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

Statistical methods for long-term follow-up of infectious diseases

Anna Törner

Karolinska Institutet

Stockholm 2014

Published by Karolinska Institutet. Printed by Universitetsservice AB.
Front cover by Anders Gunér Design 2014.

ISBN 978-91-7549-472-2

To my Mother Anna-Greta. I wish you were here.

Abstract

The overall aim of this work has been to investigate methodological issues connected to long-term follow-up of infections diseases. The work extended to prevalent cohorts in general. The common denominator for the main methodological efforts in these four papers is issues connected to selection bias. In the first three papers methods for visualizing selection bias in prevalent cohorts were explored and different approaches to adjust for this bias discussed. In the fourth paper, capture-recapture modeling was used to examine ascertainment level for liver cancer in the Swedish Cancer Register.

Study 1: In this study we investigated a novel approach to visualize and adjust for selection bias in prevalent cohorts. The method is an extension of the standard interval-based approach, where a risk estimate is calculated for disjointed time periods after inclusion in the cohort of interest. In the proposed method, observation time and events are cumulated, giving more power and more precise estimates which may be useful for studies with few events where it may be difficult to judge what is a true effect and what is random noise. The proposed method, cumulative SIR, is exemplified using data on hepatitis-C virus infection and the outcome liver cancer and non-Hodgkin lymphoma. The results using this novel approach were comparable to a standard approach with disjoint intervals. The results indicate that the method may be useful in situations with few events in the cohort. The method is only useful for cohorts where the risk of the studied outcome is fairly stable over time.

Study 2: Spurious observations have indicated that there may be a relationship between hepatitis C virus (HCV) infection and kidney cancer. In this study the relationship between HCV-infection and kidney cancer was investigated by use of disease registers. In addition the known association of HCV-infection and other forms of kidney disease was explored further. Methods for investigating selection bias explored in Paper I were used, in addition new ideas were investigated which were further developed in paper III. The relationship between HCV-infection and kidney cancer was not confirmed in this study, but the association of HCV-infection with other kidney-related diseases was investigated further.

Study 3: For cohorts that may have high hazard immediately after inclusion in the cohort, which then first decreases to later increase with follow-up time, the method of cumulative SIR must not be used. The cumulative properties will obscure the initial decrease and the method cannot give clear answers. In paper III we used restricted cubic splines to model to instantaneous failure rate (hazard). The shape of the hazard function may give an indication of the possible presence of selection bias in the cohort. The proposed method was exemplified using 1) data on HCV-infection where the outcome of interest was ‘kidney disease’ and 2) a cohort a patients with Monoclonal Gammapathy of Uncertain Significance (MGUS) and the outcome of interest ‘death’. The model was useful to study the shape of the hazard in the cohorts and the number of knots was adjusted to give a suitably flexible model, clearly showing the shape of the hazard without being too flexible.

Study 4: In this study we explored capture-recapture modeling, using a log-linear model to estimate ascertainment level of the Swedish Cancer Register (CR). We used a three-source model: CR, the National Patient Register (PR) and the Cause of Death Register (DR). Due to the limited degrees of freedom in available data, a full model can not be used. We chose to estimate a single two-way interaction between the most dependent registers (DR and PR) and a three-way interaction. This model will estimate the number of unreported cases of liver cancer to about 25% of the total number of cases in all three registers together, accounting for overlap. The analysis is likely to be biased by false positive cases identified in the PR and/or DR.

List of publications

I. Törner A, Duberg AS, Dickman P, Svensson A.
A proposed method to adjust for selection bias in cohort studies.
Am J Epidemiol. 2010 Mar 1;171(5):602-8.

II. Hofmann JN, Törner A, Chow WH, Ye W, Purdue MP, Duberg AS.
Risk of kidney cancer and chronic kidney disease in relation to hepatitis C virus infection: a nationwide register-based cohort study in Sweden.
Eur J Cancer Prev. 2011 Jul;20(4):326-30.

III. Törner A, Dickman P, Duberg AS, Kristinsson S, Landgren O, Björkholm M, Svensson Å.
A method to visualize and adjust for selection bias in prevalent cohort studies.
Am J Epidemiol. 2011 Oct 15;174(8):969-76.

IV. Anna Törner, Knut Stokkeland, Fereshte Ebrahim, Åke Svensson, Paul Dickman, Rolf Hultcrantz, Scott M Montgomery, Ann-Sofi Duberg.
Liver Cancer in Sweden. Combining information from different registers and using a log-linear model to estimate the true incidence of liver cancer.
Manuscript in preparation.

Table of contents

List of abbreviations

CR	The Swedish Cancer Register
DR	The Cause of Death Register
PR	The National Patient Register
HBV	Hepatitis B Virus
HCC	Hepatocellular Carcinoma
HCV	Hepatitis C Virus
MGUS	Monoclonal gammapathy of uncertain significance
SIR	Standardised Incidence Ratio

1. Introduction

The focus of this thesis is selection bias – how and why register based cohort studies may be biased because the cohort we study is not be representative of the population we are interested in. Usually a cohort or a register contains a subset of the population we are interested in, and different selection processes and mechanisms affect which population is actually captured by the register. There are almost no exceptions to this rule, even the birth register listing all live births in Sweden is thought to miss 0.5-3 % of all births (1). Other registers, which are thought to be complete, e.g. the cause of death register, may be complete in the sense that all or most deaths are captured, but instead the correctness of the recorded underlying causes of death may be questioned. The cancer register aims to be complete, and diagnosis of cancer is notifiable by law, but completeness may be questioned (2). The same is true for the National Patient Register, which is used routinely for all inpatient episodes; meaning that all patients who are hospitalised are recorded in this register, but the validity of the recorded diagnoses may sometimes be questioned

2. Cohorts for research

The methods discussed in the papers for this thesis have utilized epidemiological cohort data from a number of registers in Sweden. These registers are described briefly below.

HCV-cohort

For a number of infectious diseases in Sweden, it is mandatory to report newly diagnosed cases, which are recorded in national registers kept by The Public Health Agency of Sweden. Hepatitics C virus infection is among those infections that are notifiable by law. HCV was detected in 1989 and between the years 1990 and 2006, 43,000 cases of HCV-infection were reported. Before 1990, HCV was not diagnosed. Since HCV-infection often is asymptomatic or mildly symptomatic during long periods of time before being recognized, it is likely that a large number of individuals with HCV are not identified at all and also that diagnosis sometimes occurs late in the course of infection. HCV infection has been linked to cancers such as non-Hodgkin lymphoma, liver cancer and also kidney diseases (non-cancer) (3-5).

MGUS-cohort

Monoclonal gammopathy of uncertain significance (MGUS) is a condition in which an abnormal protein (monoclonal protein or M protein) is increased in the blood. Some patients with MGUS may progress to malignant disease and it is likely that mortality is elevated in this patient population (6). Since MGUS is a silent asymptomatic condition, it is likely to be diagnosed in conjunction with medical problems emerging from MGUS or other unrelated medical problems. The studied cohorts, who have been gathered in hospital clinics, are therefore not likely to be representative of a true cross-sectional selection of patients with MGUS.

The Swedish Cancer Register

The Swedish Cancer Register (CR) was founded in 1958. It is compulsory for every health care provider to report newly detected cancer cases to the registry. A report has to be sent for every cancer case diagnosed at clinical, morphological or other laboratory examinations as well as cases diagnosed at autopsy, i.e. dual reports in most cases. The basis for diagnosis is recorded. The CR does not receive information from death certificates for the purpose of ascertaining new cases. A quality study of the CR estimated the underreporting at approximately 4 percent and found the degree of underreporting was dependent on the cancer site, increased with age and diagnoses without histology or cytology were overrepresented (2). The liver and pancreas are among the cancer sites with a higher degree of underreporting.

The Cause of Death Register

The Cause of Death Register (DR) includes all deaths among Swedish residents, relies on information from death certificates (missing in only 1-2% of deaths). In the DR, the date and cause of death (underlying and contributing) and the basis for the diagnosis are recorded.

The National Patient Register

In the 1960's the National Patient Register (PR) started to collect information regarding inpatients at public hospitals. From 1984 the participation was mandatory for all county councils and from 1987 the PR includes all inpatient care in Sweden. Every inpatient episode is recorded with dates and diagnoses at discharge. The validity has been

reported to be high for most inpatient diagnoses. Patients who have never been inpatients will not be in this register. In this study we did not use the outpatient register.

Diagnoses in these three registers are coded according to the International Statistical Classification of Diseases and Related Health Problems (ICD). The cancer diagnoses in the Cancer Register are available as ICD-7 codes for the whole study period, 1975–2011. The diagnoses in the Patient Register and Death Register were classified according to ICD-8 in 1968–1986, ICD-9 in 1987–1996, and ICD-10 from 1997 and afterwards.

Statistics Sweden

Statistics Sweden keeps census data for the full population, including migration, demographic and economic data. All Swedish citizens have a unique 10-digit personal identification number (7). This identification number may be used to uniquely connect entries in different register. It may also be used to retrieve general information on demographic characteristics and migration from Statistics Sweden.

3. Selection bias

Selection bias is a concern in any cohort where the individuals in the cohort are not representative of the cohort of interest (8). Figure 1 below describes the relationship between a source population and a study population. A) The source population is the total population of all individuals with the condition of interest, diagnosed and undiagnosed. Nested within the source population is B), the diagnosed cohort: all individuals diagnosed and included in the register or cohort. Finally, nested within B) is C), the study population which is often smaller than the diagnosed cohort, because for example, individuals are excluded to avoid selection bias. The study population is seldom a random selection of the source population, which means that the possibility for selection bias must be considered (6).

