Cancer (NorTwinCan), estimate risk ratios which can be applied to estimate the risks in a population with a different baseline risk, and compare methods of calculating the risks. Our results suggest that both models provide slightly different absolute risk estimates compared to the empirical estimates, which is not surprising since these models rely on differing modeling assumptions and condition on different covariates. While estimates of heritability of cancer are high, for an unaffected individual the implications of having an affected co-twin remain relatively contained, and additional family history should continue to play a role in counseling and decision making.

In the last section, we explore cancer resistance using twin studies. It has been hypothesized that some individuals have a decreased risk of cancer, or a cancer resistance, resulting from genetic predisposition. Using simulation studies and twin study data, we explored the question of whether these studies can also be used to investigate a genetic predisposition to avoiding cancer. We first conduct simulations to assess the impact such a genetic predisposition would have on the proportions of cancer concordant MZ and DZ twins, we postulated a simple model wherein a fraction of the population carries an inherited and extreme resistance to cancer, and developed a likelihood-based approach to estimate this prevalence. We then applied our approach to the Nordic Twin Studies of Cancer (NorTwinCan), a cohort of over 200,000 individual twins from Sweden, Norway, Denmark and Finland. We estimate the prevalence of the "cancer resistance" genotypes as 1.7% (95% C.I.: $1.2, 2.1\%$) in this population. These results are obtained using the following assumptions; 1) all cancers are considered together, 2) the resistance genotype is fully penetrant, and 3) distributions for age of onset of cancer, censoring, and death are fixed. We do, however, provide a general framework under which these assumptions can be relaxed. Our results suggest that predisposition to avoiding cancer may be heritable, that twin data can provide information on this hypothesis, and that the largest twin studies allow for quantitative exploration of genetic parameters.

Contents

List of Figures

List of Tables

In memory of my grandmother, Mattie Pearl Turner, and my loved ones whose life was ended by cancer.

Acknowledgments

I would first like to thank God, whom I truly believe I would not have made it without. Thank you to my wonderful and supportive advisor, Giovanni Parmigiani who has guided me throughout this process and helped me to grow as a researcher. I would like to thank my parents and my sister Jeffrey, Patricia, and Amanda McIntosh whose support has been invaluable to helping me finish this process and in life. Thank you to my fiance, Khaden Nurse, whose love and support has helped me to finish this process. Lorenzo Trippa, Lorelei Mucci, and Nan Laird deserve a special thank you as members of my committee who continued to believe in me and encourage throughout the entire process. Thank you to members of the Nordic Twin Studies of Cancer (NorTwinCan) for allowing me to collaborate with you and use the data. Danielle Braun has been instrumental in guiding me through this process, and I am very grateful for her mentorship and friendship.

1. Analyzing Pedigrees for Association Analysis

Christina McIntosh[1], Wai-Ki Yip[1], Nan Laird[1]

[1]Department of Biostatistics, Harvard School of Public Health

1.1 Introduction

Due to genotyping technology, the Human Genome Project, and the lack of success in linkage studies for identifying genes for complex traits, the focus of gene mapping in humans is now largely based on association studies. In genetic association studies, both population based and family based tests are used to detect genes that may be associated with a particular trait such as disease. An advantage of using family based designs is that they protect against false positive findings due to population substructure (Laird and Lange, 2010).

The most simple family based design for testing association uses trios which include an affected offspring and two parents. The transmission disequilibrium test (TDT) was introduced as a test for linkage in the presence of association, but it is now widely used as a test for association (Spielman et al., 1993). The approach of the TDT is to compare the null distribution of the offspring's genotype under Mendel's laws to the offspring genotype that is actually observed in trios. A general method for testing association in family based designs, known as the family based association test (FBAT), and can be applied to testing pedigrees for association (Laird et al., 2000). This testing approach is the same as the TDT, namely to compare the observed genotype of the offspring to what is expected under Mendel's laws. In fact, when using the FBAT statistic to test for association in pedigrees, there are two ways to determine the distribution of the genotypes of the offsprings in the pedigree. The first approach is to use the parents' genotypes to determine the null distribution of the offspring genotype under no association. The second approach involves using the founders' genotypes in the pedigree to determine the genotypes of the offspring. The power of these two approaches when analyzing pedigrees for association has not yet been compared.We conducted a simulation study with three goals:

1. To compare the power of two methods of analyzing pedigrees for genetic association based on a simple Mendelian disease model. The first method involves conditioning on parents' genotypes and traits, and the second method involves conditioning on founders' genotypes and traits.

2. To compare the power of using FBAT to analyze pedigrees to the power of using FBAT to analyze trios and multiplex trios (two parents and two offspring).

3. To compare the power of using FBAT to analyze pedigrees under a complex disease model with shared factors (i.e. environmental) within families).

1.2 Using Pedigrees for Association Analysis

A trio is defined as two parents and an affected offspring, known as the proband. A nuclear family with two affected siblings is referred to as affected sib pairs (Figure 1.1a). A family pedigree not only includes genetic information about the proband and its parents (and possibly siblings), but it may include information from the proband's other relatives as well, such as the proband's aunts, siblings, children, grandparents, great-greatparents. An example of a simple family pedigree can be found in Figure 1.1c. If 4 in the pedigree represents the proband in Figure 1.1c, then 1 and 2 represent the proband's parents, while 4 and 3 are the parents of 5 in the pedigree. People in the pedigree with no parental information are known as founders in the pedigree. As an example, 1, 2, and 3 are the founders in the pedigree in Figure 1.1c. The family based association test (FBAT) is a generalization of the TDT test that can be used to determine the association between a trait and gene/s in any type of pedigree.

We consider trios (Figure 1.1a), affected sib pairs (Figure 1.1b), and mulitplex pedigrees, as in Figure 1.1c with two affected offspring. Both tests described below are based on transmissions to affected offspring. In the first approach to the calculation of the FBAT test statistic, genotypes that are observed in the offspring are compared to those expected under Mendel's law conditioning on the parental genotypes. Then, contributions of each nuclear family are combined over pedigrees.

The general form of the FBAT statistic is

$$\frac{\sum_{ijk} T_{ijk}(X_{ijk} - E(X_{ijk}|P_{jk}))}{\sum_{ijk} T_{ijk}^2 var(X_{ijk}|P_{jk})}.$$

We sum over the total number of K pedigrees, $k = 1, 2, ..., K$ and define X_{ijk} to be the

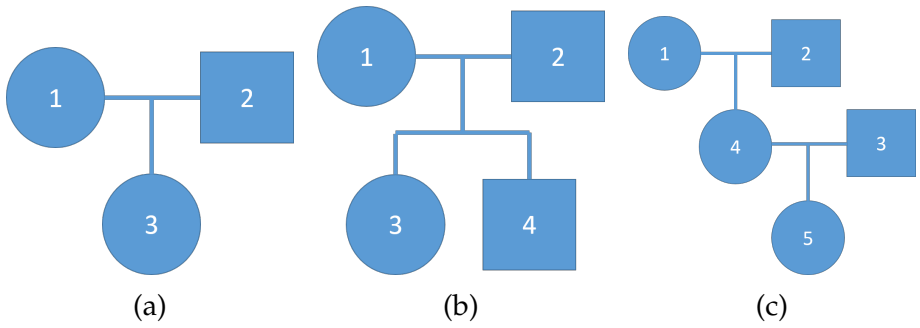


Figure 1.1: **Examples of study designs used in simulations.** A trio where 1 and 2 are parents and 3 is the offspring (a), a sib pair where 1 and 2 are the parents and 3 and 4 are the offspring (b), and a three-generation pedigree where 1,2,4 are the founders and 3 and 5 are the offspring(c).

coded genotype of the i^{th} offspring in the j^{th} family, T_{ijk} is the coded trait for the i^{th} offspring in the j^{th} nuclear family, and P_{jk} are the parental genotypes of the j^{th} nuclear family. The distribution of X_{ijk} is dependent of the genotypes of parents that contribute to the offspring genotypes. When analyzing pedigrees for association, the numerator and denominator of the FBAT statistic are calculated separately for each offspring. The numerator of the FBAT statistic for the pedigree is calculated by summing the numerators for each of the offspring, similarly, the denominator of the FBAT statistic is calculated by summing the denominators for each of the offspring. For each offspring, the expected distribution of the offspring's genotype varies with the parental genotypes that contribute to the offspring genotypes.

In the second approach of the FBAT statistic, the expected distribution of the genotypes of the offspring are compared to those expected under Mendel's laws and the founder genotypes. The general form of the statistic when conditioning on the founder genotypes is

$$\frac{\sum_{ijk} T_{ijk}(X_{ijk} - E(X_{ijk}|F_k))}{var(\sum_{ijk} T_{ijk}^2(X_{ijk}|F_k))},$$

where T_{ijk} is the same as in the FBAT statistic. The difference between the two approaches is that the expected distribution of the genotypes of the offspring is calculated conditional on the founders' genotypes in the pedigree and not the parents' genotypes as in the first approach. In order to obtain the distribution of the genotypes of the offspring using the founders, the algorithm in Rabinowitz and Laird is used (Rabinowitz and Laird, 2000).

When calculating the FBAT statistic for Figure 1.1c, one can split the pedigree into nuclear families. The first family includes 1, 2, and 3 and the second family includes 3, 4, and 5. The distribution of the genotype of 3 is determined by the genotypes of the parents, 1 and 2. Similarly, the distribution of the genotype of 5 is determined by the genotypes of the parents, 3 and 4 when using the FBAT statistic. When conditioning on the founders, the distribution of the genotype of 4 is determined by the founders 1 and 2; therefore, the null distribution of the genotype of offspring 4 is the same genotype distribution when you condition on the parents. This is not the case when the distribution of the genotype of 5 is calculated using the genotype of the founders,1, 2, and 3 in the pedigree. Therefore, using the founder genotypes to calculate the expected distribution of offspring genotypes accounts for all of the possibilities of the parental genotypes which leads to using more information in the pedigree. In the example pedigree using Figure 1.1c, all possible genotypes of offspring 4 are considered when calculating the genotype distribution for offspring.

The power of the TDT is influenced by the mode of inheritance and the number of informative mating types used in the calculation of the statistic (Rabinowitz and Laird, 2000). It is hypothesized that because conditioning on the founders in the test statistic uses more information in the pedigree, the power of the statistic will be greater compared to conditioning on the parents.

1.3 Methods

1.3.1 Simulation Design

A dichotomous trait is simulated with a single disease allele. Let A be the disease allele and B be the non-disease allele. As in Risch and Merikangas, Knapp, and Whittaker and Lewis, we assume the best-case scenario for the marker locus- that is, that the marker locus is that disease locus which contains the disease allele (Spielman et al., 1993) (Risch et al., 1996) (Knapp, 1999). A pedigree was built with three generations which include three founders and two offspring (Figure 1.1c). The pedigrees were simulated by first

calculating the probabilities of the individual pedigrees using the logit model described below.

Let Y_{ij} and g_{ij} represent the trait and genotype of the j^{th} individual in the i^{th} family. Let g_{ip} represent the genotypes of the parents in the i^{th} family. In general, the probability that both siblings are affected by a particular disease depends on the genotypes of both siblings, the genotypes of the parents, and other individual and family factors . For our first set of simulations, we assume that the probability of an offspring being affected is independent of its siblings' outcomes and genotypes as well as parental genotypes. More specifically given two offspring, we assume

$$P(Y_{i1} = Y_{i2} = 1|g_{i1}, g_{i2}, g_{ip}) = P(Y_{i1} = 1|g_{i1}, g_{i2}, g_{ip})P(Y_{i1} = 1|g_{i1}, g_{i2}, g_{ip}).$$

Using the logit model,

$$logit(P(Y_{ij} = 1)) = \beta_0 + \beta_1 g_{ij}$$

the probability that both siblings are affected is

$$P(Y_{i1} = Y_{i2} = 1|g_{i1}, g_{i2}) = \int \left(\frac{e^{\beta_0+\beta_1 g_{i1}}}{1 + e^{\beta_0+\beta_1 g_{i1}}}\right)\left(\frac{e^{\beta_0+\beta_1 g_{i2}}}{1 + e^{\beta_0+\beta_1 g_{i2}}}\right)$$

It is seen that the probability of disease is solely dependent on the genotype of the offspring which leads to the implicit assumption that the genotypes of each of the offspring are independent of each other conditional on their parents? genotypes. Therefore, we refer to this model as the independence model. This independence assumption is relaxed in the last part of the simulations.

Once the three-generation pedigrees were simulated, the FBAT test statistic conditioning on the parent genotypes and traits and the statistic conditioning on the founder genotypes were calculated for each of the pedigrees. The alpha level was calculated by determining the proportion of the statistics which have an absolute value greater than 1.96 (the cutoff point for the standard normal p-value of 0.05) when the probability of disease given each genotype is the same. Next, the power of the test is compared under the additive and recessive modes of inheritance for rare and common disease (prevalence

0.1% and 14% respectively). There were 5000 simulations and each simulation contained 100 pedigrees. The range of the frequency of the disease allele was from 0.001 to 0.5. In the second goal of the paper, the power of the two approaches for calculating the FBAT test statistic (conditioning on the parents' genotypes and conditioning on the founders' genotypes) for the pedigree are compared to the empirical power of the FBAT statistic calculated for the trio and sib pairs. The number of trios is doubled to evenly match the number of affected offspring in the pedigree and multiplex trios.

In the simulations constructed for the first goal of the paper, the trait is obtained solely based on the individual's genotype. This, however, is a reasonable assumption if the disease is Mendelian but not if other factors contribute to the disease as well. It is known that siblings and parents and their offspring share genetic and environmental effects. Consequently, simulations are conducted to account for shared effects by adding the parameter ϕ_i to the logit model used in part one of the simulations, the where i represents the family of the individual. Specifically, let Y_{ij}, g_{ij}, and ϕ_i represent the trait, genotype, and shared correlated effects of the j^{th} individual in the i^{th} family. Let g_{ip} represent the genotypes of the parents in the i^{th} family. In the simulation, the probability that both siblings are affected by a particular disease depends on the genotypes of both siblings, genotypes of the parents, and other shared effects such as environment. Given, g_{i1} , g_{i2}, g_{ip}, and ϕ_i the probability of the offspring being affected are independent. More specifically, we assume

$$P(Y_{i1} = Y_{i2} = 1|g_{i1}, g_{i2}, g_{ip}, \phi_i) = P(Y_{i1} = 1|g_{i1}, g_{i2}, g_{ip}, \phi_i)P(Y_{i1} = 1|g_{i1}, g_{i2}, g_{ip}, \phi_i).$$

Using the logit model,

$$P(Y_{i1} = Y_{i2} = 1|g_{i1}, g_{i2}, \phi_i) = \int \left(\frac{e^{\beta_0+\beta_1 g_{i1}+\phi_i}}{1+e^{\beta_0+\beta_1 g_{i1}+\phi_{i1}}}\right)\left(\frac{e^{\beta_0+\beta_1 g_{i2}+\phi_i}}{1+e^{\beta_0+\beta_1 g_{i2}+\phi_i}}\right) f(\phi_i)d\phi_i$$

The same structure can be used to account for additional variation in the three-generation pedigree structure with parent/offspring where Y_{i1} is the phenotype of 4 in Figure 1.1c (the affected parent who is also an offspring), and Y_{i2} is the phenotype of 5 in Figure 1.1c (the offspring 4 and 5 in the pedigree).

We assume $\phi_i \sim N(0, \sigma^2)$ for the straightforward calculation. The probability of both offspring and parent-offspring pairs being affected using the logit model is used to calculate the probability of the pedigrees and the multiplex trios and the recurrence risk ratios. Then, the power of the FBAT statistic when conditioning on parental genotypes and founder genotypes using the random effect are compared to the power of FBAT statistic when conditioning on parental and founder genotypes under the assumption that the trait is generated solely based on the genotypes with no shared family effects.

1.4 Results

1.4.1 Comparison of FBAT Conditioning on Parents' and Founders' Genotypes for Pedigrees

In the scenarios under the additive model, both offspring are affected, all genotypes are known, and one pedigree type is used Figure 1.1c. The alpha level was calculated to be around 0.05 on average for a nominal level of 0.05 (not shown). In Figure 1.2a, the power for FBAT while conditioning on the founders of the pedigree and while conditioning on the parents of the pedigree of a rare disease (prevalence of 0.001) and odds ratio (OR) of 1.4 and 2 are compared.

For each of the allele frequencies, the power for the FBAT when conditioning on founders? genotypes in the pedigree is minutely higher than the power when we condition on parental genotypes . In Figure 1.2b, the power is also calculated for a common disease (prevalence of 0.14). The power of the statistic is lower with a rare disease, yet there is still a very small increase in power when conditioning on the founder genotypes for the statistic.Spielman et al. (1993) showed that when ascertaining on all offspring being affected, the power of the FBAT is heavily influenced by the genotypic relative risk and proportion of informative families (those families whose variance is not 0 in the statistic) that contribute to the statistic. Therefore, the increase in power with relative risk is consistent with Spielman et al. (1993). Also, for a common disease, the power is less dependent on the relative risk compared to a rare disease, where the power is more de-

pendent. The number of pedigrees in each simulation were doubled when calculating the empirical power under the recessive mode of inheritance. However, the power was extremely low and is not shown.

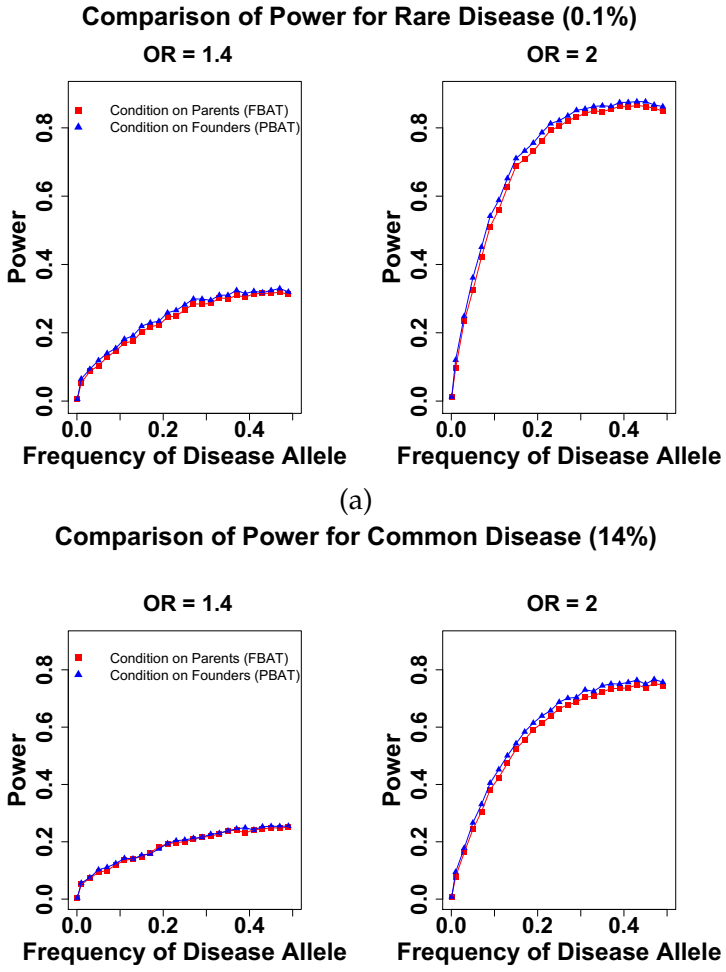
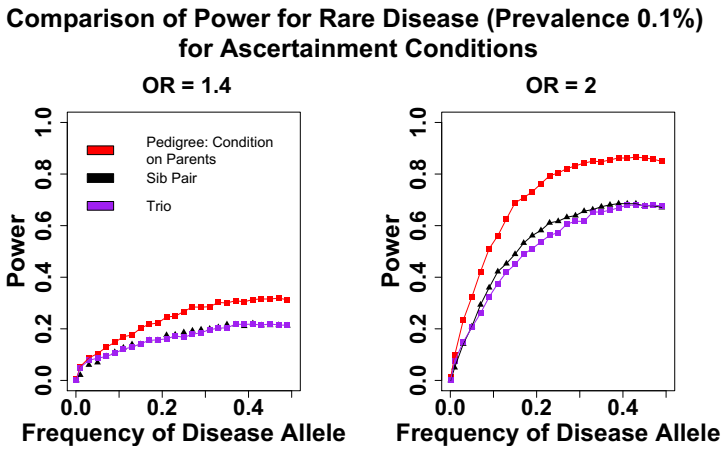


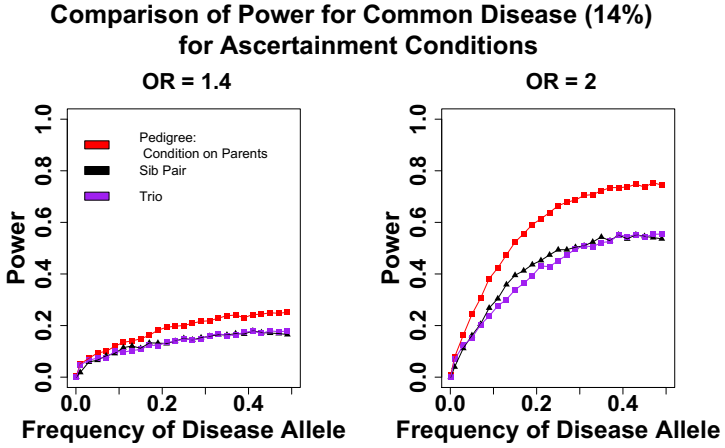
Figure 1.2: **Empirical power of the FBAT statistic for pedigrees.** The disease prevalence is 0.1% in panel (a) and 14% on in panel (b). The panel on the left in (a) and (b) represents and odds ratio of 1.4 and the panel on the right represents an odds ratio of 2. On the x-axis is the frequency of the disease allele and on the y-axis is the power. The red line represents the empirical power for the FBAT statistic of pedigrees conditional on the parents , and the blue line represents the empirical power for the FBAT statistic of pedigrees conditional on the founders.

