The Impact of Truncation and Missing Family Links in Population-based Registers on Familial Risk Estimates

Monica Leu, Kamila Czene, and Marie Reilly

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

Received for publication April 2, 2007; accepted for publication July 27, 2007.

Family history information is often incomplete in population-based disease registers because of truncation and/or missing family links. In this study, the authors simulated complete populations of related individuals with realistic age, family structure, and incidence rates. After mimicking the realities of register-based data, such as left truncation of family history and missing family links due to death, the authors explored recovery of familial association parameters from standard epidemiologic models. Truncation of family history produced almost no bias for a familial risk of 2 and 50 years of follow-up, but it had a dramatic impact when the familial risk was 10. The age distribution of disease and the magnitude of background incidence rates also affected family history loss and thus the magnitude of bias. One can safeguard against bias by starting follow-up later, with the number of registration years to be ignored in the analysis depending on the value of familial risk. The missing familial links due to death had no effect, except when there was differential mortality for cases with and without a family history of disease. In summary, truncation, and to a lesser extent missing family links, induces bias in familial risk estimates from population-based registers.

bias (epidemiology); familial risk; medical record linkage; missing data; registries

To estimate the familial contribution to the risk of diseases that aggregate in families, valuable information is provided by considering the number of affected relatives, their degree of relationship, and their age at diagnosis. Population-based registers that record such information offer follow-up cohorts with family history of disease verified medically. Such registers have made significant contributions to the study of familial aggregation of disease (1–4). Whether they cover regional or nationwide populations, these registers may be incomplete because of inclusion/exclusion criteria (5), broken family links (6), and left truncation of any disease events that occurred in family members before the start-up date of registration (7, 8). Because the available statistical methodology for modeling incomplete family data (9–12) is cumbersome to implement, researchers concerned about potential bias often simply restrict their analyses to only “safe” study cohorts, such as certain birth cohorts or recent years of registration for which data are essentially complete (13, 14). These approaches are conservative, resulting in the loss of large volumes of usable data and a consequent loss of statistical power for addressing the research hypothesis.

Many published studies use all the available data in assessing familial contributions to diseases, although such data will be subject to one or more of the limitations just mentioned. Insufficient exploratory work is available for assessing the magnitude of the biases that may result from analyses that ignore these limitations of register-based data. In our study, we simulated realistic populations of related individuals with various features of familial aggregation of disease, and, after mimicking the realities of register-based data (such as left truncation and missing family links), we explored recovery of familial association parameters from standard epidemiologic models.

MATERIALS AND METHODS

We used the R software package PopLab (15, 16) to create virtual population registers of related Swedish individuals...
for the time interval 1955–2002. We used the real Swedish age- and calendar-year-specific mortality and fertility rates, which are available online from Statistics Sweden (17). We chose female breast cancer as our primary example of a disease with familial aggregation because the age-specific incidence profile, familial risk, and case mortality ratio are representative of many cancers (18). We used the 1980 age-specific incidence rates, which are available online in the Cancer Epidemiology Database from the International Agency for Research on Cancer (19).

Creation of virtual populations

For the first year for which age- and calendar-year-specific fertility and mortality rates are available (1955), a large electronic population of related individuals is first created, as explained in detail elsewhere (15), starting with 500,000 males and 500,000 females with unknown parents, whom we refer to as “founders.” This baseline population, which consists of complete families, then evolves dynamically over time until 2002. We assume a relative risk model in which a woman is at increased risk of breast cancer from the time that her mother becomes a case; thus, age-specific rates of disease are multiplied by a constant factor (the incidence rate ratio) for daughters of affected mothers. In our models, a person affected by disease experiences a higher mortality rate than a disease-free person of the same age.

To enable extrapolation of our findings to a range of diseases, we simulated several levels of familial risk (2, 5, and 10), case mortality ratios (2, 5, and 10), background incidence rates (1980 Swedish breast cancer rates scaled by a factor of 0.5, 1, and 2), and disease age patterns (actual 1980 age-specific breast cancer rates, and these same rates shifted to younger ages by 20 years). For each combination of parameters, an independent large population was created. We also explored the stability of familial risk estimates in small populations (50,000 male and 50,000 female founders).

