Practice in analysis of multistate models using Lexis

Bendix Carstensen

Steno Diabetes Center, Copenhagen, Denmark

We shall analyse transitions in an illness-death model with states “Nephropathy” (kidney disease), “Remission” and “Death” among diabetes patients from Steno Diabetes Center. The Lexis machinery from the Epi package for representation and manipulation of multistate data will be introduced and used for visualization of states and transitions.

We shall fit both Cox-models and Poisson models to the transition rates in order assess the effect of “Remission” on mortality. Moreover, it will be of interest to assess the fraction of persons obtaining remission, so the “Death” state must be subdivided according to previous occurrence of “Remission”.

Furthermore, we will need multiple time scales in modeling transition rates, so we will make predictions of state occupancy probabilities for non-Markov models. This is algebraically intractable so we shall use simulation to compute these probabilities.

You will hopefully walk away with a functioning R-program conducting an entire analysis as well as reporting of results in tabular and graphical form.

You must have the latest version of the Epi package installed. The Lexis machinery is described in two papers [1, 2], and the ideas of multiple time scales in [3], whereas the tools for simulation from multistate models are described in a vignette (attached pdf-file) in the Epi package.

Course material will in due course appear on:
http://bendixcarstensen.com/AdvCoh/courses/Frias-2016/.

References

[1] Martyn Plummer and Bendix Carstensen. Lexis: An R class for epidemiological studies with long-term follow-up. Journal of Statistical Software, 38(5):1–12, 1 2011.

[2] Bendix Carstensen and Martyn Plummer. Using Lexis objects for multi-state models in R. Journal of Statistical Software, 38(6):1–18, 1 2011.

[3] S. Iacobelli and B. Carstensen. Multiple time scales in multi-state models. Stat Med, 32(30):5315–5327, Dec 2013.

Sensitivity analysis using lava simulation

Thomas Alexander Gerds

Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

In this tutorial I will start by introducing the simulation functions of the R-package lava (Klaus K Holst, 2006-2016, https://github.com/kkholst/lava). I will then demonstrate how to utilize these tools to perform a sensitivity analysis for a given regression analysis. The first step is to simulate data that are alike the real data under the substantive model which was fitted to the real data for the original regression analysis. To do this we specify a structural equation model for the joint distribution of all variables; in addition to the substantive model which describes the relation between outcome and predictors, this requires parametric models for the interrelationships of the predictor variables. Simulated data under the so obtained system of structural equations should be alike the real data. The second step is then to introduce various deviations from the substantive model in order to study the robustness of the results found in the real data. Which deviations are of particular interest depends very much on the subject matter question and are difficult to discuss in great generality. The participants of the tutorial will therefore work with the data and results of a published research report and perform a sensitivity analysis.

Some Aspects of the Design and Analysis of Life History Studies

Jerry Lawless

University of Waterloo

Life history cohort studies are an important source of information on human development, health, disease and related factors. Observational cohort studies often have limitations due to issues such as inclusion factors, irregular patterns of follow-up, management or treatment of subjects by indication, and missing information. Carefully planned studies with randomly selected cohorts can also face similar issues and studies of all types must deal with heterogeneity of cohort members, the complexity of life history processes, and budgetary constraints that limit cohort follow-up. These factors pose challenges for the design and analysis of studies and are potential sources of biased inference. In this talk I will review issues related to cohort selection and follow-up, the definition and measurement of key variables, the need to recognize initial conditions, and model specification as a basis for inference. Some recent developments will also be considered, including the use of large administrative databases, the design of sub-studies within large cohorts, and issues arising in genetic epidemiology.

Immortal Time Blunders: history, identification, severity

James A Hanley

McGill University, Montreal, Québec, Canada

For almost two centuries, teachers have warned against errors involving what is now called ‘immortal time.’ Despite the warnings, and many examples of how to proceed correctly, this type of blunder continues to be made in a widening range of investigations. In some instances, the consequences of the error are less serious, but in others the false evidence has been used to support theories for social inequalities; to promote greater use of pharmaceuticals, medical procedures and medical practices; and to minimize occupational hazards.

