Practice of Epidemiology

A Proposed Method to Adjust for Selection Bias in Cohort Studies

Anna Törner*, Ann-Sofi Duberg, Paul Dickman, and Åke Svensson

* Correspondence to Anna Törner, Department of Epidemiology, Swedish Institute for Infectious Disease Control, Torstekodavägen 19A, 171 82 Solna, Sweden (e-mail: anna.torner@smi.se).

Initially submitted December 1, 2008; accepted for publication December 4, 2009.

Selection bias is a concern in cohort studies in which selection into the cohort is related to the studied outcome. An example is chronic infection with hepatitis C virus, where the initial infection may be asymptomatic for decades. This problem leads to selection of more severely ill individuals into registers of such infections. Cohort studies often adjust for this bias by introducing a time window between entry into the cohort and entry into the study. This paper describes and assesses a novel method to improve adjustment for this type of selection bias. The size of the time window is decided by calculating a standardized incidence ratio as a continuous function of the size of the time window. The resulting graph is used to decide on an appropriate window size. The method is evaluated by using the Swedish register of hepatitis C virus infections for 1990–2006. The complications studied were non-Hodgkin lymphoma and liver cancer. Selection bias differed for the studied outcomes, and a time window of a minimum of 2 months and 12 months, respectively, was judged to be appropriate. The novel method may have advantages compared with an interval-based method, especially in cohort studies with small numbers of events.

bias adjustment; bias (epidemiology); cohort studies; communicable diseases; data analysis; statistical modeling

Abbreviations: HCV, hepatitis C virus; SIR, standardized incidence ratio.

Some infections are initially asymptomatic and may be undiagnosed for a considerable time, which is common with, for example, chronic viral hepatitis and human immunodeficiency virus infection. The diagnosed population therefore comprises patients for whom duration from the date of infection to the date of diagnosis is unknown—a prevalent cohort (1, 2). A prevalent cohort is defined as a cohort of individuals in whom the disease process has already begun when follow-up starts, that is, the first period of the infection is unobserved. The concept of a prevalent cohort is frequently also used for noninfectious conditions and may be used when individuals are sampled according to a cross-sectional sampling criterion (3).

Studies of prevalent cohorts present specific methodological challenges associated with the delayed diagnosis. Standard statistical methods for survival analysis fail when backward recurrence times (i.e., times of infection prior to diagnosis) are unknown (2, 4). When estimating risk of secondary complications of infection, use of “time since diagnosis” as the analysis timescale may bias the results if the underlying hazard of developing complications is not constant over time (5). Some previous work in this area has focused on human immunodeficiency virus as the prevalent infection and acquired immunodeficiency syndrome as the outcome event (1, 4, 6).

In previous work by our group, the unknown time from infection to diagnosis (delayed entry or left truncation) for patients with hepatitis C virus (HCV) infection was addressed by creating a model for estimated date of primary HCV infection (7, 8). Date of infection was modeled by using birth year, age at diagnosis of HCV, suspected route of infection, and data on how the HCV epidemic spread in Sweden during 1960–2000. The modeled date of infection was used to stratify patients according to prior duration of HCV infection. For both complications studied, non-Hodgkin lymphoma and liver cancer, the risk of disease was found to increase with longer duration of HCV infection.

Stratification, as described, may handle random left truncation, which occurs when the infection is diagnosed independent of the individual’s risk of developing a studied outcome. This assumption may not be valid because
emergence of secondary symptoms, and hence the outcome, is likely to increase the probability of diagnosing the infection. The diagnosed cohort thereby differs from the source population in that more individuals with the emerging outcome are included in the diagnosed cohort. This type of bias, where the diagnosed cohort differs from the population of interest, may be referred to as selection bias (9).

To not include individuals for whom the exposure (i.e., the infection) was diagnosed because of the studied outcome, a common method is to introduce a time window between time of entry into the diagnosed cohort and time of entry into the study population. In practice, using this method means that person-years and events occurring during this time window are excluded from the study population (10). The size of the time window needs to be chosen carefully. The objective is to avoid selection bias effects, but, not to waste information and to preserve statistical power, the time window should not be unnecessarily wide. The time window method is not applicable when an acute infection and acute outcome is studied, for example, death from meningococcal infection. In this situation, other methods to address selection bias must be used.

The objective of this study was to evaluate a cumulative method of choosing the appropriate time window by comparing it with an established interval-based method. The cumulative method has advantages regarding power and precision, especially for risk estimates in studies based on relatively few events. In this paper, the proposed method is illustrated by using data from the Swedish HCV register and the Swedish Cancer Registry.