Study cohorts

In our analyses, a study cohort was defined as women who were alive and cancer free at the beginning of follow-up (1955) or were born after 1955 and before the end of follow-up (2002). A person was followed until disease incidence, death, or end of follow-up, whichever occurred first. A positive family history of cancer (i.e., affected mother) was defined as exposure. Starting with the simulated populations, with complete family links and complete information on familial exposure, we considered two types of incompleteness: left truncation of family history of disease due to the start-up date of the Swedish Cancer Register (20) and missingness patterns seen in the Swedish Multi-Generation Register (5).

The Swedish Cancer Register has recorded primary cancers diagnosed in Sweden since 1958, based on compulsory reports from health care providers, and the completeness of the register for cancers incident after 1958 is close to 100 percent (21). To reflect this, we “hid” in our populations any maternal cancer incident before the first year of registration by assuming the mother to be cancer free. Figure 1 illustrates, for the cohort at risk, the percentage of maternal cancers recorded in the data after imposing such a start-up effect of registration.

The Multi-Generation Register is a database of individuals, called index persons, born in Sweden since 1932. They are uniquely identified by their identification number, and the identification numbers of their biologic parents have been recorded. However, family relationships are not complete, and the proportion of persons missing parental identifiers depends on the date of birth and date of death. For persons who died prior to 1991, a substantial fraction (~50 percent) have unidentified parents. We mimicked these missing family links by assigning an unknown mother with probability 0.5 to each individual who died before 1990, and with probability 0.1 to those who died between 1990 and 2002. The impact of these broken links on the loss of maternal cancers is illustrated in figure 1.

Statistical analysis

For each population created, we first analyzed the ideal (i.e., complete) data to investigate recovery of the familial relative risk from standard epidemiologic analysis. We then performed naive standard analysis of the data with the two patterns of missingness imposed, and we examined the bias resulting from ignoring the incompleteness. In the analysis of the left-truncated data, mothers whose cancers were truncated were assumed to be cancer free; in the analysis of data
with missing family links, only those persons with a known mother were included. We also created and analyzed data sets in which both forms of incompleteness (truncation and missing family links) were considered simultaneously. Additionally, we considered a sibling relative risk model in which exposure was defined as having at least one affected sister. Incidence rate ratios were estimated from Poisson regression models, and the estimates were adjusted for age.

**RESULTS**

Table 1 shows the familial risk estimates from the ideal populations for which family links and cancer information were complete. For each value of familial risk, there was excellent recovery of the true value using standard analysis. Even with smaller populations, the analyses produced valid estimates but with wider 95 percent confidence intervals, as expected.

Table 2 illustrates the effects of left truncation and missing mother’s identity, and the combined effect of these two sources of incompleteness. The bias due to truncation increased with increasing value of familial risk, with the 95 percent confidence intervals not including the true familial risk at higher values. The mortality ratio had little or no impact on the results. When we mimicked missing family links only (i.e., no truncation), the risk estimates were very close to the true values. The populations affected by both sources of incompleteness followed patterns of bias similar

### TABLE 1. Recovery of familial risk estimates from complete populations*

<table>
<thead>
<tr>
<th>Case mortality ratio and population</th>
<th>Familial risk 2</th>
<th>Familial risk 5</th>
<th>Familial risk 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IRR</td>
<td>95% CI</td>
<td>IRR</td>
</tr>
<tr>
<td>2 Small</td>
<td>1.88</td>
<td>1.68, 2.09</td>
<td>5.12</td>
</tr>
<tr>
<td>Large</td>
<td>2.02</td>
<td>1.95, 2.08</td>
<td>4.95</td>
</tr>
<tr>
<td>5 Small</td>
<td>2.03</td>
<td>1.83, 2.25</td>
<td>5.20</td>
</tr>
<tr>
<td>Large</td>
<td>2.02</td>
<td>1.95, 2.08</td>
<td>5.00</td>
</tr>
<tr>
<td>10 Small</td>
<td>1.90</td>
<td>1.71, 2.11</td>
<td>5.22</td>
</tr>
<tr>
<td>Large</td>
<td>1.99</td>
<td>1.93, 2.06</td>
<td>4.95</td>
</tr>
</tbody>
</table>

* Large populations were generated from 1 million founders and small populations from 100,000 founders (individuals with unknown parents).
† IRR, incidence rate ratio; CI, confidence interval.

