International Journal of Epidemiology, 2018, 236–245
doi: 10.1093/ije/dyx134
Advance Access Publication Date: 9 August 2017
Original article

Methods

Incidence rate estimation, periodic testing and the limitations of the mid-point imputation approach

Alain Vandormael,[1,2]* Adrian Dobra,[3] Till Bärnighausen,[1,4,5,6] Tulio de Oliveira[2,7] and Frank Tanser[1,6,7,8]

[1]Africa Health Research Institute, KwaZulu-Natal, South Africa, [2]Nelson R Mandela School of Medicine, College of Health Sciences, University of KwaZulu-Natal, South Africa, [3]Department of Statistics, Department of Biobehavioral Nursing and Health Informatics, Center for Statistics and the Social Sciences, and Center for Studies in Demography and Ecology, University of Washington, Seattle, WA, USA, [4]Heidelberg Institute for Public Health, University of Heidelberg, Heidelberg, Germany, [5]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, USA, [6]Research Department of Infection and Population Health, University College London, London, UK, [7]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa and [8]School of Nursing and Public Health, University of KwaZulu-Natal, South Africa

*Corresponding author. 719 Umbilo Road, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, 4001, South Africa. E-mail: vando026@umn.edu

Editorial decision 18 June 2017; Accepted 29 June 2017

Abstract

Background: It is common to use the mid-point between the latest-negative and earliest-positive test dates as the date of the infection event. However, the accuracy of the mid-point method has yet to be systematically quantified for incidence studies once participants start to miss their scheduled test dates.

Methods: We used a simulation-based approach to generate an infectious disease epidemic for an incidence cohort with a high (80–100%), moderate (60–79.9%), low (40–59.9%) and poor (30–39.9%) testing rate. Next, we imputed a mid-point and random-point value between the participant's latest-negative and earliest-positive test dates. We then compared the incidence rate derived from these imputed values with the true incidence rate generated from the simulation model.

Results: The mid-point incidence rate estimates erroneously declined towards the end of the observation period once the testing rate dropped below 80%. This decline was in error of approximately 9%, 27% and 41% for a moderate, low and poor testing rate, respectively. The random-point method did not introduce any systematic bias in the incidence rate estimate, even for testing rates as low as 30%.

Conclusions: The mid-point assumption of the infection date is unjustified and should not be used to calculate the incidence rate once participants start to miss the scheduled test dates. Under these conditions, we show an artefactual decline in the incidence rate towards the end of the observation period. Alternatively, the single random-point method is straightforward to implement and produces estimates very close to the true incidence rate.

Key words: incidence, imputation, mid-point, random-point, interval censoring, periodic testing, HIV

Key Messages

- Recent evidence suggests that the mid-point of the latest-negative and earliest-positive test dates—the censoring interval—can be used to infer the timing of the infection event.
- Using a simulation-based approach, we show that the infection date does not occur at the mid-point of the censored interval once participants start to miss their scheduled test dates.
- Under these circumstances, the mid-point method may lead epidemiologists to falsely conclude that the incidence rate is declining toward the end of the observation period.
- Imputation of a random infection date within the censored interval, based on a Monte Carlo approach, is straightforward to implement and produces estimates very close to the true incidence rate.

Background

The incidence rate is a fundamental concept in infectious disease epidemiology. It is used to measure the frequency at which new infection events occur per unit of person-time.[1] An important task for any incidence study is to precisely identify the timing of a new infection event.[2] But this is difficult to do because we cannot, at least in most situations, test participants on a daily basis. Instead, the current 'gold-standard' approach is to schedule test dates at fixed time intervals, say on a weekly, monthly or yearly basis.[3] In this case, we can only infer that the infection event occurred at some time-point between the latest-negative and earliest-positive test dates. The use of periodic testing to identify the infection event gives rise to the *standard* interval censoring problem.[4–8] Even if participants present at all of their scheduled test dates, we would still not know the exact amount of person-time that has been contributed since the start of the study period. The standard interval censoring problem therefore reflects an enumeration uncertainty in the denominator of the incidence rate measure.

It is intuitive that our uncertainty of the infection date will be proportional to the length of the testing interval. This uncertainty, by inference, will increase once participants start to miss their scheduled test dates. In sub-Saharan Africa, for example, reasons for missed HIV test dates have been associated with work commitments, illness, transportation costs, frequent migration and the fear of stigma or discrimination, among many others.[9–13] Irregular testing means that participants will have some probability of missing a test date that is contiguous to the interval containing the true (but unobserved) infection date. In other words, missed test dates are likely to extend the width of the censoring interval across one or more fixed testing intervals. This scenario, which we describe as *extended* interval censoring, means that we cannot definitively identify the testing interval in which the infection event truly occurs. Extended interval censoring therefore reflects an enumeration uncertainty in both the denominator and numerator of the incidence rate measure, which we illustrate with a straightforward example in Figure 1.

In recent years, a number of advanced and sophisticated methods have been designed to address the interval censoring problem.[14–27] However, there is no clear guidance on how these interval censoring methods can be used to estimate the incidence rate, and in which situations they should be applied.[28] In practice, epidemiologists are likely to treat the infection date as a missing data point for which more familiar imputation methods are available.[4] One popular *ad hoc* approach, which is the focus of this study,