For the examples given, liver cancer and non-Hodgkin lymphoma in HCV-infected individuals, the source population is the nationwide population of HCV-infected individuals (Figure 1). However, only those with a diagnosed HCV infection, the diagnosed cohort, are available for a register-based cohort study. To reduce the effect of selection bias, the study population is derived from the diagnosed cohort after exclusion of individuals with emerging outcomes occurring within a chosen time window after diagnosis of HCV infection. The observation time starts for all subjects in the study population after this time window from HCV diagnosis.

Surveillance bias, that is, a higher propensity to detect the outcome when the infection has been diagnosed, might further confound the situation by contributing to a concentration of outcome events in the early phase following diagnosis. Surveillance bias is assumed to be negligible for the examples presented here. Other types of important biases described for prevalent cohorts are onset confounding and length-biased sampling (4). These forms of bias were not investigated in our study.

MATERIALS AND METHODS

Statistical methods

The standardized incidence ratio (SIR) estimates the risk of disease or other complications in cohort studies. It is defined as the observed number of events, for example, cancer, divided by the expected number of events, the latter being based on the experience in a matched population that serves as an unexposed control group, typically the general population. The expected number of events is calculated by multiplying age-, sex-, and calendar-year-specific incidence rates from the general population by person-years accumulated in the corresponding strata in the exposed cohort. This method of standardization of incidence rates is generally referred to as an indirect method of standardization.

A standard method of detecting selection bias is to calculate the SIR for several disjoint time intervals to determine whether it is higher in intervals close to time of entry into the cohort compared with later intervals (11, 12). For each time interval, only those events and person-years accumulated in that particular interval are used for SIR estimation. A higher SIR in early intervals reveals a concentration of outcome events close to time of entry into the cohort and indicates selection bias. To adjust the SIR calculations for selection bias, a time window between time of entry into the cohort and time of entry into the study is introduced. The size of the time window corresponds to those time intervals during which the SIR is excessively increased, as judged by visual inspection.

The proposed alternative method for choosing the size of the time window is to calculate a cumulative version of the SIR. Here, the SIR for the diagnosed cohort is calculated by excluding person-years and events within a time window of successively increasing size (i) after entry into the cohort. The size of the time window is increased successively from 1 day up to a fairly long time period, for example, 730 days (2 years). The estimated SIRs are plotted on a graph against the size of the time window. Selection bias will be evidenced as higher SIR estimates for short time windows. In the graph, the SIR will gradually decrease until it levels off.

Figure 1. Relation between different populations of individuals; refer to the text for more information. A) The source population: the total cohort of all individuals infected with hepatitis C virus (HCV), diagnosed and undiagnosed. Nested within the source population is B), the diagnosed cohort: all individuals diagnosed and notified with HCV, including individuals with an emerging outcome at the time of HCV diagnosis. Finally, nested within B) is C), the study population in which individuals with an emerging outcome in close connection to HCV diagnosis have been excluded.
at least temporarily, at a time window value presumably unaffected by selection bias.

The methods were evaluated with a Swedish HCV data set (after linkage to other registers) with regard to both liver cancer and non-Hodgkin lymphoma. For the interval-based method, the SIR was calculated separately for the following time periods after HCV notification: first, second, third, fourth, fifth and sixth months, seventh through 12th months combined, and second year after notification. A month was defined as 30 days. The resulting SIRs for the different time intervals were plotted against time since HCV notification. SIRs for liver cancer and non-Hodgkin lymphoma were also calculated by using the cumulative version of the SIR. To determine whether the selection bias effect varied between groups in the cohort, a stratified analysis was performed, in which individuals diagnosed with HCV at different ages (<45 years, 45–60 years, and >60 years) were analyzed separately by using the cumulative version of the SIR. Confidence intervals for SIR estimates were calculated by using the Max method, assuming that the observed numbers of events are Poisson distributed (13).

All statistical analyses were performed with SAS version 9.1 software (SAS Institute, Inc., Cary, North Carolina). A description of how to perform the calculations for the cumulative SIR can be found at the following website: http://www.s-gem.se/about_sgem/researchers/anna_torner_en.html.

Studied cohort

The prevalence of HCV antibodies in Sweden (population 9 million) is estimated to be less than 0.5% (14). The majority of HCV-infected patients have a chronic infection, very often asymptomatic but with a highly increased risk of liver cirrhosis and liver cancer after 20–30 years with infection, as well as a doubled risk of non-Hodgkin lymphoma (7, 8). In Sweden, HCV infection is a notifiable disease, and all diagnosed cases should be reported to the Swedish Institute for Infectious Disease Control. Notification is made by both the treating physician and the laboratory confirming the diagnosis. The notification date is the earliest date the infected individual is reported in the system, and, in most cases, the notification date is identical to the date of laboratory confirmation. Studies of the Swedish surveillance system document very high sensitivity of the system. For the infections studied thus far, nearly 100% of diagnosed individuals were notified (15). In our study, the date of notification was assumed to be the date of diagnosis.