I will begin with some historical examples of immortal time issues, and point to early instances of dividing up person-time into states, rather than people into groups. Once this concept of time-division is accepted, the statistical ways to avoid immortal time blunders become clearer.

Because it is sometimes easy to go along with what seem like plausible conclusions based on immortal time, I will describe some ways to identify situations that might lead into the immortal time trap. To estimate the amount of bias that can be induced by this trap, I will describe the 'random but invisible lottery' idea Turnbull introduced for the heart transplant data, and that I have since borrowed.

I will end with an attempt, not yet successful, to quickly convert corrected hazard ratios into corrected durations of life extension.

Late entry and immortal time bias in cancer patient cohorts - implications for calculations of survival probabilities

Michael Hoffmeister

German Cancer Research Center (DKFZ), Heidelberg, Germany

The DACHS cohort of colorectal cancer patients was used to investigate the role of late entry and immortal time bias in cancer patient cohort studies. The DACHS study is one of the world's largest case-control studies on colorectal cancer with long-term follow-up of patients (currently more than 5,000 patients). The study is located in the Rhine-Neckar region and is ongoing since 2003. In the DACHS study, detailed information is available about lifestyle factors, medical and family history, comorbidities, therapy of colorectal cancer, recurrence and cause of death up to 10 years after diagnosis of colorectal cancer.

Follow-up time in cancer cohort studies usually starts with date of diagnosis but only few patients are recruited at the day of diagnosis. The longer the time between diagnosis and study entry, the higher is the probability that patients will be missed due to cancer-specific mortality or for other reasons. In cooperation with the Institute for Medical Biometry and Statistics, we used a published example and re-analyzed the association of statin use with survival after colorectal cancer to explore if late entry bias had an impact on the survival estimates in the published study and how adequate control for late entry is secured.

In another example from the DACHS study, the role of immortal time bias will be highlighted in an analysis on survival of colorectal cancer patients after occurrence of metastasis. To date, both synchronous metastatic disease (i.e., diagnosed along with primary colorectal cancer) and a short time interval between CRC diagnosis and the occurrence of distant organ metastases are considered indicators of aggressive tumor biology and poor prognosis. We demonstrate that available studies did not consider immortal time appropriately in their analyses which leads to drastically different results and interpretations.

Beta blockers and cancer prognosis - The role of immortal time bias: A systematic review and meta-analysis

Janick Weberpals[1], Lina Jansen[1], Prudence R Carr[1], Michael Hoffmeister[1], Hermann Brenner[1,2,3]

[1] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.
[2] Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany
[3] German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.

Background: Findings from experimental and observational studies have suggested beneficial effects of beta blocker (BB) use on cancer survival. Nevertheless, results have been inconclusive and there have been concerns that the observed associations might have resulted from immortal time bias (ITB). We conducted a systematic review and meta-analysis to summarize existing evidence, paying particular attention to this potential source of bias.

Methods: A systematic literature search was performed in PubMed and Web of Science. Studies investigating the association between BB use and overall or cancer-specific survival were included. Summary estimates were derived from meta-analyses using random effects models. The potential influence of ITB was investigated.

Results: We identified 30 eligible studies including 88,026 cancer patients in total. We deemed 11 studies to be at high or unclear risk of ITB. Including all studies in the meta-analysis, BB users had a significantly better overall (hazard ratio (HR) 0.88, 95% CI 0.79-0.97) and cancer-specific (HR 0.75, 95% CI 0.64-0.88) survival. Excluding the studies deemed to be prone to ITB resulted in HRs (95% CIs) of 1.00 (0.93-1.07) and 0.90 (0.83-0.98), respectively. Analyses on cancer site and BB type did not show beneficial associations besides overall survival among melanoma patients. However, melanoma-specific survival was not improved.

Conclusion: We found no clinically meaningful evidence for an association between BB use and survival after excluding studies with a possible ITB. Our results support suggestions that the proposed beneficial effect of BBs on cancer survival might be based on ITB.