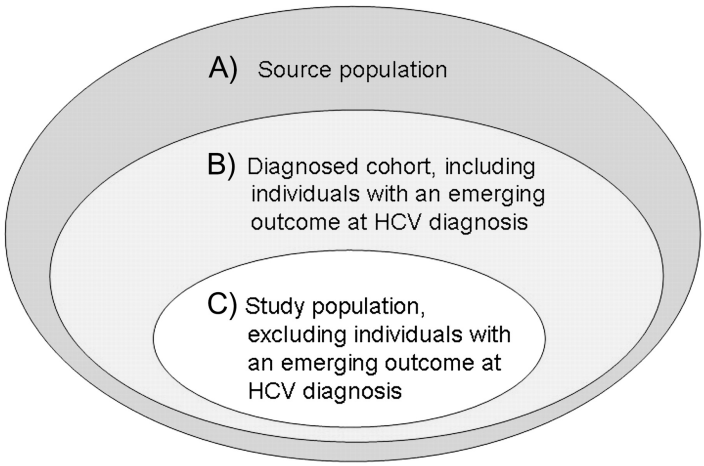

Figure 1. An example showing the relationship between source population, diagnosed cohort and study population for patients with HCV-infection.

Selection bias is a concern in both incident and prevalent cohorts, but the mechanisms may be different and also adjustment for this bias needs to be addressed in different ways. In this thesis, only selection bias for prevalent cohorts is discussed.

Prevalent versus incident cohort

A *prevalent* cohort may be described as a cohort where individuals are sampled according to a cross-sectional sampling criterion (9). This means that the first period of the disease is usually unobserved. The length of this unobserved period may differ between individuals. In contrast, an *incident* cohort is a cohort where individuals are included at the onset of disease and followed forwards. For some diseases the onset is asymptomatic or very mildly symptomatic. These diseases are rarely diagnosed at onset but rather at some later time point; either through screening programs or when the disease is reaching a symptomatic stage or as complications to the condition arise or as a coincidental finding when patients seek medical attention for other, unrelated health care problems.

The possibility for *selection bias* is obvious when the underlying disease develops under long time periods and patients reaching a symptomatic stage of the disease have a higher probability of being diagnosed with the disease. When the probability of diagnosis of the initial disease, and thereby inclusion in the cohort, is related to the risk of the studied outcome, the cohort may be affected by *selection bias.*

The term prevalent cohort was used frequently in the early HIV/AIDS era to discuss and model scenarios for HIV developing into AIDS (10-12). For the cohorts and

conditions studied in this research project; HCV and MGUS, the terms prevalent cohort and the theory surrounding prevalent cohorts are equally valid since these conditions are rarely diagnosed at the onset of disease. Both MGUS and HCV are asymptomatic or mildly symptomatic before eventually turning into a more severe disease (13).
When the first time period of the disease or infection is unobserved in a prevalent cohort, this may lead to additional challenges. Usually, the length of the unobserved period of the initial disease varies between individuals and the length of this period is usually unknown. Sometimes the length of the unobserved period may be related to some specific covariate and if this specific covariate is not taken into account this may result in a specific form of bias called onset confounding (10). The first unobserved time for individuals in prevalent cohorts is termed backward recurrence time (14). An additional challenge is left-truncation, i.e. patients may have experienced the outcome of interest, i.e. HCV-infection followed by liver cancer before being included in the cohort (15).
If the length of backward recurrence times are ignored in epidemiological studies, it would mean that all individuals are evaluated on the same time scale; follow-up time. If the interest is focused on the natural course of disease, ignoring left-censoring and left-truncation will give non-informative risk estimates. Sometimes covariates, such as CD4-count in HIV infection, may give information on the duration of the unobserved period of disease. If all information about the length of the unobserved time period could be captured by covariates, left-censoring of data would not be important, the information would be captured by the covariate.
Another way to capture information on the length of the unobserved time period of disease is to use available epidemiologic information to model the unobserved time period (3, 4). For our previous work on HCV-infection and non-Hodgkin lymphoma and liver cancer respectively, we constructed a model taking into account development of the HCV-infection epidemic in Sweden, age and route of transmission of infection. The subsequent analysis may then use this information, either to stratify results by prior time of disease or as part of the statistical analysis. Some statistical methods are better suited to handle left-truncated data (14).
The presented methods to investigate and adjust for selection bias in this thesis may be applicable to a wider range of cohorts. In any situation where the method or process for gathering (identifying) members of the cohort is positively correlated with the risk for the studied outcome, the proposed methods may be applicable. This means that in a situation where there is a common cause for both the studied outcome and probability

of inclusion in the cohort, there is a possibility for selection bias. However for conditions and diseases where the condition confers an immediate risk of the studied outcome, like influenza or meningitis, a higher risk of the studied outcome (e.g. death) early after inclusion in the cohort may not be described as selection bias and the methods discussed here are not applicable. These are incident cohorts and selection bias must be handled in other ways.

4. Correcting for selection bias

Selection bias in prevalent cohorts is well known and a method, which is sometimes used to adjust for selection bias, is simply to remove some observation time (months to years) and any events occurring during this first time period of observation. When the cohorts are large with many events, this is a reasonable approach.

A simple method to investigate selection bias in cohort studies is to calculate effect estimates for several disjoint time periods after inclusion in the cohort (16). If the planned statistical method for the analysis is Standardized Incidence Ratio (SIR), this SIR is calculated for several separate time periods from the start of the observation time. This means that one estimate is calculated for the first 6 months the individuals (patients) are members of the cohort, another SIR estimate for month 7-12, month 13-18, month 19-24 etc. If these estimates are higher (not necessarily statistically significantly) during some of the earlier time periods compared to later time periods, this may indicate selection bias of the type discussed here. The judgment is informal and from a statistical point of view, the effect estimates for the time periods are very seldom statistically significantly different due to low power. Judgment is based on observing the point estimates and using medical knowledge and information on how the cohort has been gathered. After removing person years and events occurring during these early time periods, the analysis is conducted as planned.

The increase in the effect estimate during the first few time periods may be due to selection bias, but there are also other possibilities. Another explanation may be surveillance bias; an observed higher incidence of the studied outcome during the first time period after inclusion in the cohort due to increased surveillance for the studied outcome. An increased surveillance for the studied outcome may lead to an initial higher incidence of the studied outcome and the incidence will then decrease as the initial depletion of outcomes takes effect. If the true incidence is increasing with follow-up time in the cohort, this could theoretically give a “double peak” in the

studied outcome. For the outcomes studied in the two papers on selection bias in this thesis, non-Hodgkin lymphoma, liver cancer, death and hospitalisation due to kidney related disease, surveillance bias has been assumed to be negligible (17, 18). The reason for this is the severity of the studied outcomes, making it very likely these outcomes would be diagnosed approximately at the same time, regardless of diagnosis of the underlying condition.

The distinction between selection bias and surveillance bias may be subtle. The selection bias described above may also be understood as a surveillance bias if the focus is diagnosis of HCV-infection or MGUS. Patients presenting with medical problems, maybe emerging liver problems or symptoms relating to a failing immune system, will be subject to further medical examinatons and an HCV-infection or MGUS is more likely to be diagnosed for these patients compared to patients who are well. If one chooses this perspective, this bias may be described as surveillance bias, i.e. increased surveillance for HCV/MGUS. In the projects presented in these papers, we choose to view the problem from another perspective, and that is how representative the cohort is for the source population. Which terminology is the most appropriate to describe a bias not always clear and another expression which is sometimes mixed up with selection bias is confounding by indication (19). What matters is to understand the mechanism of bias and to be able to describe how this bias arises and to understand how this bias may be avoided or adjusted for.

5. Selection bias using casual inference and DAG-theory

In discussions of confounding and selection bias, directed acyclic graphs (DAGs) may be used to graphically describe and analyse how selection bias enters a cohort study (20). Possible mechanisms for selection bias in prevalent cohorts is described in Figure 2 using DAGs.

The objective is to study if the asymptomatic or mildly symptomatic disease (X) causes the outcome (Y), this is denoted X → Y. The disease X may cause Y either directly or by first developing into a more severe symptomatic stage (S) before progressing to Y. Only patients diagnosed with the disease (X) can enter a diagnosed cohort (D), which forms the basis for the cohort study. Since the disease is asymptomatic or mildly asymptomatic, not all patients with disease X are diagnosed. Patients who have entered a symptomatic stage of the disease (S) will more easily be identified and diagnosed as

having X. This increased probability of diagnosis of patients with symptomatic disease will lead to selection bias; that is, patients in the diagnosed cohort D will not be representative of the total study base (source population). Patients with a more severe stage of disease X, already progressing towards Y, will be over-represented among the diagnosed. In the situation described in the directed acyclic graph, we are interested in the relation whereby X causes Y (X→Y), either directly or over the symptomatic stage S . Conditioning on D, that is being diagnosed with the condition X, which we must do because we may only study patients who have been diagnosed and are part of the diagnosed cohort, opens a path from X to Y over the inverted fork D, leading to the selection bias described. Quantitative estimates of the risk of Y for patients with the disease X, may be biased toward higher estimates because the population in D is not representative of the total population with the disease X. Other unknown or unmeasured common causes (here, named U) that both cause the outcome Y and lead to diagnosis of X (i.e., D = 1) may introduce further selection bias into the analysis. An example of an unknown, unmeasured confounder would be in the MGUS-example presented in paper 3, where patients with a severe illness are more readily diagnosed with the condition X (MGUS) and also have and increase risk of dying, which was the study objective: to investigate if MGUS patients have increased mortality. Estimates of mortality for the MGUS-cohort may therefore be biased towards showing excessively high mortality.