During the time period 1990–2006, after excluding patients also infected with hepatitis B virus, 39,450 individuals were notified with HCV. This cohort formed the diagnosed cohort (Figure 1). The HCV register was linked to the Swedish Cancer Registry for information on incident cancers and to the registers at Statistics Sweden to provide dates of death and migration. The linkage was performed by using the personal identification number, which provides a unique identifier for each individual and is used for all health care contacts and in national registers. The data set was anonymized after the linkage was completed. The Swedish Cancer Registry also provided data on cancer incidence in the general population. The study was approved by the Stockholm Ethics Review Board.

RESULTS

In this cohort of HCV-infected individuals, 362 liver cancers and 70 non-Hodgkin lymphomas had been diagnosed from the date of HCV notification until the end of the study period (December 31, 2006). Any individuals with liver cancer or non-Hodgkin lymphoma occurring before diagnosis of HCV were not included in the studied cohort and the statistical analyses. However, among all notified, HCV-infected individuals, 62 liver cancers and 35 non-Hodgkin lymphomas occurred less than 3 months prior to diagnosis of HCV, making it likely that the HCV infection was diagnosed as a consequence of the cancer diagnosis.

Liver cancer

Liver cancer is a rather common late complication in patients with HCV. The number of observed cases of liver cancer in a specified time interval is therefore relatively high. Figure 2 shows that the SIR is significantly higher in the early time intervals, approximately up to 1 year, decreasing later to approximately 34 (95% confidence interval: 22, 49) in the second year after HCV diagnosis. This finding indicates that the incidence of liver cancer is estimated to be 34 times higher in the study population compared with an age- and sex-matched reference population (here, the general population). The large number of cancers makes it possible to investigate the SIR for a number...
of time intervals with reasonable precision. In addition, the cumulative version of the SIR indicates that it will be considerably higher if the time period close to diagnosis of HCV is included in the analysis. The estimated SIR starts at about 40, when time of entry into the study is defined as time of diagnosis of HCV; that is, there is no time window. The SIR drops to about 33 (95% confidence interval: 29, 38) when the first year after diagnosis is excluded from the analysis (Figure 3).

Non-Hodgkin lymphoma

For non-Hodgkin lymphoma, the estimated SIRs are shown in Figure 4 (disjoint time intervals) and Figure 5 (cumulative estimate). The SIR for the disjoint time intervals is estimated as 18 (95% confidence interval: 6, 41) during the first month after diagnosis of HCV and in the following time intervals during the first year as 7, 0, and approximately 4. In the second year after diagnosis of HCV, the SIR is estimated as 2.0 (95% confidence interval: 0.8, 4.4). Visually and numerically, the estimated SIRs are higher at least during the first 1 or 2 months after diagnosis of HCV. However, only 15 non-Hodgkin lymphoma cases occur during the first year, and the confidence intervals overlap for all adjacent intervals. It is difficult to differentiate between random noise and a selection bias effect, apart from the first month. The cumulative version of the SIR shows that the SIR will be slightly higher if there is no time window between time of diagnosis of HCV and time of entry into the study. However, the SIR drops to 2.1 (95% confidence interval: 1.6, 2.7) if a time window of 2 months is introduced. The selection bias effect is much less evident for non-Hodgkin lymphoma than for the liver cancer example.

Stratified analysis

Individuals diagnosed with HCV through screening should be less susceptible to selection bias because diagnosis is independent of the severity of the underlying disease. The elderly and others outside high-risk groups are not normally screened for HCV. In such patients, infection may be diagnosed either by chance during a contact with the health care system or when complications secondary to the HCV infection emerge. The selection bias effect may therefore be different across age groups, where age may be interpreted as a surrogate marker for primary route of diagnosis. An age-stratified analysis was performed in which individuals diagnosed with HCV at different ages (<45 years, 45–60 years, and >60 years) were analyzed separately. This analysis showed a prominent selection bias effect in the 2 older age groups, where there is a clear decrease in the SIR if the time window close to diagnosis of HCV is excluded from the analysis. For individuals diagnosed with HCV before the age of 45 years, there is no evidence of selection bias, and the SIR is remarkably constant at about 20, regardless of the size of the time window introduced between time of diagnosis and time of entry into the study (Figure 6). An interesting finding in the age-stratified analysis is that the SIR is substantially higher for the middle-age group compared with the older age group. The middle-age group has been infected for a considerably shorter time than the older age group, evaluated by the date-of-infection model (8).
DISCUSSION

Selection bias is a concern in cohort studies where recruitment of patients into a cohort could be related to the outcome of interest. Regarding the HCV and liver cancer example, patients presenting with advanced liver cirrhosis close to diagnosis of the infection should be removed from the study population. However, for register-based research, this level of individual information is usually not available. Therefore, methods that may identify these patients must be based on the relation in time between diagnosis of the infection and the outcome studied. A higher-than-expected number of outcomes that occur close to diagnosis of the infection will be interpreted as the type of selection bias described here.