1.4.2 Comparison of ascertainment conditions

Next, the ascertainment of pedigrees is compared with nuclear families. Because the power of the FBAT statistic is very close when conditioning on parents and founders, only the power for conditioning on parents is shown in Figure 1.3. Under the additive mode of inheritance with an odds ratio of 1.4 and 2, the power of the FBAT used with pedigrees is higher than the power of the sib pairs and the trios is the lowest disease allele frequency (Figure 1.3). However, the power difference between sib pairs and trios is extremely minimal for both common and rare disease. In fact, the power of ascertaining pedigrees is the highest for both rare and common disease (Figure 1.3).



(a)



(b)

Figure 1.3: **Comparison of ascertainment conditions for simple disease.** The disease prevalence is 0.1% in panel (a) and 14% in panel (b). The panel on the left in (a) and (b) represents and odds ratio of 1.4 and the panel on the right represents an odds ratio of 2. On the x-axis is the frequency of the disease allele and on the y-axis is the power. The red line represents the empirical power for the FBAT statistic of pedigrees conditional on the parents , the black line represents sib-pairs, and the blue line represents the power for trios.

1.4.3 Comparison of ascertainment conditions for "complex disease"

In the first part of the simulations, the disease trait is simulated only based on the genotype. We make the assumption that there are other factors that contribute to the outcome of disease using a logit model with a random effect. The random effect represents shared factors between family members. These shared factors can be additional genetic

components or environmental components. By increasing the variance component of the shared familial effect, one also increases the recurrence risk ratio (RRR). The sibling RRR is defined as the probability of a person being affected given his/her sibling is affected divided by the probability the sibling is affected. The RRR can also be calculated for parent/offspring. The RRR shown in Figure 1.4 is the sibling RRR, the sibling RRR is very close in magnitude to the parent/offspring RRR. Figure 1.4 has three lines that represent the change in power for the FBAT statistic when applied to pedigrees, sib pairs, and trios. In all three situations, as the RRR increases, the power decreases. When $\sigma^2 = 0$, the sibling RRR is 1.01. Because there is an additional component that contributes to the outcome, the genetic effect is attenuated causing a decrease in power. In addition, as σ^2 increases, the recurrence risk ratio for siblings and parent-offspring increases. Our results show that the power of family based designs decrease because shared genetic factors increase heritability. Our results are consistent with Ferreira et. al who show that inclusion of shared factors by families (i.e. environmental) decrease the power (Ferreira et al., 2007). When including shared familial factors, families with a high value of (which means more shared familial components) are more likely to be selected even if there are no disease alleles in the family. However, for smaller values of the RRR, ascertaining pedigrees still has higher power than ascertaining sib pairs and trios. As the RRR gets large, the difference between the power of the designs is very small.

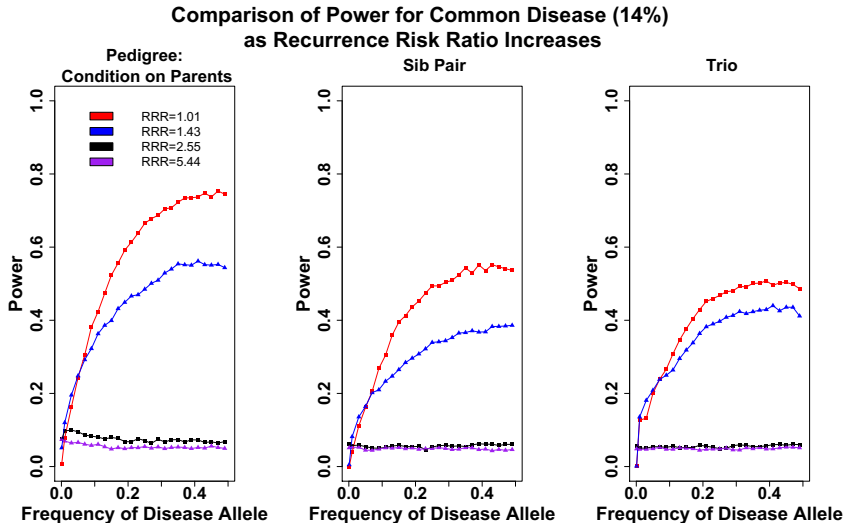


Figure 1.4: **Comparison of ascertainment conditions for complex disease.** Compares the empirical power of the FBAT used to analyze the pedigree for association, PBAT used for the pedigree analysis, sibpairs, and trios for the additive mode of inheritance and prevalence of 0.14.

1.5 Data Application

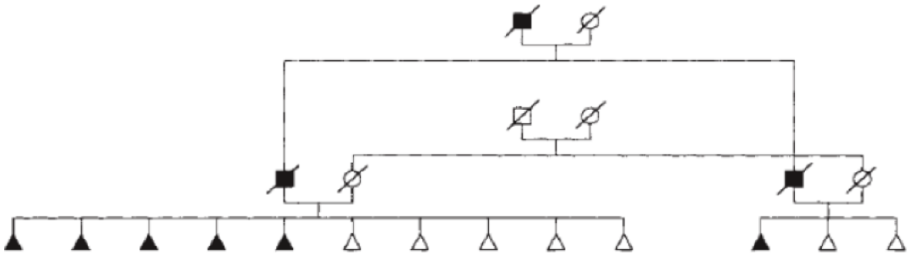


Figure 1.5: **Data application: APP gene.** Pedigree representing a family with the APP gene which is known to be associated with early onset Alzheimer's disease (Goate et al., 1991). Black indicates a carrier of the gene and slashes indicate the person is dead. The triangles are used to preserve the anonymity of a family.

We compare the power of the test statistic when conditioning on the parents' genotypes and conditioning on the founders' genotypes in the pedigree in which early-onset AD is inherited as an autosomal dominant disorder in Figure 1.5. Although we do not have the actual genotypes in the pedigree, the genotypes can be inferred using the dominant mode of inheritance. The pedigree in Figure 1.5 was analyzed for association using

the FBAT, conditioning on the parents and the founders, assuming one marker at the DSL and autosomal dominant inheritance. When conditioning on the parents, the value of the FBAT statistic is $\chi^2 = 8(p = 0.001)$. When conditioning on the founders, the FBAT statistic is $\chi^2 = 5.67(p = 0.0345)$. We notice that conditioning on the parents leads to a more significant statistic. This example further illustrates the small difference between conditioning on the parents' genotypes and conditioning on the founders' genotypes in the pedigree.

1.6 Discussion

In this simulation study, we have shown the following:

1. In our scenarios, conditioning on founders has a very small power advantage over conditioning on parents.

2. Multiplex pedigrees can have a large power advantage over trios and sib pairs when the causal locus is a substantial risk factor.

3. For diseases with a large sib RRR due to shared family factors, the power decreases as RRR increase for all ascertainment designs. Yet, multiplex pedigrees still have larger power over trios and sib pairs for smaller RRR.

In this study, we have shown that there is minimal difference in the power of FBAT, conditioning on founders and parents in a pedigree. In fact, the power of conditioning on founders is slightly higher in the study, but in the data application, the statistic for condition- ing on parents is more significant than the statistic computed by conditioning on founders. To understand this, consider that the power of the statistic depends on the degree of heterozygosity in the family. As the number of heterozygotes increases, the power of the TDT and FBAT statistic increases. Figure 9 shows an example where conditioning on founders is more powerful than conditioning on parents. Let A be the disease allele. When conditioning on the parents 1.1c, offspring 5 (with parents 3 and 4) does not contribute to the test statistic because both parents are homozygous. On the

other hand, when the statistic is calculated conditioning on the founders (1, 2, and 3), offspring 5 contributes to the statistic because both possible genotypes for offspring 4 are considered (AB and BB), which increases the sample space of the genotype of offspring 5 (AB and BB) 1.1c. Therefore, conditioning on founders provides more information about potential outcomes. For complex disease, the power reduces as the recurrence risk ratio increases for trios, sibpairs and multiplex pedigrees. As ϕ increases, the RRR increases which means families that do not have the disease allele are more likely to enter the sample. As a result, the frequency of the disease allele decreases and so does the power of the statistic. This result is consistent with the results from Ferreira et al. (2007). In addition, our study shows that pedigrees still have a power advantage over ascertaining trios and sib pairs for a complex disease model. As the strength of the familial factors increases, though, the power advantage decreases as well. These results suggest that pedigrees are powerful in detecting association where additional familial factors such as environment and polygenic effects do not strongly influence the disease risk. Comparing the FBAT statistic when conditioning on founders versus parent genotypes, conditioning on the founders was more powerful for the additive and recessive inheritance models and over varying allele frequencies, but the power difference was very slight. When comparing using association tests for pedigrees to using the association statistic in other designs, it is important to consider shared familial factors which decrease the power of the FBAT statistic when applied to trios, sib pairs, and pedigrees. Analyzing multiplex pedigrees for association has a large power advantage compared to ascertaining trios and sib pairs for Mendelian diseases. The power advantage is still there when analyzing association for complex disease for smaller values of the RRR. As the RRR gets large, the power advantage diminishes. Although our results show that conditioning on founders is slightly but consistently more powerful than conditioning on parents, in many cases the genotypes of founders are not available and power may be lost when using methods to reconstruct the missing family genotypes (Laird et al., 2000). In the scenario presented in the paper, all of the genotypes are known. Future work includes comparison of the power when genotypes are missing. The scenario included in the paper is a three-generation pedigree,

but one could look at the effect of different pedigree structures on the power of the FBAT statistic. Furthermore, one could explore the implication of the power difference on the cost effectiveness of genotyping when comparing ascertainment conditions.

Acknowledgements

This work was supported by grants from the National Institutes of Health (ES007142, ES000002, ES016454, CA134294, CA160736, CA016672, DA032581) and the Ford Fellowship Foundation.

Cancer Risk Assessment in Twins

Christina McIntosh[1,2], Danielle Braun[1,2], Jacob v. B. Hjelmborg [3], Soren Moller [3], Jaakko Kaprio[4,5], Kamila Czene[6], Jennifer Harris[7], Lorelei Mucci[8], Giovanni Parmigiani [1,2]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[2]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute
[3]Department of Biostatistics, University of Southern Denmark, J.B. Winslowsvej 9B, DK-5000 Odense, Denmark
[4]Department of Genetic Epidemiology, University of Helsinki
[5]Institute for Molecular Medicine, Finland
[6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet
[7]Department of Genes and Environment, Norwegian Institute of Public Health
[8]Department of Epidemiology, Harvard T.H. Chan School of Public Health

2.1 Introduction

Twin studies provide important information about familial risk of cancer by using the unique genetic relationships of monozygotic (MZ) and dizygotic (DZ) twin pairs. Generally, twin studies are analyzed to quantify heritability, or partition phenotypic variation among genetic and other sources (Kempthorne and Osborne, 1961). Heritability is the proportion of the total variation of a trait that is due to genetic factors in a population. Lichtenstein et al. (2000) reported comprehensive heritability estimates of cancer using twin registries from Sweden, Denmark, and Finland. For breast cancer, heritability was estimated to be 27%, 42% for prostate cancer, and 35% for colorectal cancer. The estimates for other common cancers were not interpretable. Mucci et al. (2016) reported updated heritability estimates using the Nordic Twin Studies of Cancer (NorTwinCan), which includes twins from nationwide registers in Denmark, Finland, Norway, and Sweden, and presently constitutes the largest twin study in the world. Twin pairs are followed for an average of 32 years for cancer incidence and mortality. The authors estimate the heritability of cancer overall to be $33\%[30\%, 37\%]$. They also report estimates of concordance; the frequency with which twins within a pair experience the same health history. The familial risk is defined as the risk of cancer in a twin given his/her co-twin was diagnosed with the same cancer. Mucci et al. (2016) found an excess cancer risk in twins whose co-twin was diagnosed with cancer. After age 65, the familial risk of any cancer by age 100 in these twins was $37\%[36\%, 38\%]$ for DZ twin pairs and almost $46\%[44\%,\ 48\%]$ for MZ pairs.

When one considers the trait defined as the onset of an age-dependent disease, as is the case in cancer, heritability is estimated by creating an artificial, unobserved, trait whose variance components are inferred statistically (Wright, 1934). This makes the concept of heritability somewhat abstract and only indirectly relevant for clinical practice. Concordance is a more intuitive concept, but it also falls short of full clinical usability, as most clinical decisions are best framed within the context of a specific time period, say 10 or 20 years.

The first aim of this paper is to expand analyses of heritability and concordance by

providing absolute risk estimates conditional on the affection status of the co-twin, for use in practical cancer prevention settings. For example, consider a twin who does not yet have cancer and would like to be counseled about his/her future risk of cancer. Based on knowledge of the cancer status of his/her twin, we use data from NorTwinCan to compute the absolute 5-year, 10-year, 20-year, and 30-year cancer risk. These absolute risks in MZ and DZ twins provide a practical tool that can be used, for example, in assisting decisions about targeted prevention strategies. DZ twins share approximately half of their genetic make-up which makes them similar to full-siblings, as such the risks for DZ twins approximates the risks for full siblings.

As a second aim, we also evaluate risk ratios, comparing MZs and DZs twins' risk to the baseline cancer risk without conditioning on the co-twin. We can then use these risk ratios to obtain the risk conditional on a co-twin's status in a new population, by multiplying these ratios by the baseline cancer risk of the new population. For example, we apply the risk ratios estimated from NorTwinCan to data from the Surveillance, Epidemiology, and End Results program (SEER) (Howlader et al., 2013), to calculate the risks conditional on the co-twin's affection status in the U.S. population.

Several alternative methods can be used for estimating risk in this context. Unlike heritability, absolute risk estimates can be derived by direct empirical estimation of the cumulative incidence curve, using the Aalen-Johansen estimator (Aalen and Johansen, 1978) to account for competing risks (in this case mortality). This approach does not require any assumptions about latent traits, their distributions or the distribution of the failure times. It is interesting to contrast these direct estimates to those implied by state of the art methods for heritability analyses, which also allow one to derive risk estimates. We consider two: the semi-parametric random effects model developed by (Scheike et al., 2010), which can be used to estimate the joint probability of cancer for twin pairs in the presence of competing risks, left-truncation, and right-censoring; and the liability threshold model typically used to estimate heritability. Holst et al. (2016) expanded the standard liability threshold model to include inverse probability of censoring weighting of complete observations, which leads to consistent estimates of heritability and concordance

estimates in the presence of censoring. The final aim of the paper is to provide estimates of the absolute risks using all three approaches, allowing us to investigate how the assumptions of each model impact the estimates of risk.

2.2 Goals

Using the NorTwinCan cohort, we compute the absolute 5, 10, 20, and 30-year risk estimates using three methods:

1. The empirical conditional cumulative incidence curve in the stratum of interest, using the Aalen-Johansen estimator. This is done directly on the observed twin data from NorTwinCan without simulations.

2. The semi-parametric random effects model of Scheike et al. (2010). This is done directly on the observed twin data from NorTwinCan without simulations using the estimates of concordance and cumulative incidence produced using the semi-parametric random effects model.

3. The Holst et al. (2016) version of the liability threshold model for right censored data. We fit the model to NorTwinCan data, simulate twin data based on the model parameters, and then empirically calculate the cumulative incidence curves based on the simulated data.

2.3 Methods

2.3.1 Study Populations

The NorTwinCan cohort is the largest twin cohort in the world. It includes $357,377$ twin individuals, and is comprised of both MZ, same-sex DZ, and opposite-sex DZ twins. Our analysis considers the $202,868$ MZ and same-sex DZ twin pairs where both twins were alive at the start of follow-up. We excluded twin pairs whose zygosity was unknown. Individuals in NorTwinCan were followed prospectively until: cancer diagnosis,

death from other causes, emigration during follow-up, or time of last follow-up. The median follow-up time in NorTwinCan is 32 years, and there were $27,156$ incident cancers in the cohort. See Mucci et al. (2016) for further description of the cohort.

Once we estimate risk ratios for cancer risk conditional on a co-twin's affection status, we use these risk ratios to estimate the absolute risks of cancer in the United States by using estimates of baseline risk of cancer from the SEER program registry. SEER began in 1973 and collects data on cancer cases from locations throughout the United States (Howlader et al., 2013).

2.3.2 Definitions

Let K be the number of twin pairs. We consider two events; cancer and mortality. For the i^{th} twin, $i = 1, 2$ in the k^{th} twin pair, $k = 1, 2, \ldots, K$, let the event time be T_{ki} and the event type be ϵ_{ki}. Let $\epsilon_{ki} = 1$ if the event that is observed is cancer and $\epsilon_{ki} = 2$ if the event is death. Let δ_{ki} equal to 0 for a censored individual. Thus, we observe $\tilde{\epsilon}_{ki} = \epsilon_{ki}\delta_{ki}$, which is 0 for a censored individual, 1 for an individual diagnosed with cancer, and 2 for a deceased individual. Additionally, let $\mathbf{X_{ki}} = (1, X_{ki,1}, ..., X_{ki,p})^T$ indicate a vector of covariates for the i^{th} twin in the k^{th} twin pair.

For our analyses, twins are considered to be classified in three ways: 1) diagnosed with cancer of any type, 2) dead without cancer diagnosis, or 3) alive without cancer diagnosis or lost to follow-up. Based on the country of birth for the twins, we defined the time of entry into the study and end of follow-up. We assume the same censoring in pairs, which we believe is reasonable since censoring in this type of data is often be attributed to administrative causes. Hence, identifying the bivariate censoring distribution can be reduced to estimating the marginal censoring distributions (Scheike et al., 2014). We test the assumption of equivalent cumulative incidence for MZ and DZ twins using the Fine-Gray regression model. We found that the cumulative incidence for MZ twins was around 1% higher than DZ twins.