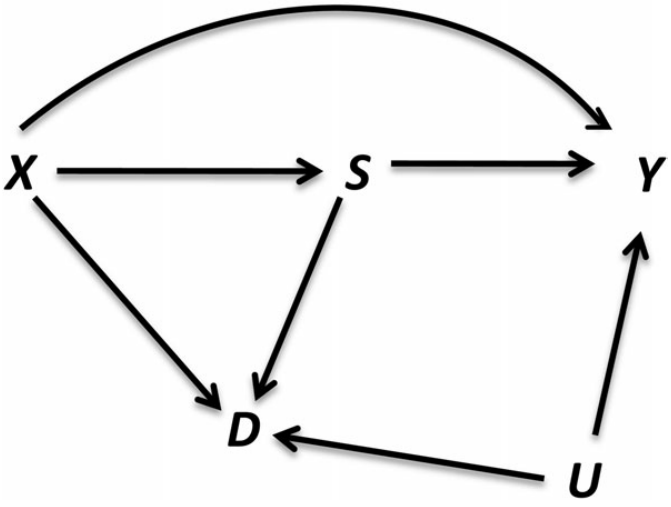

Figure 2. Description of selection bias using Directed Acyclic Graphs.

The processes described here should ideally be viewed as processes over time, something that may be quite complicated by using causal inference and directed acyclic graph theory. Selection bias caused by diagnosis of patients who have reached a symptomatic phase of the disease (S) may be dealt with by excluding patients who very soon after diagnosis of X reach the outcome Y. In practice this means that we choose to disregard the first time period (and outcomes Y) occurring soon after diagnosis of X. This will weaken the path between S and D. When the common cause U leading to the outcome Y and diagnosis of X (i.e. D=1) is an acute illness, the selection bias effect of this common cause is likely to be more pronounced just after diagnosis of the acute illness.

Theoretically we could close the backdoor path by either conditioning on S (conditioning will close this node) or by *not conditioning* on D (this will close an inverted fork). However, we may only study patients who are actually diagnosed, i.e. D=1, and there is no straightforward way to correct for S because it is only possible to study individuals that are actually diagnosed. A better approach is to weaken the arrow between S and D by removing time and events after diagnosis of X as described above. If this arrow is weakened or abolished by removal of observation time and events, the backdoor path over S is closed.

6. Capture-recapture models to quantify ascertainment level in registers

Introduction

Capture-recapture modelling was originally known from ecological research to quantify wild-life population sizes. The principle is easy and the basic idea is to capture a number of animals in a defined area/population, mark these animals in some way, let them go, and then make a second capture. Based on the information of the number of marked animals in the second capture, it is possible to estimate the total population size. This simple model assumes independence between the two sources (captures) of animals. In medical epidemiological research the situation is more complex, partly because independence between sources cannot be assumed and also because other factors beyond statistics may influence the validity of the results to an even greater extent. The capture-recapture methods have been extended to handle more complex scenarios with dependent sources and covariates. These methods are now used in epidemiology to estimate ascertainment level in medical registers (21, 22).

For a two-source model as described above, the data that will be available for modelling has three degrees of freedom and for this reason independence between the two sources is assumed; there is not enough information to model a possible dependence between sources (23). For models using three or more sources, more information will be available and it is possible to model dependence between sources. In the capture-recapture modelling presented in this thesis (paper IV) we have used the following sources (captures); the Swedish Cancer Register (CR), the National Patient Register (PR) and the Cause of Death Register (DR) in an attempt to estimate the true liver cancer incidence in Sweden and to explore the usefulness of capture-recapture models in this setting. These sources are clearly not independent as reporting into the various registers are highly dependent processes, e.g. a cancer being diagnosed in the hospital during an inpatient episode will increase the probability of this cancer also being reported to the CR. Further, a patient with a liver cancer reported to the CR who later dies is more likely to be recorded in the DR, with liver cancer either as underlying cause of death or contributing cause of death, compared to a patient with an undiagnosed liver cancer.

Models for capture-recapture

The full model

Capture-recapture modelling may be done in different ways, but a simple straightforward approach is log-linear modelling (24). A full (saturated) log-linear model for a three source model would be on the following form where the three sources are named A, B and C. The model will need coefficients for the three separate sources (β), there are also three separate two-way interactions (γ), one three-way interaction (θ) and in addition the intercept (α); in total 8 parameters.

$$\log N = \alpha + \beta_1 \cdot A + \beta_2 \cdot B + \beta_3 \cdot C + \gamma_1 \cdot AB + \gamma_2 \cdot BC + \gamma_3 \cdot AC + \theta \cdot ABC$$

Data from three different registers as described above and the information on which cases are registered in two or three of these registers will provide 7 degrees of freedom (number in each separate register, number in all combinations of two registers and number of cases present in all three registers). This means that a fully saturated model is not possible to fit, at least one of the parameters in the model needs to be omitted. In

some previous studies, the three-way interaction has been set to zero (22, 25-27). This may be a major assumption. When the three-way interaction is set to zero, the intercept is simply adjusted to fit with three two-way interactions and a no three-way interaction. For our modelling, all two-way interaction terms have negative coefficients since the registers in our example are all positively correlated, i.e. the overlap is greater as would be expected for independent sources. Therefore, in our modelling, to accommodate three negative two-way interactions, the intercept will take a large positive value when the three-way interaction is fixed to zero. The number of "missing" cases, i.e. cases captured by any of the three registers, is estimated as exp (α), and therefore an inflated intercept will give a high number of unreported cases. Also, we have no means of evaluating model fit since we cannot check the assumption of a zero three-way interaction (28). A log-linear model may also include covariates, which may to some extent compensate for dependence between sources (21).

Modelling options

Since a full model including a three-way interaction is not possible, we explored the following modelling options:

A. A two-source model, created by joining the two most dependent sources (registers).
B. A three-source model with three separate two-way interactions and no three-way interaction.
C. A three-source model with a single two-way interaction for the two most dependent registers (sources) and a three-way interaction.

Model A

Joining the two most dependent source into one source is a useful approach if two sources are heavily dependent and if independence between this joined source and the remaining source may be assumed (29). Of the models discussed here, model (A) would be expected to be slightly conservative since the process of joining two sources and then assuming independence would be expected to underestimate the number of unidentified cases. Model A may therefore be considered to provide a lower bound for the number of unascertained cases (30). By positive dependence we understand that the sources overlap to a greater extent than would be expected from independent sources. If we fail to compensate for a 'net positive dependence' we risk underestimating the

number of undiagnosed cases and also vice versa; if we fail to compensate for 'net negative dependence' we may overestimate the number of undiagnosed cases (31).

Model B

In our case, model B turns out to give a very large number of unreported cases, but it is not possible to say that a model without a three-way interaction will always over-estimate the number of unreported cases. In a full log-linear model as described above, the three-way interaction doesn't have an intuitive value or interpretation. Therefore, when fitting a model with a three-way interaction set to zero, we have limited means of investigating if this assumption is reasonable, it may be difficult to know if it is conservative assumption or if we risk overestimating the number of unreported cases. As described previously the intercept will be adjusted to account for three separate two-way interactions, which in the application investigated for our paper will inflate the intercept and give a very high number of unascertained cases.

Model C

Model C, which was the model we chose for our application, involves fitting a two-way interaction for the two most dependent sources and a three-way interaction. For the studied application of liver cancer, the overlap between the PR and DR was the greatest overlap and therefore this interaction term was estimated. The suggested model takes into consideration the strongest dependence between two of the sources, while ignoring the weaker dependencies. Therefore, all other assumptions being valid, this model should also be slightly conservative, i.e. underestimate the number of unreported cases. The most important feature of model C compared to model A and model B is that a three-way interaction is estimated. Also other models involving a three-way interaction are possible to fit. An obvious example would be a model with two separate two-way interactions and a three-way interaction (in addition to an intercept and a coefficient for each source). Such a model will give a slightly higher estimate of unreported cases compared to a model with one two-way interaction. The reason for our preference for the simpler model with one two-way interaction is simplicity and the desire to choose a slightly conservative model. Most dependencies will be captured by one two-way interaction and a three-way interaction. To the non-statistician, the concept of adjusting for the two most dependent sources should be straightforward. From a methodological perspective we consider this model to be superior to the model with the three-way

interaction set to zero since the assumptions we are making may be more readily examined.

Capture recapture models – assumptions

There are a number of underlying assumptions and conditions for capture recapture estimates to be valid (32).

1) Firstly, the cases need to be defined and ascertained in the same way for all sources. This is an important condition, which may be violated for epidemiological studies employing different sources where it is possible that the diagnosis was made in different ways and where bases for diagnosis differ between sources.
2) Within a source, each case needs to have the same probability of being diagnosed by that source. However, all sources do not have to capture the same number of cases or have the same probability of capturing a case as another source. In formal language, the requirement is homogenous catchability.
3) The methods are applicable for closed populations. If individuals move in and out of the cohort, they are available to be captured for shorter time periods, implying they will have a lower chance of being captured since they are available in the cohort for a shorter time period, i.e. a special case of variable catchability as described under 2).
4) For epidemiological applications the time window must not be to short, the 'case' needs time to be available for catching by the different sources (31). This fourth criterion aligns with criteria 2) and 3) concerned with variable catchability.

Ascertainment of cases

In a register-based study, identifying cancers from three different registers, it is evident that the bases for diagnosis between these sources may differ considerably. Validity of data is important in capture-recapture modelling and there should not be misdiagnosed cases in any of the sources. Misdiagnosis, false positive and false negative cases, may lead to over- or under estimating the number of missing cases (33, 34).
For the registers studied in paper four there were 13,749 patients with a diagnosis of liver cancer in the three registers. Of these 13,749 patients, 6,439 were reported to the

CR and 12,475 patients were reported to the PR and/or DR. A large number of patients were reported to the PR and/or DR with liver cancer, but not to the cancer register. One reason for the large number of patients reported to the PR but not to the CR with liver cancer may be the increasing use of non-invasive imaging methods such as CT and MR for diagnosis. Doctors may be reluctant to report a liver cancer diagnosis to the CR without PAD, or it may simply be forgotten. Another possible explanation may be that some of the liver cancer diagnoses reported to the PR were not so certain. It may be that liver cancer was the diagnosis available for the PR at that time, but later when more information was available, either retrieved from earlier records or complimentary investigations, the final cancer diagnosis reported to CR was a different cancer diagnosis.