Our proposed method may adjust for factors located on the causal path from HCV infection to liver cancer, such as cirrhosis. Unknown or unmeasured potential risk factors, such as intravenous drug use or alcohol consumption, correlated to the exposure of interest (HCV) but not located on the causal path from HCV to liver cancer, are difficult to control for. These unknown or unmeasured risk factors are generally called confounders (16).

Our proposed method assumes an asymptomatic early infection or an asymptomatic noninfectious condition that, over long time periods, may result in cancer or other outcomes. Another assumption inherent to the model is that the estimated SIR, after exclusion of the initial period with selection bias, is fairly constant over time. If the SIR is decreasing with increasing follow-up time, it may be difficult to find a natural time window since the method is based on selecting a time point when the decrease in the SIR has leveled off. On the other hand, if the SIR is increasing significantly with infection time, an initially increased incidence due to selection bias may be difficult to detect. For the outcomes studied here, liver cancer and non-Hodgkin lymphoma, the estimated SIR levels off, at least temporarily, to one that is fairly constant.

For the diagnosed cohort of HCV-infected individuals and the studied outcome of liver cancer, selection bias proved to be considerable. Methods to counter selection bias in cohort studies are to either arbitrarily define a time period thought to be sufficiently long (17) or to calculate the SIR for different time intervals after entry into the study cohort (11, 12). Then, based on visual inspection, estimate the time point when the SIR is not influenced by selection bias. When the number of events is large, such as liver cancer for patients with HCV infection, it is possible to determine whether the SIR is excessively increased in several disjoint intervals. For relatively common outcomes, selection bias is therefore easily evaluated by using either the interval-based method or the proposed cumulative version. The objective for both methods is the same: finding a compromise that will maximize validity while minimizing loss of precision. The information provided by the graphs is slightly different, however. The interval-based calculations show what the estimated SIR actually is in different time periods after diagnosis of HCV; the graph of the cumulative SIR shows the actual impact of choosing a specific time window.

Figure 5. The cumulative standardized incidence ratio for non-Hodgkin lymphoma for individuals notified with hepatitis C virus (HCV). Stippled lines, 95% confidence interval.

Figure 6. The cumulative standardized incidence ratio (SIR) for liver cancer for individuals notified with hepatitis C virus (HCV), stratified by age at diagnosis of HCV infection. The upper curve is the SIR for individuals notified with HCV at age 45–60 years, the middle curve is for individuals notified at an age older than 60 years, and the lower curve is the SIR for individuals notified younger than age 45 years. In this figure, confidence intervals are omitted for visual clarity. At all time points, the SIR is significantly higher for the middle age group compared with the younger and older age groups.
Selection bias may be related to how individuals have been recruited into the cohort. In a stratified analysis, in which individuals were classified into groups according to age at diagnosis, a selection bias effect is present for only the 2 older age groups: 45–60 years and older than 60 years. In the younger age group, younger than 45 years, the SIR appears to be independent of choice of time of entry into the study. A plausible explanation is that younger people are more likely to be active intravenous drug users, a risk group for which screening programs are routinely conducted (14), but also that HCV-related liver complications are uncommon before the age of 50 years, that is, they are seldom the cause for an HCV test before age 45 years. In older age groups, diagnosis may sometimes be made through screening, but a larger number of patients are probably diagnosed because of emerging health care problems related to HCV infection.

Selection bias may also vary with calendar time for cohorts assembled over long time periods. Regarding HCV infection, diagnosis was introduced in 1990, and it is likely that screening and testing was more selective in the early 1990s compared with later years of the cohort.

In this paper, we have discussed selection bias in the context of prevalent cohorts of individuals diagnosed with infectious diseases. However, the proposed method should be applicable in any situation in which the probability of inclusion in a study cohort is related to an emerging outcome. The cumulative SIR may have advantages compared with an interval-based method, especially in cohort studies with small numbers of events.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Swedish Institute for Infectious Disease Control, Solna, Sweden (Anna Törner); Department of Infectious Diseases, Örebro University Hospital, Örebro, Sweden (Ann-Sofie Duberg); Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna, Sweden (Paul Dickman); and Department for Mathematical Statistics, Stockholm University, Stockholm, Sweden (Åke Svensson).

This work was financially supported by the Swedish Institute for Infectious Disease Control.

The authors thank Sean McGuigan for proofreading the manuscript.

Conflict of interest: none declared.

REFERENCES


Am J Epidemiol 2010;171:602–608