2.3.3 Calculating Risk Using The Empirical Distribution

The Aalen-Johansen estimator is a nonparametric maximum likelihood estimator of the cumulative incidence of cancer at age a, denoted by $P(T_{ki} \leq a, \epsilon_{ki} = 1)$ (Aalen and Johansen, 1978). Under some regularity conditions, the Aalen-Johansen estimator converges with probability 1 to the cumulative incidence function and is asymptotically normally distributed (Aalen and Johansen, 1978). Using the Aalen-Johansen estimator, we compute the absolute risks for counseling an unaffected individual based on his/her cotwin's history. Let $i = 2$ indicate the unaffected twin that is being counseled, and let $i = 1$ indicate the other twin. Let a_1 be the current age of the twins, and t_r be the time interval of interest for risk evaluation, say $t_r = 5, 10, 20, 30$.

1. If twin $i = 1$ is unaffected, we wish to estimate the probability that twin $i = 2$ develops cancer in the next t_r years, given that both twins are alive and unaffected at their current age a_1:

$$P(T_{k2} \leq a_1 + t_r, \epsilon_{k2} = 1 | T_{k1} > a_1, T_{k2} > a_1). \tag{2.1}$$

2. If twin $i = 1$ is affected, we wish to estimate the probability that twin $i = 2$ will develop cancer in the next t_r years, given that he/she is alive and unaffected, and twin $i = 1$ was diagnosed in the period between a_0 and a_1 (where a_0 is the beginning of the time period that is used to define the conditioning event for twin $i = 1$):

$$P(T_{k2} \leq a_1 + t_r, \epsilon_{k2} = 1 | a_0 \leq T_{k1} \leq a_1, T_{k2} > a_1, \epsilon_{k1} = 1). \tag{2.2}$$

To calculate these risk estimates, we must first create strata as follows.

1. To compute (3.2), where twin $i = 1$ is unaffected: Identify twin pairs where both twins are unaffected at age a_1.

2. To compute (3.2), where twin $i = 1$ is affected: Identify twin pairs where exactly one twin has cancer between the ages of a_0 and a_1 and the other twin is cancer free at age a_1.

Once the appropriate stratum is created, we use the Aalen-Johansen estimator to estimate that cumulative incidence function within each strata. The unconditional risks are calculated without conditioning on the co-twin's affection status, $P(T_{k2} \leq a_1 + t_r, \epsilon_{k2} = 1|T_{k2} > a_1)$. We then use these curves to estimate the absolute 5, 10, 20, and 30-year risks. We calculate 95% confidence intervals using the bootstrap method (Monaco et al., 2005).

2.3.4 Calculating Risk Using a Semi-parametric Random Effects Model for Multivariate Competing Risks Data

Scheike et al. (2010) propose a semi-parametric random effects model that accounts for left-truncation and right-censoring in clustered survival data. In our case, clusters are twin pairs. The model allows one to estimate the marginal cumulative incidence functions, and the associations between cause-specific failure times within a pair are modeled through dependence parameters of copula functions. Using the cumulative incidence functions and the estimates of the associations between the failure times, one can calculate the cumulative incidence for a specific pair, which in turn allows estimation of the joint probability of both twins experiencing an event by a particular time. Using the joint probabilities, we can then estimate the risk for a twin experiencing an event given his/her co-twin has already experienced an event.(Scheike et al., 2010).

Let θ_k denote the random effects, and assume they are independently distributed random variables, one for each pair, that capture variability in risk from pair to pair. Individual twins within pairs are independent conditional on θ_k. We also consider gender and country as covariates ($\mathbf{X_{ki}}$) in the model in order to estimate the dependence parameters. The model assumes a cumulative incidence function of the form

$$P_c^*(T_{ki} \leq a, \epsilon_{ki} = 1|x_{ki}) = 1 - exp(-\theta_k \Psi_{v_k}^{-1}[-\eta(a)^T x_{ki}]), \tag{2.3}$$

where $\eta(a)$ is a $(p+1)$-dimensional vector of regression functions that vary with age, a, and $\Psi_{v_k}(a) = E_{\theta_k}[exp(-\theta_k a)|x_{ki}]$ is the Laplace transform of the random effects θ_k. In our simulation, we choose $\Psi_{v_k}(a)$ to be the Laplace transform of a gamma distribution with mean 1 and variance v_{MZ} for MZ twins, and v_{DZ} for DZ twins which is the standard

choice. Here, v_{MZ} and v_{DZ} control the association between the ages of cancer diagnosis.

The marginal cumulative incidence function is then modeled using the generalized semi-parametric additive model.

$$-log[1 - P_c(T_{ki} < a, \epsilon_{ki} = 1|x_{ki})] = \eta(a)^T x_{ki}. \tag{2.4}$$

Once we have obtained $\eta(a)$ from 2.4, we can also evaluate the bivariate cumulative incidence function, P_{11} for two arbitrary ages a and a':

$$\begin{aligned} P_{11}(a, a') = P(T_1 \leq a, \epsilon_1 = 1, T_2 \leq a', \epsilon_1 = 1, \epsilon_2 = 1) = \\ 1 - \exp\{-\eta(a)^T X_{k1}\} - \exp\{-\eta(a')^T X_{k2}\} \\ + \Psi_{v_k}\left[\Psi_{v_k}^{-1}(\exp\{-\eta(a)^T X_{k1}\}) + \Psi_{v_k}^{-1}(\exp\{-\eta(a')^T X_{k2})\}\right]. \end{aligned} \tag{2.5}$$

Using the bivariate cumulative incidence function and the marginal cumulative incidence, we can calculate the absolute risks conditional on a co-twin's affection status. For example, to calculate the t_r-year risk of twin $i = 2$ getting cancer conditional on twin $i = 1$ being affected, we can evaluate:

$$\begin{aligned} P(T_{k2} \leq a_1 + t_r, \epsilon_{k2} = 1 \mid T_{k1} \leq a_1, T_{k2} > a_1, \epsilon_{k1} = 1) = \\ \frac{P_{11}(a_1, a_1 + t_r)}{P_c(a_1)} - \frac{P_{11}(a_1, a_1)}{P_c(a_1)}. \end{aligned} \tag{2.6}$$

Without taking into account truncation, we can over estimate the dependence (Scheike et al., 2014). The probabilities of being censored were estimated using the the Kaplan-Meier method stratified by zygosity and country.

2.3.5 Calculating Risk Using the Liability Threshold Model for Right-Censored Data

We use the liability threshold model to estimate heritability on the liability scale by classifying subjects as cases or non-cases. In our simulation, we divide the twins into cancer cases and non-cancer cases. The liability threshold model postulates a latent outcome (or liability) which is assumed to have a bivariate normal distribution. We then assume a cancer case is observed if the corresponding latent outcome is above a defined threshold,

and not observed if the outcome is below the threshold. A full model without the latent variable structure would be a practical choice, and in order to compare MZ and DZ twins, one could constrain twin 1 and twin 2 and the MZ and DZ twin marginals to be the same.

The model estimates the disease covariance between MZ and DZ twin pairs to decompose the variation of cancer liability into additive genetic effects (A), dominant genetic effects (D), common environmental effects (C), and unique environmental effects (E). The variances of each of the terms are, $\sigma_A^2, \sigma_D^2, \sigma_C^2, \sigma_E^2$ respectively. The within-pair covariance of the liability for MZ twins is $\sigma_A^2 + \sigma_C^2$ and the within-pair covariance of the liability for DZ twins is $\frac{1}{2}\sigma_A^2 + \sigma_C^2$. The unique environmental effects (E) do not contribute to the within-pair covariance of the twin pairs and it is assumed to be independent within twin pairs.

The model proposed by Holst et al. (2016) extends the classical liability threshold model to include the inverse probability of censoring weighting of complete observations to correct for right censoring in the data. In the twin registry, we believe that censoring can often be attributed to administrative causes; therefore, we assume that twins are censored at the same time. Therefore, identifying the bivariate censoring distribution can be reduced to estimating the marginal censoring distributions. The Kaplan-Meier method stratified by zygosity and country is used to estimate the probabilities of being censored. Left truncation is also present in these data because both twins must be alive at the beginning of follow-up in order to be included in the analyses. In the liability threshold model, it is assumed that everyone is followed until age a (Holst et al., 2016). In this time period, twins are classified as having cancer or not having cancer before age a.

Let η_{ki}^A be the liability component for the additive genetic effects, and η_k^C that for the shared environmental effect for the k^{th} twin pair. These can also be interpreted as random effects. The model assumes:

$$P(T_{ki} \leq a, \epsilon_{ki} = 1 | \eta_{ki}, X_{ki}) = \Phi(\beta^T X_{ki} + \eta_{ki}^A + \eta_{ki}^C). \tag{2.7}$$

Here Φ is the cumulative standard normal distribution. The E component, unique environmental effect, is indirectly modeled through the inverse link function, and the heri-

tability H^2 may be defined as

$$H^2 = \frac{\sigma_A^2 + \sigma_D^2}{\sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2}. \tag{2.8}$$

The A, D, and C components of the model cannot be estimated simultaneously due to statistical identifiability issues.

As a result, we analyzed a series of models, ACE, ADE, AE, CE, and E models. The estimates for the genetic effects are similar for each model. We compared the models using the AIC. We estimate the concordance and the absolute risks through the following steps:

1. Fit the liability threshold model using the twin data from NorTwinCan to estimate the model parameters.

2. Use the parameters obtained in step 1 to simulate twin survival data for the same number of twin pairs as included in NorTwinCan.

3. Create strata of twin pairs based on the affection status of the twins.

4. Calculate the empirical absolute risks in the simulated cohort, and calculate the 95% confidence intervals using the bootstrap method.

2.4 Results

2.4.1 Risk Estimates based on the Empirical Distribution

We consider a twin who is alive and seeking counseling about his/her risk. Figure A.7 presents the empirical estimates of the 5-year, 10-year, 20-year, and 30-year risks of cancer calculated using the Aalen-Johansen estimator. Although the cumulative incidence estimates in MZ and DZ twins differs by approximately 1%, we believe it is a reasonable assumption to assume the marginals for MZ and DZ twins are the same. Estimates depend on whether the co-twin is affected with cancer (MZ red line, DZ blue line), or not (MZ green line, DZ orange line). We also present the unconditional risks estimates as a black line. As an example of how to interpret the risk estimates, we focus on the 10-year

risk in the second panel from the top. If the counseled twin is 50 years of age, and the MZ co-twin is affected, we refer to the blue line and see that he/she has a $4.61\%[4.40\%, 4.82\%]$ probability of developing cancer by the age of 60. At a counseling age of 70, the 10-year risk of cancer for an MZ twin with an affected co-twin is $9.76\%[7.88\%, 11.88\%]$ and is $8.36\%[7.03\%, 9.83\%]$ for a DZ twin with an affected co-twin. Appendix Figures 1, 2, and 3 show the risk if the co-twin developed cancer in the preceding 5, 10, or 20 years. The risk of cancer does not vary substantially depending on the interval considered.

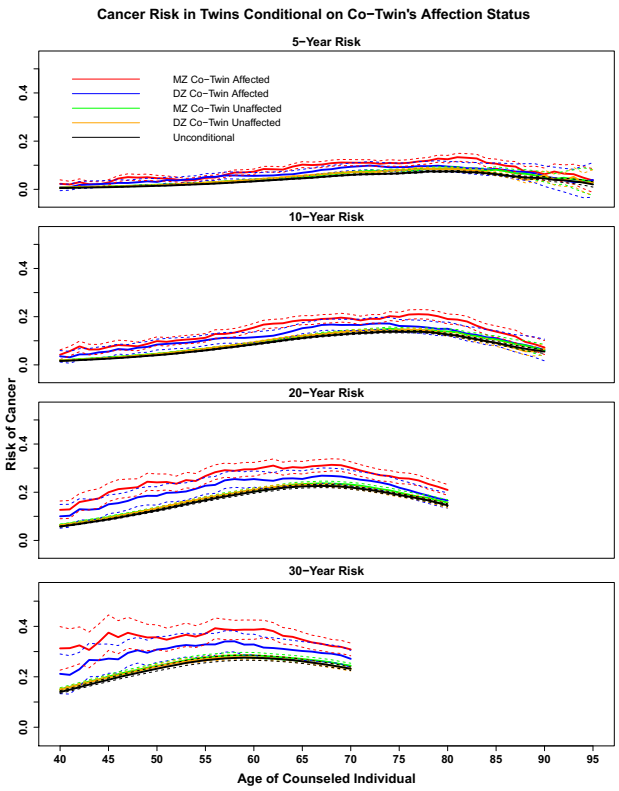


Figure 2.1: **Empirical Absolute Risk Estimates.** The figure shows the 5-year, 10-year, 20-year, and 30-year risks (arranged from top to bottom) of having cancer conditional on whether or not the co-twin had cancer previously . The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin. The risk is estimated empirically using the cumulative incidence curve. Red represents the risk for MZ twins with an affected co-twin and blue represents the risk for DZ twins with an affected co-twin. The purple represents the risk conditional on having an unaffected co-twin. The dashed lines represent the 95% confidence intervals which are calculated using the bootstrap method.

A general conclusion from Figure A.7 is that the risks for MZ and DZ twins are clearly

separated in the direction expected based on prior observation of heritability of cancer. However, the absolute risk curves are not far and the confidence intervals overlap. Only when the risk assessment horizon is 20 or 30 years the difference is of a magnitude likely to affect clinical management. We notice a drop in risk after age 80. This drop in risk is comparable the drops in risk seen in the SEER data. Furthermore, when we look at the 20-year risk, we see that the confidence intervals are wide at age 40 then get smaller. This is because we look at the probability of cancer within the range of 40 to 60 years of age. As cancer is mostly a late onset disease, more cancer cases are observed at the later ages.

The cumulative incidence of cancer in MZ twins estimated using the Aalen-Johansen estimator is 33% and the cumulative incidence of cancer in DZ twins is similar and is estimated to be 32%. Over time, MZ twins of affected twins have a higher risk of cancer when compared to DZ twins conditional on having an affected co-twin. When we compare the unconditional estimates (black line) to the risk having an unaffected twin (MZ green, DZ orange), we see that the lines overlap. This suggests that the risk of cancer for a twin with an unaffected co-twin is similar to the risk for a person for whom we do not have information about the twin's affection status.

In the appendix, Figure A.3 presents the 5, 10, 20, and 30-year risks of cancer for MZ twin compared to DZ twins when the co-twin has cancer in the past 5 years. More specifically, we estimate 10-year risk of developing cancer for an unaffected individual whose twin was diagnosed with cancer in the previous 5 years. Similarly, Appendix Figures 5 and 6 presents the 5, 10, 20, and 30-year risks of cancer for MZ twin compared to DZ twins when the co-twin has cancer in the past 10 and 20 years respectively.

Figure 2.2 presents the risk ratio estimates for the 5-year, 10-year, 20-year, and 30-year risk estimates for an MZ/DZ twin with an affected/unaffected co-twin compared to the unconditional risk estimates. The risk ratios can be used to estimate the absolute risk in a different population with a different baseline risk. We apply these risk ratios to SEER data in Figure 2.3b to calculate the absolute risks conditional on a co-twin's cancer status. We also compare the unconditional risk estimates from SEER (purple) and from NorTwinCan (black) (Figure 2.3a), and we see that SEER has a higher baseline of risk.

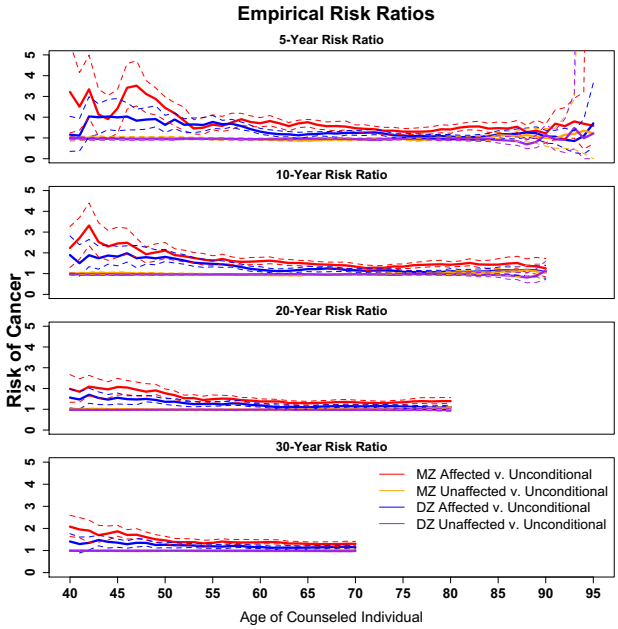
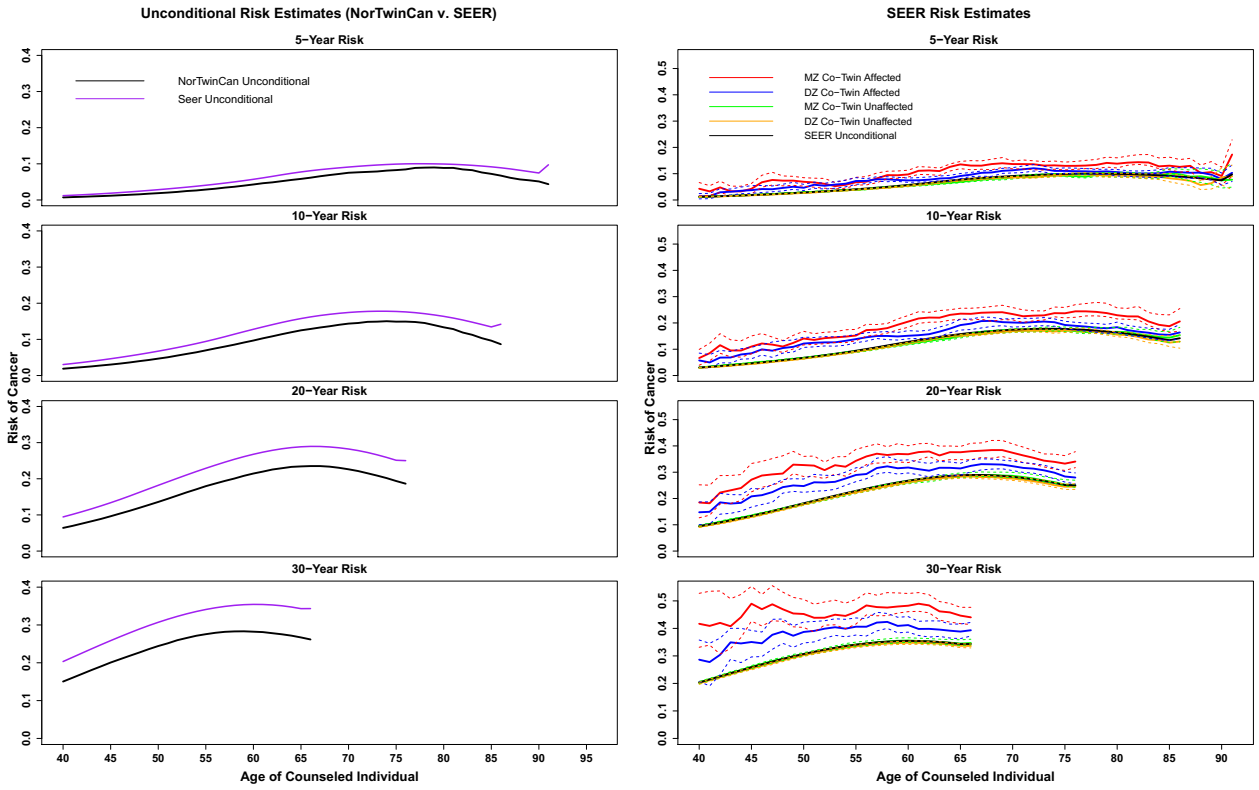


Figure 2.2: **Empirical Absolute Risk Ratio Estimates.** The figure shows the ratio of the 5-year,10-year, 20-year, and 30-year risk of having cancer conditional on having an affected/unaffected co-twin compared to the unconditional risk estimates. The red line represents the risk ratio for an MZ twin that has an affected co-twin compared to the unconditional risk. The blue line represents the risk ratio for DZ twin with an affected co-twin compared to the unconditional risk estimates. The orange and purple lines represent the risk ratio for MZ (orange) and DZ (purple) twins with an unaffected co-twin compared to the unconditional risk estimates. The dashed lines represent the 95% confidence intervals which are calculated using the bootstrap method.