Absolute risk prediction using nested case-control data

Bénédicte Delcoigne[1], Agus Salim[2], Kamila Czene[1] and Marie Reilly[1]

[1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[2] Department of Mathematics and Statistics, La Trobe University, Victoria, Australia

Background: Nested case-control studies are traditionally analyzed with conditional logistic regression. However, such analyses have several limitations. For example, while hazards ratios can be estimated, the cumulative risk cannot be calculated due to the "matching away" of time in the risk sets. In the last two decades, alternative methods have been developed to analyse such data in which the matching between the cases and their controls is broken. The hazard ratios are estimated by maximizing a weighted partial likelihood, using weights that are calculated with basic information from the underlying cohort. While the literature has focused on the assessment of relative risk, these weighted analyses offer the possibility to estimate absolute risk.

Methods: We developed an estimator of absolute risk and simulated 500 independent cohorts of 50,000 subjects each. Within each of these we generated several nested case-control studies. The cumulative baseline hazard estimates obtained under the weighted method were compared to the cohort estimates and were unbiased.

Results: We applied the weighted method to analyze data from a case-control study of lung cancer nested within a cohort of breast cancer patients. We calculated the absolute risk of developing lung cancer given the radiation dose received for breast cancer. This risk increased steadily with increasing radiotherapy dose in smokers, with no such effect in non-smokers.

Conclusion: Absolute risk can be validly estimated from nested case-control data. In an era of growing interest in risk prediction, this can offer significant savings in cost and resources compared to cohort studies.

Practical aspects of prediction in multistate models

Bendix Carstensen

Senior Statistician, Steno Diabetes Center, Copenhagen, Denmark

Life history studies will typically require that you set up a multistate model. In clinical studies states typically represent different kinds of treatment and complications. Some studies will exclusively focus on comparison of subsets of the transition rates, in which case traditional modeling of rates and comparison via hazard ratios will be the focus: For example, we could compare the cancer occurrence rate between persons with and without diabetes, ignoring both the occurrence rates of diabetes as well as the mortality rates. But it will be essential to model the entire multistate process if we want study concepts such as the life-time risk of cancer/ diabetes / both diseases. If the model is fairly simple with a limited number of states, and in particular if only one time scale is used, we can obtain closed-form analytic expressions of this type of probabilities as a function of the transition rates.

But if we have:

- many states and possible transitions
- multiple timescales
- time since entry to a particular state
- time at entry to a particular state

. . . then calculation of risks quickly becomes analytically intractable.

The solution is then to simulate persons' path through the states using the models for the transitions. This can produce a large simulated cohort with no censorings, and hence any probability relation to state occupancy can be computed by simple enumeration of persons in the simulated cohort.
I shall describe the `simLexis` machinery from the `Epi` package that does this type of job and give an example of calculations of the cumulative risk of kidney complications from a clinical epidemiology study at Steno diabetes center.
And I shall argue that models with smooth effects of timescales are both conceptually and computationally preferable over the much used non-parametric approaches.

[1] S. Iacobelli and B. Carstensen. Multiple time scales in multi-state models. Stat Med, 32(30):5315{5327, Dec 2013.

Missing values in cardiac arrest studies

Christian Torp-Pedersen

Department of Health Science and Technology, University of Aalborg, Aalborg, Denmark

Thomas Alexander Gerds

Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

We will present the Danish Cardiac Arrest Register. This register is based on case record forms filled by ambulance personnel in all cases of cardiac arrest outside a hospital in Denmark from 2001 to now. Given the highly variable circumstances missing data are common. In the typical cardiac arrest registry study, we are not mainly concerned with the power of our statistical analysis (because of the large sample size) but mainly with the risk of bias incurred into our main results by potential inappropriate treatment of missing values. For this presentation we will select one study from our ongoing collaborative research and in that context illustrate the standard ways to handle missing values, including complete case analysis, the worst case principle, and multiple imputation. We will then review these methods from the theoretical statistical point of view and also from the applied statistical point of view (software solutions). Specifically, we will illustrate the following repeatedly seen scenarios:

- Problem 1. Many covariates with few missing values but many patients have at least one missing value.
- Problem 2. Main analysis was stratified, e.g., by whether cardiac arrest was in a home or in a public place. How to handle missing values in the strata variable.
- Problem 3. Outcome is time to event. How to handle right censored data and competing risks in the multiple imputation.

