

# Combined analysis of retrospective and prospective occurrences in cohort studies: HIV-1 serostatus and incident pneumonia

Stephen R Cole,<sup>1\*</sup> Haitao Chu,<sup>1</sup> Paul D Allison<sup>2</sup> and Stephen J Gange<sup>1</sup>

Accepted 7 July 2006

**Background** The authors show how information collected on retrospective occurrence times may be combined with prospective occurrence times in the analysis of recurrent events from cohort studies.

**Methods** We demonstrate how the observed data can be expanded from one to two records per participant and account for the within-individual dependence when estimating variances. We illustrate our methods using data from the Women's Interagency HIV Study, which recorded 384 retrospective and 352 prospective occurrences of pneumonia in 9478 retrospective and 7857 prospective person-years among 2610 adult women.

**Results** The hazard of non-*Pneumocystis carinii* pneumonia among the 2056 HIV-1 infected women was 2.24 times (95% confidence limits: 1.74, 2.89) that of the 554 uninfected women, independent of age. This hazard ratio was homogeneous across retrospective and prospective occurrences ( $P$  for interaction = 0.96) and combining occurrence types increased the precision by reducing the standard error by about a fourth.

**Conclusions** As expected, HIV-1 infection increases the hazard of pneumonia, with more precise inference obtained by combining information available on bidirectional occurrences. The proposed method for the analysis of bidirectional occurrence times will improve precision when the estimated associations are homogeneous across occurrence types, or may provide added insight into either the data collection or disease process when the estimated associations are heterogeneous.

**Keywords** Cohort studies, HIV-1, survival analysis, recurrent events

Prospective cohort studies are often used to explore possible determinants of recurrent events. Examples of events that may recur abound in epidemiological research, including myocardial infarction, depression, influenza and pneumonia. It is common for investigators to follow cohort members for the first occurrence of the event of interest, especially when recurrences are infrequent.

In addition to the prospective ascertainment of an occurrence, investigators often quantify (to varying degrees) the number and timing of occurrences prior to enrolment into the

cohort. Such information on the history of occurrences is typically used as eligibility criteria, or summarized as covariates, which may be thought to confound or modify the relation between the exposure of interest and the prospective occurrence.

Here, we show how information obtained on retrospective occurrence times may be used to improve inference on the relative hazard of the recurrent event. In particular, we describe methods that allow one to combine information from retrospective and prospective occurrence times into models for bidirectional occurrence times. The essential advantage of the approach is to increase precision when effects of exposures are homogeneous across retrospective and prospective occurrences, or point investigators towards insights regarding data collection or disease process when exposure effects are heterogeneous. We demonstrate the method using data from a prospective cohort study of pneumonia among women.

<sup>1</sup> Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA.

<sup>2</sup> Department of Sociology, University of Pennsylvania, PA, USA.

\* Corresponding author. Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, 615 North Wolfe Street, Room E7640, Baltimore, MD 21205 USA. E-mail: scole@jhsph.edu

## Methods

### Study population

Beginning in 1994 the Women's Interagency HIV Study<sup>1</sup> enrolled 2628 women in New York, Chicago, Los Angeles, San Francisco, and Washington DC. At enrolment and every 6 months thereafter, women completed an extensive interviewer-administered questionnaire and provided a blood sample for the determination of HIV-1 serostatus. HIV-1 serostatus was determined by ELISA with confirmation by Western Blot. Institutional review boards approved the protocol and informed consent forms, and each patient gave written informed consent. Analyses presented here are limited to the 2610 or 99% of women on whom HIV-1 serostatus was constant over follow-up. Date of birth was obtained from each participant upon enrolment.

### Endpoints ascertainment

The retrospective study endpoint was number of years from the date of study enrolment backwards to the most recent prior occurrence of non-*Pneumocystis carinii* pneumonia. The date of retrospective pneumonia occurrence was reported to the nearest month and was assumed to occur at the midpoint of the given month. We administratively censored retrospective occurrences of pneumonia at 4 years prior to enrolment to limit recall bias.