(a) (b)

Figure 2.3: **SEER Risk Estimates.** (a) Unconditional risk estimated using 2011-13 SEER data (purple) and the unconditional estimates from NorTwinCan data (black). (b) The figure shows the 5-year, 10-year, 20-year, and 30-year risks (arranged from top to bottom) of having cancer conditional on whether or not the co-twin had cancer previously based on SEER baseline risks . The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin. The risk is estimated empirically using the cumulative incidence curve. Red represents the risk for MZ twins with an affected co-twin and blue represents the risk for DZ twins with an affected co-twin. The purple represents the risk conditional on having an unaffected co-twin. The dashed lines represent the 95% confidence intervals which are calculated using the bootstrap method.

2.4.2 Risk Estimates based on the Semi-parametric Random Effects Model for Multivariate Competing Risks Data

Next we repeat this analysis using estimates from the semi-parametric random effects model using equation 2.8. The estimates of the dependence parameters are $\hat{v}_{MZ} = 1.03$ with standard error 0.40 for MZ twin pairs and $\hat{v}_{DZ} = 0.84$ with standard error 0.30 for DZ twin pairs. The larger estimate for MZ twins is reflective of the stronger association

between failure times for MZ twins compared to DZ twins. Using these estimates, we calculate the risks using equation 3.3.

In order to calculate the 5, 10, 20, and 30-year risk estimates, we use $\hat{v_k}$ for MZ and DZ twins (1.03 and 0.84 respectively) and 2.8 . Figure 3.4 shows the 5, 10, 20, and 30-year risk of cancer conditional on the co-twin having cancer using the data is simulated from the semi-parametric random effects model.

At a counseling age of 50, there is a $5.01\%[4.52\%,\ 5.33\%]$ probability that an unaffected individual with an affected MZ co-twin (blue line) will get cancer by the age of 60. The 10-year risk of cancer for an MZ twin with an affected co-twin at 70 is $20.86\%[17.32\%,\ 24.54\%]$ and $17.52\%[14.22\%, 21.47\%]$ for a DZ twin with an affected co-twin.

Similar to the empirical risk calculations, risk estimates are higher for MZ twins (red line) and DZ twins (blue line) at each age. Even though the risk for twins with unaffected co-twins are similar, the risk estimates for MZ and DZ twins are higher when we condition on having an affected co-twin when we compare to the empirical estimates. Over time, however, there appears to be a greater separation of the risks conditional on a twin having an affected co-twin and the risk of having an unaffected twin.

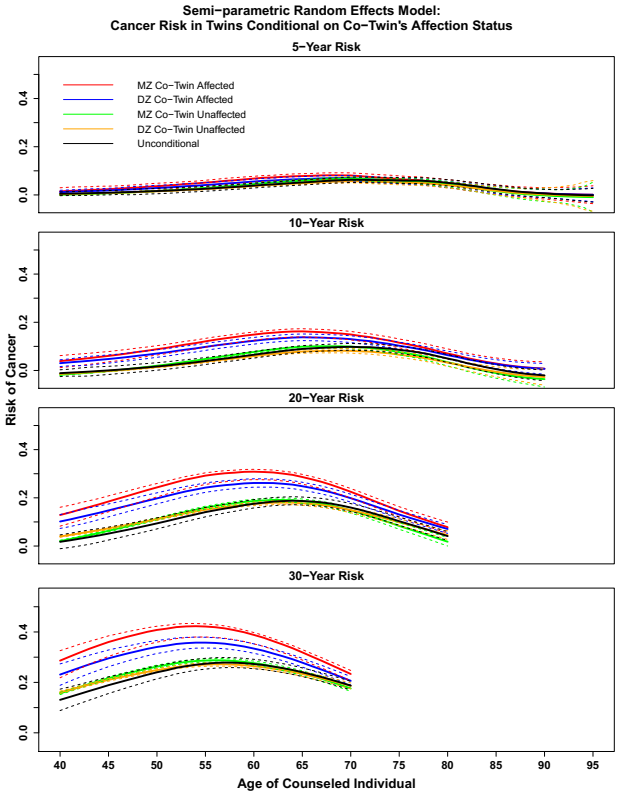


Figure 2.4: **Semi-parametric Random Effects Model Absolute Risk Estimates.** The figure shows the 5-year, 10-year, 20-year, and 30-year risks (arranged from top to bottom) of having cancer conditional on whether or not the co-twin had cancer previously based on the semi-parametric random effects model. The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin. The risk is estimated empirically using the cumulative incidence curve. Red represents the risk for MZ twins with an affected co-twin and blue represents the risk for DZ twins with an affected co-twin. The purple represents the risk conditional on having an unaffected co-twin. The dashed lines represent the 95% confidence intervals which are calculated using the bootstrap method.

2.4.3 Risk Estimates based on the Liability Threshold Model

Our results are consistent with Mucci et al. (2016) who estimated there to be no shared environmental effect for cancer overall in the ACE model and additive genetic effects estimated as $33\%[30\%, 37\%]$. Therefore, the most parsimonious model and the model with the lowest AIC was the AE model which estimates for additive genetic effects and unique environmental effects after we corrected for bias due to censoring. We use twin data simulated using the parameters estimated from the AE liability threshold model adjusting

for land and sex and correcting for censoring, to estimate the risk by calculating the empirical cumulative incidence curve. The results from the liability threshold model are of the same general magnitude as the empirical risks using the twin data from NorTwinCan (Figure 3.5). However, the risk tends to be greater at older ages compared to the empirical risks. Figure 5 shows the 5-year, 10-year, 20-year, and 30-year risk estimates for the data simulated using the liability threshold model.

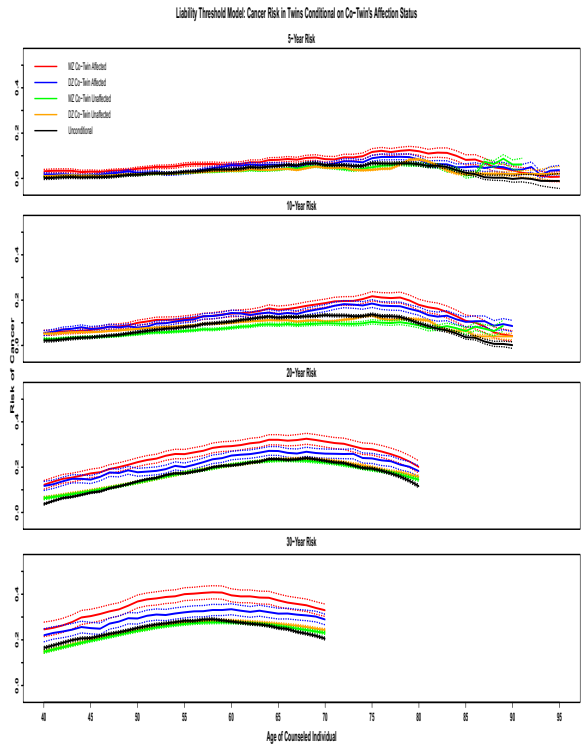


Figure 2.5: **Liability Threshold Model Absolute Risk Estimates.** The figure shows the 5-year, 10-year, 20-year, and 30-year risks (arranged from top to bottom) of having cancer conditional on whether or not the co-twin had cancer previously based on data simulated from a liability threshold model. The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin. The risk is estimated empirically using the cumulative incidence curve. Red represents the risk for MZ twins with an affected co-twin and blue represents the risk for DZ twins with an affected co-twin. The purple represents the risk conditional on having an unaffected co-twin. The dashed lines represent the 95% confidence intervals which are calculated using the bootstrap method.

2.4.4 Comparison of the Models

In this section, we compare the empirical risk estimates to the risk estimates we obtained from the liability threshold and semi-parametric random effects model. Figure 2.6 shows the 5-year, 10-year, 20-year, and 30-year risks calculated empirically, using the semi-parametric random effects model, and using the liability threshold model. Generally the curves show comparable patterns. Yet, several differences are worth noting. The semi-parametric random effects model gives higher estimates than the empirical estimates at earlier counseling ages, and then later crosses the empirical estimates at later ages. In contrast, the curve based on the liability threshold model is close to the empirical curve. However, at the 20-year risk, the estimates from the liability threshold model are slightly lower than the random effects model and the empirical model at earlier ages. It is not surprising that these models provide slightly different absolute risk estimates compared to the empirical estimates, since modeling assumptions differ, but the estimates are still relatively close.

When we estimate the risks empirically, we are imposing no assumptions on the correlation of the outcomes for twins in a twin pair. For the semi-parametric random effects model, we make the assumption that the frailty distribution for the random effects is gamma distributed with mean 1 and variance v_k and allow different degrees of association for MZ and DZ twins. The random effects structure for the liability threshold model was estimated using the probit scale and adjusted for censoring using inverse probability weighting. The random effects associated with the twin pair in the liability threshold model are assumed to be correlated and normally distributed. These different distributional assumptions could affect the estimates of the risks. However, when we look at the risk ratios comparing the risks for MZ and DZ twins with an affected co-twin we see that each of the models produce estimates of the risk ratio that are above 1 indicating the risk is higher for MZ twins. We see however at the much later ages of the 5-year and 10-year risks (beyond age 90), the liability threshold falls below 1, but the confidence intervals are very wide. Additionally, the risk ratios for the 5-year risk at earlier ages and at later

ages are very unstable because of the minimal amount of cancer observations in those intervals.

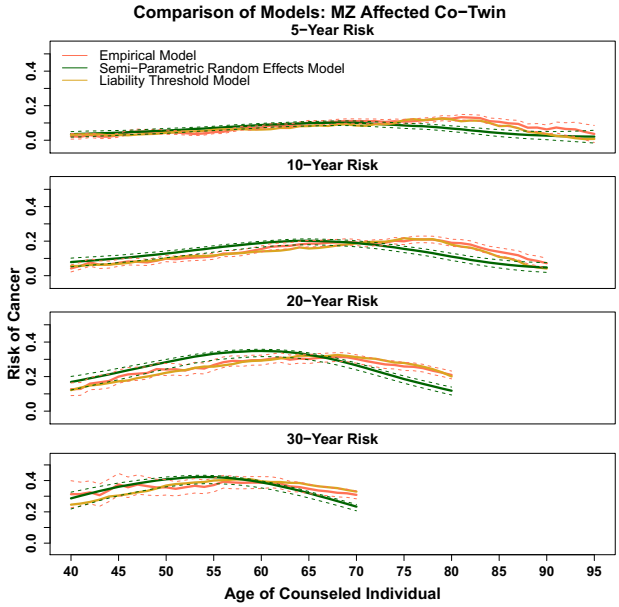


Figure 2.6: **Model Comparison of Absolute Risk Estimates.** Comparison of 5, 10, 20, and 30-year risk estimates for a twin with an affected MZ co-twin. The empirical risks are calculated using the using NorTwinCan data (pink), estimates the semi-parametric random effects model (green), and simulated data from the liability threshold model (orange). 95% confidence intervals are calculated using the bootstrap method.

2.5 Discussion

The purpose of this manuscript is three-fold. First, we provide 5, 10, 20, and 30-year estimates of the risk of cancer conditional on the knowledge of a co-twin's affection status. Next, we provide risk ratios that allow for calculations of the risks conditional on a co-twin's affection status in a population with a different baseline risk. Finally, we compare risk estimates calculated empirically, to estimates obtained using a semi-parametric random effects model and a liability threshold model.

When looking at the differences in risk for unaffected individuals whose twin has cancer, MZ twins have a higher risk compared to DZ twins, with risk ratios consistently above 1 across counseling ages. The risk ratios are of the order of 1.2 to 1.3 for most

counseling ages. This is consistent with the well-known heritability of cancer which suggests that MZ twins have a slightly higher risk of cancer than DZ twins. However, using the risks directly addresses counseling needs unlike heritability estimates. The risk ratio comparing the risk estimates for a twin with an unaffected co-twin to the unconditional risk estimates are very close to 1. This indicates that the risk of getting cancer for a twin that has an unaffected twin is practically equivalent to the risk when we do not know the co-twin's cancer status.

The unconditional risk estimate is calculated using all of the twins in the NorTwinCan cohort, which is a very unique population. This cohort covers birth years as early as 1895, while risks estimates in SEER include data only from the years 2011- 2013. When estimating the absolute risk in SEER, we multiplied risk ratio estimated from NorTwinCan to the baseline estimates in SEER. These risk estimates may be more comparable if we used cancer diagnoses during the same time period.

To assess the extent to which prospective risk estimates are affected by modeling assumptions, we used three different approaches to estimate the risks; an empirical estimate using the Aalen-Johansen estimator; a semi-parametric random effects model, and the liability threshold model. The model-based estimates are generally consistent with the empirical estimates. However, important differences exists: the random effects model appears to underestimate the risk at older ages. As mentioned in Scheike et al. (2014), different choices for the random effects distribution lead to different types of dependence. We chose to model the dependence using gamma distributions, however, further research needs to be conducted in order to measure the goodness of fit for the distributions used to estimate the dependence (Scheike et al., 2014).

Overall, the risk differences and risk ratios between MZ and DZ twins can serve as practical metrics of the degree to which genetic factors affect cancer onset. These differences are generally not very large. However, combining all cancers in the risk evaluation, as we did, can dampen the effect. It is known that certain cancers, such as prostate or breast cancer, have higher heritability than others (Mucci et al., 2016). A natural next step would be to conduct similar studies for separate cancers individually. These could possi-

bly reveal much higher risk ratios for cancers that are known to have strong heritability.

Acknowledgements

This work was supported by grants from the National Institutes of Health (ES007142, ES000002, ES016454, CA134294).

3. Exploring Cancer Resistance in Twins

Christina McIntosh[1,2], Danielle Braun[1,2], Lorelei Mucci[3], Giovanni Parmigiani [1,2]

[1]Department of Biostatistics, TH Chan Harvard School of Public Health
[2]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute
[3]Department of Epidemiology, TH Chan Harvard School of Public Health

3.1 Introduction

The NIH defines cancer susceptibility as "an increased risk of cancer due to genetic predisposition" (National Institutes of Health, 2017). Cancer resistance is commonly thought of as the lack of cancer susceptibility, or the lack of genetic mutations that contribute to a higher risk of cancer. However, Klein (2009, 2014) presents arguments about why cancer resistance should be considered as its own entity, including work in mice which identified multiple specific tumor resistance genes, leading to lower tumor incidence (Nagase et al., 1995; Manenti et al., 1996). Furthermore, a study conducted by Abegglen et al. (2015) explores the possibility of genetic adaptations that suppress cancer risk in elephants. The authors found that aging elephants seem to have low cancer rates and concluded that extra copies of the TP53 gene carried by all elephants may contribute to their cancer-free longevity. Klein (2009) discusses various mechanisms of resistance; the two most powerful mechanisms, he states, are genetic and intracellular. The genetic mechanisms ensure that DNA is replicated and repaired properly. The intracellular mechanisms can occur when apoptosis is triggered due to DNA damage or the improper activation of oncogenes. These mechanisms are hypothesized to be the most important because people with genetic defects in the genes involved in the relevant biological pathways have an increased frequency of tumors (Klein, 2009). It is becoming increasingly clear that cancer resistance can be due to different mechanisms. The work of Gorbunova and colleagues (Gorbunova et al., 2012) on the naked mole rat showed at least two different mechanism: an early contact inhibition, not seen in mice or humans, and a high MW hyaluronic acid (HA) that appears to have evolved to make the connective tissue malleable so that the animal can force its way through narrow tunnels. It prevents cancer transformation at the cellular level. On a totally different note, we have the Laron dwarfs that are missing a receptor for growth hormone and are cancer resistant (Jacobs et al., 1976). With this in mind, we define resistance as a decreased risk of cancer due to genetic predisposition.

Twin studies have been traditionally used to study the contributions of genes to can-

cer susceptibility. Comparisons of monozygotic (MZ) twins and dizygotic (DZ) twins allow one to discern the genetic and environmental contributions of cancer, because MZ twins share approximately 100 % of their genes and DZ twins share approximately 50 % of their genes. Typically, concordance risk estimates and estimates of heritability are used to quantify cancer susceptibility Mucci et al. (2016). Concordance risk for a specific cancer is the probability that a twin develops cancer conditional on his/her co-twin having that cancer. A higher concordance risk in MZ twins compared to DZ twins would indicate that there is a genetic component that contributes to that cancer. Concordance estimates can be used to estimate excess familial risk of cancer, for example as the ratio of the concordance risk to the cumulative incidence of cancer in the population.

Heritability is defined as the proportion of the variation of a trait, or the corresponding liability for binary and censored traits, that is due to genetic variation in the population Falconer (1965). Lichtenstein et al. (2000) provided estimates of concordance and heritability for specific cancers in twins in the Denmark, Sweden, and Finnish cancer registries. Recently, Mucci et al. (2016) provided updated estimates of concordance and heritability using the largest twin cohort in the world, the Nordic Twin Studies of Cancer (NorTwinCan). They estimated the familial risk for cancer overall for DZ twins as $37\%[36\%, 38\%]$ and for MZ twins as $46\%[44\%, 48\%]$. They estimated the heritability of cancer overall to be $33\%[30\%, 37\%]$.

While estimates of cancer susceptibility are fairly well established, discerning the genetic contribution to cancer resistance remains challenging. In the general population, cancer remains a relatively late onset event. Therefore, effects of genes on cancer resistance require analysis of individuals of high longevity. This reduces the sample sizes available, even in the largest cohorts. It also raises the question of how to discern genetic components that have a direct effects on cancer resistance from those which have effects on longevity and on competing causes of mortality (Christensen et al., 2012). It is reasonable to question whether it is possible at all to identify empirically any direct effect of genes on cancer resistance.

The goal of this paper is to lay the foundation for the quantitative investigation of

inherited cancer resistance in twin studies. We use the concept of latent genometype, as introduced in (Roberts et al., 2012). A genometype is a class of genomes. Individuals with genomes in this class have the same predispositions for the traits of interest. We consider the possible existence of a genometype that completely protects its carriers from cancer. We then ask two questions. First, we ask: if this genometype existed, what would the implication be for data on MZ and DZ twins? We explore this question conducting a simulation study to establish to what degree this genometype can be identified from twin registry data. Next, we propose a likelihood based approach using a mixture model to estimate the prevalence of this genometype using twin studies. We conduct simulations to evaluate the performance of this approach, and apply our approach in the NorTwinCan cohort.

3.2 Methods

3.2.1 NorTwinCan Cohort

The NorTwinCan cohort is the largest twin cancer registry in the world. It includes $357,377$ twin pairs from Denmark, Finland, Norway, and Sweden and is comprised of both MZ, same-sex DZ, and opposite-sex DZ twins. In our analysis we focus on the $202,868$ MZ and same-sex DZ twins where both twins were alive at the start of follow-up. We excluded twin pairs whose zygosity was unknown. Individuals in NorTwinCan were followed prospectively until: cancer diagnosis, death from other causes, emigration during follow-up, or end of study. The median years of follow-up in NorTwinCan is 32 years, and there were 27,156 incident cancers in the cohort . The dates of entry and follow-up are are presented by age in Table 3.1 (see Mucci et al. (2016) for further description of the cohort).