This hypothesis is supported by the fact that a large proportion of patients with liver cancer diagnosis in the PR were reported to the CR, but with other cancer diagnoses, either occurring before the current diagnosis of liver cancer in the PR or after. Of the 7,310 patients with liver cancer reported to the PR and/or DR, which were not reported to the CR, 3,925 of these (54 %) were found in the cancer register with some other cancer diagnosis, reported to the CR either before the liver cancer diagnosis or after. About 1,390 of these patients were reported with ICD-code 199, cancer at unspecified site. It is plausible that a proportion of the liver cancers in the PR and/or DR are not liver cancer, but metastases to other cancers. Also, some patients who were reported to the PR with liver cancer in conjunction with hospitalisation were later reported to the CR with other cancer diagnoses. The large proportion of patients reported with liver cancers in the PR and some other cancer diagnosis in CR, indicate that this was an important source of bias. Since the capture-recapture modelling relies on the overlap (and non-overlap) of registers, a large number of liver cancers in the PR and/or DR that are not true liver cancers will seriously inflate the estimate of unascertained cases. The present results indicate that the assumption of equal criteria for diagnosis/reporting of a case may not have been fulfilled.

Variable catchability and closed population

An underlying assumption for capture-recapture models to be valid is the assumption of homogenous catchability (32, 35). In essence this means that all cases (cancers) should have equal probability of being captured by each source (register). Each register doesn't have to "catch" an equal number of cases, but each case should have equal

probability of being captured by a given source. This condition is obviously violated for register-based studies, e.g. all cancers are not fatal, meaning that all cancers cannot be captured by the DR. Also, it may be easily shown that patients with liver cancer in the PR and/or DR only, differ with regards to demographic characteristics to patients reported to the CR, one example being age; patients in the CR were in our study an average of 69 years at the time of reporting of liver cancer while patients with liver cancer in PR/DR only, were on average 7 years older. In modelling, if we ignore variable catchability, this will give a conservative bias, i.e. we will underestimate the number of unreported cases (28, 36). Variable catchability may be accounted for, at least to some extent, by including covariates in the log-linear modelling. However, in register-based research, most of the interesting covariate information, such as disease severity, will not be available. In our modelling we have chosen to ignore variable catchability. The justification for this is that 1) the bias is conservative (which is more acceptable than non-conservative bias), 2) most covariates that would be of interest to adjust for catchability are not available and finally; 3) other sources of bias a likely to play a larger role.

7. Conclusion

The papers dealing with selection bias are just scratching the surface (17, 18). Selection bias related to study populations not being truly representative for the intended cohorts is probably part of most epidemiological research projects. Refining methods for analysing and presenting data in epidemiological research can never fully compensate for systematic errors in the form of bias. Systematic errors may not be analysed away using large samples, we need to deal with these sources of errors with more intelligent tools. At the start of this research project, the approach for the three methodological papers was aimed at statistical approaches and the objective was to explore useful methodologies. Working with these projects made it clear that in depth understanding of data and how data has been collected is the basis for an adequate analysis. Selection bias may never be interpreted by simply looking at curves in graphs. We need to understand the characteristics of the population and the unique features of the disease and mechanism whereby patients are identified. Therefore, there is no one solution to analyse and adjust for selection bias. Probably some of the most useful tools are graphs of hazards, DAG and other visual methods, whereby we can understand the

mechanisms for how bias may enter the cohort. By understanding we may adjust our analyses to compensate for this bias.

Regarding the capture-recapture modelling for epidemiological research using disease registers, there are some papers focused on modelling and comparing models by statistical criteria. In our work it became clear that some of the basic conditions for the use of these models might not be fulfilled, at least not for the Swedish registers. If we cannot be sure that we don't have large numbers of large false positives in one of the registers, adding log-linear modelling to estimate ascertainment level makes little sense. Again, understanding data and the basis for collecting and interpreting data is the fundament for an intelligent analysis.

Equally important is to challenge the models and the assumptions we make. For any model we chose to use, we should question this model; by considering alternative models (which then preferably should give a similar result) and by examining assumptions. For the selection bias papers (17, 18) we had the possibility to compare different approaches, which gave concordant results. For capture-recapture models we may examine the validity of the model by comparing with simpler models providing lower bounds for the number of missing cases. Regarding the validity of underlying assumptions these assumption may often be clarified by examining data in more detail. For the capture-recapture models discussed in the fourth paper, there are reasonable solutions for modelling that can handle interactions and covariates. However, for the issues related to differences in case ascertainment by different sources (registers) there are no clear-cut solutions that can adjust for this. To adequately analyse data we must truly understand data.

Sometimes questions are as important as answers. The methodological papers in this thesis probably create more questions than providing answers. However, questions are in some way also answers, if we know enough to question processes, such as selection of individuals into cohorts, we can interpret results in light of these uncertainties and limitations.

8. Summary of papers

Paper I

Motivation

The idea behind the proposed method in the first paper was to improve the "interval based method" for investigating selection bias in prevalent cohort studies. The effect measure investigated was Standardized Incidence Ratio (SIR). A common method to investigate possible selection bias in a cohort is to calculate SIR for several disjoint intervals, where follow-up time and events are divided by follow-up time from inclusion in the cohort. If the SIR estimates for the first intervals after inclusion are significantly higher (also by the non-statistical meaning of the word) these time periods with higher effect estimates are discarded from analysis. It is difficult to give guidance on what is "higher", but generally in prevalent cohorts, there is no obvious reason for why the risk of the outcome interest would decrease immediately after inclusion in the cohort. The observed risk estimates needs to be judged in light of the known characteristics of the medical condition. However, often the number of outcomes in each disjoint interval is low (unless very common outcomes are studied). This means that the precision and power in each interval is poor with regards to judging what is a truly elevated effect measure and what is random noise. The idea in the first paper was to develop this interval-based method into a cumulative measure that would have better power to reveal how long time periods after inclusion in the cohort that should be removed to avoid selection bias. The idea was also to visualize more clearly what is the impact of removing a certain time period and the events associated with this time period.

Material

To illustrate the use of the proposed method we used data derived from the Swedish Institute for Infectious Disease Control and the Swedish Cancer register. All individuals with a reported HCV-diagnosis during the time period 1990-2006, after exclusion of patients with concomitant HBV-infection, were connected by use of the unique personal identification number to the Swedish Cancer register. All patients with a diagnosis of liver cancer or non-Hodgkin lymphoma were identified. Data from Statistics Sweden was used to correct the cohort for death and emigration. Standardized

Incidence Rates were calculated using age-, sex- and calendar year specific incidence data for the general population from the Cancer Register.

Methods

The proposed method of cumulative SIR is based on estimating SIR for the cohort repeated many times, removing one day at a time. First the SIR is estimated for the full cohort, follow-up starting at day of inclusion in the cohort. In the next step SIR is estimated starting from day 2, i.e. the first day with "person years" and events occurring during this first day is removed. This procedure is repeated by removing one additional day (with person years and events) for each reiterated calculation. This procedure is repeated a large number of times, e.g. 730 days (2 years) and the resulting estimates are plotted in a graph against time since inclusion in the cohort, see Figure 3 below showing cumulative SIR for liver cancer in a HCV-cohort. If the estimated SIR decreases with time since inclusion in the cohort, this may indicate selection bias. In a prevalent cohort where time of inclusion in the cohort is not related to an emerging outcome, no specific decrease in hazard immediately after inclusion in the cohort would be expected.

Based on the graph, a suitable cut-off is chosen, where the estimated SIR has levelled off to a constant level. If the risk of the studied outcome increases considerably with time, the method of cumulative SIR may be difficult to use. The reason for this is that this method cumulates observation time and events occurring from day (i) and forward. If the risk is higher for long follow-up times, this higher incidence will obscure the initial decrease in the SIR estimate. This method is therefore only applicable when the SIR levels off to some constant value after the selection bias has worn off.

Results

The proposed method was illustrated using a cohort of patients notified with HCV-infection. The studied outcomes were liver cancer and Non-Hodgkin lymphoma, one frequent cancer (liver cancer) in this patient population of HCV-infected and one not so frequent cancer (non-Hodgkin lymphoma) in this cohort.

Liver cancer

For liver cancers the number of observed cancers is very high and the calculated SIR was in the range SIR=33-40 depending on how much observation time (and events)

after inclusion in the cohort was removed, Figure 3. The estimated SIR was high immediately after inclusion in the cohort and then decreased during the first year down to approximately 34. This indicated that at least a year of observation time and events after inclusion in the cohort should be removed to correct for selection bias. Since the number of observed events was high, removing observation time and events had little impact on the precision of the final results, i.e. width of confidence intervals. SIR was estimated to around 34 if at least a year was removed.

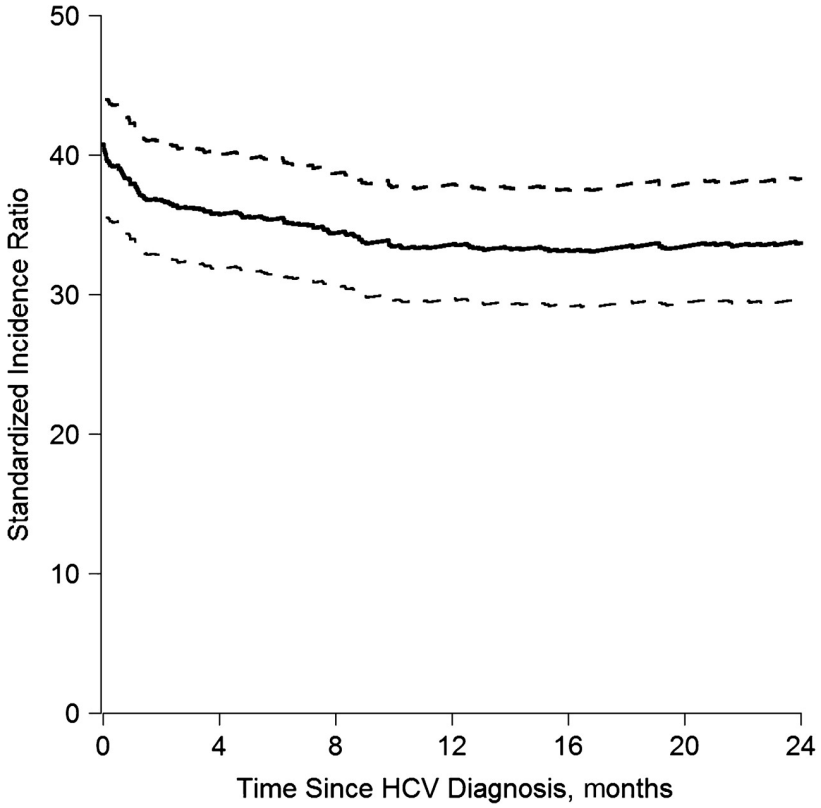

Figure 3. The cumulative standardised incidence ratio for liver cancer for individuals notified with hepatitis C virus (HCV). Dashed lines, 95% confidence interval.