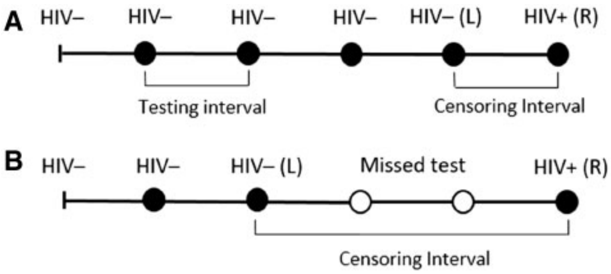

Figure 1. An example of Standard (Panel A) and Extended (Panel B) interval censoring. In Panel A, the participant is successfully tested at each scheduled test date, represented by the solid circles. We know that the infection event occurs somewhere between the latest-negative (L) and earliest-positive (R) test date. But we do not know the exact amount of person-time that should be contributed to the denominator of the incidence rate measure for the last time interval. In Panel B, the participant misses two scheduled test dates, as represented by the hollow circles. This makes it difficult to determine if the true infection event occurs in the 3rd or 4th or 5th time interval. In this case, there is an enumeration uncertainty in both the denominator and numerator of the incidence rate measure for each of these time intervals.

is to impute the infection date at the mid-point of the participant's censored interval.[29–42] There is some evidence that the mid-point method can give a reasonable approximation of the incidence rate if the standard interval censoring assumption is satisfied.[43–45] However, to the best of our knowledge, the performance of the mid-point method has not been systematically evaluated for incidence studies once participants start to miss their scheduled test dates. To learn more about the mid-point method, we used a simulation-based approach to generate an infectious disease epidemic for an incidence cohort with a high (80–100%), moderate (60–79.9%), low (40–59.9%) and poor (30–39.9%) testing rate. Our work has implications for infectious diseases studies that use the mid-point method to address the interval censoring problem.

Methods

Study design

This study is motivated by the low and irregular testing rate that we have observed in one of sub-Saharan Africa's largest HIV seroconverter cohorts.[46,47] Despite annual household visits by trained field-workers, an average censored interval length of 3.2 years has made it difficult to infer the timing of the HIV infection event. For this reason, we use the case of missed HIV test dates to systematically investigate the limitations of the mid-point method for incidence rate estimation. To do this, we used an epidemic model to generate HIV infection events for an incidence cohort (in either an open or closed system) with a fixed number of scheduled test dates. We then varied the rate at which participants missed their scheduled test dates and imputed a mid-point and a random-point value within each participant's censored interval. With this approach, we could then compare the incidence rate derived from these imputed values with the true incidence rate generated from the epidemic model.

Incidence cohort

Consider a cohort of $i = 1, \ldots, N$ study participants who are enrolled into a longitudinal survey or a randomized controlled trial. In the former study design, a single cohort of participants are followed over time; in the latter study design, participants are randomized to either a treatment or control cohort and followed over time. Let j denote the $j = 1, \ldots J$ intervals between the scheduled test dates for the observation period. For both study designs, participants must be HIV-uninfected when they enter into the study, so that their survival times start at the beginning of the first interval for a closed cohort or at the beginning of their entry interval for an open cohort. Survival time stops at the earliest HIV-positive date or at the end of the observation period if they remain HIV-negative. The test date could occur on any day within the testing interval. For this analysis, we scaled j on the unit interval [0, 1] so that the length of the testing interval was invariant to the unit of calendar time (i.e. month, half-year or year, etc.) between the scheduled test dates.

Epidemic model

We used a Susceptible-Infected-Recovered (SIR) model to generate the exact infection dates, denoted by T, over the J intervals of the observation period. The system of differential equations for the SIR model is given as:

$$\frac{dS}{dt} = -\lambda \; SI + bN, \quad \frac{dI}{dt} = \lambda \; SI, \text{ and } \frac{dR}{dt} = \nu I \qquad (1)$$

which represents the rate at which participants transition from a susceptible (S) to an infected (I) to a recovered (R) compartment. Known as the force of infection, λ is given by $\frac{\beta c}{N}$, where β is the probability of HIV transmission per contact, c is the rate of contact and N is the population size for the jth interval. The SIR model also includes a parameter b, which is the entry rate for participants into the study, where $b = 0$ for a closed cohort, and the parameter ν, which is the recovery rate for infected participants.

We used realistic parameter values for the SIR model, based on earlier HIV studies that have been undertaken in the sub-Saharan Africa context. To this extent, we varied c within the range of 50 to 120 sexual acts per year based on data collected from serodiscordant couples across eastern and southern African sites.[48–50] Previous research has shown considerable heterogeneity in the probability of HIV transmission per sexual contact, largely due to factors associated with the viral load level, genital ulcer disease, stage of HIV progression, condom use, circumcision and use of antiretroviral therapy (ART).[48–51] Following a systematic review of this topic by Boily *et al.*,[51] we selected values for β within the range of 0.003–0.008. Further, we based the recovery rate (ν) on the potential for ART to reduce the virologic suppression level of the infected population. The concentration of HIV RNA in the blood or genital tract is highly correlated with the onward sexual transmission of the virus.[52,53] Here, we chose values for ν within the range of 0.15–0.35, which are slightly conservative, but supported by population-based estimates from the sub-Saharan African context.[54,55]

For the longitudinal survey, we selected parameter values to generate a truly stable, increasing and decreasing incidence rate across 5, 10 and 15 testing intervals. For the randomized controlled trial, we selected an intervention efficacy E to reduce the HIV transmission rate for the treatment cohort when compared with the control cohort. We used the *EpiModel* package of Jenness *et al.*[56] to implement the SIR model and performed all remaining calculations with R software (version 3.3.3). Further details of the SIR model and the parameter values are provided in Section 1.1 of the Supplementary Data, available as Supplementary Data at *IJE* online.

Standard and extended interval censoring

Our next task was to simulate a testing rate over the observation period. For this analysis, we considered a successful HIV test date to be an independent random variable with a Bernoulli distribution. We denoted this random variable by H and the probability of a successful test date by $\Pr(H=1)=p$ for $(0 \leq p \leq 1)$.