	Denmark	Finland	Norway	Sweden	Total
Birth Cohort	1870-2004	1875-1957	1915-1979	1886-2000	
End of Follow-up	2009	2010	2008	2009	
Number Twins	$68,248$	$24,438$	$23,572$	$86,610$	$202,868$
Number Female Twins	$33,339$	$12,507$	$12,824$	$45,869$	$104,539$
N Uncensored MZ/DZ pairs at Follow-up	$1,300/2,456$	$388/819$	$231/298$	$1,632/2,843$	$3,551/6,416$
Median Years of Follow-up	41.6 years	34.7 years	27.9 years	25.0 years	32.2 years
Number of Incident Cancers	$8,218$	$3,811$	$2,592$	$10,463$	$25,084$

Table 3.1: **Study population.** Characteristics of the Nordic Twin Studies of Cancer

3.2.2 Notation

For twin $k = 1, 2$ in the i^{th} twin pair, let Y_{ik} represent age of diagnosis, D_{ik} represent age of death, and C_{ik} be age at censoring. Age is measures in years, and is discrete. In our analyses we group all cancers together so that Y represents the earliest cancer experienced by an individual. The technical development also applies to alternative definitions of Y focusing on specific classes of cancer. Let $\omega_{ik} = 1$ if $Y_{ik} \leq D_{ik}$, and $\omega_{ik} = 2$ if $Y_{ik} > D_{ik}$. We define $T_{ik} = min(Y_{ik}, D_{ik}, C_{ik})$, and $\delta_{ik} = I(T_{ik} \leq C_{ik})$. Also, $\epsilon_{ik} = \omega_{ik}\delta_{ik}$ indicate the event type, such that $\epsilon_{ik} = 0$ corresponds to censoring, $\epsilon_{ik} = 1$ corresponds to cancer, and $\epsilon_{ik} = 2$ corresponds to death. For a twin pair, we observe $(T_{i1}, T_{i2}, \epsilon_{i1}, \epsilon_{i2})$.

We denote the joint survival function for the two twins in pair i as: $P(T_{i1} > t_{i1}, T_{i2} > t_{i2}) = S(t_{i1}, t_{i2})$, while $S(t_{ik})$ is the overall survival for twin k in the i^{th} pair. The joint survival function can be expressed as; $S(t_{i1}, t_{i2}) = S(t_{i1}|t_{i2})S(t_{i2})$. Let $S_j(t_{ik}) = P(T > t_{ik}|\epsilon_{ik} = j)$ indicate the conditional survival distribution for a failure of type $\epsilon_{ik} = j$. The probability of developing cancer at age t_{ik} can be represented by $f(t_{ik}, \epsilon_{ik} = 1) = P(t \leq T_{ik} < t + 1, Y_{ik} \leq D_{ik})$. Similarly, the probability of dying at age t_{ik} can be represented by $f(t_{ik}, \epsilon_{ik} = 2) = P(t \leq T_{ik} < t + 1, Y_{ik} > D_{ik})$.

3.2.3 Model Assumptions and Likelihood Function

We assume three genometypes in the population: a wild type, a susceptible type and a resistant type. Individuals with the resistant type are simply immune to cancer. Individuals with the susceptible type develop cancer at an age consistent with that of the major autosomal dominant syndromes such as the Lynch or the BRCA syndrome. Clearly, this

is a drastic oversimplification. As quantitative investigation of resistance is in its infancy, our goal is to consider the simplest model that can shed light on the question. Consideration of the susceptible genometype is critical as we need to distinguish resistance from the decreased susceptibility that occurs in an aging population from excluding susceptible individuals. In reality, a range of susceptibility genes of varying penetrance and prevalence exist and more advance model should consider this complexity.

Let γ indicate the probability of being a carrier for the resistance type and let β indicate the probability of being a carrier for a susceptibility type. Let $g_{ik} = 1$ if the person is a carrier of the resistance gene and $g_{ik} = 0$ if the person is not a carrier of the resistance gene; therefore, $\gamma = P(g_{ik} = 1)$. Similarly, let $b_{ik} = 1$ if the person is a carrier for a susceptibility gene and $b_{ik} = 0$ if the person is not a carrier for a susceptibility gene; hence, $\beta = P(b_{ik} = 1)$. We propose a likelihood approach to estimate β and γ.

The likelihood can be written as a mixture, with components defined by the genometypes. Assuming twin pairs are independent we can take the product over n twin pairs of the joint density of the observed events for each twin pair (t_{i1}, t_{i2}, ϵ_{i1}, ϵ_{i2}). The joint density is then factorized in Equation (3.2), where the choice of twin 1 and 2 is arbitrary.

$$L = \prod_{i=1}^{n} f(t_{i1}, t_{i2}, \epsilon_{i1}, \epsilon_{i2}) \tag{3.1}$$

$$= \prod_{i=1}^{n} f(t_{i1}, \epsilon_{i1} | t_{i2}, \epsilon_{i2}) f(t_{i2}, \epsilon_{i2}) \tag{3.2}$$

In the appendix, we provide an expansion of the likelihood, to incorporate censoring and factorize the likelihood into the different combinations of the observed event types (See Equation 3.2).

In the likelihood above, the joint probability of the carrier statuses for both twins can be written in terms of β and γ. For example, under the assumptions that MZ twins share 100% of their genes and that the probability of the carrier status for susceptibility is independent of the probability of the carrier status for resistance, the joint probability that both twins are not susceptible and both twins are resistant can be written as:

$$
\begin{aligned}
& P(b_{i1} = 0, b_{i2} = 0, g_{i1} = 1, g_{i2} = 1) \\
&= P(b_{i1} = 0, b_{i2} = 0) P(g_{i1} = 1, g_{i2} = 1) \\
&= P(b_{i1} = 0 | b_{i2} = 0) P(b_{i2} = 0) P(g_{i1} = 0 | g_{i2} = 0) P(g_{i2} = 0) \\
&= 1 \times (1 - \beta) \times 1 \times \gamma \\
&= (1 - \beta)\gamma
\end{aligned}
$$

The likelihood in Equation 3.2 is then maximized to obtain estimates $\hat{\gamma}, \hat{\beta}$. The joint probability of the carrier status for the twin pairs can vary for MZ and DZ twins. Also, the likelihood assumes independent censoring of the twins in a twin pair, but this assumption can be relaxed.

3.3 Simulations

3.3.1 Data Generation

We generate a twin cohort that mimics the NorTwinCan cohort. We generate $n = 101,434$ total twin pairs, $n_d = 61,674$ DZ twin pairs, and $n_m = 39,760$ MZ twin pairs, the same number of same-sex MZ and DZ twin pairs in NorTwinCan. For each individual twin, we first generate an independent censoring age, C_{ik} and independent death age, D_{ik}. The censoring age, C_{ik}, is generated by sampling from a density function representing the censoring distribution in the NorTwinCan cohort, which is estimated using Kaplan-Meier estimates for censoring in the cohort. Death ages are similarly generated by sampling from a density function which is estimated using Kaplan-Meier estimates for death in the cohort.

Next, we generate carrier status for the resistance genometype, g_{ik} and the susceptibility genometype, b_{ik} for each individual twin. For each twin pair, we begin by generating the resistance and susceptibility carrier status for the first twin in a twin pair (where the ordering of the twins is arbitrary). The carrier statuses are generated from a Bernoulli distribution with γ for resistance and β for susceptibility. For different simulation scenarios, we vary the value of γ and β. Once the carrier status for the first twin is sampled, we sample the carrier status for the second twin by assuming that MZ twins share 100% of

their genes and DZ twins share 50% of their genes. Therefore, if the first twin in a MZ twin pair is a carrier (non-carrier) then the second twin is also a carrier (non-carrier). If the first twin in a DZ twin pair is carrier then the second twin is a carrier with 50% probability. If the first twin in a DZ twin pair is a non-carrier, then the second twin is a carrier with probability $\frac{0.5\gamma}{1-\gamma}$ for resistance. (We calculate the probability that the second twin is a non-carrier given the first twin is a carrier by solving for $P(g_{i2} = 0|g_{i1} = 1)$ in the equation $P(g_{i2} = 1) = P(g_{i2} = 1|g_{i1} = 1)P(g_{i1} = 1) + P(g_{i2} = 1|g_{i1} = 0)P(g_{i1} = 0)$. Substituting γ into the equation, $\gamma = 0.5\gamma + P(g_{i2} = 1|g_{i1} = 0)(1 - \gamma)$ and $P(g_{i2} = 1|g_{i1} = 0) = \frac{0.5\gamma}{1-\gamma}$).

We begin by generating the age of cancer Y_{i1} for the first twin conditional on their carrier status. We assume that the resistance genometype is fully penetrant. Therefore, carriers of the resistance genometype, $g_{i1} = 1$, will not get cancer. If an individual is not resistant but is susceptible, $g_{i1} = 0$ and $b_{i1} = 1$, the age of cancer, Y_{i1}, is generated by sampling from the BRCA2 penetrance for breast cancer using the BayesMendel package in R Chen et al. (2004). If an individual is not a carrier for either genes, $g_{i1} = 0$ and $b_{i1} = 0$, the age of cancer, Y_{i1}, is generated from the distribution of ages of cancer, $f(y_{i1}|g_{i1} = 0, b_{i1} = 0)$, estimated from NorTwinCan as follows. Since $g_{i1} = 0$ is not observed, we assume $f(Y_{i1} = y_{i1}) = f(Y_{i1} = y_{i1}|g_{i1} = 0)$ which is the rare gene assumption. Let $f(Y_{i1} = y_{i1}|g_{i1} = 0)$ indicate the probability distribution of twin 1 having cancer at age y_{i1} given that they are not a carrier for the resistance genometype.

$$f(Y_{i1} = y_{i1}|g_{i1} = 0) = \beta f(y_{i1}|g_{i1} = 0, b_{i1} = 1) + (1 - \beta)f(y_{i1}|g_{i1} = 0, b_{i1} = 0).$$

Therefore,

$$f(y_{i1}|g_{i1} = 0, b_{i1} = 0) = \frac{f(Y_{i1} = y_{i1}|g_{i1} = 0) - \beta f(y_{i1}|g_{i1} = 0, b_{i1} = 1)}{(1 - \beta)},$$

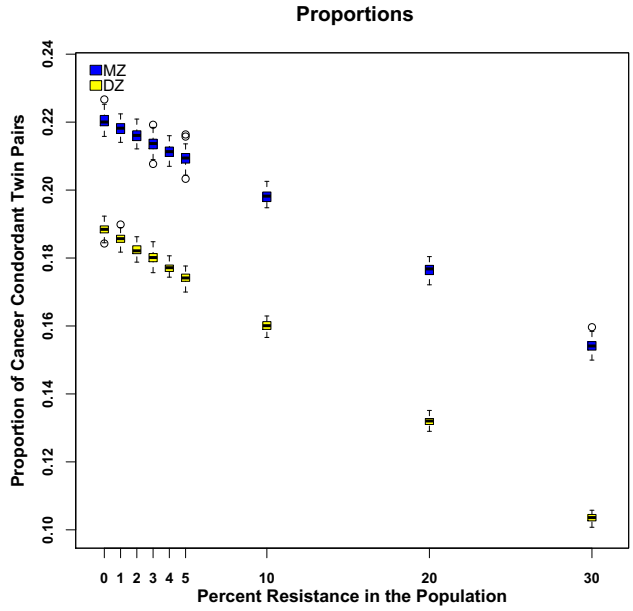
where we assume $f(y_{i1}|g_{i1} = 0, b_{i1} = 1)$ is the distribution for an individual who has the susceptibility genometype, and is the BRCA2 penetrances for breast cancer, and $f(Y_{i1} = y_{i1}|g_{i1} = 0)$ is estimated in NorTwinCan.

For twin pairs where neither twin is a carrier for the resistance genometype (and are either carriers or non-carriers of the susceptibility genometype), the age of cancer for twin

2 is generated from a normal distribution with a mean that is equal to the age of cancer for twin 1 and standard deviation that is estimated from NorTwinCan. The standard deviation for MZ twins and DZ twins is estimated separately using twin pairs where both twins are concordant for having cancer. For twins who are carriers of the resistance genometype, Y_{i2} is set to 1000. We generate 100 replications of the data.

3.3.2 Effects of Resistance on Disease Concordance

Our first aim in these simulations is to study the to what degree cancer resistance can be estimated using twin cohort data. To answer this question, we conduct simulations with varying γ, while fixing $\beta = 10\%$. We evaluate differences in proportions of concordance twins for MZ and DZ twins, and how these differences vary across varying γ. We first conduct simulations assuming no censoring (Figure A.7). On the left, we present the proportion of MZ and DZ twin pairs who are both concordant for cancer, for varying levels of resistance. On the right, we present the difference in these proportions for MZ and DZ twins. We observe that as resistance increases, as expected, the proportion of both MZ and DZ twins where both get cancer decreases. The proportion of MZ twins is consistently greater than the proportion of DZ twins even as the amount of resistance increases. Importantly, the difference between the proportion of MZ twin pairs and DZ twin pairs increases as resistance increases. In fact, the median difference of the proportion of cancer concordant twins between MZ and DZ twins is 0.032 when there is no resistance, and 0.051 when there is 30% resistance. Results on proportion of twin pairs who both die cancer-free are shown in Appendix Figure A.9. As expected, as the level of resistance increases, the proportion of MZ and DZ twin pairs that both die cancer-free increases. We also see that the difference in the proportions increase as resistance increases.



(a)

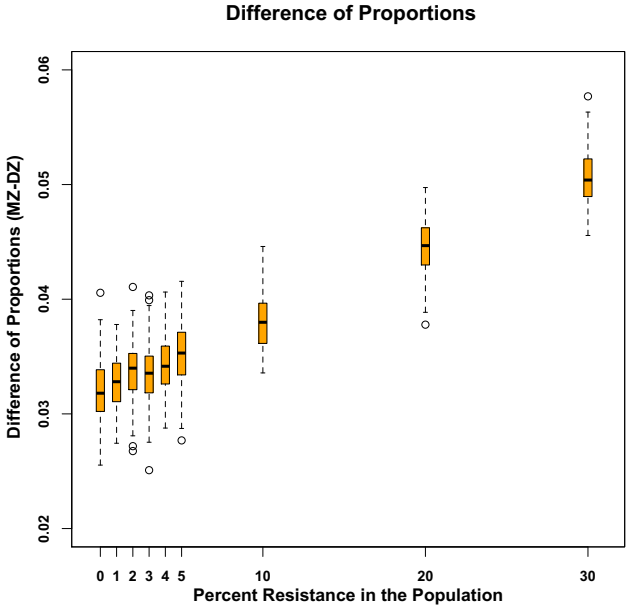
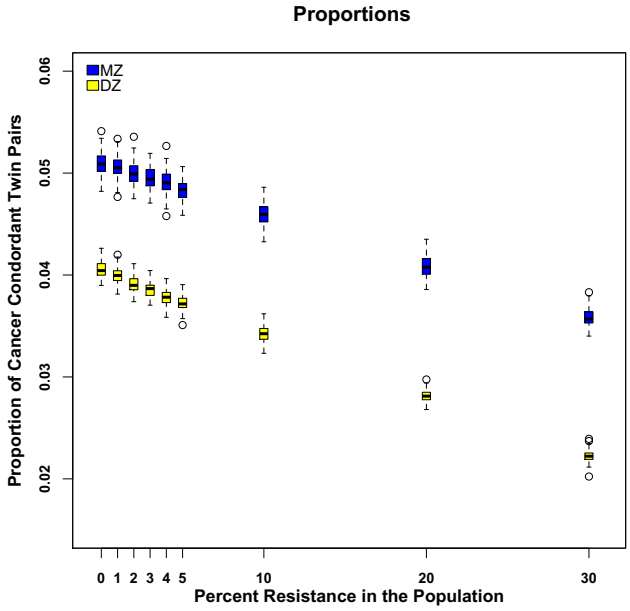
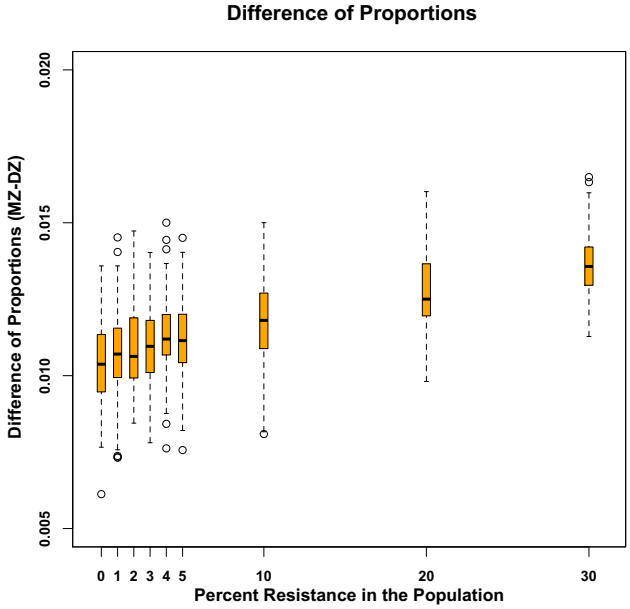


Figure 3.1: **Concordant cancer twin pairs with no censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with no censoring and 10% susceptibility. Proportion of cancer concordant twin pairs as we vary levels of resistance in the population (left). Differences in proportions of cancer concordant MZ and DZ twins as we vary levels of resistance (right). Results are based on 100 data generations.

We repeat these simulations after we introduce censoring in our simulations, which is a more realistic representation of twin studies. Because censoring is estimated to be around 70% in NorTwinCan, we induce 70% censoring in our simulations (Figure 2.2). As we saw with no censoring, in both MZ and DZ twins, the proportion of twin pairs where both are cancer concordant decreases as resistance increases. However, these differences are far smaller and are subject to greater variability. The median difference in proportions for cancer concordant twin pairs is 0.011 with no resistance and 0.013 with 30% resistance. These results indicate that in the presence of censoring it would be difficult to identify resistance by looking at the differences in concordance proportions in MZ and DZ twin pairs, as these differences are small compared to their variability, even in a study of the size of the largest twin cohort.



(a)



(b)

Figure 3.2: **Concordant cancer twin pairs with** 70% **censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with no censoring and 10% susceptibility. Proportion of cancer concordant twin pairs as we vary levels of resistance in the population (left). Differences in proportions of cancer concordant MZ and DZ twins as we vary levels of resistance (right). Results are based on 100 data generations.

We also assess differences of proportions of concordant MZ and DZ twin pairs by age. Figure 3.3 presents results without censoring while Figure 3.4 considers 70% censoring. When there is no censoring, at ages 50 and 60, the differences in proportions of MZ and DZ twin pairs which both have cancer is virtually the same across all levels of resistance. At ages 70 and 80, the differences increase as resistances increases, but these differences are not reliable as all the boxplots overlap across varying resistance levels. In Figure 3.4, we see that censoring greatly attenuates the differences at all ages and the estimates are very similar across varying resistance levels. In the presence of censoring at both low and high resistance levels, it is difficult to detect a difference in the proportion of cancer concordant MZ and DZ twin pairs. Results looking at proportion of twin pairs who both die cancer-free are shown in Appendix Figures 2 and 3. These results confirm that the differences in proportions are difficult to detect especially in the presence of censoring.