Non-Hodgkin lymphoma

Non-Hodgkin lymphoma is a rare cancer in the HCV-cohort, compared to liver cancer, and in a previous study investigating if HCV may increase the risk of non-Hodgkin lymphoma, the number of cancers was just sufficient to show this relationship with statistical significance (3). In this situation it was difficult to motivate removing long time periods and many observed events, unless it really could be shown that selection bias was a problem. Also from a medical perspective, these cancers occurring very soon after HCV-diagnosis are likely to be related to the HCV-infection, but the only

reason these patients were diagnosed with HCV-infection, and included in the cohort, was that HCV was diagnosed as a consequence of symptoms of the emerging outcome (cancer). For the data of non-Hodgkin lymphoma, SIR was calculated for disjoint intervals after inclusion in the cohort showed high SIR immediately after inclusion in the cohort, but the SIR varied between intervals, and it was difficult to set an absolute cut-off, Figure 4. The cumulative SIR shows that SIR was slightly elevated if no time after inclusion in the cohort was removed, but the impact of selection bias was small if 3-6 months of observation time and events were removed, Figure 5.

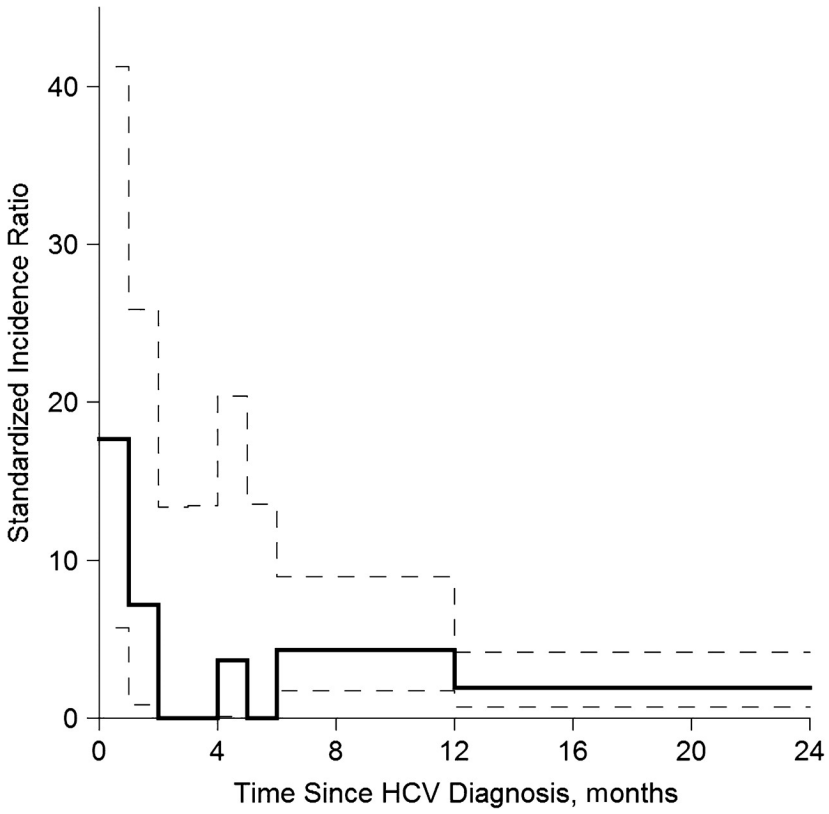

Figure 4. The standardized incidence ratio for non-Hodgkin lymphoma for individuals notified with hepatitis C virus (HCV). Interval based method, in which the standardised incidence ratio is calculated separately for different time intervals after diagnosis of HCV. Dashed lines, 95% confidence intervals.

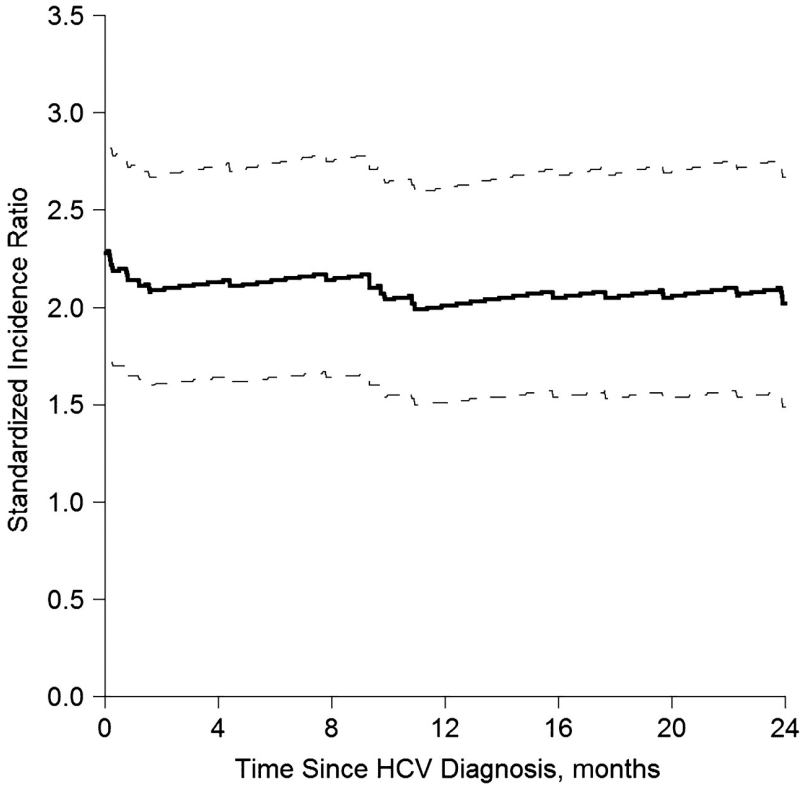

Figure 5. The cumulative standardised incidence ratio for non-Hodgkin lymphoma for individuals notified with hepatitis C virus (HCV). Dashed lines, 95% confidence interval.

Conclusion

The examples of HCV-infection and liver cancer and lymphoma show that the cumulative method works and that for some scenarios it may have advantages over a standard approach; calculating SIR for disjoint intervals. This is especially true if the number of events are few and it is therefore important not to discard observation time and events unnecessarily.

The impact of removing a specific time period and events belonging to this time period may be read directly from the graph.

However, the proposed method of cumulative SIR has several important limitations. Firstly, this concept of selection bias as it is discussed here is only relevant for *prevalent cohorts*, i.e. when inclusion of the patient in the cohort is not directly connected to the time point when the patient contracts the infection of the disease which is the denominator for the cohort. HCV-registers are good examples of prevalent cohorts. The infection may be symptomatic at infection and the patient may carry the infection for years or decades before finally being diagnosed, either in conjunction with random screening, because of other medical problems or because the patient is presenting with an emerging outcome of the infection, i.e. liver cancer. For incident

cohorts, e.g. the study of influenza and mortality, decreasing hazard after inclusion in the cohort may have other more likely interpretations such as declining risk of death with time.

Another important limitation of the method is that use of the method assumes that the hazard is high immediately after inclusion in the cohort (selection bias) and then declines to a steady state level, or at least that the hazard is increasing very slowly with time. If the hazard increases steeply with follow-up, the graphing will not work, as increasing hazard with follow-up will affect the early parts of the graph, because data is cumulated. It is therefore important to map out how the overall hazard for the outcome of interest appears in the cohort before applying the proposed method.

Paper II

Motivation

Chronic hepatitis C virus (HCV) infection is an established cause of liver cancer, and studies have also suggested a link with kidney cancer (4). HCV-infection has also been shown to increase the risk of other kidney related diseases and this association was further explored in this paper.

Material

The Swedish nationwide register of HCV-infected individuals during the years 1990-2006 compromising 43,000 individuals was connected to the Swedish Cancer Register to identify patient with kidney cancer diagnoses (ICD7 180.0 and 180.9). Data on hospitalisation discharge diagnoses were extracted from the National Patient Register, using the following ICD-codes:

1) Glomerular diseases (ICD-9 codes 580–583, ICD-10 codes N00-N08)
2) Renal failure (ICD-9 codes 584–586, ICD-10 codes N17-N19
3) Other kidney diseases (ICD-9 codes 587–589 and 593, ICD-10 codes N25-N28).

Events were defined as the first hospitalisation per subject for which each of the listed conditions above were registered as the principal diagnosis code.

Statistics Sweden provided information on death and date of emigration (if applicable).

Methods

The Standardized Incidence Ratio (SIR) was calculated using the HCV-cohort linked with the Cancer Register and the data on the incidence of kidney cancer in the general

population. Indirect standardization was performed with regards to calendar year, age (5-year intervals) and sex. To investigate and adjust for selection bias, the cumulative SIR method was used and SIRs were calculated with varying lag times (no lag-time up to two years) (18).

For the Cox-regression, a control cohort matched for age, sex and county of residence at time of HCV-infection was used. For each HCV-infected individual, five controls were selected.