Using this definition, we could then vary the testing rate for the incidence cohort by selecting a value for p. For standard interval censoring, we set $p=1.0$ to ensure that all participants would be successfully tested at each of their scheduled dates. For extended interval censoring, we set $p<1.0$ so that some participants would miss one or more of their scheduled test dates. As an example, a probability $p=0.6$ means that participants would be successfully tested at their scheduled dates 60% of the time. We considered a high testing rate to range from 80% to 100%, a moderate testing rate to range from 60% to 79.9%, a low testing rate to range from 40% to 59.9% and a poor testing rate to range from 30% to 39.9%.

Due to periodic testing, the infection event is known only to occur within the censored interval. For both standard and extended forms of interval censoring, the censored interval has non-zero length and bounds the infection date so that $L_i < T_i < R_i$, where L_i and R_i are observable random variables that denote the latest-negative and earliest-positive test dates of the ith participant. For each participant, we obtained the censoring dates with $L_i = max(H_{ij}{:}H_{ij} < T_i)$ and $R_i = min(H_{ij}{:}H_{ij} \geq T_i)$. Apart from the observed L_i and R_i test dates, the censored interval does not provide any extra information on the timing of the participant's infection event.[5]

Imputation of the infection dates

For the mid-point approach, we imputed an infection date for the ith participant using $t_i^m = (L_i + R_i)/2$. Alternatively, the mid-point can be obtained by sampling t_{ik}^r dates with replacement from the set $t \in [L_i,\ R_i]$, where $k = 2, \ldots,\ K$; and then taking the average of these dates, denoted by $\overline{T}$. To show this, let the probability density function of a uniform distribution be $f(t) = 1/(R - L + 1)$ with mean $\mu_T = (L + R)/2$. According to the Law of Large Numbers, the sample mean T of $T_1, \ldots,\ T_K$ random variables converges to μ_T in probability as K increases in size, where $\overline{T} \rightarrow^p\ \mu_T$.[57] For the single random-point approach, we set $k = 1$ and sampled a value t_{i1}^r from a uniform distribution bounded by $[L_i,\ R_i]$.

Calculating the incidence rate

We used the infection dates (T) generated by the SIR model to calculate the true incidence rate, denoted by θ. Using the standard formula,[1] θ_j is the number of new infection events (E) divided by the person-time (PT) contributed for the jth interval. Thus,

$$\theta_j = \frac{\sum_{i=1}^{N} E_{ij}}{\sum_{i=1}^{N} PT_{ij}} \times 100, \tag{2}$$

where $E_{ij} = 1$ if T occurs within the jth interval (otherwise $E_{ij} = 0$). We express θ_j as a rate per 100 person-units, since j is scaled on the unit interval [0, 1]. Equation (2) can also be described as an instantaneous incidence rate because it is calculated at fixed time points over the observation period.[58] The numerator of Equation (2) makes it clear that the infection events are being counted over the length of the jth testing interval. In some instances, the length of j will be less than the length of the aggregating interval: e.g. when test dates are scheduled on a monthly basis but the infection events are counted over 1-year intervals. When calculating this instantaneous measure, we assumed that the length of the testing interval j was always equal to the length of the aggregating interval.

We also calculated the cumulative incidence rate from the start of the observation period to the end of the jth interval, changing the notation slightly so that $j = [1,\ j]$. Boily *et al.*[58] have shown that the cumulative incidence rate ratio (CIRR) is a more appropriate measure for evaluating the intervention efficacy of a randomized controlled trial. We calculated the CIRR by dividing the cumulative incidence rate of the treatment cohort by the cumulative incidence rate of the control cohort, so that $\hat{\theta}_j^{CIRR} = \hat{\theta}_j^{\text{Inter}}/\hat{\theta}_j^{\text{Ctrl}}$. For the cumulative incidence rate, we note that the length of the aggregating interval [1, j] will always be greater than the length of the testing interval for $j > 1$.

We estimated the incidence rate after imputing an infection date within each participant's censored interval. Because the testing rate is a function of a stochastic process (i.e. H has a Bernoulli distribution), it was necessary to obtain more than one incidence rate estimate in order to quantify the uncertainty introduced by our simulation-based approach. Let $\hat{\theta}_j$ denote the estimated incidence rate for the jth time interval. To calculate $\hat{\theta}_j$, we right censored the data at the imputed values and indexed the resulting dataset with (d). We then obtained $\hat{\theta}_j^{(d)}$ for $d = 1, \ldots, D$ datasets using the standard formula, so that $E[\hat{\theta}_j] = \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_j^{(d)}$. For this analysis, we set $D = 1000$.

Measures of accuracy

To evaluate the accuracy of the mid-point and single random-point methods, we calculated the deviation between the estimated and true incidence rate for the *j*th interval. We used two principal measures for this purpose: the bias or error, which is given by $\epsilon[\hat{\theta}_j] = E[\hat{\theta}_j] - \theta_j$, and the mean-square error, $MSE[\hat{\theta}] = E[\hat{\theta} - \theta]^2$. Using these two measures, we also calculated the mean absolute percentage error as $\mathrm{MPE}[\hat{\theta}_j] = \frac{100}{J} \sum_{j=1}^{J} |\frac{E[\hat{\theta}_j]-\theta_j}{\theta_j}|$ and the root mean-square deviation as $\mathrm{R}MSD[\hat{\theta}_j] = \sqrt{MSE}$. The mean percentage error (MPE) and RMSD give a single measure of accuracy for each imputation method over the entire observation period.