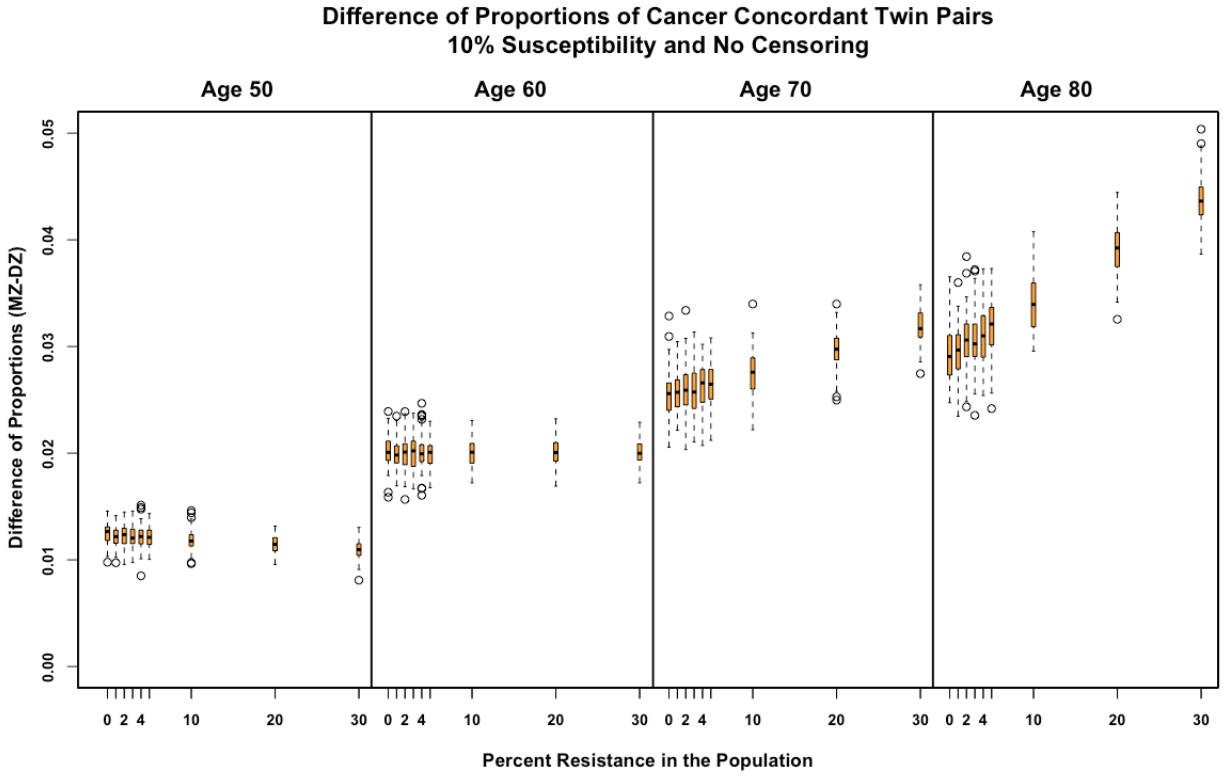


Figure 3.3: **Concordant cancer twin pairs by age with no censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with no censoring and 10% susceptibility. Differences in proportions of cancer concordant MZ and DZ twins as we vary levels of resistance, across different age groups (50, 60, 70, 80). Results are based on 100 data generations.

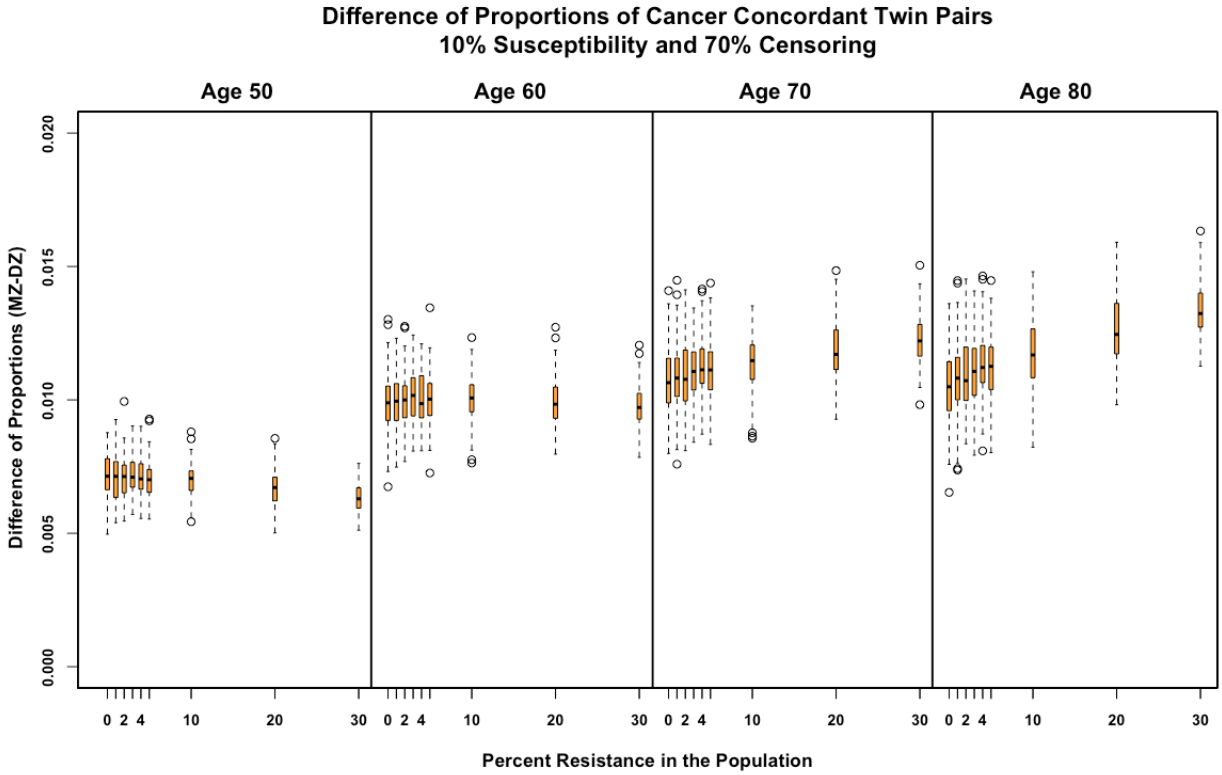


Figure 3.4: **Concordant cancer twin pairs by age with 70% censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with 70% censoring and 10% susceptibility. Differences in proportions of cancer concordant MZ and DZ twins as we vary levels of resistance, across different age groups (50, 60, 70, 80). Results are based on 100 data generations.

3.3.3 Likelihood Estimation

Next, we apply the proposed likelihood approach described in Section 3.2.3 to simulations, to evaluate its performance. The likelihood approach has the potential for greater power compared to simple comparisons of proportions, as long as the model it assumes is not far from reality. For the simulations, we fix β at 10%, and vary γ. When maximizing the likelihood in Equation 3.3, we make the following assumptions; 1) independent censoring, 2) being a carrier for resistance is fully penetrant, so individuals with cancer, $\epsilon_{ik} = 1$ are not carriers for the resistance genometype, 3) being a carrier for resistance is independent of being a carrier for susceptibility 4) the ages of death are independent. These are the same assumptions we make when generating the data. We maximize the likelihood to obtain maximum likelihood estimates for $\hat{\gamma}, \hat{\beta}$. In Figure 3.5 we present es-

timates $\hat{\gamma}, \hat{\beta}$ compared to the true γ, β used to generate the data with 70% censoring. The likelihood estimates are very close to the true estimates, indicating we are able to recover the true data generating parameters.

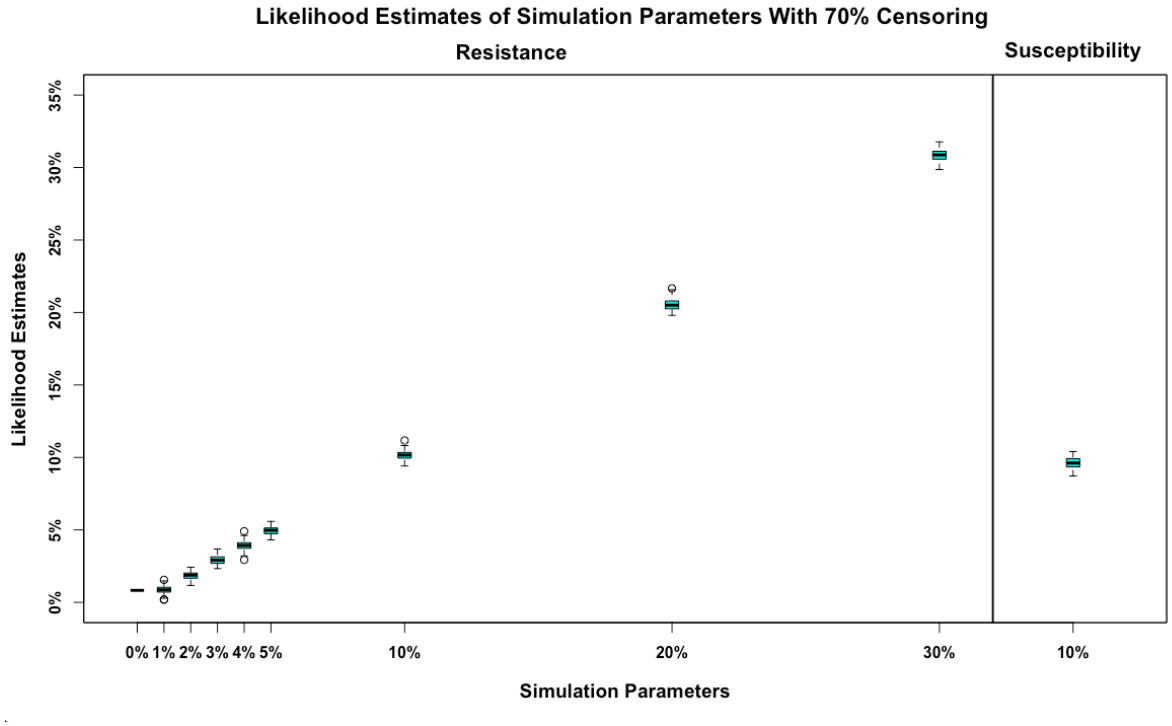


Figure 3.5: **Likelihood Based Approach**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with 70% censoring and 10% susceptibility. We compare true parameters used to generate the data (actual estimates), to those estimated using our proposed likelihood based approach.

3.4 Results in NorTwinCan

Lastly, we apply the likelihood based approach of Section 3.2.3 to the NorTwinCan cohort. When maximizing the likelihood we make the same assumptions as we do when applying this proposed approach the assumptions described in Section 3.3.3. We estimated the confidence intervals for β and γ by bootstrapping NorTwinCan data 100 times. We obtain the following estimates; $\hat{\beta} = 0.0901[0.089, 0.0931]$ and $\hat{\gamma} = 0.017[0.012, 0.0214]$.

Both our simulations and data application are implemented using strong assumptions. Technically, our approach can be extended to relax some of these assumptions. In

practice, we explores sensitivity to some of our modeling choices, and found that estimates of γ and β in the NorTwinCan cohort are sensitive to the distributions chosen for cancer age of onset of both susceptible and wild type individuals. For example, when we used estimates from the Surveillance, Epidemiology and End Results program (Howlader et al., 2013), γ and β were estimated to be larger $\beta = 0.11$ and $\gamma = 0.03$. The SEER distribution of cancer is higher than that NorTwinCan, a fact which may be responsible for the increased estimate of susceptibility.

3.5 Discussion

Exploring cancer resistance as its own entity is challenging. We defined cancer resistance as a counterpart of cancer susceptibility, that is a genometype that protects a person from cancer. Using this definition, our simulations show that, as expected, the percent of resistance influences the proportion of concordant MZ and DZ twin pairs. However, these differences are small, and in the presence of censoring, hard to detect because of high variability, even when resistance is as common as 30%. There is 70% censoring in NorTwinCan which would make differences of proportions between MZ and DZ twins very difficult to detect among varying levels of resistance. Given these results, it is not surprising that pervious studies have not been able to detect or quantify resistance.

We propose a novel approach to estimate cancer resistance and show that this approach performs well in simulations. In our simulations and data application we make the following assumptions; 1) independent censoring, 2) being a carrier for resistance is fully penetrant, so individuals with cancer, $\epsilon_{ik} = 1$ are not carriers for the resistance genometype, 3) being a carrier for resistance is independent of being a carrier for susceptibility, 4) the ages of death are independent for each twins, 5) the competing risks of death and cancer and censoring are independent.

Under the assumption of a fully penetrant resistance genometype, the estimate of resistance is (1.7%). This is likely to be a lower bound on the proportion of people whose genomes offer some overall resistance. For example in a scenario where the resistance

genometype is not fully penetrant, then the estimate of resistance would likely to increase because we would potentially capture a greater percentage of carriers who have a decreased risk of cancer as opposed to an absolute chance of not getting cancer.

Including different penetrances for different cancers will also lead to more accurate estimates of susceptibility and resistance, and could be the subject of future work. In addition, our proposed approach is flexible and would allow for the inclusion of other diseases into the likelihood (for example, Alzhiemer's or cardiovascular disease), if one believed risk of these diseases is related to a cancer resistant genometype.

We estimate that one to two percent of individuals of Scandinavian origin may carry genomes that make them virtually resistant to cancer, at least during the course of a lifetime. While this is likely a lower bound for inherited resistance as a whole, we are able to quantify this proportion for the first time. We hope that our results will motivate future research in cancer resistance, eventually leading to the exploration of mechanisms of resistance and drug therapies.

Acknowledgements

This work was supported by grants from the National Institutes of Health (ES007142, ES000002, ES016454, CA134294). We would like to acknowledge the Nordic Twin Studies of Cancer group and the NIH, Ford foundation, and GEM fellowships for funding.

Bibliography

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, pages 141–150.

Abegglen, L. M., Caulin, A. F., Chan, A., Lee, K., Robinson, R., Campbell, M. S., Kiso, W. K., Schmitt, D. L., Waddell, P. J., Bhaskara, S., et al. (2015). Potential mechanisms for cancer resistance in elephants and comparative cellular response to dna damage in humans. *JAMA*, 314(17):1850–1860.

Chen, S., Wang, W., Broman, K. W., Katki, H. A., and Parmigiani, G. (2004). Bayesmendel: an r environment for mendelian risk prediction. *Statistical applications in genetics and molecular biology*, 3(1):1–19.

Christensen, K., Pedersen, J. K., Hjelmborg, J. v. B., Vaupel, J. W., Stevnsner, T., Holm, N. V., and Skytthe, A. (2012). Cancer and longevity?is there a trade-off? a study of cooccurrence in danish twin pairs born 1900–1918. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 67(5):489–494.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76.

Ferreira, M. A., Sham, P., Daly, M. J., and Purcell, S. (2007). Ascertainment through family history of disease often decreases the power of family-based association studies. *Behavior genetics*, 37(4):631–636.

Goate, A., Chartier-Harlin, M.-C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra,

L., Haynes, A., Irving, N., James, L., et al. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial alzheimer's disease. *Nature*, 349(6311):704–706.

Gorbunova, V., Hine, C., Tian, X., Ablaeva, J., Gudkov, A. V., Nevo, E., and Seluanov, A. (2012). Cancer resistance in the blind mole rat is mediated by concerted necrotic cell death mechanism. *Proceedings of the National Academy of Sciences*, 109(47):19392–19396.

Holst, K. K., Scheike, T. H., and Hjelmborg, J. B. (2016). The liability threshold model for censored twin data. *Computational Statistics & Data Analysis*, 93:324–335.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Neyman, N., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., et al. (2013). Seer cancer statistics review, 1975-2010.[based on the november 2012 seer data submission, posted to the seer web site, april 2013.]. *Bethesda, MD: National Cancer Institute*.

Jacobs, L., Sneid, D., Garland, J., Laron, Z., and Daughaday, W. (1976). Receptor-active growth hormone in laron dwarfism. *The Journal of Clinical Endocrinology & Metabolism*, 42(2):403–406.

Kempthorne, O. and Osborne, R. H. (1961). The interpretation of twin data. *American journal of human genetics*, 13(3):320.

Klein, G. (2009). Toward a genetics of cancer resistance. *Proceedings of the National Academy of Sciences*, 106(3):859–863.

Klein, G. (2014). Evolutionary aspects of cancer resistance. *Seminars in cancer biology*, 25:10–14.

Knapp, M. (1999). A note on power approximations for the transmission/disequilibrium test. *The American Journal of Human Genetics*, 64(4):1177–1185.

Laird, N. M., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genetic epidemiology*, 19(S1):S36–S42.

Laird, N. M. and Lange, C. (2010). *The fundamentals of modern statistical genetics*. Springer Science & Business Media.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal Medicine*, 343(2):78–85.

Manenti, G., Gariboldi, M., Elango, R., Fiorino, A., De Gregorio, L., Falvella, F. S., Hunter, K., Housman, D., Pierotti, M. A., and Dragani, T. A. (1996). Genetic mapping of a pulmonary adenoma resistance locus (par1) in mouse. *Nature genetics*, 12(4):455–457.

Monaco, J., Cai, J., and Grizzle, J. (2005). Bootstrap analysis of multivariate failure time data. *Statistics in medicine*, 24(22):3387–3400.

Mucci, L. A., Hjelmborg, J. B., Harris, J. R., Czene, K., Havelick, D. J., Scheike, T., Graff, R. E., Holst, K., Möller, S., Unger, R. H., McIntosh, C., Nuttall, E., Brandt, I., Penney, K. L., Hartman, M., Kraft, P., Parmigiani, G., Christensen, K., Koskenvuo, M., Holm, N. V., Heikkilä, K., Pukkala, E., Skytthe, A., Adami, H.-O., Kaprio, J., and Nordic Twin Study of Cancer (NorTwinCan) Collaboration (2016). Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA : the journal of the American Medical Association*, 315(1):68–76.

Nagase, H., Bryson, S., Cordell, H., Kemp, C. J., Fee, F., and Balmain, A. (1995). Distinct genetic loci control development of benign and malignant skin tumours in mice. *Nature genetics*, 10(4):424–429.

National Institutes of Health (2017). Genetics home reference. `https://ghr.nlm.nih.gov/primer/mutationsanddisorders/predisposition`.

Rabinowitz, D. and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human heredity*, 50(4):211–223.

Risch, N., Merikangas, K., et al. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

Roberts, N. J., Vogelstein, J. T., Parmigiani, G., Kinzler, K. W., Vogelstein, B., and Velculescu, V. E. (2012). The predictive capacity of personal genome sequencing. *Science translational medicine*, 4(133):133ra58–133ra58.

Scheike, T. H., Holst, K. K., and Hjelmborg, J. B. (2014). Estimating heritability for cause specific mortality based on twin studies. *Lifetime data analysis*, 20(2):210–233.

Scheike, T. H., Sun, Y., Zhang, M. J., and Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks data. *Biometrika*, 97(1):133–145.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215.

A.1 Cancer Risk Assessment in Twins

A.1.1 Additional Analyses

We show the absolute cancer risks for a DZ twin with an affected co-twin and θ_k in the 5th, 25th, 50th, 60th, 75th, and 95th percentile. The same estimates can be calculated for MZ twins (not shown). Higher values of θ_k indicate increased association between the failure times of twins in the same cluster. We can see that as the percentile increases, the conditional risks increase as well. Additional information on specific exposures or behaviors could potentially lead to more accurate risk prediction by capturing this additional variability.