The prospective study endpoint was number of years from the date of study enrolment to the first occurrence of non-*Pneumocystis carinii* pneumonia, as reported by women at semi-annual study visits. The date of prospective pneumonia occurrence was reported to the day. We administratively censored prospective occurrences of pneumonia at 4 years after enrolment to preclude affects of highly active antiretroviral therapies, which were introduced in 1996 but gained widespread use in the Women's Interagency HIV Study around 1998. The time-scale for both retrospective and prospective endpoints was time on study; alternate analyses with age as the time-scale provided materially similar results.

### Statistical methods

Let  $S_i$  be the time, in years, from study enrolment backwards to the (i) last retrospective occurrence of pneumonia prior to study enrolment or (ii) administrative censoring at 4 years prior to study enrolment for woman  $i$ , where  $i = 1$  to  $N$  and  $N = 2610$ ; and  $E_i = 1$  if participant  $i$  had an uncensored retrospective occurrence, 0 otherwise. Let  $S_i^*$  be the time, in years, from study enrolment to (i) the first prospective occurrence of pneumonia, (ii) dropout, or (iii) administrative censoring at 4 years after study enrolment for participant  $i$ ; and  $D_i = 1$  if participant  $i$  had a prospective occurrence before dropout or censoring, 0 otherwise. We assume both dropout and administrative censoring are independent of occurrence, within levels of measured covariates. Let  $\mathbf{X}_i$  be a row vector of the measured covariates (i.e. age at enrolment and HIV-1 serostatus) for participant  $i$ .

We propose expanding<sup>2-4</sup> the observed data set from  $N = 2610$  to  $2N = 5220$ , where each subject contributes two records, one retrospective and one prospective. The retrospective record contains data on the occurrence time

$T_i = S_i$ , occurrence indicator  $\delta_i = E_i$ , covariates  $\mathbf{X}_i$ , and indicator of retrospective occurrence type  $R_i = 1$ . The prospective record contains data on the occurrence time  $T_i = S_i^*$ , occurrence indicator  $\delta_i = D_i$ , covariates  $\mathbf{X}_i$ , and indicator of retrospective occurrence type  $R_i = 0$ .

Extending the work of Allison,<sup>5</sup> the hazard or instantaneous risk of a bidirectional occurrence takes its standard form, conditional on occurrence type  $R$ , as  $h_R(t) = \lim_{\Delta t \rightarrow 0} \Pr(t \leq T < t + \Delta t | R = r) / \Delta t$ . Then a bidirectional Cox proportional hazards regression model<sup>6</sup> combining retrospective and prospective occurrence information is  $h_{R,\mathbf{X}}(t) = h_{R0}(t) \times e^{\mathbf{X}\alpha + R\mathbf{X}\beta}$ , where  $h_{R0}(t)$  is the occurrence type stratified hazard further conditioned to reflect the reference group (i.e.  $\mathbf{X} = 0$ ). The vector of unknown regression parameters  $\alpha$  represents the log hazard ratios relating components of  $\mathbf{X}$  to the hazard of a prospective occurrence, while the unknown regression parameters  $\beta$  represent the addition to the log hazard ratios for components of  $\mathbf{X}$  for the hazard of retrospective occurrence. Therefore, Wald tests  $\beta = 0$  are tests of the homogeneity of the effects of covariates  $\mathbf{X}$  on prospective and retrospective recurrences. When the Wald test does not reject the null hypothesis  $\beta_j = 0$  for a given  $j$ th covariate of  $\mathbf{X}$ , one may remove the corresponding term for the interaction of the covariate and  $R$  and, thereby, simplify the model to assume a homogeneous effect across occurrence types for that covariate. This approach assumes that, conditional on measured covariates, the retrospective and prospective hazards are not dependent on time since the last occurrence.<sup>5,7</sup>

