Published in final edited form as:

Stat Med. 2008 July 30; 27(17): 3442–3445. doi:10.1002/sim.3199.

Precision of Rate Estimation under Uniform Interval Censoring

Neal Alexander^{1,*†}

¹Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine, London, U.K.

SUMMARY

Medical events are often recorded not at their time of occurrence but rather at a subsequent examination. In survival analysis this is called interval censoring. This paper presents a relation between the sampling error of the estimated event rate and the number of examinations, which are assumed to be evenly spaced throughout the study. This relation should help balance the cost of additional examination rounds against the precision gained.

Keywords

interval censoring; survival analysis; clinical trials; hookworm

1. INTRODUCTION

Time to an event is a common endpoint in medical studies. In some situations this time is not measured precisely. In particular, the event may be one which is not immediately perceived by study participants so, in the absence of continuous monitoring, is only detected *post hoc* when they are next examined (assuming that the evidence of the event is still present). For example, infection with many pathogens, including HIV, is not always recognised immediately by the person infected. In such circumstances, the event time can only be specified as between consecutive examinations. In survival analysis, this is called interval censoring.

When analysing interval-censored data, Lindsey comments that ‘a common *ad hoc* approach is to assume that the event occurred at the end (or beginning or midpoint) of each interval, and then apply methods for standard time-to-event data’[1]. This ignores the imprecision in the observed times associated with interval censoring. Methods for interval estimation which do account for censoring are usually fitted iteratively by generalized linear model theory or the expectation maximization (EM) algorithm, or are complicated functions of the events occurring at each observation time[1]. These methods are flexible in that they allow for arbitrary variation between subjects in the observation times. By contrast, the current paper will obtain a simple analytic expression for the standard error of the rate, at the cost of assuming that the rate is constant and events observed at uniformly spaced times (‘uniform interval censoring’).

The imposition of time periods defined by the investigators (rather than defined by events) is a feature which the current work has in common with the life table method (also known as actuarial or Cutler-Ederer survival)[2]. These periods are usually chosen to have equal length, and the outcome is usually expressed in terms of the proportion surviving at any

*Correspondence to: Tropical Epidemiology Group, Infectious Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom. .

[†]neal.alexander@lshtm.ac.uk

given time, and its standard error[3]. By contrast, the current approach will consider the event rate (as opposed to proportion), a single value of which is assumed to pertain throughout the study, or within each arm of a trial. This is obviously less general than the life table's accommodation of an arbitrary rate within each period. Moreover, losses to follow-up will not be considered. On the other hand, there is no need for the life table method's assumption that each period is short compared to the mean time to event[4], which for the present purpose is not tenable. A relation will be obtained between the number of examination rounds and the precision of the rate estimation, to help investigators decide how many rounds would be worth scheduling. The work is motivated by the design of baseline studies and trials within the Human Hookworm Vaccine Initiative[5].

2. METHODS AND RESULTS

Consider a study of n people at risk of an event which occurs at a constant rate λ . Suppose there are m examination rounds scheduled uniformly throughout the study, at times iT/m for $i=1 \dots m$, where T is the study duration. People who experience the event are no longer considered at risk, representing situations such as permanent seroconversion, infections which are not cleared (e.g. HIV), and people who are treated and removed from the study when their first infection is found.

The probability of remaining free of the event till the i th examination round is $e^{-\lambda T/m}$, while the probability of having the event between rounds $i-1$ and i is $e^{-(i-1)\lambda T/m}(1-e^{-\lambda T/m})$. If d_i people have the event between rounds $i-1$ and i , then these probabilities yield the following multinomial likelihood:

$$L = \frac{n!}{(n - \sum_{j=1}^m d_j)!} e^{-\lambda T(n - \sum_{j=1}^m d_j)} \prod_{i=1}^m \frac{1}{d_i!} \left(e^{-(i-1)\lambda T/m} (1 - e^{-\lambda T/m}) \right)^{d_i}$$

So the log-likelihood is (omitting terms not involving λ):

$$\log(L) = - \left(n - \sum_{j=1}^m d_j \right) \lambda T + \sum_{i=1}^m \left(-d_i (i-1) \lambda \frac{T}{m} + d_i \log(1 - e^{-\lambda T/m}) \right)$$

The Fisher information is obtained by differentiating twice with respect to λ :

$$\begin{aligned} \frac{\partial \log L}{\partial \lambda} &= - \left(n - \sum_{j=1}^m d_j \right) T + \sum_{i=1}^m \left(-d_i (i-1) \frac{T}{m} + \frac{d_i T}{m} \frac{1}{e^{\lambda T/m} - 1} \right) \\ \frac{\partial^2 \log L}{\partial \lambda^2} &= - \sum_{i=1}^m \left(\frac{d_i T^2}{m^2} \frac{e^{\lambda T/m}}{(e^{\lambda T/m} - 1)^2} \right) \end{aligned}$$