To avoid selection bias, subjects were excluded from these analyses if they had ever been hospitalised with kidney disease as the principal diagnosis from 1969 until one year after HCV-notification or the corresponding reference dates for non-HCV-infected subjects.

Subjects who were hospitalised for any condition in the year prior to HCV-notification were also excluded, leaving 25,412 HCV-infected subjects and 198,124 controls in these analyses.

Restricted cubic splines were fitted to investigate the change in hazard over time in the cohort. These analyses revealed a fairly sharp decrease in hazard over the first year and then a slowly decreasing hazard from one year onwards (37). This slowly decreasing hazard was judged to reflect frailty in a population consisting of patients with different underlying intensity for hospitalisation for kidney-related disease. Conducting an analysis based on time to first event will gradually attenuate the population with regards to patients with high propensity for hospitalisation, and the hazard will therefore decrease slowly with time. To further control for selection bias, related to the sharp decrease in hazard during the first year, follow-up started one year after notification of HCV-diagnosis, Figure 6, paper III.

Results

The calculated SIR for kidney cancer in the HCV-cohort was 1.2 (95% CI: 0.8-1.7) using a one year lag-time. Therefore the results of this study do not indicate a relationship between HCV-infection and kidney cancer. Chronic HCV-infection was associated with a significantly increased risk of hospitalisation for other non-cancer kidney diseases. Risk of all non-cancer kidney diseases differed significantly by sex ($P_{interaction} = 0.045$). Hazard ratios of 3.9 (95% CI 3.2–4.8) and 5.8 (95% CI 4.2–7.9) were observed among men and women, respectively. Among men, risk estimates ranged from 2.9 (95% CI 2.0–4.3) for glomerular diseases to 4.6 (95% CI 3.7–5.8) for

renal failure. Among women, risk estimates ranged from 5.1 (95% CI 3.0–8.8) for glomerular diseases to 10 (95% CI 2.7–37) for other non-cancer kidney diseases. The interaction is likely to reflect a lower absolute risk level in women and a poor fit of a multiplicative model for both sexes.

Treatment of end-stage renal disease may involve blood transfusions and haemodialysis, which are potential risk factors for HCV-transmission. Reverse causation bias was therefore a potential concern in the analyses of hospitalisations for non-cancer kidney diseases. However, the impact of this type of bias on the reported results is likely minimal because all individuals with prior diagnoses of kidney diseases since 1969 were excluded from these analyses. Also, risk estimates were similar after excluding subjects for whom transfusion of blood/blood products or nosocomial infection was the suspected route of HCV-transmission.

Conclusion

The present study did not confirm that HCV-infection increases the risk of kidney cancer. The known relationship of HCV-infection and other kidney diseases such as renal failure and glomerular disease was further confirmed and described in this study.

Paper III

Motivation

The objective of the method presented in the second paper was similar to that of the first paper; to explore methods for visualizing, and thereby providing an opportunity to correct for selection bias in prevalent cohorts.

The proposed method is based on modelling the instantaneous failure rate using a restricted cubic spline model. The basic idea is to investigate how and if the hazard for the outcome of interest changes as a function of time since inclusion in the study. Theoretically, in a cohort where date of inclusion in the cohort is unrelated to any emerging symptoms of the outcome studied, no specific change of hazard in relation to time of inclusion in the cohort would be expected. However, if the hazard decreases immediately after inclusion in the cohort, to later stabilize or increase again, this may indicate selection bias.

Material

The proposed method was applied on two data sets used for previous studies.

1. The HCV-cohort linked to the Swedish Cancer Register described in the second paper, studying the association of HCV-infection with hospitalisation for kidney related disease (5).
2. Data from a study examining the excess mortality for patients with MGUS, monoclonal gammapathy of uncertain significance (6).

Methods

The hazard for the cohort is modelled as a function of time since inclusion in the cohort. In practice any method for modelling the hazard may be used, as long as the model is flexible and the estimates are fairly smooth. The particular advantages of the suggested restricted cubic spline model are however several.

The model is flexible, but not too flexible, where the number of knots (connecting the cubic polynomials) may be varied from a few knots up to 10 knots to give a sufficiently flexible model. If the model is extremely flexible it may be difficult to discern what is true change in hazard and what is random noise. If the model is not flexible enough, the model may be too rigid to model the hazard adequately. Since the model allows for a range of number of knots, flexibility may be varied to verify that different models give approximately the same result with regards to follow-up time that may be affected by selection bias.

The calculations are easy to perform, using just a few lines of code and the package stmp2 in STATA. The resulting graphs are easily interpretable and may give additional information on e.g. how the hazard for the studied outcome develops over time (37).

Results

The examples shown; MGUS and HCV with the studied outcomes death and hospitalisation due to kidney related disease, respectively, show that this method works in practice to identify a time period after inclusion in the cohort possibly affected by selection bias.

For HCV-infection and kidney related disease, the hazard was very high immediately after inclusion in the cohort and decreases steeply up to a year after inclusion in the cohort. The hazard continued to decrease also after about one year, but very slowly.

The first steep phase was interpreted as selection bias – there was no other obvious cause for the hazard to decrease immediately after inclusion in the cohort for a prevalent condition. The second phase with very slowly decreasing hazard was interpreted as frailty as discussed for the second paper above. This may also explain why the risk of the studied outcome did not increase with time, Figure 6.

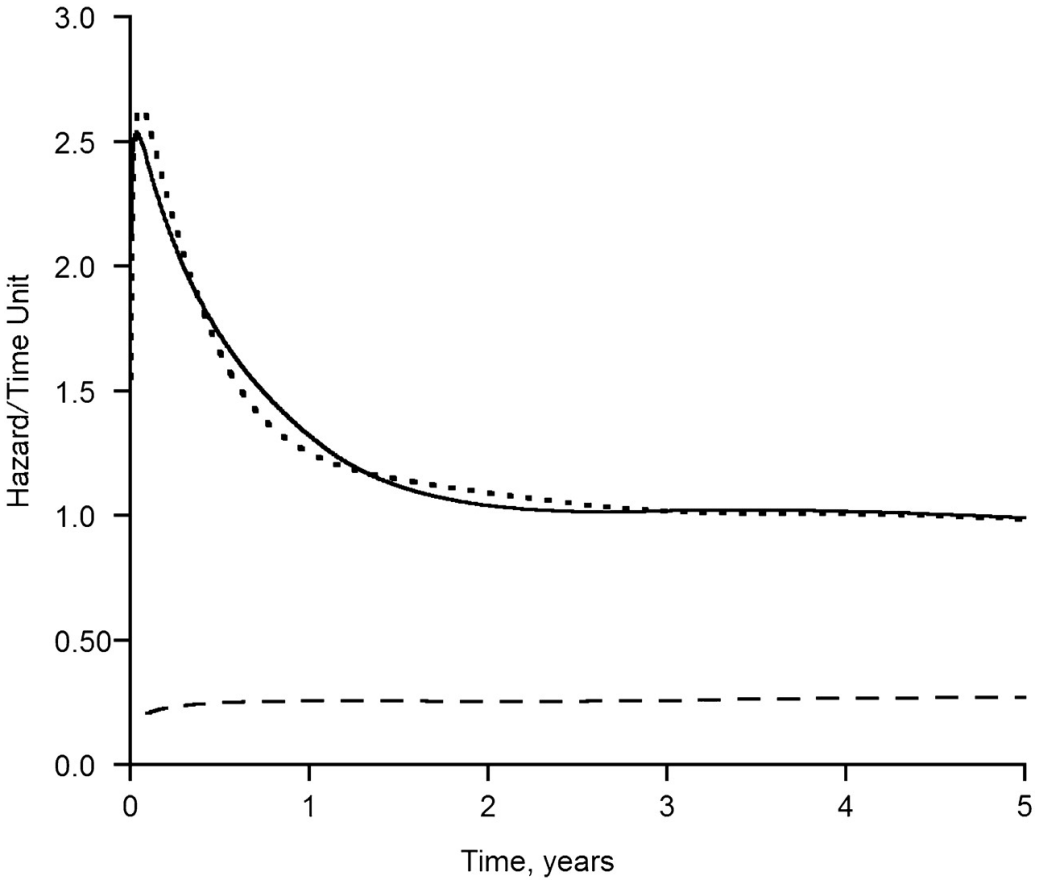

Figure 6. The hazard for the outcome "first hospitalisation for kidney-related disease" was modelled by using a restricted cubic spline model for survival data. The timescale on the *x* axis is time since hepatitis C virus (HCV) diagnosis in years. The solid line denotes the hazard for the HCV-cohort modelled with 3 knots; the dotted line, 5 knots; and the dashed line, the hazard for the matched control cohort (3 knots).