For these problem definitions their corresponding descriptive statistics, main analyses and sensitivity analyses, we will set up a lively discussion with a focus on the necessary compromise between the medical doctor's aim and the statistician's preferred approach.

Cardio-protection by alcohol: pro- versus retrospective analyses

Kirsten Mehlig and Lauren Lissner on behalf of the INTERGENE collaboration

Dept. of Community Medicine and Public Health, Institute of Medicine, University of Gothenburg, 405 30 Gothenburg, Sweden

Retrospective, case-control studies of self-reported exposures as risk factors for disease might be influenced by knowledge of disease among cases. We compared retrospective versus prospective associations between alcohol consumption and coronary heart disease (CHD) in the same population.

Analyses were based on data from the INTERGENE research programme that consists of both a case-control study for CHD and a cohort with more than 10 years of follow-up. Logistic and Cox proportional hazard regression were used to assess the association of CHD with alcohol. Ethanol intake was dichotomized at the sex-specific median (10 g/day for men, 5 g/day for women), with non-drinkers constituting a separate category. Mortality due to causes other than CHD was considered as a competing event.

Results from the case-control study (595 patients, 3200 controls from the cohort)[1] showed significantly reduced odds of CHD for higher versus lower ethanol intake, OR = 0.72 (0.59 – 0.88), and no differences in abstainers. Prospectively, cohort members with higher compared to lower ethanol intake were also protected from incident CHD, HR = 0.69 (0.52, 0.93). Regarding competing risks, they also had decreased risk of total and non-cardiovascular mortality.

Both study designs showed a negative association between higher intake of ethanol and CHD suggesting that coronary patients reported their habitual alcohol consumption with similar accuracy as members of the population cohort. Thus, we could not confirm the “knowledge-of-disease bias” attributed to case-control studies involving alcohol[2]. Reduced risk of CHD for high ethanol intake could not be explained by increased mortality in this group.

[1] K Mehlig, E Strandhagen, PA Svensson et al., *CETP TaqIB genotype modifies the association between alcohol and coronary heart disease: the INTERGENE case-control study*, Alcohol **48** (2014) 695-700.

[2] E Giovannucci, MJ Stampfer, GA Colditz et al., *Recall and selection bias in reporting past alcohol consumption among breast cancer cases*, Cancer Causes and Control **4** (1993) 441-448.

Using multiple imputation for missing smoking data to avoid bias in an intergenerational analysis

Edwin Amalraj Raja[1], Sohinee Bhattacharya[2], Sam Tweed

[1]Research Fellow and [2]Senior Lecturer, University of Aberdeen, Scotland, UK

The Aberdeen Maternity and Neonatal Databank contains obstetric histories for all mothers giving birth in Aberdeen Maternity Hospital since 1950, offering a unique opportunity to study intergenerational effects. Where exposure data is systematically missing complete case (CC) analysis results in the exclusion of subjects reducing power and incurring bias. Multiple imputation allows all eligible subjects to be used in the analysis.

We aimed to assess the relationship between in utero exposure to maternal smoking and daughter's pregnancy outcomes. For all women born on or before 31st Dec 1975, birth records were linked with any records of their own pregnancy using unique identifiers, providing multilevel data of offspring nested within mothers. The data were analysed using (complete case analysis and combined dataset analysis after multiple imputation(MI) using REALCOMIMPUTE program. Relative increase in variance (RVI) and Fraction of missing Information (FMI) were used to determine whether or not the imputations were reasonable.

Maternal smoking was available for 30, 475 (35.5%) pregnancies. The missingness of smoking was related to other covariates and was assumed to be MAR. The ORs (95% CI) of smoking were 1.37 (1.23, 1.53) vs 1.24 (1.11, 1.39) for live births, 1.22 (0.76,1.96) vs 1.29 (0.77, 2.15) for still births and 1.39 (1.25,1.55) vs 1.30 (1.19, 1.43) for miscarriage using CC and MI analyses respectively. MI utilised all available women resulting in narrower confidence intervals although the direction of effect remained same. The highest estimated RVI (65%) and FMI (41%) are associated with maternal smoking indicating improvement by increasing number of imputations to 65.