When one or more covariates are constrained to have homogeneous associations across occurrence types, this approach suffers from an induced positive statistical dependence between the  $N$  pairs of records. One can think of this induced dependence as being due to unmeasured time-fixed covariates, which represent individual propensities to recur, or frailties. Our interest here is in estimating population-averaged, rather than subject-specific, effects. Therefore, rather than estimating subject-specific log hazard ratios (e.g. using a mixed effects accelerated failure time model), we estimate population-averaged log hazard ratios (i.e. using a conventional Cox regression model for the combined sample) with both robust 'sandwich' variance estimates<sup>8</sup> and non-parametric bootstrap variance estimates.<sup>9</sup> For the former, we use a robust variance estimate (such as implemented by the COVS option in the PHREG procedure in SAS). For the latter, we resampled the 2610 women with replacement  $K = 500$  times and substituted the standard deviation of the resampled  $\alpha_k$  and  $\beta_k$  as the standard error of  $\alpha$  and  $\beta$  in the typical formula for the 95% confidence limits (CL). Proportionality of effects over follow-up time was assessed by including an interaction term between exposure and time and by visual inspection of log-log survival plots. SAS version 9 (SAS Institute, Cary, NC) was used for all analyses.

## Results

The 2610 participants had a median age of 36 years at enrolment into the study in 1994 (interquartile range: 31–41 years) and 2056 of 2610 or 79% were infected with HIV-1.

### Pneumonia occurrence

Of the 2610 women 352 or 13% of them reported a retrospective occurrence of pneumonia within 4 years of enrolment; during which there were 9478 person-years yielding an incidence of 3.7 cases per 100 person-years (95% CL 3.4, 4.1). Of the 2610 women 384 or 15% of them reported a prospective occurrence of pneumonia within 4 years of enrolment; during which there were 7857 person-years yielding an incidence of 4.9 cases per 100 person-years (95% CL 4.4, 5.4). Of the 2610 women 642 or 24% of them dropped out of the study before administrative censoring at 4 years after enrolment. The pneumonia rate was lower in the retrospective data (rate ratio = 0.76; 95% CL 0.66, 0.88).

Table 1 displays the Cox regression estimates for a (homogeneous) model in which prospective and retrospective effects are constrained to be the same and two (heterogeneous) models in which the effects are allowed to differ. In the homogeneous model, for every additional decade of age, the hazard of pneumonia increased by ~25%. This elevated hazard was similar for the 352 retrospective and 384 prospective pneumonia events ( $P$ -value for homogeneity = 0.33). Combining the retrospective and prospective data yielded a notably more precise estimate of the age association, with the standard error about 3/4 (= 0.0444/0.0617) the size of the standard error for the prospective data alone. There was no strong evidence that the association of age with pneumonia changed over follow-up ( $P$ -value for proportional hazards = 0.19).

As expected, the hazard of pneumonia among the 2056 women infected with HIV-1 was more than twice the hazard among the 554 not infected with HIV-1, independent of age (Table 1). This elevated hazard was similar for the 352 retrospective and 384 prospective pneumonia events ( $P$ -value for homogeneity = 0.96). Combining the retrospective and prospective data again yielded a notably more precise estimate, with the standard error again about 3/4 (= 0.1287/0.1656) the size of the standard error for the prospective data alone. There was no appearance of effect measure modification of the association of HIV-1 with pneumonia by age ( $P$ -value for interaction = 0.94). But there was a suggestion that the

magnitude of the association of HIV-1 with pneumonia decreased over study follow-up ( $P$ -value for proportional hazards = 0.07). For all measured associations the CL based on the robust and bootstrap variance estimates were similar. Figure 1 depicts the retrospective and prospective cumulative incidence proportions over time for HIV-1 infected and uninfected women.