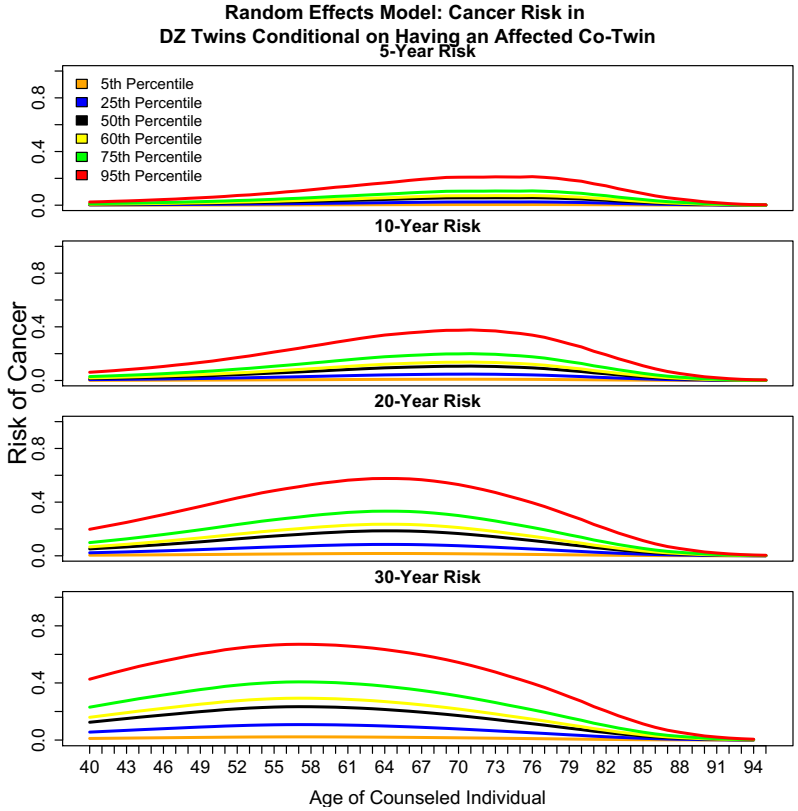


Figure A.1: **Semi-parametric Random Effects Model: Risk Varies by Percentile of Random Effect.** The graph shows the 5th, 25th, 50th, 60th, 75th, and 95th percentiles of the 5-year, 10-year, 20-year, and 30-year risk of having cancer conditional on the DZ co-twin (MZ not shown) having cancer for data using the semi-parametric random effects model. The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin.

Using the liability threshold model, we can calculate the case-wise concordance for a fixed age, a. This is the probability that twin 2 has cancer before age a given that the other twin had cancer before age a,

$$\frac{P(T_1 \leq a, \epsilon_1 = 1, T_2 \leq a, \epsilon_2 = 1)}{P(T_1 \leq a)}.$$

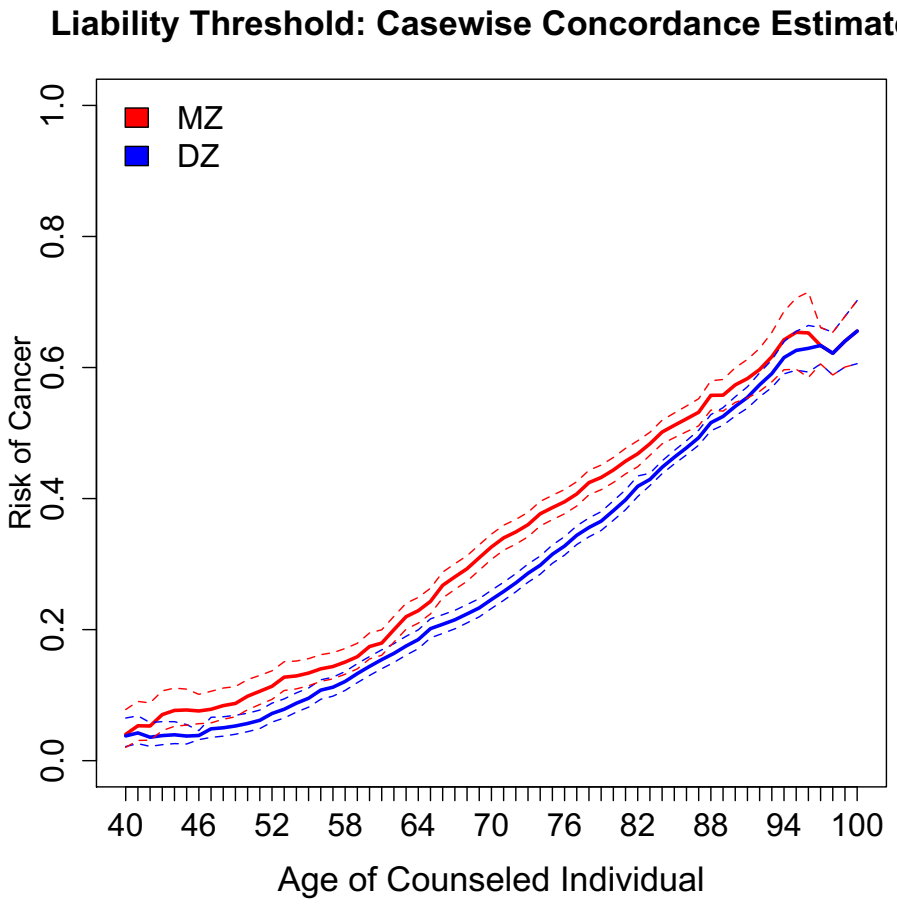


Figure A.2: **Liability Threshold Model: Casewise Concordance Estimates.** The graph shows the casewise concordance estimates of MZ and DZ twins estimated from the liability threshold model. The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the casewise concordance. The red represents the MZ twin and the blue represents the DZ twin. The dotted lines are the 95% confidence intervals.

In the figures below, we present empirical risk estimates for cancer in twins whose co-twin was affected in the past 5 (Figure A.3), 10 (Figure A.4) and 20 years (Figure A.5).

It looks as if the estimates are comparable to having a co-twin that is affected with cancer at any time in the past.

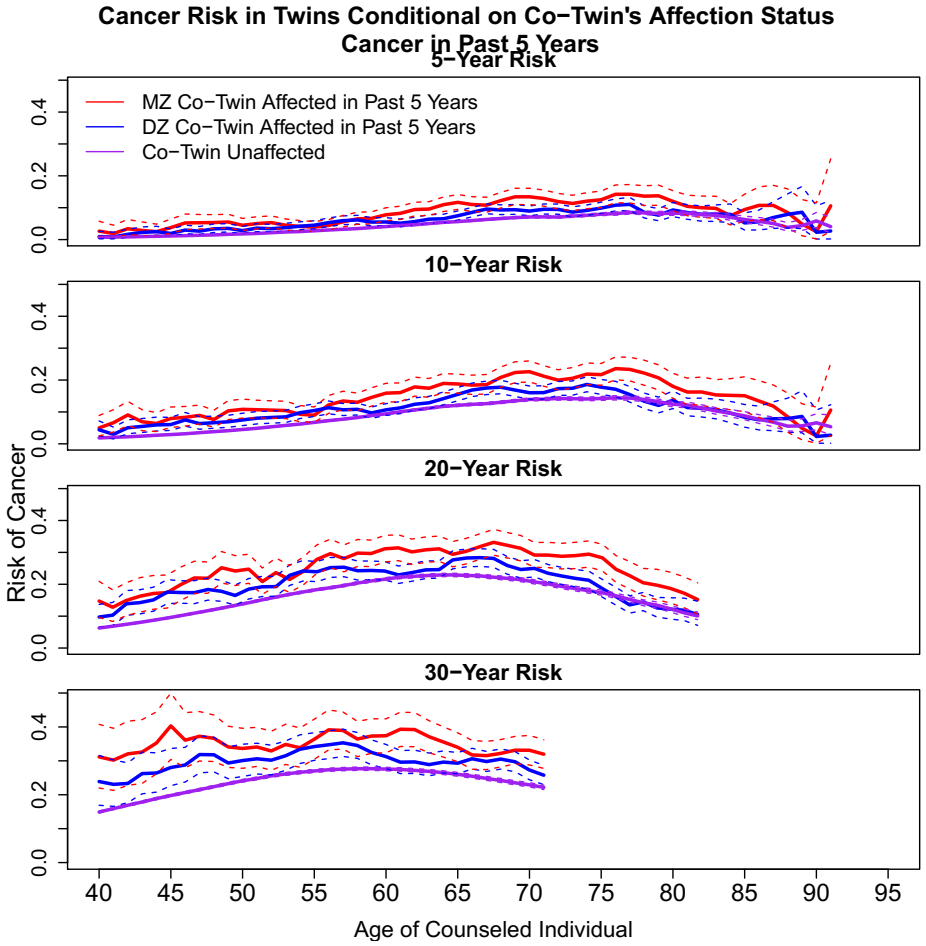


Figure A.3: **Empirical Absolute Risks for Twin with Co-Twin Affected in Past 5 Years.** The graph shows the 5-year, 10-year, 20-year, and 30-year risk of having cancer conditional on whether the co-twin has cancer in the past 5 years or the co-twin doesn't have cancer. The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin. The risk is estimated empirically using the cumulative incidence curve. The red represents the MZ twin with an affected co-twin and the blue represents the DZ twin with an affected co-twin. The purple is representative of the risk conditional on having an unaffected co-twin. The 95% confidence intervals are calculated using the bootstrap method.

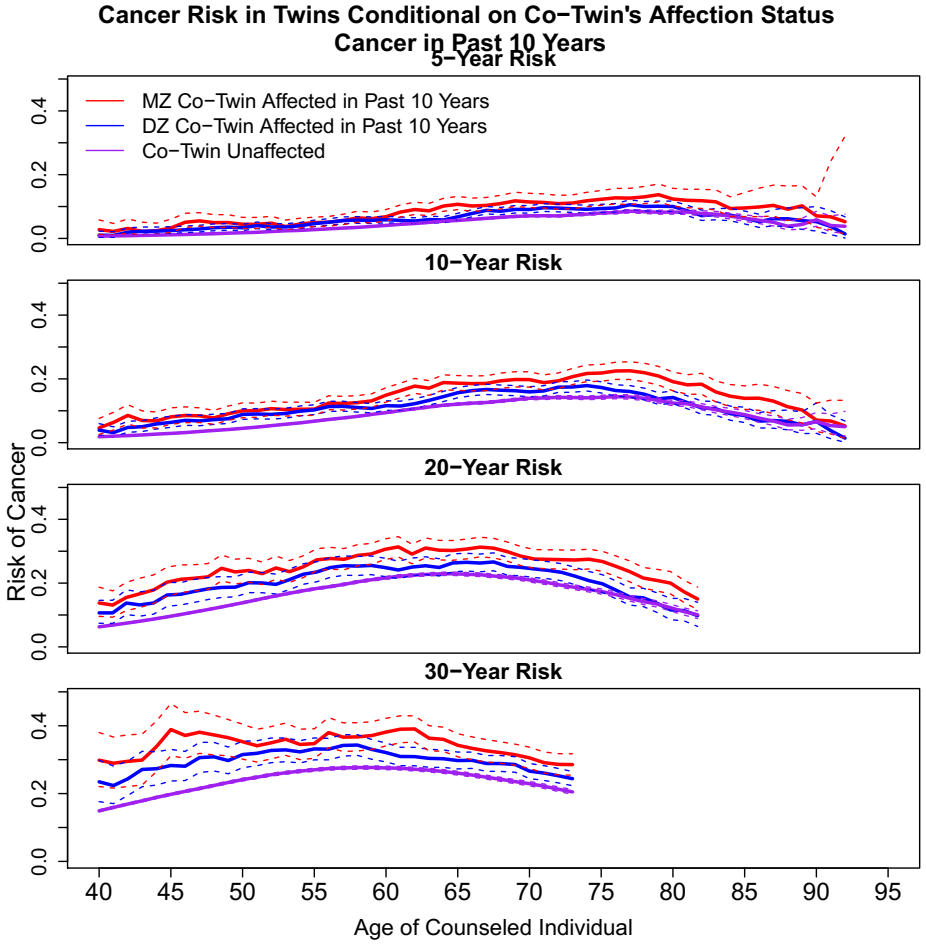


Figure A.4: **Empirical Absolute Risks for Twin with Co-Twin Affected in Past 10 Years.** The graph shows the 5-year, 10-year, 20-year, and 30-year risk of having cancer conditional on whether the co-twin has cancer in the past 10 years or the co-twin doesn't have cancer. The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin. The risk is estimated empirically using the cumulative incidence curve. The red represents the MZ twin with an affected co-twin and the blue represents the DZ twin with an affected co-twin. The purple is representative of the risk conditional on having an unaffected co-twin. The 95% confidence intervals are calculated using the bootstrap method.

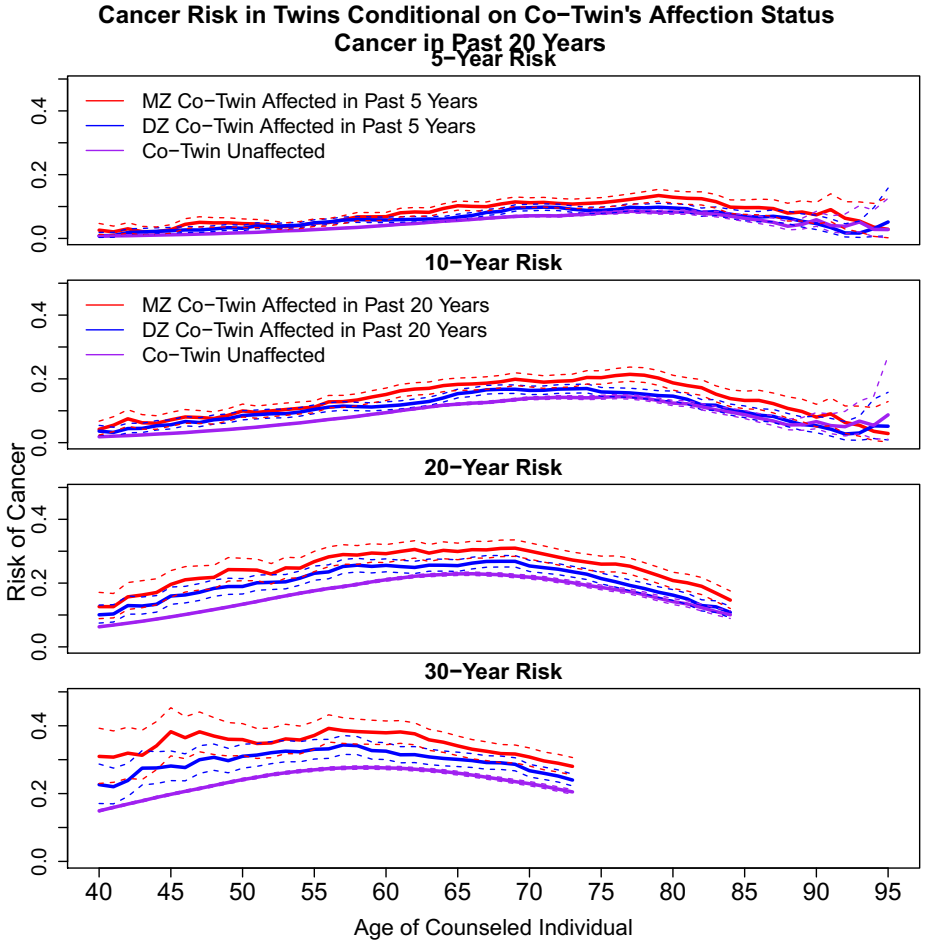
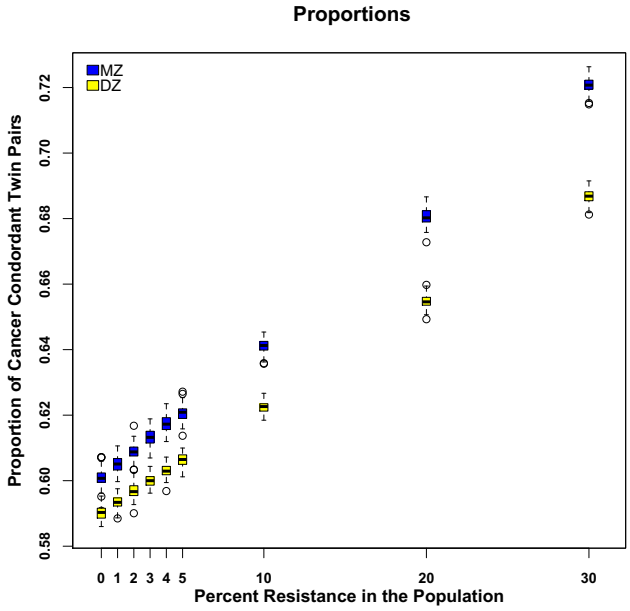


Figure A.5: **Empirical Absolute Risks for Twin with Co-Twin Affected in Past 20 Years.** The graph shows the 5-year, 10-year, 20-year, and 30-year risk of having cancer conditional on whether the co-twin has cancer in the past 20 years or the co-twin doesn't have cancer. The x-axis is the age of the unaffected twin that is being counseled. The y-axis represents the risk of cancer for the unaffected twin. The risk is estimated empirically using the cumulative incidence curve. The red represents the MZ twin with an affected co-twin and the blue represents the DZ twin with an affected co-twin. The purple is representative of the risk conditional on having an unaffected co-twin. The 95% confidence intervals are calculated using the bootstrap method.

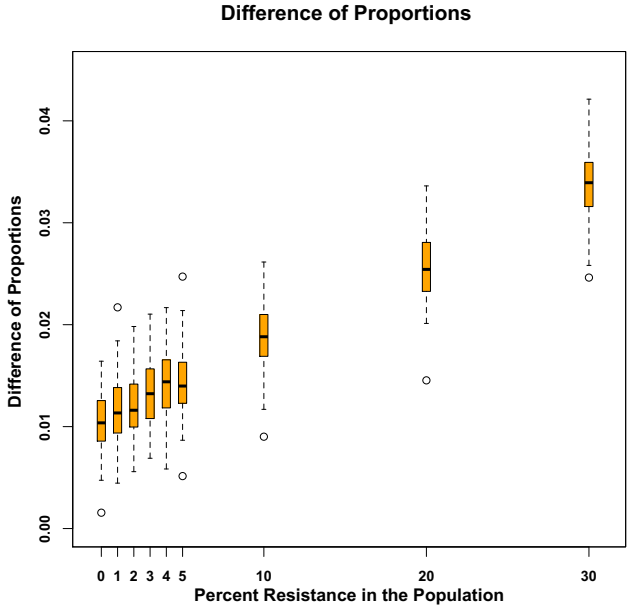
A.2 Cancer Resistance Using Twin Studies

A.2.1 Additional Simulation Results

We present additional simulation results referenced in Section 3.3.2. In Figure A.6a, we observe that as the percent of resistance increases with no censoring present, the proportion of both MZ and DZ twin pairs where both twins die free of cancer increases. Figure A.6b shows the difference in the proportions of MZ and DZ twin pairs where both twins die free of cancer. The median difference for a population simulated with no resistance is 0.01 while the median difference for a population simulated with 30% resistance is approximately 0.025.