Real-world example

To empirically demonstrate the performance of the mid-point and random-point methods, we used data from a

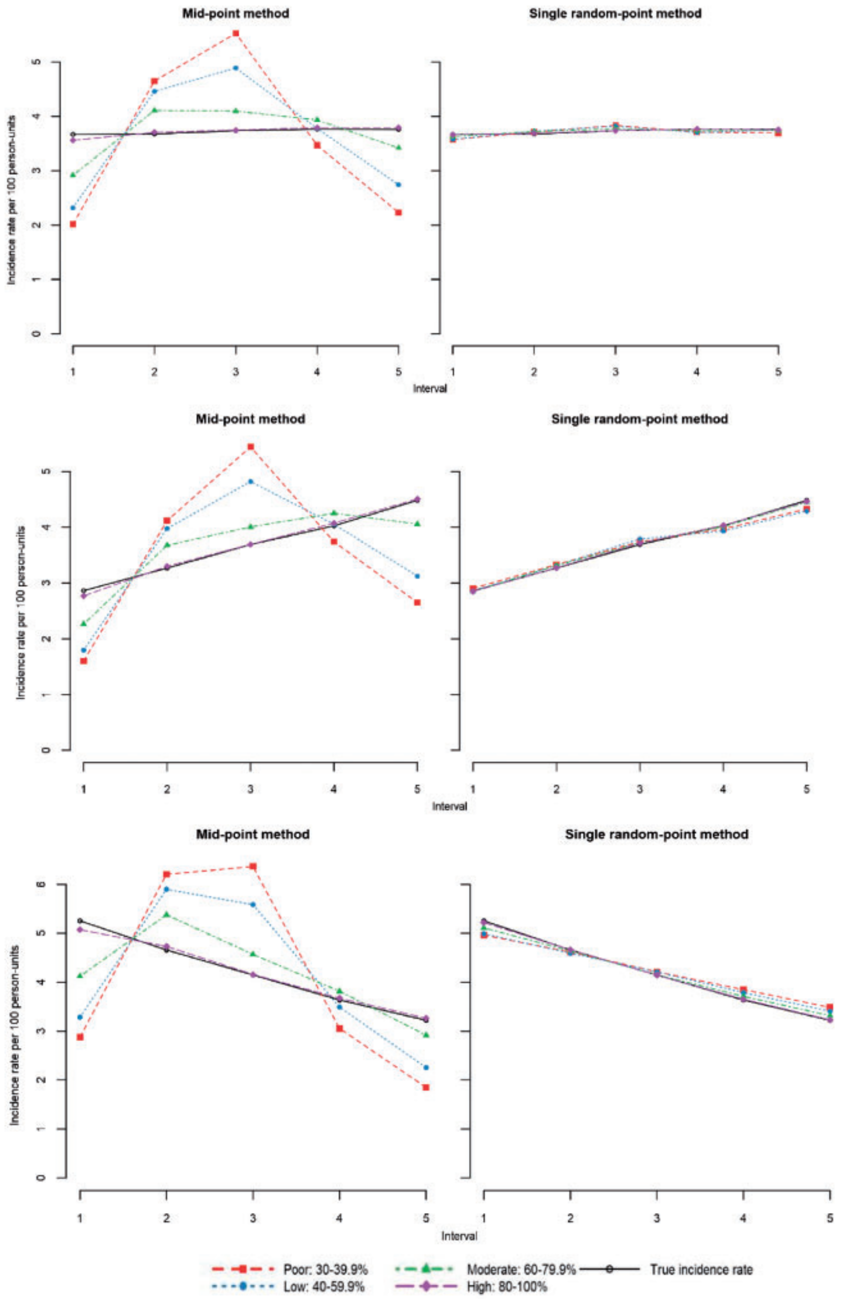

Figure 2. Compares the performance of the mid-point method (left column) against the single-random point method (right column) for a longitudinal survey with 5 testing intervals. The solid line is the true incidence rate and the non-solid lines represent the estimated incidence rates for a high (80–100%), moderate (60–79.9%), low (40–59.9%), and poor (30–39.9%) testing rate. We show that the mid-point incidence rate artefactually increases in the early stages, and then decreases in the later stages, of the observation period once the testing rate drops below 80%. Details of the epidemic models are discussed in Section 1.1 of the Supplement.

Table 1. Shows the percentage bias results for the mid-point (MP) and single random-point (SRP) methods

	Testing rate							
	High (80–100%)		Moderate (60–79.9%)		Low (40–59.9%)		Poor (30–39.9%)	
	MP	SRP	MP	SRP	MP	SRP	MP	SRP
Longitudinal survey								
1	–2.95	–0.21	–20.51	–1.02	–36.77	–2.46	–45.07	–2.52
2	0.81	0.38	11.81	1.19	21.32	1.35	26.5	0.95
3	0.29	0.00	9.68	0.98	30.88	2.31	47.93	2.64
4	0.70	0.11	4.53	–0.84	0.37	–1.58	–7.88	–1.42
5	0.84	–0.14	–9.05	–0.49	–27.07	–0.29	–40.63	–1.75
Randomized controlled trial								
1	–3.60	–0.96	–21.97	–6.65	–38.99	–10.81	–45.97	–12.55
2	–0.82	–0.23	–3.46	–1.89	–6.80	–3.00	–9.11	–4.31
3	–0.29	0.00	1.46	–0.18	6.09	0.03	8.77	–0.75
4	–0.02	0.14	2.01	0.56	4.77	1.29	5.21	0.90
5	0.33	0.33	1.73	1.73	3.20	3.20	3.11	3.11

The upper panel results correspond with the incidence rates presented in Row 1 of Figure 2. We do not include the remaining results from Figure 2 due to limitations of space. The lower panel results correspond with the CIRRs presented in Figure 3. Overall, the MP method gives a higher percentage bias for lower testing rates when compared with the SRP method.