Analysing baseline covariates in studies with delayed entry using a joint model

Regina Stegherr, Tobias Bluhmki, Jan Beyersmann

Institute of Statistics, Ulm University

The time of study entry into a randomized clinical trial is a natural choice for "time zero", but in other life history cohorts study entry may happen after time origin, leading to left-truncated data. E.g., in observational health services data on diabetes patients a possible time origin would be diagnosis of diabetes. Some patients may enter the study upon diagnosis, but others may have a known date of diagnosis before start of data collection. A relevant baseline covariate would be Body Mass Index (BMI), the problem being that such data may be measured upon study entry and, hence, not at baseline for those with a delayed study entry. The problem has been succinctly summarized in a letter by N Keiding and M Knuiman, Statistics in Medicine, Vol. 9, 1221-1222 (1990). The aim of the present work is to investigate whether a joint model for a longitudinal covariate such as BMI and time to disease event may be used to analyse the impact of baseline covariates, possibly unmeasured due to delayed study entry, on the time to event. This is in contrast to standard use of a joint model where the typical aim is to investigate the impact of the current value of the longitudinal marker, possibly unmeasured because of non-continuous updating, on the hazard of an event.

Dynamic frailty models for recurrent events data

Theodor Adrian Balan and Hein Putter

Dept. of Medical Statistics and Bioinformatics, Leiden University Medical Center

Recurrent events data are increasingly common in clinical studies. Within the framework of survival analysis, the frailty model has been commonly employed to account for population heterogeneity due to individual-specific unmeasured factors. The frailty is a random effect generally assumed to be constant in time. Considering the longitudinal character of the data, this is a real limitation. In other words, it is implicitly assumed that the unmeasured factors that the frailty accounts for do not change in time, nor do their effects change in time. This assumption may not always be justifiable. We adapt the model of Putter and van Houwelingen (2015) to accommodate recurrent events data. They proposed replacing the frailty by a dynamic frailty process which can change in time. A time-dependent frailty can be seen as accounting for unmeasured time-dependent covariates or unmeasured covariates with time-dependent effects. In many cases, this might be more plausible than a time-constant frailty model. Furthermore, the proposed method allows for a flexible autocorrelation structure, in addition to accommodating a large family of distributions for the random effects, including the gamma and the positive stable distributions.

We discuss the implications of this model for parameters of interest, such as the baseline intensity of the recurrent events process and the covariate effects. The interpretation, advantages and limitations of the dynamic frailty process approach are discussed in the light of a simulation study. Finally, the proposed methods are illustrated on a data set comprising recurrent infections on patients of chronic granulotomous disease.

Generalization from self-selected epidemiological studies

Niels Keiding

University of Copenhagen, Denmark

Low front-end cost and rapid accrual make web-based surveys and enrollment in studies attractive. Participants are often self-selected with little reference to a well-defined study base. Of course, high quality studies must be internally valid (validity of inferences for the sample at hand), but web-based sampling reactivates discussion of the nature and importance of external validity (generalization of within-study inferences to a target population or context) in epidemiology. A classical epidemiological approach would emphasize representativity, usually conditional on important confounders. An alternative view held by influential epidemiologists claims that representativity (in a narrow sense) is irrelevant for the scientific nature of epidemiology. Against this background, it is a good time for statisticians to take stock of our role and position regarding surveys and observational research in epidemiology. The central issue is whether conditional effects in the study population may be transported to desired target populations. This will depend on the compatibility of causal structures in study and target populations, and will require subject matter considerations in each concrete case. Statisticians, epidemiologists and survey researchers should work together to develop increased understanding of these challenges and improved tools to handle them.

Reference

Keiding, N. & Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussion). J.Roy.Statist.Soc. A **179,** 319-376.