### TABLE 2. Effects of truncation and missing family links on estimated IRRs*

<table>
<thead>
<tr>
<th>Case mortality ratio</th>
<th>Familial risk 2</th>
<th>Familial risk 5</th>
<th>Familial risk 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IRR</td>
<td>95% CI</td>
<td>IRR</td>
</tr>
<tr>
<td>Truncated†</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.96</td>
<td>1.87, 2.06</td>
<td>4.36</td>
</tr>
<tr>
<td>5</td>
<td>1.88</td>
<td>1.79, 1.98</td>
<td>4.33</td>
</tr>
<tr>
<td>10</td>
<td>1.98</td>
<td>1.89, 2.08</td>
<td>4.27</td>
</tr>
<tr>
<td>Missing mother’s identity‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.04</td>
<td>1.96, 2.11</td>
<td>5.00</td>
</tr>
<tr>
<td>5</td>
<td>2.01</td>
<td>1.94, 2.09</td>
<td>5.09</td>
</tr>
<tr>
<td>10</td>
<td>2.03</td>
<td>1.95, 2.11</td>
<td>5.02</td>
</tr>
<tr>
<td>Combined§</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.03</td>
<td>1.93, 2.13</td>
<td>4.51</td>
</tr>
<tr>
<td>5</td>
<td>1.98</td>
<td>1.87, 2.09</td>
<td>4.64</td>
</tr>
<tr>
<td>10</td>
<td>2.09</td>
<td>1.98, 2.20</td>
<td>4.61</td>
</tr>
</tbody>
</table>

* IRRs, incidence rate ratios; CI, confidence interval.
† Maternal cancer diagnosis was truncated before 1955.
‡ Based on daughter’s year of death.
§ Combination of truncation and missing mother’s identity.
to those for the populations affected by left truncation only, with the magnitude of such biases somewhat smaller. As with the truncated populations, the 95 percent confidence intervals covered the true value for an incidence rate ratio of 2 but not for the larger values.

To further explore the bias due to truncation, we investigated the contribution of calendar time, background incidence, and age distribution. Since we have already shown that case mortality ratio had no impact on the bias, we used a mortality ratio of 5 throughout. To investigate how the recovery of familial risk estimates changed with time after start-up of registration, we present in figure 2 a plot of period-specific estimates for 10-year intervals illustrating that recovery time depends on the strength of the familial association. For a small value of familial risk (i.e., incidence rate ratio = 2), the estimates assessed from follow-up cohorts 20 years after registry initiation were close to the true value. For a large value of familial association (i.e., incidence rate ratio = 10), the magnitude of bias also decreased with time after registry initiation, but a substantial bias still remained after the 40 years studied.

The contributions of the background incidence rates to the magnitude of the bias are presented in table 3. We examined the effects of left truncation on familial risk estimates from a population simulated with a disease incidence in which age-specific rates were half those of the Swedish 1980 breast cancer rates and another population simulated with double the 1980 rates. When the true familial risk was 2, the estimates were close to the true value and were not affected by incidence rates. For higher values of familial risk, the bias became apparent and increased with background incidence rates.

We investigated the contribution of the age distribution of disease and age structure of the population to the bias from left truncation. We showed recovery of familial risk estimates from populations simulated by using a theoretical disease incidence, in which the rate applied to each age group was the breast cancer incidence rate for women 20 years older; that is, the age distribution was shifted 20 years toward younger ages. With the resulting substantial incidence rates at young ages, the extent of bias was larger compared with that for a population with the same value for familial risk but delayed incidence, especially when familial risk was

![FIGURE 2. Familial risk (incidence rate ratio (IRR)) estimates and standard errors recovered from statistical analyses of truncated populations stratified by calendar time for three different levels of IRR (dotted lines).](image)

**TABLE 3. Effects of truncation on estimated IRRs* for various background incidence rates and age distributions**

<table>
<thead>
<tr>
<th>Incidence of disease</th>
<th>Familial risk 2</th>
<th>Familial risk 5</th>
<th>Familial risk 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IRR 95% CI*</td>
<td>IRR 95% CI</td>
<td>IRR 95% CI</td>
</tr>
<tr>
<td>brc:2</td>
<td>1.82 1.65, 2.01</td>
<td>4.43 4.17, 4.71</td>
<td>8.29 7.94, 8.66</td>
</tr>
<tr>
<td>brc</td>
<td>1.88 1.79, 1.98</td>
<td>4.33 4.20, 4.47</td>
<td>7.47 7.31, 7.64</td>
</tr>
<tr>
<td>brc:2</td>
<td>1.88 1.83, 1.93</td>
<td>3.80 3.73, 3.86</td>
<td>5.86 5.78, 5.94</td>
</tr>
<tr>
<td>Age distribution</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shifted brc</td>
<td>1.85 1.79, 1.90</td>
<td>3.82 3.76, 3.89</td>
<td>5.87 5.79, 5.95</td>
</tr>
<tr>
<td>brc restricted§</td>
<td>2.05 1.89, 2.23</td>
<td>4.86 4.61, 5.13</td>
<td>9.16 8.80, 9.52</td>
</tr>
<tr>
<td>Shifted brc restricted§</td>
<td>1.96 1.89, 2.04</td>
<td>4.35 4.24, 4.47</td>
<td>7.09 6.96, 7.23</td>
</tr>
</tbody>
</table>