Table 2. Mean percentage bias results for the mid-point (MP) and single random-point (SRP) methods

	Longitudinal survey						RCT	
	Stable Incidence Rate		Increasing Incidence Rate		Decreasing Incidence Rate		Cumulative Incidence Rate Ratio	
Testing Rate	MP	SRP	MP	SRP	MP	SRP	MP	SRP
High	1.12	0.17	1.21	0.40	1.54	0.32	1.01	0.33
Moderate	11.12	0.90	11.42	0.81	12.31	1.65	6.13	2.20
Low	23.28	1.60	24.13	2.2	26.56	3.57	11.97	3.67
Poor	33.6	1.86	33.12	1.93	38.11	4.42	14.43	4.33

Shows the mean percentage bias results for the mid-point (MP) and single random-point (SRP) methods. Results correspond with the estimates presented in Figures 2 and 3 (for five scheduled test dates). We show that the MP method introduces a greater degree of bias into the incidence rate estimates once participants start to miss their scheduled test dates.

population-based HIV surveillance programme based in the northern KwaZulu-Natal province of South Africa.[59] Since 2004, trained field-workers have visited over 10 000 households annually and repeatedly tested 17 400 adults (>15 years of age) for HIV antibodies. We calculated the annual HIV incidence rate for this cohort using the methodology described above.

Results

We observed that the mid-point method did not give accurate incidence rate estimates once the testing rate dropped below 80%. The poor performance of the mid-point method can be clearly seen in Figure 2, which shows the results for a longitudinal survey with an open cohort of size $N = 1000$. Here, the mid-point imputed incidence rate artefactually increases in the early stages, and then artefactually decreases in the later stages, of the observation period once participants start to miss their scheduled test dates. We report similar mid-point incidence rate results for sample sizes >500 participants, for both open and closed cohorts, and for 10 and 15 scheduled test dates (shown in Supplementary Figures 1 and 2, available as Supplementary Data at *IJE* online).

Table 1 shows the percentage errors for both imputation methods when compared with the truly stable incidence rate presented in Row 1 of Figure 2. For example, in the fifth testing interval, the decline in the mid-point estimate is in error of 9.05%, 27.07% and 40.63% for a moderate, low and poor testing rate, respectively (see Row 5 of the upper panel in Table 1). Table 2 shows the MPE results for the incidence rate estimates presented in Figure 2. For example, the mid-point MPE is in the range of 23.28–38.11% for a low and poor testing rate, when compared with a range of 1.60–4.42% for the single random-point method (see Rows 3 and 4 of Table 2). The MPE results for 10 and 15 scheduled test dates are presented in Supplementary Table 1, available as Supplementary Data at *IJE* online; see also Supplementary Table 2, available as Supplementary Data at *IJE* online, for the RMSD results.

Figure 3 shows the CIRRs for a randomized controlled trial in which $N = 2000$ participants were assigned to either a control or treatment cohort. The mid-point method significantly overestimates the efficacy of the treatment

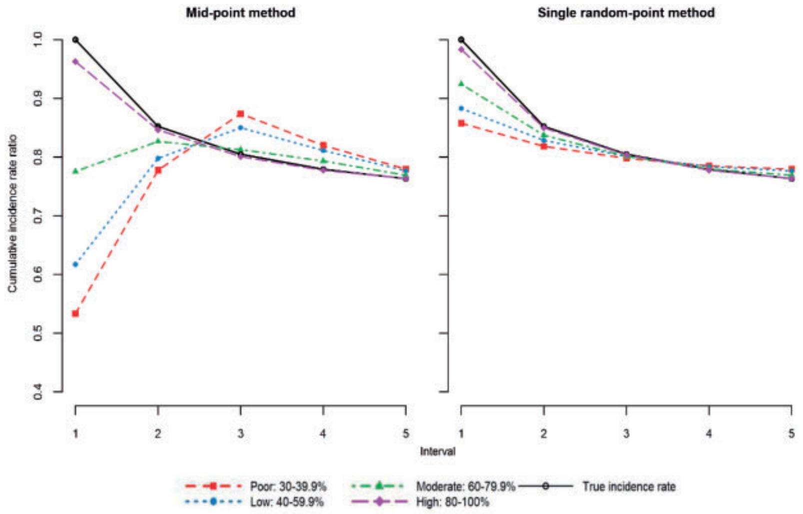

Figure 3. Compares the performance of the mid-point method (left column) against the single-random point method (right column) for a randomized controlled trial with 5 scheduled test dates. The solid line is the true cumulative incidence rate ratio (CIRR) and non-solid lines are the estimated CIRRs for a high (80–100%), moderate (60–79.9%), low (40–59.9%), and poor (30–39.9%) testing rate. No treatment effect is represented by a CIRR = 1. We show that the mid-point method significantly overestimates the treatment effect at the beginning of the observation period, although deviations from the true CIRR are attenuated at the last scheduled test date. Details of the epidemic models are discussed in Section 1.1 of the Supplement.

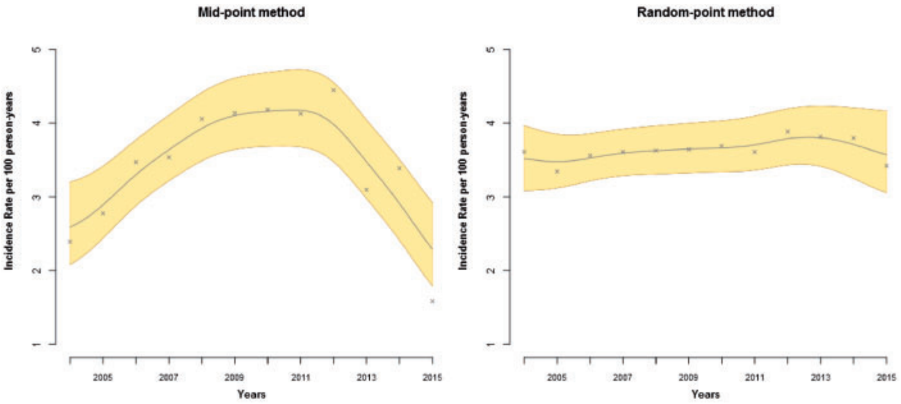

Figure 4. Compares the HIV incidence rates for the mid-point method (left) and single randompoint method (right) using data from a population-based HIV surveillance program (N ~ 17 400) in the KwaZulu-Natal province of South Africa. The dramatic difference in the estimates is due to a wide censoring interval (on average 3.2 years), which exposes the limitations of the mid-point method. This is because the mid-point method concentrates the imputed infection events at the middle of the observation period once participants start to miss their scheduled test dates. In this case, we would falsely conclude that the incidence rate rapidly increased in the beginning and then sharply decreased toward the end of the observation period. As our simulation results demonstrate, the single-random point is a far more accurate method for incidence rate estimation, which shows that the HIV incidence rate in our study population has been relatively stable over the last 10 years.