### Discussion

In summary, we demonstrated a method to combine information from prospective and retrospective occurrence times to improve inferences. In the presented example of non-*Pneumocystis carinii* pneumonia occurrence in women, HIV-1 infection and age increased the hazard of pneumonia independent of each other in both retrospective and prospective data. As illustrated with the example, combined inference is warranted when the effects of a given covariate are homogeneous across the occurrence types. The bidirectional analysis will yield gains in precision over standard analysis of prospective occurrences in cases where there is no marked heterogeneity between prospective and retrospective estimates of association. When such estimates are heterogeneous, likely causes of the heterogeneity include (i) recall bias in the retrospective occurrence times or (ii) time dependency in the effect of the covariate. Such heterogeneity may provide interesting clues into the data collection or disease process for further examination. However, one cannot determine whether an observed heterogeneity in effect is due to a data collection or disease process from the data alone. In the face of heterogeneity, inference should be centred on the (typically) more rigorous prospective occurrence times. It must be noted that an observed homogeneous effect may be due to recall bias offsetting a true time dependency. Therefore, an apparent gain in precision due to pooling retrospective and prospective data may be misleading when true but masked heterogeneities are present. A similar issue arises whenever pooling data, such as in meta-analyses. While parametric methods have been developed that allow the retrospective hazard to depend on time since last occurrence,<sup>10</sup>

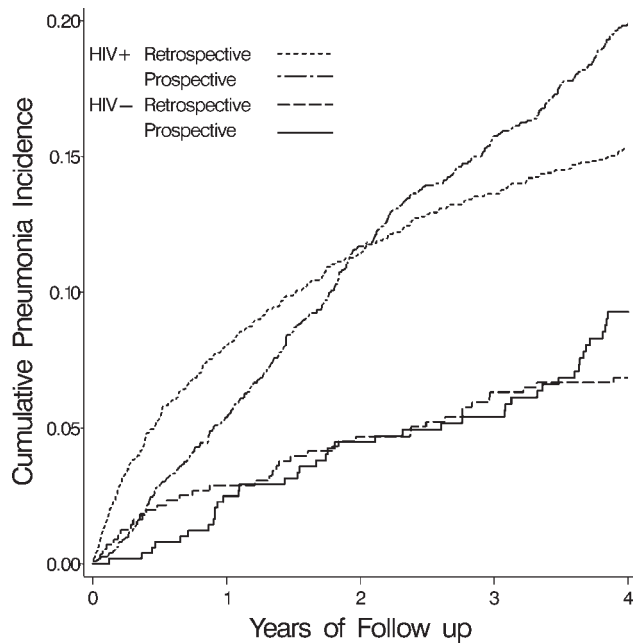
**Table 1** Bidirectional Cox regression models for pneumonia occurrence,  $N = 2610$

|  | Hazard ratio | 95% confidence limits   | Bootstrapped 95% confidence limits | Homogeneity $P$ -value* |
|--|--------------|-------------------------|------------------------------------|-------------------------|
| Homogeneous effects model                |              |                         |                                    |                         |
| Retrospective = Prospective              |              |                         |                                    |                         |
| Age, per 10 years                        | 1.25         | 1.14, 1.36 <sup>b</sup> | 1.15, 1.36                         | 0.33                    |
| HIV-1 infected                           | 2.24         | 1.74, 2.89 <sup>b</sup> | 1.72, 2.91                         | 0.96                    |
| Heterogeneous effects model <sup>a</sup> |              |                         |                                    |                         |
| Retrospective                            |              |                         |                                    |                         |
| Age, per 10 years                        | 1.19         | 1.05, 1.35              | 1.06, 1.33                         | –                       |
| HIV-1 infected                           | 2.26         | 1.61, 3.17              | 1.56, 3.28                         | –                       |
| Prospective                              |              |                         |                                    |                         |
| Age, per 10 years                        | 1.30         | 1.15, 1.47              | 1.16, 1.46                         | –                       |
| HIV-1 infected                           | 2.23         | 1.61, 3.09              | 1.62, 3.08                         | –                       |

\*  $P$ -value for test of equality of prospective and retrospective hazard ratios.