* IRRs, incidence rate ratios; CI, confidence interval.
† "brc" represents the Swedish 1980 breast cancer incidence rates; "brc:2" and "brc:2" are half and double these rates, respectively.
‡ Breast cancer incidence rates were shifted 20 years toward younger ages.
§ Persons in the study cohort were born after 1932.
high. We also mimicked the real Swedish family data context, that is, by restricting the populations to persons born after 1932. The bias again increased with the true value of familial risk but was of a lesser magnitude than the bias resulting from analyzing the unrestricted populations.

We continued to explore the impact of left truncation and unidentified mothers by considering the scenario in which there was differential mortality for cases with and without a family history of disease. For cases without a family history, we assigned a mortality ratio of 2 relative to the background population, while exposed cases experienced a mortality ratio of 10. Table 4 illustrates the estimated familial risks from these populations. Differential mortality had little or no effect on the magnitude of bias due to left truncation of family history. However, the magnitude of the bias due to missing family links was substantial in contrast to the lack of bias when mortality was nondifferential, and this bias appeared to be independent of the magnitude of familial risk.

In all our investigations, left truncation of family history of cancer induced the most dramatic bias in familial risk estimates. We explored the effect of such incompleteness in a sibling relative-risk model for populations simulated by using the 1980 breast cancer incidence rates and a non-differential mortality ratio of 5. The patterns of bias were similar to those seen for the maternal relative-risk populations but were of smaller magnitude: the incidence rate ratio estimates were 2.06 (95 percent confidence interval: 1.96, 2.17), 4.79 (95 percent confidence interval: 4.63, 4.96), and 8.94 (95 percent confidence interval: 8.69, 9.19) for a familial risk of 2, 5, and 10, respectively. Figure 3 illustrates the differential loss of cancer history for mothers and sisters due to registration start-up in 1955 for maternal and sibling models using an incidence rate ratio of 2.

**DISCUSSION**

Our aim in this study was to assess the reliability of familial risk estimates when analyzing register-based data, exemplified by the Swedish Multi-Generation Register (5) and the Swedish Cancer Register (20). We performed a simulation study, using the R software package PopLab (16), to create realistic population registers in which all parameters are fully known, and we observed the magnitude of bias from analyzing such populations when subjected to realistic patterns of incompleteness. Our illustrations included levels of familial risk and case mortality ratios that span the range for many cancers. We chose female breast cancer as the example of an incident disease in our population because of a well-acknowledged contribution of family history to this malignancy and a middle-range magnitude of familial risk (~2.0) among those cancers that aggregate in families (18). For diseases whose incidence and age profiles differ from those for cancer, similar investigations can be conducted wherever register data are available.

To understand the contribution of each register to the bias in familial risk estimates, our first step was to retain all family information in our ideal population but to truncate cancer history to reflect start-up of cancer registration. Naive
statistical analyses were then performed by using just the available complete data. Since familial risk was represented by the proportion of cases with a positive family history, it is not surprising that the loss of family history due to truncation affected these risk estimates. Left truncation resulted in a downward bias for all models studied, thus yielding conservative estimates of risk. The bias was more serious at high levels of familial risk and with large background incidence rates. For example, the incidence rate ratio estimates corresponding to the three levels of familial risk were 1.88 (resulting in 6 percent bias), 4.33 (13 percent bias), and 7.47 (25 percent bias), respectively, for the model with 1980 breast cancer incidence rates and a mortality ratio of 5.

It is clear that, in the early years of registration, the information from the cancer register cannot give valid estimates of familial risk. However, as time passes, family history is more faithfully recorded; thus, study designs that use, for example, only recent years from a longer registry accrual time can produce valid estimates (13, 22). To minimize loss of statistical power, it is of interest to estimate how long it takes a register to “recover” from the truncation at start-up in a given application. Figure 2 illustrates that, to “safeguard” the analysis against bias, the number of initial registration years that should be discarded depends on the magnitude of familial risk. We noted that, for cancers with a low familial risk, reasonable estimates of familial risk can be achieved in a relatively short time (~20 years). However, when this risk is higher, there is a more long-lasting bias and thus slower recovery time.