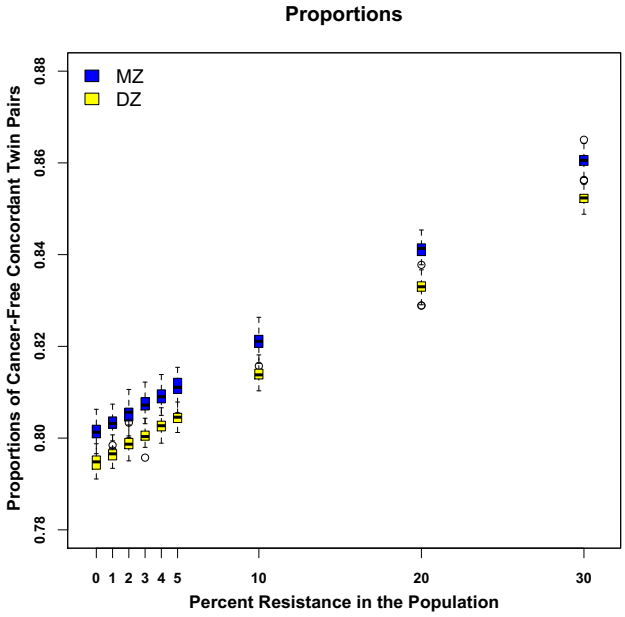
(a)



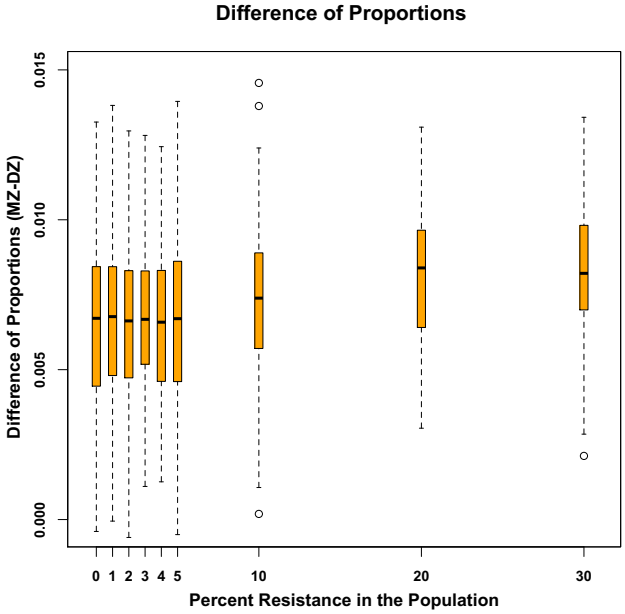
(b)

Figure A.6: **Concordant cancer-free twin pairs with no censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with no censoring and 10% susceptibility. Proportion of cancer-free concordant twin pairs as we vary levels of resistance in the population (left). Differences in proportions of cancer concordant MZ and DZ twins as we vary levels of resistance (right). Results are based on 100 data generations.

In Figure A.7a we present the results for concordant twin pairs where both twins die free of cancer with 70% censoring in Figure A.7a. We must use caution when interpreting the concordance for cancer-free twin pairs because we are not including twin pairs where twins were censored. We only consider those twins where both are observed to have died without obtaining cancer. We see that in the presence of censoring, we observe the same patterns of concordance in MZ and DZ twins where the concordance of cancer free twin pairs increases as resistance increases. However, the difference in the proportions for MZ and DZ twins are very small across the varying levels of resistance, and the boxplots from the 100 replications of the data overlap.



(a)



(b)

Figure A.7: **Concordant cancer-free twin pairs with 70% censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with 70% censoring and 10% susceptibility. Proportion of cancer-free concordant twin pairs as we vary levels of resistance in the population (left). Differences in proportions of cancer concordant MZ and DZ twins as we vary levels of resistance (right). Results are based on 100 data generations.

We then look at the difference in proportions for MZ and DZ twin pairs where both die cancer-free by age for simulations with no censoring (Figure A.8) and with 70% censoring (Figure A.9). We find that with censoring, across all ages, the differences between the proportions of MZ and DZ twins are virtually indistinguishable. Even without censoring, it is still difficult to determine a difference between MZ and DZ concordant cancer-free deaths for varying levels of censoring. This may be because most people die cancer free whether or not they are resistant to cancer.

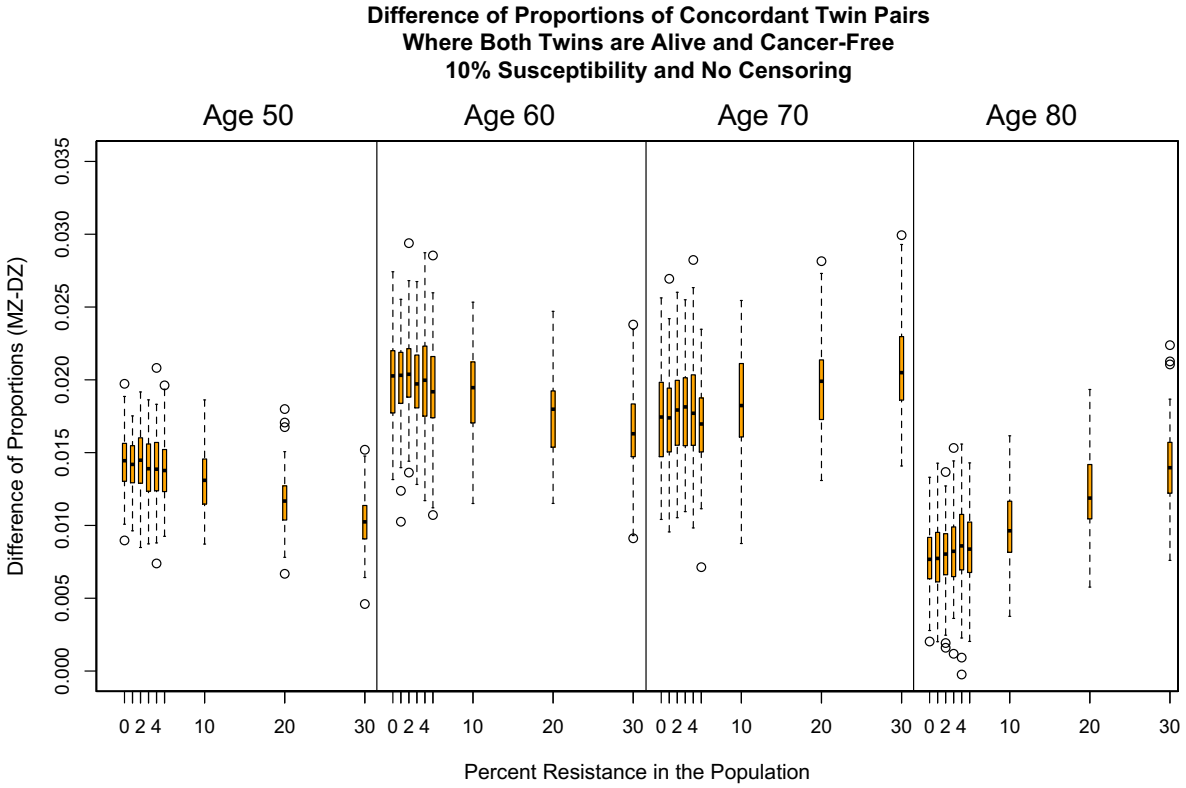


Figure A.8: **Concordant twin pairs where both twins are alive and cancer-free by age with no censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with no censoring and 10% susceptibility. Differences in proportions of concordant MZ and DZ twin pairs that are alive and cancer-free, as we vary levels of resistance, across different age groups (50, 60, 70, 80). Results are based on 100 data generations.

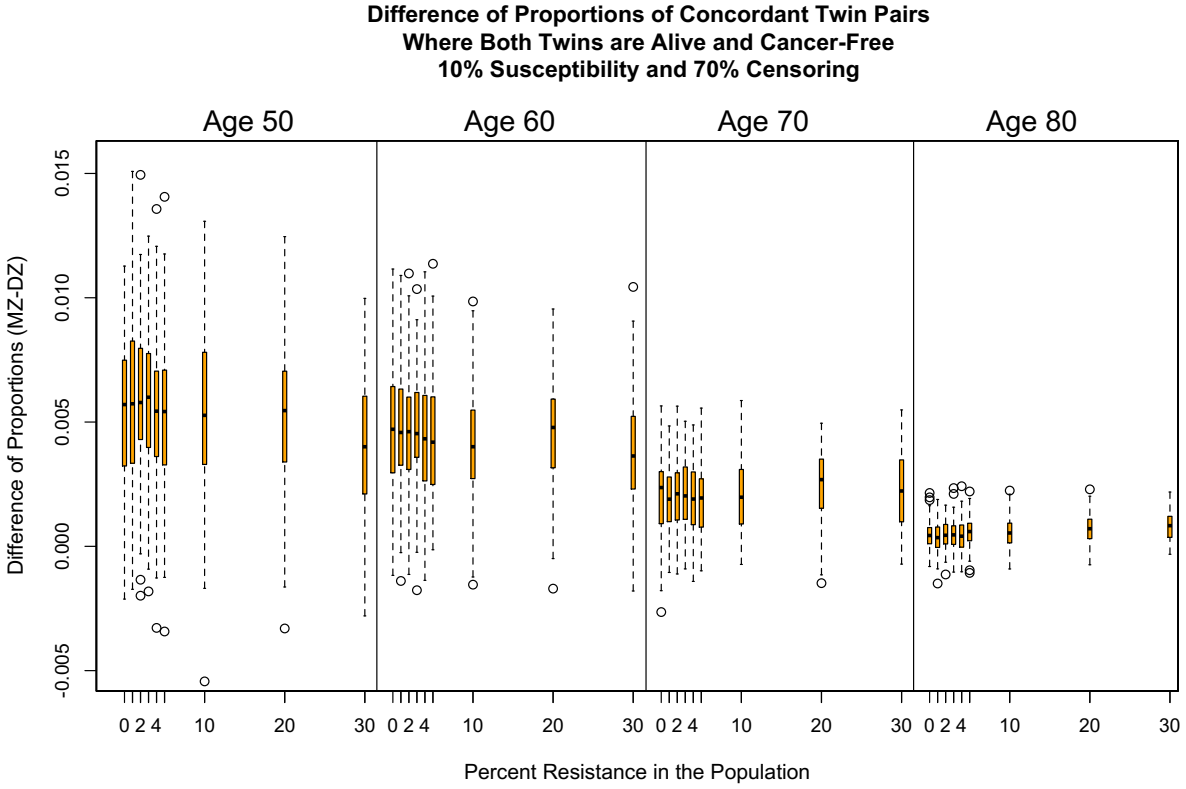


Figure A.9: **Concordant twin pairs where both twins are alive and cancer-free by age with 70% censoring**. Simulation results based on $79,520$ MZ twins and $123,348$ DZ twins with 70% censoring and 10% susceptibility. Differences in proportions of concordant MZ and DZ twin pairs that are alive and cancer-free, as we vary levels of resistance, across different age groups (50, 60, 70, 80). Results are based on 100 data generations.

In order to further explore the impact of censoring, we plot the difference of proportions of cancer-free concordant twin pairs and vary the levels of censoring (20%, 50%, 80%) (Figure 10.10). We can see consistently that the differences attenuate at the levels of censoring increases. Although there are differences in the proportion of MZ and DZ twin pairs where both have cancer, the differences are small. In fact, the differences are even smaller in the presences of censoring, so although there may be resistance in the population, it may not be detectable using the proportion of concordant twin pairs.

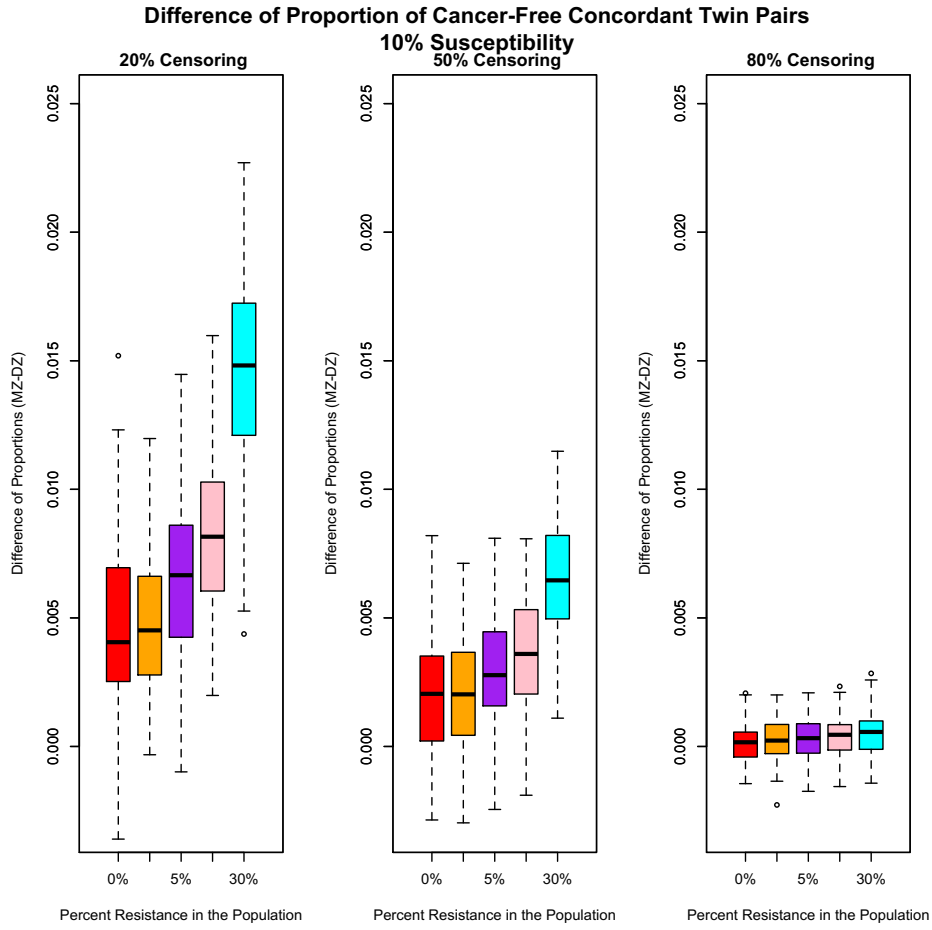


Figure A.10: **Difference of the proportions of MZ and DZ concordant cancer twin pairs with varying levels of censoring and 10% susceptibility.** The left graph are the results from 20% censoring, the middle is 50% censoring and the right graph are the results from 80% censoring.

A.2.2 Additional Details on Likelihood Function

.

We provide a detailed expression for the likelihood referenced in Section 2.3 of the main text.

$$L = \prod_{i=1}^{n} [f(t_{i1}, \epsilon_{i1} | t_{i2}, \epsilon_{i2})]^{\delta_{i1}} [S(t_{i1}, \epsilon_{i1} | t_{i2}, \epsilon_{i2})]^{1-\delta_{i1}} [f(\epsilon_{i2}, \epsilon_{i2})]^{\delta_{i2}} S(t_{i2}, \epsilon_{i2})^{1-\delta_{i2}} \tag{3.1}$$

$$\begin{aligned} = \prod_{i=1}^{n} & [f(t_{i1}, \epsilon_{i1} = 1 | t_{i2}, \epsilon_{i2} = 1) f(t_{i2}, \epsilon_{i2} = 1)]^{I(\epsilon_{i1}=1, \epsilon_{i2}=1)} \times \\ & [f(t_{i1}, \epsilon_{i1} = 2 | t_{i2}, \epsilon_{i2} = 2) f(t_{i2}, \epsilon_{i2} = 2)]^{I(\epsilon_{i1}=2, \epsilon_{i2}=2)} \times \\ & [f(t_{i1}, \epsilon_{i1} = 1 | t_{i2}, \epsilon_{i2} = 2) f(t_{i2}, \epsilon_{i2} = 2)]^{I(\epsilon_{i1}=1, \epsilon_{i2}=2)} \times \\ & [f(t_{i1}, \epsilon_{i1} = 2 | t_{i2}, \epsilon_{i2} = 1) f(t_{i2}, \epsilon_{i2} = 1)]^{I(\epsilon_{i1}=2, \epsilon_{i2}=1)} \times \\ & [S(t_{i1}, \epsilon_{i1} = 0 | t_{i2}, \epsilon_{i2} = 1) f(t_{i2}, \epsilon_{i2} = 1)]^{I(\epsilon_{i1}=0, \epsilon_{i2}=1)} \times \\ & [S(t_{i1}, \epsilon_{i1} = 0 | t_{i2}, \epsilon_{i2} = 2) f(t_{i2}, \epsilon_{i2} = 2)]^{I(\epsilon_{i1}=0, \epsilon_{i2}=2)} \times \\ & [f(t_{i1}, \epsilon_{i1} = 1 | t_{i2}, \epsilon_{i2} = 0) S(t_{i2}, \epsilon_{i2} = 0)]^{I(\epsilon_{i1}=1, \epsilon_{i2}=0)} \times \\ & [f(t_{i1}, \epsilon_{i1} = 2 | t_{i2}, \epsilon_{i2} = 0) S(t_{i2}, \epsilon_{i2} = 0)]^{I(\epsilon_{i1}=2, \epsilon_{i2}=0)} \times \\ & [S(t_{i1}, \epsilon_{i1} = 0 | t_{i2}, \epsilon_{i2} = 0) S(t_{i2}, \epsilon_{i2} = 0)]^{I(\epsilon_{i1}=0, \epsilon_{i2}=0)} \end{aligned} \tag{3.2}$$

The likelihood can be further rewritten by conditioning on the resistance carrier status, g_{ik}, and the susceptibility, b_{ik} carrier status, for both twins (Equation 3.3). We can then substitute in $\gamma = P(g_{ik} = 1)$ and $\beta = P(b_{ik} = 1)$ and maximize the likelihood with respect to γ and β.

$$
\begin{aligned}
L = \prod_{i=1}^{n} \Bigg(& \sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) f(t_{i1}, \epsilon_{i1} = 1 | t_{i2}, \epsilon_{i2} = 1, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& f(t_{i2}, \epsilon_{i2} = 1 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=1, \epsilon_{i2}=1)} \Bigg) \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) f(t_{i1}, \epsilon_{i1} = 1 | t_{i2}, \epsilon_{i2} = 2, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& f(t_{i2}, \epsilon_{i2} = 2 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=1, \epsilon_{i2}=2)} \Bigg) \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) f(t_{i1}, \epsilon_{i1} = 2 | t_{i2}, \epsilon_{i2} = 1, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& f(t_{i2}, \epsilon_{i2} = 1 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=2, \epsilon_{i2}=1)} \Bigg) \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) f(t_{i1}, \epsilon_{i1} = 2 | t_{i2}, \epsilon_{i2} = 2, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& f(t_{i2}, \epsilon_{i2} = 2 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=2, \epsilon_{i2}=2)} \Bigg) \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) S(t_{i1}, \epsilon_{i1} = 0 | t_{i2}, \epsilon_{i2} = 1, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& f(t_{i2}, \epsilon_{i2} = 1 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=0, \epsilon_{i2}=1)} \Bigg) \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) S(t_{i1}, \epsilon_{i1} = 0 | t_{i2}, \epsilon_{i2} = 2, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& f(t_{i2}, \epsilon_{i2} = 2 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=0, \epsilon_{i2}=2)} \Bigg) \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) f(t_{i1}, \epsilon_{i1} = 1 | t_{i2}, \epsilon_{i2} = 0, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& S(t_{i2}, \epsilon_{i2} = 0 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=1, \epsilon_{i2}=0)} \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) f(t_{i1}, \epsilon_{i1} = 2 | t_{i2}, \epsilon_{i2} = 0, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& S(t_{i2}, \epsilon_{i2} = 0 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=2, \epsilon_{i2}=0)} \Bigg) \times \\
& \Bigg(\sum_{b_{ik}=0}^{1} \sum_{g_{ik}=0}^{1} \Big[P(b_{i1}, b_{i2}, g_{i1}, g_{i2}) S(t_{i1}, \epsilon_{i1} = 0 | t_{i2}, \epsilon_{i2} = 0, b_{i1}, b_{i2}, g_{i1}, g_{i2}) \\
& S(t_{i2}, \epsilon_{i2} = 0 | b_{i1}, b_{i2}, g_{i1}, g_{i2}) \Big]^{I(\epsilon_{i1}=0, \epsilon_{i2}=0)} \Bigg)
\end{aligned}
\tag{3.3}
$$