intervention in the early stages of the observation period. For example, the attributed efficacy is in error of 45.97% and 38.99% in the first of five scheduled test dates under a poor and low testing rate, respectively (see Row 1 of the lower panel in Table 1). However, the mid-point estimates converged to the true incidence rate at the end of the observation period. Overall, the MPE for the mid-point method is in the range of 1.01–14.43% when compared with a range of 0.33–4.33% for the single random-point method (see Columns 7 and 8 of Table 2).

We show, in Figure 4, the results for the two imputation methods using data from our population-based HIV surveillance programme. The estimates from the mid-point method are consistent with our simulation results. We see an increase and then a decrease in the HIV incidence rate at the beginning and end of the observation period, respectively. These annual estimates can be compared with the random-point method, which suggests that the HIV incidence rate has been relatively stable over the 2004–15 period.

Discussion

Our results show that the infection event does not occur at the mid-point of the censored interval once participants start to miss their scheduled test dates. Under these conditions, the mid-point method gives systematically biased incidence rate estimates. Importantly, we found that the instantaneous incidence rate artefactual increased in the early stages, and then artefactually decreased in the later stages, of the observation period. This pattern became more extreme as we systematically increased the probability of missing a scheduled test date, e.g. the decline in the incidence rate was in error of 9%, 27% and 41% for a moderate (60–79.9%), low (40–59.9%) and poor (30–39.9%) testing rate, respectively, in the later stages of the observation period. We observed this trend irrespective of a truly stable, increasing or decreasing incidence rate, for a closed and open cohort, for a range of sample sizes and for a different number of scheduled test dates.

An important limitation of the mid-point method is that it clusters the imputed the infection events at the middle of the observation period. This is because there are more left (latest-negative) and right (earliest-positive) test date combinations that give a mid-point in the middle interval of the observation period than all other combinations for the remaining testing intervals. We provide a simple and intuitive example of this mid-point behaviour in Section 2.1 of the Supplementary Data, available as Supplementary Data at *IJE* online. A better approach, based on the Monte Carlo methodology, would be to impute a single random infection date within the participant's censored interval, obtain an estimate from the resulting dataset, repeat this procedure several times and then take the average of the estimates for each interval. We show in this paper that the single random-point method approach makes less restrictive assumptions about the infection date when compared with the mid-point method (even for testing rates as low as 30%).

A number of advanced interval censoring methods have been developed within a survival analysis and Cox proportional hazards framework.[14,15,17–19,22,23,25,60] Some of these interval censoring methods can be found in statistical software programs such as *SAS*, *Stata* and *R*.[27,61–64] But these programs do not directly or intuitively estimate an incidence rate for continuous or discrete time periods as far as we can tell. We do acknowledge an approach by Hsu *et al*.,[60] who used the auxiliary information of participants to identify a set of nearest neighbours and then imputed multiple HIV infection times from a non-parametric distribution based on this neighbourhood.[60] Importantly, their method produced more accurate survival rates and hazard ratios when compared with the single random-point method. However, the authors did not directly extend their approach to estimate the incidence rate over time. Their method could be adapted for such a purpose; however, a potential improvement in accuracy would have to be traded for the convenience of the single random-point method.

We comment on the findings of Skar *et al*.,[45] who concluded that mid-point dating is a valid approach for population-based HIV incidence studies with regular testing intervals. Here, they are describing the performance of the mid-point method under the standard interval censoring assumption. But missed test dates are an unavoidable consequence of the periodic testing for an infectious disease. The surprising finding of our analysis is that participants need to be tested more than 80% of the time to produce accurate mid-point incidence rate estimates. If a high testing rate cannot be achieved, then we discourage use of the mid-point method for incidence rate estimation. Indeed, this method would lead us to falsely conclude that the HIV incidence rate in our study area has been dramatically declining over the last 3 years (as shown Figure 4). In contrast, results from the random-point method suggest a stable incidence rate over time, which are confirmed by the findings of an external phylodynamic analysis using HIV sequence data from the same incidence cohort.[65] In conclusion, if an *ad hoc* imputation method is to be considered, then the single random-point method, as described in this paper, is straightforward to implement and produces estimates close enough to the true incidence rate.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

A.V., T.D. and F.T. were supported by the South African Medical Research Council (SA MRC) Flagship grant (MRC-RFA-UFSP-01–2013/UKZN HIVEPI). F.T. was supported by two National Institute of Health (NIH) grants (R01HD084233 and R01AI124389) as well as a UK Academy of Medical Sciences Newton Advanced Fellowship (NA150161). T.B. was supported by the Alexander von Humboldt Foundation through the endowed Alexander von Humboldt Professorship funded by the German Federal Ministry of Education and Research, as well as by the Wellcome Trust, the European Commission, the Clinton Health Access Initiative and the National Institutes of Health's Fogarty International Center (D43-TW009775). Funding for the Africa Health Research Institute's Demographic Surveillance Information System and Population-based HIV Survey was received from the Wellcome Trust. The funders had no role in the design of the study, data analysis, interpretation of results or writing of the manuscript.

Conflict of interest: The authors have no conflicts of interest to declare.