On using non-representative samples in observational epidemiological studies

Costanza Pizzi, Lorenzo Richiardi

Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Italy

It is common to use non-representative source populations (i.e. those that are not based on the general population of a defined geographical area) in observational epidemiologic studies, but there has been considerable debate about whether this affects validity of the study results. We have been involved in this discussions a number of times since 2005, when we started an internet-based birth cohort in Italy (NINFEA cohort, www.progettoninfea.it). Internet-based recruitment has advantages in terms of feasibility, costs and possibilities of reaching traditionally understudied populations, however, this approach is often criticized on the basis of its consequent lack of representativeness of the general population.

We discuss these issues using the example of the NINFEA cohort focusing on two main criticisms which we have received against using non-representative populations in internet-based birth cohorts: lack of heterogeneity and the potential for bias. We also explore the situation in which the probability of being a member of the study depends on an intermediate variable in the causal pathway between the exposure and the outcome and discuss the consequences of non-representativeness in relation to non-collapsibility.

Our view is that representativeness is typically not a necessary requirement for validity unless the aim of the study is descriptive. Conversely, when the focus is on scientific inference, lack of representativeness may often enhance the validity of the study.

Statistical projection in the analysis of mortality using survey data: Improving weighted estimates by survival information

André Moser[1,2*], Matthias Bopp[3], Marcel Zwahlen[2] for the Swiss National Cohort

(1) Department of Geriatrics, Inselspital, Bern University Hospital, and University of Bern, Switzerland
(2) Institute of Social and Preventive Medicine (ISPM), University of Bern, Finkenhubelweg 11, CH 3012 Bern, Switzerland
(3) Epidemiology, Biostatistics and Prevention Institute, University of Zürich

In 2010 Switzerland changed from a full enumeration census to a register-based and rolling census system. The traditional system with an almost complete census (coverage of 98.6%) allowed for almost unbiased mortality analyses after a probabilistic record linkage of mortality records [1-4]. The new Swiss census system is based on a yearly register-based census, which collects information from harmonized community registers from the whole Swiss population (STATPOP), and a structural enquiry (SE) sample of roughly 3.5% of the resident population, collecting information which is not available in the STATPOP census (e.g. attained educational level). The Swiss Federal Statistical Office provides calibrated survey weights for the analysis of the SE samples, allowing for estimates which should be representative for the whole Swiss population [5]. However, because of possible selection bias due to nonresponse of certain population subgroups (response rate of SE 2010 was 87.1% [6]), mortality rates (MR) for SE respondents could not represent the whole Swiss population [7]. In this study we aimed to correct for selection bias by using an inverse probability (IP) weighting approach incorporating known sampling covariates as well as survival information. We used 6.7 million individuals from the STATPOP 2010 population aged 15 years or older, of which 317,079 were sampled within the SE 2010. We calculated and compared unweighted, survey weighted and IP weighted sex and age-group specific MR. We found that MR were underestimated in unweighted or survey-weighted analysis in all age-groups higher than age 40, whereas IP-weighted MR were close to official MR.

[1] Renaud A. Methodology report—Coverage Estimation for the Swiss Population Census 2000. Neuchâtel, Bundesamt für Statistik, 2004.

[2] Spoerri A, Zwahlen M, Egger M, Bopp M. The Swiss National Cohort: a unique database for national and international researchers. Int J Public Health. 2010;55:239-42.

[3] Schmidlin K, Clough-Gorr KM, Spoerri A, et al. Impact of unlinked deaths and coding changes on mortality trends in the Swiss National Cohort. BMC Med Inform Decis Mak. 2013;13:1.

[4] What does your neighbourhood say about you? A study of life expectancy in 1.3 million Swiss neighbourhoods. Moser A, Panczak R, Zwahlen M, et al. J Epidemiol Community Health. 2014;68:1125-1132.