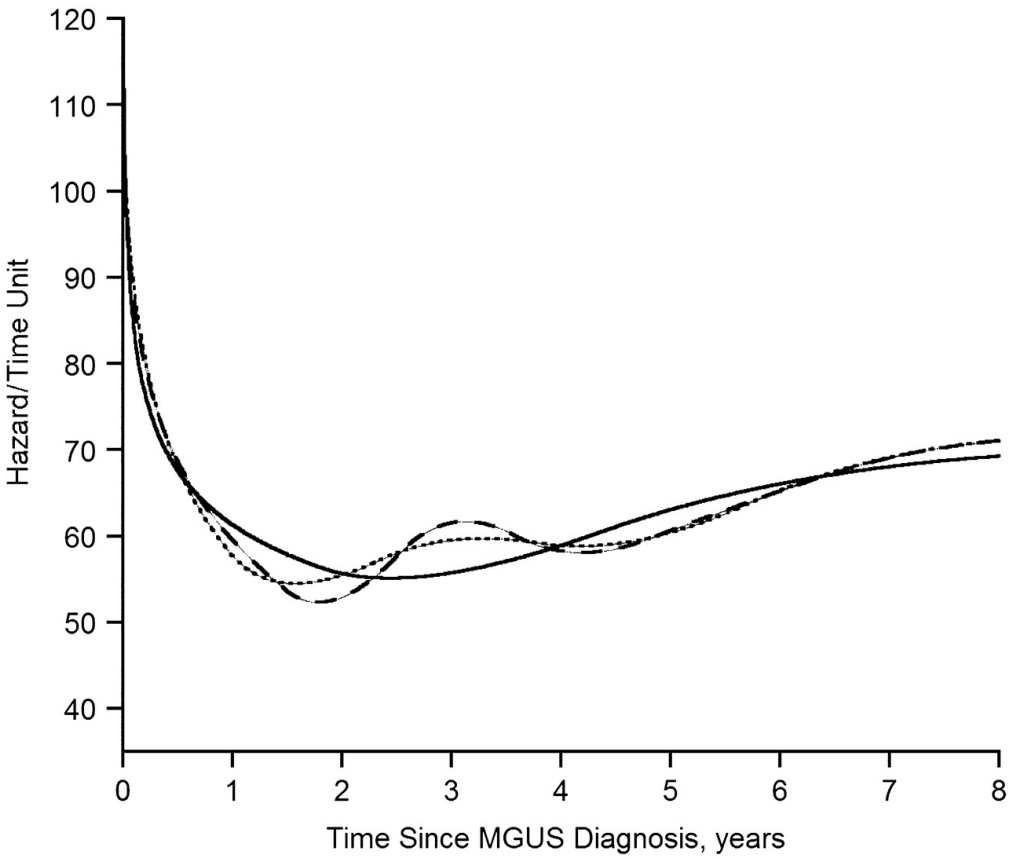

Figure 7. The hazard for the outcome death was modelled by using a restricted cubic spline model for survival data. The timescale on the *x* axis is time since diagnosis of monoclonal gammopathy of uncertain significance (MGUS) in years. The solid line denotes a model with 2 knots for the baseline hazard; the small dotted line, 5 knots; and the dashed line, 7 knots.

For MGUS, which is basically an asymptomatic disease in early stages, there was a steep decrease in hazard during the first two years after inclusion in the cohort and then the hazard started to increase with time, indicating that the hazard (risk) of dying for MGUS-patients was increasing with time, Figure 7.

Conclusion

As for the method cumulative SIR, the method proposed in this paper is not applicable for the study of acute diseases where the risk of the studied outcome would be expected to be highest immediately after contraction of the disease. In this case a decreasing hazard for the outcome studied, immediately after inclusion in the cohort is likely to have other explanations. The method is suited for prevalent cohorts where the patients are identified as members of cohort at different time points after contraction of the initial disease.

The method using restricted cubic splines to model hazard has several advantages over the method proposed in the first paper. The method of cumulative SIR assumes that the hazard reaches a steady state level to which the estimated SIR may level off towards, otherwise the early estimations of selection bias are influence by later higher SIR. Another advantage of the restricted cubic spline model is that it gives a good view over the hazard for the event of interest in the cohort over time, which may provide additional information on how the risk of the studied outcome changes over time in the cohort.

Paper IV

Motivation

The aim of this project was to investigate the completeness of the Swedish Cancer Register (CR) for hepatocellular carcinoma (HCC) diagnoses and to investigate if the low and declining incidence of HCC in Sweden represents a true decrease, or may be a result of changed diagnostic procedures. Another objective was to produce a more accurate estimate of the true incidence of primary liver cancer in Sweden by combining information from the CR, the Cause of Death Register (DR) and the Patient Register (PR). A secondary objective of the study was to investigate whether a capture-recapture model can be used to reliably estimate the number of cancers not reported to any of the registers.

Material

For the descriptive graphs of liver cancer over time, cancer incidence data was derived from the CR for the years 1970-2011. In addition, all individual patients with a cancer diagnosis 1550 (primary liver cancer) or 156 (liver cancer, unspecified) in the CR were extracted for the years 1975-2011. From the Cause of Death Register, for 1997-2012, patients with code C22.0, C22.9 and C22.99 (depending on year), either as underlying cause of death or as contributing cause of death, were extracted. The same codes were used for extraction from the PR. For the capture-recapture modelling and estimation of the true incidence of liver cancer, data from 1997-2012 was used.

Methods

Liver cancer incidence over the time period 1970-2011 derived from the cancer register was presented descriptively, as were the bases for liver cancer diagnoses during the

same time period. The number of patients with liver cancer in the CR, the DR and the PR were graphed in proportional Venn diagrams showing the number of patients in each register separately and also the number of patients present in more than one of the registers. The Venn diagrams were fitted so that areas were proportional to the number of patients, as closely as possible. For comparison, Venn diagrams for colon cancer and lymphoma were graphed and presented, Figure 8.

For the time period 1997 – 2011 data from the CR, DR and the PR was used to model the number of liver cancers not reported at all. A log-linear capture recapture model was used to estimate the number of liver cancers not reported at all based on the number of patients in the CR, PR and DR. A saturated model would need 8 degrees of freedom, however, available data has only 7 degrees of freedom so a full model with coefficients for each source, all two-way interactions and a three-way interaction is not possible to fit. Different models were explored and the model chosen was a model with a parameter for each source, one two-way interaction used for one of the possible two-way interactions and a three-way interaction. Different options for the two-way interaction were considered but naturally the model with the two-way interaction for the two most overlapping registers was be most appropriate and yielded the lowest Akaike value.
To investigate a lower bound for the number of unreported cases, a two-source model was constructed by joining the two most dependent sources (registers), the PR and the DR. For a two-source model, independence between these sources must be assumed. A two-source-model assuming independence should underestimate the number of unreported cases, all other assumptions being correct. Also, a three-source model with three different two-way interactions and no three-way interaction was also explored since this is an option other researchers have chosen.

Results

The number of liver cancers in the CR has decreased steadily over since 1970. In the same time period, the bases for diagnosis have changed. The proportion of patients with a diagnosis based on non-invasive imaging techniques has increased during the later years, but biopsy confirmed diagnosis is the most common basis for diagnosis in recent time.

The graphing using Venn-diagrams showed a large overlap between the registers, which would be expected since reporting to different registers is a highly dependent process, Figure 8 (38).

For liver cancer the overlap was largest between the PR and the DR. Noticeably, there were a large number of patients reported with liver cancer to the PR, who were not reported to the CR. For other cancer investigated, colon cancer and lymphoma, a different pattern emerged, with much stronger dependence (overlap) between the PR and the CR. This was judged to reflect the different reporting processes where diagnosis of liver cancer in an inpatient setting may be set based on non-invasive imaging techniques while diagnosis of colon cancer and lymphoma commonly is based on histologic confirmation. For these cancers, the diagnosing laboratory, as well as the clinic, sends reports to the CR.

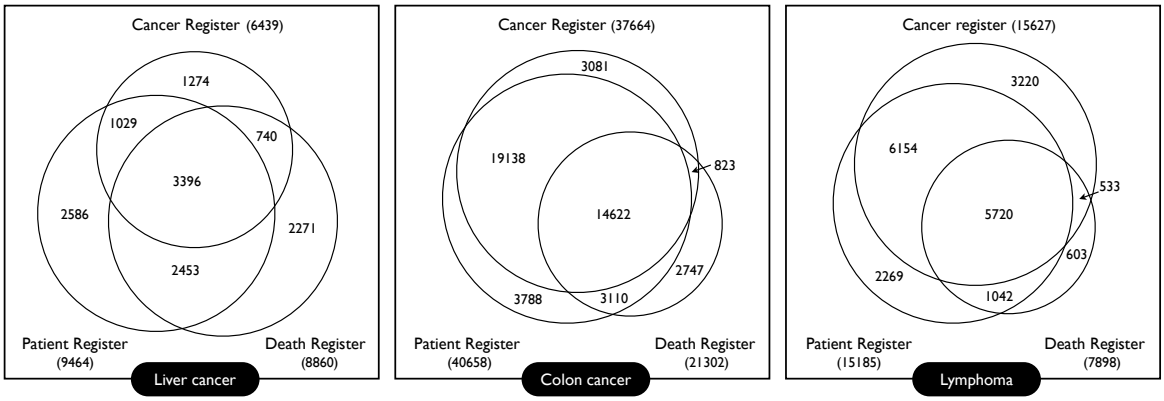

Figure 8. Venn diagrams showing number of reported cancers to different registers and combination of registers. Areas adjusted to fit number of cases as closely as possible.

The index year was defined as the first year a patient appeared in any of the three registers. The total number of liver cancers in all three registers with index year 2008-2010 was 13,749.

When studying the patients with liver cancer diagnoses from the DR and PR, not reported to the CR, it was found that as many as 3,925 (54%) of these 7,310 patients were reported to the CR for other cancer diagnoses. Of these, 1,390 patients were reported with ICD-code 199, i.e. malignant neoplasm without specification of site. This may indicate that at least part of the liver cancers identified only in PR and/or DR, were liver metastases related to other cancers. Making the drastic assumption that *all* 3,925 patients registered in the cancer register for other cancer diagnoses had received their liver cancer diagnosis erroneously in the PR and DR, this will bring down the total

number of liver cancers to 9,824 for the time period 1998-2010. Also, the completeness of the CR would then be estimated to 66%, assuming that all liver cancers were captured by at least one of the registers, i.e. not accounting for unreported cancers.

Sensitivity analysis

The two-source model gave an estimated number of liver cancers not reported to any register corresponding to about 13% of the total number of liver cancers in the three registers together, the corresponding figure for the three-source model with a single two-way interaction was about 25%. The three-source model, not accounting for a three-way interaction gave very high estimates of the number of unreported cases.

Conclusion

The CR is likely to underestimate the true incidence of liver cancer in Sweden. A large number of cases were reported to the PR and/or DR only. The reason for this may be the recommended use of non-invasive diagnostic methods based on imaging technologies for liver cancer diagnosis, resulting in a clinical diagnosis which is not always reported to the CR.