References

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Lippincott Williams & Wilkins, 2008.
2. Jajosky RA, Groseclose SL. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC Public Health* 2004;**4**:1.

3. Van Beneden C, Olsen S, Skoff T *et al*. Active, population-based surveillance for infectious diseases. In: M'ikanatha NM, Lynfield R, Van Beneden CA, de Valk H (eds). *Infectious Disease Surveillance*. Blackwell, 2008, pp. 32–43.
4. Lindsey JC, Ryan LM. Methods for interval-censored data. *Stat Med* 1998;**17**:219–38.
5. Zhang Z, Sun J. Interval censoring. *Stat Methods Med Res* 2010; **19**:53–70.
6. Sun J. Interval censoring. In: Armitage P, Colton T (eds). *Encyclopedia of Biostatistics*. Wiley InterScience, 1998.
7. Sun J. *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer Science & Business Media, 2007.
8. Leung K-M, Elashoff RM, Afifi AA. Censoring issues in survival analysis. *Annu Rev Publ Health* 1997;**18**:83–104.
9. Tanser F, Bärnighausen T, Vandormael A *et al*. HIV treatment cascade in migrants and mobile populations. *Current Opinion in HIV and AIDS* 2015;**10**:430–8.
10. Visser MJ, Makin JD, Vandormael A *et al*. HIV/AIDS stigma in a South African community. *AIDS Care* 2009;**21**:197–206.
11. Dobra A, Bärnighausen T, Vandormael A *et al*. Space-time migration patterns and risk of HIV acquisition in rural South Africa: a population-based cohort study. *AIDS* 2017;**31**: 137–45.
12. Cawley C, Wringe A, Isingo R *et al*. Low rates of repeat HIV testing despite increased availability of antiretroviral therapy in rural Tanzania: findings from 2003–2010. *PLoS One* 2013;**8**: e62212.
13. Kelly JD, Weiser SD, Tsai AC. Proximate context of HIV stigma and its association with HIV testing in Sierra Leone: a population-based study. *AIDS Behav* 2016;**20**:65–70.
14. Boruvka A, Cook RJ. A Cox-Aalen model for interval-censored data. *Scand J Stat* 2015;**42**:414–26.
15. Xue H, Lam K, Cowling BJ *et al*. Semi-parametric accelerated failure time regression analysis with application to interval-censored HIV/AIDS data. *Stat Med* 2006;**25**:3850–63.
16. Deng L, Diggle PJ, Cheesbrough J. Estimating incidence rates using exact or interval-censored data with an application to hospital-acquired infections. *Stat Med* 2012;**31**:963–77.
17. Seaman S, Bird S. Proportional hazards model for interval-censored failure times and time-dependent covariates: application to hazard of HIV infection of injecting drug users in prison. *Stat Med* 2001;**20**:1855–70.
18. Goggins WB, Finkelstein DM, Schoenfeld DA *et al*. A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* 1998;**54**:1498–1507.
19. Kooperberg C, Clarkson DB. Hazard regression with interval-censored data. *Biometrics* 1997;**53**:1485–94.
20. Joly P, Commenges D, Helmer C *et al*. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002;**3**:433–43.
21. Pan W. A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* 2001;**57**:1245–50.
22. Bebchuk JD, Betensky RA. Multiple imputation for simple estimation of the hazard function based on interval censored data. *Stat Med* 2000;**19**:405–19.
23. Goggins WB, Finkelstein DM. A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* 2000;**56**:940–3.
24. Finkelstein DM, Goggins WB, Schoenfeld DA. Analysis of failure time data with dependent interval censoring. *Biometrics* 2002;**58**:298–304.
25. Sun L, Kim Y-j, Sun J. Regression analysis of doubly censored failure time data using the additive hazards model. *Biometrics* 2004;**60**:637–43.
26. Odell PM, Anderson KM, D'Agostino RB. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 1992;**48**:951–9.
27. Zeng D, Mao L, Lin D. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* 2016;**103**:253–71.
28. Singh RS, Totawattage DP. The statistical analysis of interval-censored failure time data with applications. *Open Journal of Statistics* 2013;**3**:155–66.
29. Biraro S, Ruzagira E, Kamali A *et al*. HIV-1 transmission within marriage in rural Uganda: a longitudinal study. *PLoS One* 2013; **8**:e55060.
30. Braunstein SL, van de Wijgert JH, Vyankandondera J *et al*. Risk factor detection as a metric of STARHS performance for HIV incidence surveillance among female sex workers in Kigali, Rwanda. 2012;**6(Suppl 1: M8)**:112–21.
31. Feldblum PJ, Enosse S, Dubé K *et al*. HIV prevalence and incidence in a cohort of women at higher risk for HIV acquisition in Chokwe, southern Mozambique. *PLoS One* 2014;**9**: e97547.
32. Dubé K, Zango A, van de Wijgert J *et al*. HIV incidence in a cohort of women at higher risk in Beira, Mozambique: prospective study 2009–2012. *PloS One* 2014;**9**:e84979.
33. Cawley C, Wringe A, Slaymaker E *et al*. The impact of voluntary counselling and testing services on sexual behaviour change and HIV incidence: observations from a cohort study in rural Tanzania. *BMC Infect Dis* 2014;**14**:1.
34. Gray R, Kigozi G, Kong X *et al*. The effectiveness of male circumcision for HIV prevention and effects on risk behaviors in a post-trial follow up study in Rakai, Uganda. *AIDS (London, England)* 2012;**26**:609.
35. Kiwanuka N, Ssetaala A, Nalutaaya A *et al*. High incidence of HIV-1 infection in a general population of fishing communities around Lake Victoria, Uganda. *PLoS One* 2014;**9**:e94932.
36. Naicker N, Kharsany AB, Werner L *et al*. Risk factors for HIV acquisition in high risk women in a generalised epidemic setting. *AIDS Behav* 2015;**19**:1305–16.
37. SPARTAC Trial Investigators. Short-course antiretroviral therapy in primary HIV infection. *N Engl J Med* 2013;**2013**: 207–17.
38. Ramjee G, Wand H, Whitaker C *et al*. HIV incidence among non-pregnant women living in selected rural, semi-rural and urban areas in Kwazulu-Natal, South Africa. *AIDS Behav* 2012; **16**:2062–71.
39. Wagman JA, Gray RH, Campbell JC *et al*. Effectiveness of an integrated intimate partner violence and HIV prevention intervention in Rakai, Uganda: analysis of an intervention in an existing cluster randomised cohort. *The Lancet Global Health* 2015; **3**:e23–33.