[5] Assoulin D. Methodology report. Methodische Grundlage zu den ersten Hochrechnungsresultaten der Strukturerhebung 2010 vom 19. Juni 2012. Neuchâtel, Bundesamt für Statistik, 2012. www.bfs.admin.ch/bfs/portal/en/index/infothek/erhebungen__quellen/blank/blank/rs/03.Documet.188098.pdf

[6] Qualité L. Methodology report . Auswahl und Gewichtung der Stichprobe der schweizerischen Strukturerhebung. Neuchâtel, Bundesamt für Statistik, 2014. www.bfs.admin.ch/bfs/portal/en/index/infothek/erhebungen__quellen/blank/blank/rs/03.Documet.188101.pdf

[7] Bopp M, Braun J, Faeh D. Variation in mortality patterns among the general population, study participants and different types of nonparticipants: evidence from 25 years of follow-up. AJE 2014;180(10):1028-1035

Left truncation in cohorts of migrant populations due to return-migration to the country of origin

Odile Sauzet, Oliver Razum

Department of Epidemiology & International Public Health, Bielefeld University, School of Public Health

Studies trying to make international comparisons of health outcomes among migrant populations often ignore the various sources of bias that differing patterns of missing information for each countries contribute to. For example, a cohort of work migrants recruited during a defined period in the past will only include those migrants who are still living in the country at recruitment. A migrant who has return-migrated before recruitment will not be part of the cohort. This selection process may be differential with regard to health: migrant workers may choose to return-migrate to their country of origin when they retire, and migrant workers with ill health could be more likely to stay in the country of migration in the hope of better health-care. While the event of interest is not in itself an exclusion factor as in the usual definition of left truncation, the competing event return-migration is. Under the hypothesis that return-migration has predictors that also predict the outcome of interest, left truncation due to return-migration may lead to bias in the estimate of prevalence of diseases or mortality.

Using simulations, we estimate the bias induced by a range of return-migration models and we investigate ways to reduce this bias a competing risk analysis or by using known model for return-migration

Because the countries compared have migrant population of different origins and therefore different remigration pattern, we show that the interpretation given for differences found in prevalence or mortality rates may be erroneous and propose possible ways to control for such bias.

Using prevalence studies for risk factor analyses of hospital-acquired infections

Martin Wolkewitz

Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

Prevalence studies are often used to understand risk factors of hospital-acquired (nosocomial) infections. In contrast to cohort studies, this design is easy to conduct as it requires less time and resources. However, it is well known that long-stay patients are more likely to be sampled (length-biased sampling); a patient who stays 10 days has the double chance to get sampled than a patient who stays 5 days. As a result, very sick patients are overrepresented in prevalence studies due to the design.

It is also well known that long stayers are at higher risk for nosocomial infections than short-stayers. And further, patients with nosocomial infections stay longer in hospital.

In this talk, we study how these facts influence prevalence studies which are used for risk factor analyses. As a reference, we have real cohort-style data with full information (i.e. exact date of the nosocomial infection) and consider covariates measured on admission to ensure that risk factor precedes the infection. We mimic a one-day prevalence study by picking a sample from this cohort study. In particular, we study how prevalence odds ratios can be used as an association measures in contrast to other metrics (risk odds ratios or hazard ratios). A multi-state model which accounts for discharge/death as competing events helps us to display relationships between these association measures.

Impact and consequences of incomplete follow-up information in cohort studies

Martin Schumacher

Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

We start with some general considerations on the conditions under which valid statistical inference is possible in studies with incomplete follow-up information. We will especially highlight the problem of competing risks that preclude observing the onset of a disease of interest and that of interval censoring, i.e. if dates of onset cannot exactly be observed. A special situation arises in those studies where information on disease onset can only be retrieved in those participants who are alive at a particular follow-up visit. Ad-hoc methods of analysis are prone to the so-called truncation-by-death bias that leads to an underestimation of disease incidence rates. Here, we show that the illness-death model plays a central role in judging the magnitude of the bias (using the "Joly-formula") as well as serving as a basis for a proper analysis of studies with such incomplete follow-up information.

We illustrate the problems and possible solutions with concrete studies from environmental epidemiology, aging research and investigations on pediatric cancer survivors. Finally, we provide a detailed discussion on how the truncation-by-death bias might affect the results of a recently published study on secular trends in the incidence of dementia.

Incidence analyses in the SAPALDIA cohort study: inverse probabilty weighting (present standard) vs. multi-state modeling (future standard?)

Christian Schindler

Swiss Tropical and Public Health Institute, Basel, Switzerland.