The expected value of d_i is $ne^{-(i-1)\lambda T/m}(1-e^{-\lambda T/m})$. Substituting this for d_i gives the Fisher information (I):

$$I = \frac{nT^2}{m^2} \frac{1}{e^{\lambda T/m} - 1} \sum_{i=1}^m e^{-(i-1)\lambda T/m}$$

The sum is a geometric progression with m terms, the first of which is 1, with ratio $e^{-\lambda T/m}$. So the sum equals $(1-e^{-\lambda T})/(1-e^{-\lambda T/m})$, and the Fisher information simplifies to:

$$I = \frac{n(1 - e^{-\lambda T})}{\frac{m^2}{T^2}(e^{\lambda T/m} - 1)(1 - e^{-\lambda T/m})}$$

As m increases, the events are observed with increasing precision in time, and the denominator of the above expression tends to λ^2 , so that, as expected, the Fisher information tends to that of the exponential distribution, i.e. $n(1 - e^{-\lambda T})/\lambda^2$. More specifically, including the next term in the Taylor expansion of the denominator gives $\lambda^2(1 + (\lambda T/m)/2)$. There is no term of first order in $\lambda T/m$, so the convergence to λ^2 with increasing m is fairly fast. The speed of this convergence is illustrated by the figure, in terms of the ratio of the sampling variance (reciprocal of the Fisher information) for the interval censored data to that of the limiting exponential distribution. This is also the ratio by which the sample size should be increased, as a result of interval censoring, in order to maintain a standard error obtained under continuous surveillance. The figure also shows that the effect of interval censoring is small when the event rate is less than 1 per study duration.

3. CONCLUSIONS

This paper has considered events which are not immediately perceived by study participants, but which are detectable at subsequent examinations, e.g. infection with HIV or an intestinal parasite. An explicit relation has been derived between the number of follow-up examinations and the sampling error of the event rate. Although relatively straightforward, this relation does not seem to have previously appeared in the literature. The most important application of the current work would probably be for sample size estimation for rates. Previous work on this topic has concentrated on allowing for factors such as non-uniform accrual, non-uniform event rates, non-compliance and losses to follow-up[6] but not, apparently, interval censoring. Conversely, a limitation of the present work is that ignores losses to follow-up. If such losses are thought likely to be substantial, it may be worth increasing the number of examination rounds to reduce the potential for bias.

In principle, the equations presented here could be used as a route to maximum likelihood estimation of the event rate, once the data have been collected. However, in practice, the follow-up times would not be exactly evenly spaced, the rate may not be constant over the study, and there would be losses to follow-up. So, for analysis, existing methods[1] would probably be preferred. The results presented here should be more useful at the planning stage of studies, when there is a need to balance the number of follow-up examination rounds against the number of trial participants.

Acknowledgments

This work was supported financially by the Human Hookworm Vaccine Initiative of the Sabin Vaccine Institute, which receives funding from the Bill and Melinda Gates Foundation. I am grateful to Lyda Osorio, Yin Bun Cheung and an anonymous referee for helpful comments.

REFERENCES

1. Lindsey JC, Ryan LM. Tutorial in biostatistics: methods for interval-censored data. *Stat Med.* 1998; 17:219–38. [PubMed: 9483730]
2. Collett, D. *Modelling Survival Data in Medical Research*. Chapman and Hall; London: 1994.
3. Ederer F. A parametric estimate of the standard error of the survival rate. *Journal of the American Statistical Association.* 1961; 56:111–118.
4. Cox, DR.; Oakes, D. *Analysis of Survival Data*. Chapman and Hall; London: 1984.

5. Bottazzi ME, Miles AP, Diemert D, Hotez PJ. An ounce of prevention on a budget: a nonprofit approach to developing vaccines against neglected diseases. *Expert Review of Vaccines*. 2006; 5:189–198. [PubMed: 16608419]
6. Wittes J. Sample size calculations for randomized controlled trials. *Epidemiol Rev*. 2002; 24:39–53. [PubMed: 12119854]

Figure.

The vertical axis is the sampling variance of the event rate λ from interval-censored data to that from continuous surveillance. The horizontal axis is the number of examination rounds made at equally spaced times throughout the study. Various values of the rate are shown, in terms of events per study duration (equivalently, the study duration is set to be 1).