<sup>a</sup> Allows all estimated prospective and retrospective hazard ratios to differ.

<sup>b</sup> Calculated using a robust variance.



**Figure 1** Retrospective and prospective cumulative pneumonia incidence proportions (1—survival) as a function of follow-up time for 2056 HIV-infected and 554 HIV-uninfected women

here we assume that both the retrospective and prospective hazards do not depend on time since last occurrence, given measured covariates.<sup>5,7</sup> The present analysis of recurrent pneumonia is not subject to length-bias sampling<sup>11</sup> because the distance from the last pneumonia episode was not considered at study enrolment and participants were not removed from the risk set following a retrospective occurrence.

Here we considered the instantaneous risk or hazard as the measure of occurrence and used Cox proportional hazards models. One could alternatively consider the average incidence rate (or incidence density) and employ Poisson models. Using our notation, the rate would be  $\sum_{i=1}^N E_i / \sum_{i=1}^N S_i$  for retrospective occurrences,  $\sum_{i=1}^N D_i / \sum_{i=1}^N S_i^*$  for prospective occurrences, and  $(\sum_{i=1}^N E_i + \sum_{i=1}^N D_i) / (\sum_{i=1}^N S_i + \sum_{i=1}^N S_i^*)$  for bidirectional occurrences or  $2N / (\sum_{i=1}^N S_i + \sum_{i=1}^N S_i^*)$  assuming no censoring. From the perspective of inter-occurrence waiting times, the denominator for this incidence rate is downwardly biased by 1/2, assuming enrolment is taken at random between occurrences. While the incidence rates presented here are twice the inter-occurrence incidence rate, rate ratios would be unbiased whenever this bias cancels between compared groups. Hazards (being instantaneous risks) are not subject to this bias.

In epidemiological inquiry one must address possible alternate explanations for observed findings, including confounding, selection, and information biases. We observed no notable differences in the estimated association of HIV-1 serostatus after adjustment for age. This worked example did not intend to explore possible remaining common causes (i.e. confounders) of HIV-1 serostatus and pneumonia occurrence, which when properly measured and appropriately accounted for in analysis may alter our results. Adjustment for time-fixed covariates (e.g. sex, genotype) is straightforward. But adjustment for time-varying covariates (e.g. CD4 cell

count, depression) requires measurement of the time-varying covariate prior to the initiation of the retrospective risk period,<sup>5</sup> which may not often be undertaken in epidemiological studies. In our example, 24% of the women were lost to follow-up for the prospective occurrence, which suggests emigrative selection bias (i.e. informative dropout) as a possible explanation for these findings. However, we observed 14% (91 of 642) of prospective dropouts to have a retrospective occurrence and a similar 13% (261 of 1968) of those who did not dropout had a retrospective occurrence, which lends support to our assumption of non-informative dropout. Age and HIV-1 serostatus are unlikely to be highly misclassified. We assume that HIV-1 serostatus measured at enrolment is constant over the preceding 4 years. This assumption is tenable on average, because the median CD4 count among HIV-1 infected women at study entry was  $\sim 350$  cells/mm<sup>3</sup>, which is the median cell count expected after 6 years of HIV-1 infection.<sup>12,13</sup> Owing to self-report of pneumonia, both the retrospective and prospective occurrences may be misclassified. Relations between measured variables and the prospective occurrence times are probably subject only to non-differential misclassification. Specifically, measurement of variables at study entry is unlikely to depend upon future disease status, because future disease status is unknown to both participants and investigators at study entry. However, the retrospective occurrence times may be substantially differentially misclassified owing to recall bias. Here, the similarity of prospective and retrospective associations suggests that the retrospective occurrences were not subject to substantial differential misclassification. We administratively censored the retrospective pneumonia occurrences at 4 years prior to enrolment to limit recall bias and administratively censored the prospective pneumonia occurrences at 4 years after enrolment to preclude the effects of highly active antiretroviral therapy on our results.