The SAPALDIA study (Swiss Cohort Study on Air Pollution and Lung and Heart Diseases in Adults) was initiated in 1991 in eight areas of Switzerland representing a wide spectrum of environmental and geographic conditions. The baseline survey included 9651 subjects aged 18 to 60 years and involved lung function and allergy tests and a detailed interview on respiratory diseases, allergies and lifestyle. The cohort was reassessed in 2002, additionally including cardiovascular and chronic disease endpoints, and an extensive biobank was established. Of the 8047 subjects participating in the second survey almost 6500 underwent physical examinations. A third assessment in 2010 / 2011 involved additional measurements of carotid intima media thickness and pulse wave velocity. While 6088 subjects provided at least basic health information, more than 5000 underwent a detailed interview. We tentatively identified 689 deaths having occurred before the third survey. Participants' address histories were regularly updated between surveys and were used to estimate individual profiles of air pollution and noise exposure. SAPALDIA has extensively published on predictors of change in lung function and respiratory symptoms and of incidence of COPD (chronic obstructive pulmonary disease), partly in collaboration with other European cohort studies. In order to address participation bias, inverse probability weighting was used. However, as incidence of chronic diseases is moving into focus, multistate modeling will be introduced.

Incorporating loss to follow-up in multistate models for analyzing cohort study data with intermittent observation

David Bläsi[1], Nadine Binder[1,2], Martin Schumacher[2]

[1] Freiburg Center for Data Analysis and Modeling, University of Freiburg
[2] Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg

Intermittent observation in longitudinal studies leads to missing life history information. For instance, in aging cohorts, missing disease information due to death is a relevant problem often leading to bias in standard disease incidence or hazard ratio estimates [1, 2], and approaches have been suggested towards an unbiased analysis [3]. Lawless [4] describes a similar bias in disease incidence estimation, but for missing observation of disease onset due to loss to follow-up (LTF). This is done by incorporating a status 'LTF' into a multistate model. However, to the best of our best knowledge, efforts for a more accurate analysis than simply right-censoring LTF cases at the last observed visit are scarce. Barrett et al. [5] propose an extension of the illness-death model with two additional states for LTF and base corresponding risk factor analyses on Weibull transition hazards.

By including the status 'death' in Lawless' model, we can analyze the bias in situations, where both, truncation due to death and LTF are present. We propose a correction of the biased standard Nelson-Aalen estimate of the cumulative disease hazard. Building on the work by Barrett et al., we investigate the robustness of their model in simulation scenarios. The proposed methods are applied to a real data example.

[1] Joly P, Commenges D, Helmer C, Letenneur L (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. Biostat 3(3):433-43.

[2] Binder N, Schumacher M (2014). Missing information caused by death leads to bias in relative risk estimates. J Clin Epidemiol 67(10):1111-20.

[3] Leffondré K, Touraine C, Helmer C, Joly P (2013). Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? Int J Epidemiol 42(4):1177-86.

[4] Lawless JF (2013). Armitage lecture 2011: the design and analysis of life history studies. Stat Med 2013(32):2155-72.

[5] Barrett JK, Siannis F, Farewell VT (2011). A semi-competing risks model for data with interval-censoring and informative observation: an application to the MRC cognitive function and ageing study. Stat Med 30(1):1-10.

Illness-death regression based on interval censored time to illness

Harriet Sommer (1), Randi Grøn (2), Thomas A. Gerds (2)

(1) Institute for Medical Biometry and Statistics, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Germany
(2) Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

In an illness death model framework, we propose a direct binomial regression method for the risk of illness for the situation where the time to illness is interval censored and the status of illness at the time of death or censoring is unknown. Our approach extends the jackknife pseudo value regression method which has been explored for right-censored data. We propose an algorithm to obtain a non-parametric approximation of the transition probability for the transition from initial state to state of illness. This non-parametric approximation is used in the computation of the jackknife pseudo values. With our proposed method it is possible to obtain estimates for the association parameters of a direct binomial regression model which describes the effects of covariates on the risk of illness. The methods are illustrated by a real data example and simulations assess the performance.