The large number of patients with liver cancer reported in the PR and/or DR raises concern. It is likely that at least some of these reported cancers are other cancers and not liver cancers. When using the capture-recapture model, although mathematically correct, the results must therefore be interpreted cautiously. In spite of these concerns, it is likely that the incidence of liver cancer is underreported in the Swedish CR.

The Venn-diagrams for liver cancer compared to the Venn-diagrams for colon cancer and lymphoma indicate differences in reporting routines. The diagnoses of colon cancer and lymphoma are usually given on the basis of laboratory histologic confirmation. For these cancers the CR appears to be much more complete with about 80% of all cancers reported to the CR.

Acknowledgements

My research described in this thesis was mainly carried out at the Swedish Institute for Infectious Disease Control (SMI) and at the Department of Medical Epidemiology and Biostatistics (MEB) between the years of 2008 and 2014. I thank SMI and Professor Annika Linde for providing the opportunity and the resources for my thesis work. I am also very grateful to Henrik Grönberg, Nancy Pedersen and Hans-Olov Adami at MEB for providing the inspiring research environment and for welcoming me to your PhD program. Thank you.

I also thank my wonderful supervisors:

Åke Svensson. Without you there wouldn't be any thesis for me. For a professor in mathematical statistics you have an amazing ability to view statistics from an applied perspective. Sharing an office with you for many years was fun (you have a great sense of humour) and a great learning experience. I am proud to be your student.

Paul Dickman. To me you are the perfect statistician. When you lecture, you make hazard ratios come alive and statistics become twice as fun and exciting. I still remember how happy and grateful I was when you agreed to be my supervisor. You have added the final touch to my work for which I am very grateful.

Ann-Sofi Duberg, you are my opposite and my perfect complement; you're an expert at everything I don't do so well. You fearlessly question everything I do and improve my work by a 100 % over and over again. Without you, my papers would just be average level statistics, nothing more.

Many other people have also contributed to this work in different ways:

Kasia Wikström and Hans Gaines, together you created the practical and economic opportunity for this thesis work and encouraged me to take the opportunity. I will always be grateful for you support.

Thanks to my all my co-authors for good collaboration on the papers in this thesis.

Eva Skovlund. You inspired me to become a statistician. ST101, which you taught, was the most interesting course I ever took. After completion of this course I knew I was "destined" for statistics – Thank you!

Arvid Sjölander, thank you for being my group partner during Yudi's likelihood course. One of my life's greatest compliments was when you said you'd be willing to work with me during the next course in probability theory in Uppsala. Also, your course on Causal Inference was fantastic.

Yudi, I rate your course as one of my worthwhile achievements. A watershed moment in my statistics life.

Johan Giesecke and Karl Ekdahl. To me, your time at Smittskyddsinstitutet was the "Golden Era". You showed that research was fun and creative.

Camilla Ahlqvist, your knowledge about study requirements, administrative procedures and kind reminders has been life saving. I will forever be grateful for your kind help.

Helen Turnell, you have faithfully read and corrected my English in all my papers and in my thesis. Thank you for adding this extra quality to my work! We are now more than even for the ice hockey handbook you wrote in Swedish.

My Dream Team at SDS, thank you for your support with the final work for this thesis (practicalities, listening, graphs, proofreading, miscellaneous). You make every working day a pleasure.

Einar, you have always supported my ambitions. This is also your accomplishment. I could not have done it without your support.

Axel, Ylva and Katarina, being a parent is life's most important job. You put everything in perspective.

My friends, thank you for always supporting me in (almost) everything I do. Without you, life would be boring and unloving.

References

1. Kallen B KK. The Swedish Medical Birth Register - a summary of content and quality. Rapport, Socialstyrelsen. 2003:30.

2. Socialstyrelsen. Cancer Incidence in Sweden. Rapport, Socialstyrelsen. (2011).

3. Duberg AS, Nordstrom M, Torner A, Reichard O, Strauss R, Janzon R, et al. Non-Hodgkin's lymphoma and other nonhepatic malignancies in Swedish patients with hepatitis C virus infection. Hepatology. 2005;41(3):652-9.

4. Strauss R, Torner A, Duberg AS, Hultcrantz R, Ekdahl K. Hepatocellular carcinoma and other primary liver cancers in hepatitis C patients in Sweden - a low endemic country. Journal of viral hepatitis. 2008;15(7):531-7.

5. Hofmann JN, Torner A, Chow WH, Ye W, Purdue MP, Duberg AS. Risk of kidney cancer and chronic kidney disease in relation to hepatitis C virus infection: a nationwide register-based cohort study in Sweden. European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP). 2011;20(4):326-30.

6. Kristinsson SY, Bjorkholm M, Andersson TM, Eloranta S, Dickman PW, Goldin LR, et al. Patterns of survival and causes of death following a diagnosis of monoclonal gammopathy of undetermined significance: a population-based study. Haematologica. 2009;94(12):1714-20.

7. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A. The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. European journal of epidemiology. 2009;24(11):659-67.

8. Gail M. Selection bias. Encyclopediaof Biostatistics. 2nd ed. Armitage P CT, editor: Hoboken,NY: JohnWiley & Sons; 2005. pp. 4869–70

9. Brookmeyer R, Gail MH. Biases in prevalent cohorts. Biometrics. 1987;43(4):739-49.

10. Alcabes P, Pezzotti P, Phillips AN, Rezza G, Vlahov D. Long-term perspective on the prevalent-cohort biases in studies of human immunodeficiency virus progression. American journal of epidemiology. 1997;146(7):543-51.

11. Brookmeyer R, Gail MH. Methods for projecting the AIDS epidemic. Lancet. 1987;2(8550):99.

12. Brookmeyer R, Gail MH, Polk BF. The prevalent cohort study and the acquired immunodeficiency syndrome. American journal of epidemiology. 1987;126(1):14-24.

13. Thomas DL, Seeff LB. Natural history of hepatitis C. Clinics in liver disease. 2005;9(3):383-98, vi.

14. Howards PP, Hertz-Picciotto I, Poole C. Conditions for bias from differential left truncation. American journal of epidemiology. 2007;165(4):444-52.

15. Cain KC, Harlow SD, Little RJ, Nan B, Yosef M, Taffe JR, et al. Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. American journal of epidemiology. 2011;173(9):1078-84.

16. Hansson LE, Nyren O, Hsing AW, Bergstrom R, Josefsson S, Chow WH, et al. The risk of stomach cancer in patients with gastric or duodenal ulcer disease. The New England journal of medicine. 1996;335(4):242-9.

17. Torner A, Dickman P, Duberg AS, Kristinsson S, Landgren O, Bjorkholm M, et al. A method to visualize and adjust for selection bias in prevalent cohort studies. Am J Epidemiol. 2011;174(8):969-76.

18. Torner A, Duberg AS, Dickman P, Svensson A. A proposed method to adjust for selection bias in cohort studies. Am J Epidemiol. 2010;171(5):602-8.

19. Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. American journal of epidemiology. 1999;149(11):981-3.

20. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.

21. Tilling K, Sterne JA. Capture-recapture models including covariate effects. American journal of epidemiology. 1999;149(4):392-400.

22. McClish D, Penberthy L. Using Medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. Medical care. 2004;42(11):1111-6.

23. Schouten LJ, Straatman H, Kiemeney LA, Gimbrere CH, Verbeek AL. The capture-recapture method for estimation of cancer registry completeness: a useful tool? International journal of epidemiology. 1994;23(6):1111-6.

24. Robles SC, Marrett LD, Clarke EA, Risch HA. An application of capture-recapture methods to the estimation of completeness of cancer registration. Journal of clinical epidemiology. 1988;41(5):495-501.

25. Song M, Cho IS, Li ZM, Ahn YO. Completeness of cancer case ascertainment in Korea radiation effect and epidemiology cohort study. Journal of Korean medical science. 2012;27(5):489-94.

26. Kim DS, Lee MS, Kim DH, Bae JM, Shin MH, Lee CM, et al. Evaluation of the completeness of cancer case ascertainment in the Seoul male cohort study: application of the capture-recapture method. Journal of epidemiology / Japan Epidemiological Association. 1999;9(3):146-54.

27. Heraud-Bousquet V, Lot F, Esvan M, Cazein F, Laurent C, Warszawski J, et al. A three-source capture-recapture estimate of the number of new HIV diagnoses in children in France from 2003-2006 with multiple imputation of a variable of heterogeneous catchability. BMC infectious diseases. 2012;12:251.

28. Kiemeney LA, Schouten LJ, Straatman H. Ascertainment corrected rates. International journal of epidemiology. 1994;23(1):203-4.

29. Crocetti E, Miccinesi G, Paci E, Zappa M. An application of the two-source capture-recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy. European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP). 2001;10(5):417-23.

30. Hook EB, Regal RR. Capture-recapture estimation. Epidemiology. 1995;6(5):569-70.

31. McCarty DJ, Tull ES, Moy CS, Kwoh CK, LaPorte RE. Ascertainment corrected rates: applications of capture-recapture methods. International journal of epidemiology. 1993;22(3):559-65.

32. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiologic reviews. 1995;17(2):243-64.

33. Orton H, Rickard R, Miller L. Using active medical record review and capture-recapture methods to investigate the prevalence of Down Syndrome among live-born infants in Colorado. Teratology. 2001;64 Suppl 1:S14-9.

34. Brenner H. Effects of misdiagnoses on disease monitoring with capture--recapture methods. Journal of clinical epidemiology. 1996;49(11):1303-7.

35. Hook EB, Regal RR. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. American journal of epidemiology. 1993;137(10):1148-66.

36. Coull BA, Agresti A. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. Biometrics. 1999;55(1):294-301.

37. Lambert P RP. Further Development of Flexible Parametric Models for Survival Analysis. The Stata Journal. 2009;9(2):pp. 265-90.

38. Rodgers P, Stapleton G, Flower J, Howse J. Drawing area-proportional Euler diagrams representing up to three sets. IEEE transactions on visualization and computer graphics. 2014;20(1):56-69.