Owing to the limited exploration of potential confounding, caution requires that interpretation of these results is descriptive of the proposed method and not necessarily substantively correct. In conclusion, the analysis of bidirectional occurrence times will improve precision, when the estimated associations are homogeneous across occurrence types, or may provide added insight into either the data collection or disease process, when the estimated associations are heterogeneous.

## Acknowledgements

We would like to thank the Editor and reviewers for helpful comments on an earlier version of this manuscript. Drs Cole, Chu, and Gange were supported in part by the National Institutes of Health through the data coordinating centre for the Women's Interagency HIV Study (U01-AI-42590). Data in this manuscript were collected by the Women's Interagency HIV Study (WIHS) Collaborative Study Group with centres (Principal Investigators) at New York City/Bronx Consortium (Kathryn Anastos); Brooklyn, NY (Howard Minkoff); Washington DC Metropolitan Consortium (Mary Young); The Connie Wofsy Study Consortium of Northern California (Ruth Greenblatt); Los Angeles County/Southern California Consortium (Alexandra Levine); Chicago Consortium (Mardge Cohen); Data Coordinating Center (S.J.G.). The WIHS is funded

by the National Institute of Allergy and Infectious Diseases with supplemental funding from the National Cancer Institute, and the National Institute on Drug Abuse (UO1-AI-35004, UO1-AI-31834, UO1-AI-34994, UO1-AI-34989, UO1-AI-34993, and UO1-AI-42590). Funding is also provided by the National Institute of Child Health and Human Development (UO1-HD-32632) and the National Center for Research Resources (MO1-RR-00071, MO1-RR-00079, MO1-RR-00083).

## References

- <sup>1</sup> Barkan SE, Melnick SL, Preston-Martin S *et al.* The Women's Interagency HIV Study. WIHS Collaborative Study Group. *Epidemiology* 1998;**9**:117–25.
- <sup>2</sup> Cole SR, Ananth CV. Regression models for unconstrained, partially or fully constrained continuation odds ratios. *Int J Epidemiol* 2001;**30**:1379–82.
- <sup>3</sup> Cole SR, Allison PD, Ananth CV. Estimation of cumulative odds ratios. *Ann Epidemiol* 2004;**14**:172–78.
- <sup>4</sup> Choi T, Cole SR. A family of ordered logistic regression models fit by data expansion. *Int J Epidemiol* 2004.
- <sup>5</sup> Allison PD. Survival analysis of backwards recurrence times. *JASA* 1985;**80**:315–22.
- <sup>6</sup> Cox DR. Regression models and life tables. *J R Statist Soc B* 1972;**34**:187–220.
- <sup>7</sup> Zhang Z, Sun J, Sun L. Statistical analysis of current status data with informative observation times. *Stat Med* 2005;**24**:1399–407.
- <sup>8</sup> Lin DY, Wei LJ. The robust inference for the proportional hazards model. *JASA* 1989;**84**:1074–78.
- <sup>9</sup> Efron B, Tibshirani R. *An Introduction to the Bootstrap*. London: Chapman Hall, 1993.
- <sup>10</sup> Yamaguchi K. Accelerated failure-time mover-stayer regression models for the analysis of last-episode data. *Sociol Methodol* 2003;**33**:81–110.
- <sup>11</sup> Ali MM, Marshall T, Babiker A. Analysis of incomplete durations with application to contraceptive use. *J R Stat Soc A* 2001;**164**:549–63.
- <sup>12</sup> Margolick JB, Munoz A, Donnenberg AD *et al.* Failure of T-cell homeostasis preceding AIDS in HIV-1 infection. The Multicenter AIDS Cohort Study. *Nat Med* 1995;**1**:674–80.
- <sup>13</sup> Lau B, Gange SJ, Phair JP, Riddler SA, Detels R, Margolick JB. Use of total lymphocyte count and hemoglobin concentration for monitoring progression of HIV infection. *J Acquir Immune Defic Syndr* 2005;**39**:620–25.