40. Bärnighausen T, Tanser F, Gqwede Z *et al*. High HIV incidence in a community with high HIV prevalence in rural South Africa: findings from a prospective population-based study. *AIDS* 2008; **22**:139–44.
41. Reniers G, Slaymaker E, Nakiyingi-Miiro J *et al*. Mortality trends in the era of antiretroviral therapy: evidence from the Network for Analysing Longitudinal Population based HIV/AIDS data on Africa (ALPHA). *AIDS* 2014;**28**:S533–42.
42. Tanser F, Bärnighausen T, Hund L *et al*. Effect of concurrent sexual partnerships on rate of new HIV infections in a high-prevalence, rural South African population: a cohort study. *Lancet* 2011;**378**:247–55.
43. Remis RS, Palmer RW. Testing bias in calculating HIV incidence from the Serologic Testing Algorithm for Recent HIV Seroconversion. *AIDS* 2009;**23**:493–503.
44. White EW, Lumley T, Goodreau SM *et al*. Stochastic models to demonstrate the effect of motivated testing on HIV incidence estimates using the serological testing algorithm for recent HIV seroconversion (STARHS). *Sex Transm Infect* 2010;**86**:506–11.
45. Skar H, Albert J, Leitner T. Towards estimation of HIV-1 date of infection: a time-continuous IgG-model shows that seroconversion does not occur at the midpoint between negative and positive tests. *PloS One* 2013;**8**:e60906.
46. Tanser F, Hosegood V, Bärnighausen T *et al*. Cohort Profile: Africa centre demographic information system (ACDIS) and population-based HIV survey. *Int J Epidemiol* 2008;**37**: 956–62.
47. Larmarange J, Mossong J, Bärnighausen T *et al*. Participation dynamics in population-based longitudinal HIV surveillance in rural South Africa. *PloS One* 2015;**10**:e0123345.
48. Hughes JP, Baeten JM, Lingappa JR *et al*. Determinants of per-coital-act HIV-1 infectivity among African HIV-1–serodiscordant couples. *J Infect Dis* 2012;**205**:358–65.
49. Gray RH, Wawer MJ, Brookmeyer R *et al*. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. *The Lancet* 2001;**357**: 1149–53.
50. Wawer MJ, Gray RH, Sewankambo NK *et al*. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J Infect Dis* 2005;**191**:1403–9.
51. Boily M-C, Baggaley RF, Wang L *et al*. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *The Lancet Infectious Diseases* 2009;**9**:118–29.
52. Quinn TC, Wawer MJ, Sewankambo N *et al*. Viral load and heterosexual transmission of human immunodeficiency virus type 1. *New Engl J Med* 2000;**342**:921–9.
53. Cohen MS, Chen YQ, McCauley M *et al*. Prevention of HIV-1 infection with early antiretroviral therapy. *New Engl J Med* 2011;**365**:493–505.
54. Jain V, Liegler T, Kabami J *et al*. Assessment of population-based HIV RNA levels in a rural east African setting using a fingerprick-based blood collection method. *Clin Infect Dis* 2012; **56**:598–605.
55. Tanser F, Vandormael A, Cuadros D *et al*. Effect of population viral load on prospective HIV incidence in a hyper-endemic rural South African community: a population-based cohort study. Under review, 2017.
56. Jenness SM, Goodreau SM, Morris M. *EpiModel: Mathematical Modeling of Infectious Disease*. R Package Version 102, 2014.
57. DeGroot MH, Schervish MJ. *Probability and Statistics*, 4th edn. Pearson, 2012.
58. Boily M-C, Mâsse B, Alsallaq R *et al*. HIV treatment as prevention: considerations in the design, conduct, and analysis of cluster randomized controlled trials of combination HIV prevention. *PLoS Med* 2012;**9**:e1001250.
59. Vandormael A, Newell M-L, Bärnighausen T *et al*. Use of antiretroviral therapy in households and risk of HIV acquisition in rural KwaZulu-Natal, South Africa, 2004–12: a prospective cohort study. *Lancet Glob Health* 2014;**2**:e209–15.
60. Hsu C-H, Taylor JM, Murray S *et al*. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Stat Med* 2006;**25**:3503–17.
61. Royston P. *STPM: Stata Module to Fit Flexible Parametric Models for Survival-Time Data*. Statistical Software Compo nents, 2014.
62. SAS. *PROC LifeReg: SAS/STAT(R) 9.2 User's Guide, Second Edition*. Cary, NC, USA: SAS Institute Inc., 2008.
63. Gentleman R, Vandal A. *Icens: NPMLE for Censored and Truncated Data*., 2010. R package version 1.48.0. https://www.bioconductor.org/packages/release/bioc/html/Icens.html (17 July 2017, date last accessed).
64. Fay MP, Shaw PA. Exact and asymptotic weighted logrank tests for interval censored data: the interval R package. *J Stat Softw* 2010;**36**:i02.
65. Rasmussen D, Wilkinson E, Stadler T *et al*. External introductions helped drive and sustain the high incidence of HIV-1 in KwaZulu-Natal, South Africa. In progress, 2017.