The age pattern of disease also has an impact on the magnitude of bias due to left truncation. In our examples using the merged Swedish Cancer Register and Multi-Generation Register, follow-up began in 1955; thus, cohort members affected by loss of family history were in their twenties (the Multi-Generation Register records only those persons born after 1932), and many had parents younger than 50 years of age at start-up. Thus, when the studied disease affects mainly older persons (for breast cancer, substantial incidence occurs after the age of 50 years), left truncation of maternal cancer will have only a modest impact on familial risk estimates since these mothers are unlikely to have developed the disease before registration start-up. However, when the disease affects younger individuals, loss of family history due to truncation can be substantial, resulting in serious biases in familial risk estimates (table 3).

Left truncation of cancer diagnoses in siblings biased the estimates of familial risk to a lesser extent compared with the loss of maternal cancer diagnoses. Overall, the age difference between sisters was considerably smaller than the gap for mother-daughter pairs. Thus, for incident cases, many more mothers compared with sisters were diagnosed in the distant past, which explains the greater loss of family history for mothers (figure 3).

To assess the potential for bias due to missing family links, our second step was to retain in our ideal simulated population all cancer history information but to break the family links (e.g., hide the mother’s identity) by using patterns of missingness similar to the Swedish Multi-Generation Register. For all studied levels of familial risk and mortality ratio, there was little or no bias when mortality for familial and nonfamilial cases was the same. However, it is not unreasonable to assume that, for some diseases, familial cases can experience a differential mortality, such as carriers of the BRCA1 mutation for breast cancer who are generally considered to have a poorer prognosis than sporadic cases (23). When we simulated populations with a higher mortality for familial cancer cases, we noted a bias whose magnitude depended on the value for familial risk. This higher mortality leads to preferential exclusion of exposed cases from the study cohort, when they do not survive to the time point when good-quality parental information is available (1991 in our example). Thus, the differential mortality would be expected to lead to a downward bias in the familial risk estimates, as we observed.

After separately studying left truncation of family history and missing maternal identity, we constructed the study cohorts that would result from using real register data subject to both of these sources of incompleteness. The bias in the familial risk estimates followed patterns similar to the bias resulting from left truncation only but was of a somewhat lesser magnitude (table 2). This is a predictable consequence of some age groups (mainly older persons) being subject to exposure truncation and exclusion from the study cohort because of an unidentified mother. Consequently, persons whose family history would otherwise be truncated are excluded from the study cohort, so that the bias of familial risk estimates is mitigated.

We have demonstrated how a freely available and simple-to-use simulation tool can be used to mimic the realities of research data available from population registers and to investigate the validity of familial risk estimates from such data. Choosing the size of simulated populations depends on the purpose. While exploring the potential bias for various familial risk parameters and selected models of familial aggregation, relatively small populations can produce valid results (table 1). However, for period-specific estimates of familial risk or less prevalent diseases, consistent confidence intervals demand the creation of larger populations.

To make our illustrative analyses less complicated, we used an age-specific disease incidence that was constant over time, although calendar-specific rates can be easily accommodated. Although we presented the results from simple relative risk models only, the simulation software can also accommodate age-dependent relative risks or odds ratio models. We investigated here sources of missingness in a homogeneous, closed population, without considering immigration or emigration. Although emigration data are easily accommodated in defining correct censoring dates, immigration data are more complex. Immigrant subpopulations may well have different background incidence rates and a different age structure than their host population (24, 25), and left truncation is no longer a fixed point in calendar time but is specific to the person’s immigration date. We have shown that the bias in familial risk estimates depends on all of these factors, but understanding the nature of the dependence for migrant populations requires further research.

We studied the impact on familial risk estimates of left truncation of family history of disease and of missing parental identity, and we showed that truncation induces considerable bias. The magnitude of such bias and the time
needed for the register to recover are specific to each study because they depend on the value of familial association, background incidence rates, and the mortality mechanism for cases. They also depend indirectly on disease age pattern and family relationship through loss of family history at start-up. Our simulation tool is available for free download to researchers interested in studying the biases in population-based register data used in their studies.

ACKNOWLEDGMENTS

This study was performed with financial support from the Swedish Cancer Society (contract 030635).
Conflict of interest: none declared.

REFERENCES