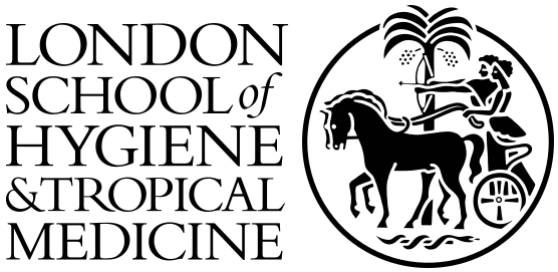

International Journal of Epidemiology, 2020, 1–11
doi: 10.1093/ije/dyaa012
Original article

Original article

Statistical inference in matched case–control studies of recurrent events

Yin Bun Cheung [1,2,3]* Xiangmei Ma,[2] K.F. Lam,[2,4] Jialiang Li,[5] Chee Fu Yung,[6] Paul Milligan[7] and Grant Mackenzie[8,9,10,11]

[1]Signature Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore 169857, [2]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore 169856, [3]Center for Child Health Research, University of Tampere and Tampere University Hospital, Tampere 33520, Finland, [4]Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong, China, [5]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, [6]Infectious Disease Service, KK Women's and Children's Hospital, Singapore 229899, [7]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK, [8]Medical Research Council Unit The Gambia at London School of Hygiene & Tropical Medicine, PO Box 273, Fajara, The Gambia, [9]New Vaccines Group, Murdoch Children's Research Institute, Melbourne, Victoria 3052, Australia, [10]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK and [11]Department of Paediatrics, University of Melbourne, Parkville, VIC 3010, Australia

*Corresponding author. Center for Quantitative Medicine, Duke-NUS Medical School, Singapore, 20 College Road, Singapore 169856. E-mail: yinbun.cheung@duke-nus.edu.sg

Editorial decision 13 January 2020; Accepted 16 January 2020

Abstract

Background: The concurrent sampling design was developed for case–control studies of recurrent events. It involves matching for time. Standard conditional logistic-regression (CLR) analysis ignores the dependence among recurrent events. Existing methods for clustered observations for CLR do not fit the complex data structure arising from the concurrent sampling design.

Methods: We propose to break the matches, apply unconditional logistic regression with adjustment for time in quintiles and residual time within each quintile, and use a robust standard error for observations clustered within persons. We conducted extensive simulation to evaluate this approach and compared it with methods based on CLR. We analysed data from a study of childhood pneumonia to illustrate the methods.

Results: The proposed method and CLR methods gave very similar point estimates of association and showed little bias. The proposed method produced confidence intervals that achieved the target level of coverage probability, whereas the CLR methods did not, except when disease incidence was low.

Conclusions: The proposed method is suitable for the analysis of case–control studies with recurrent events.

Key words: Concurrent design, logistic regression, incidence density sampling, matched case–control study

Key Messages

- Case–control studies of recurrent events using the concurrent sampling design can be analysed by unconditional logistic regression with adjustment for time variables to obtain estimates of association highly comparable to conditional logistic regression.
- Together with the robust standard error for clustered data, this analytic approach produces confidence intervals that achieve the intended coverage probability.

Introduction

In an important article on case–control studies, Rodrigues and Kirkwood described three designs to sample the participants: the exclusive design, the inclusive design and the concurrent design.[1] The interpretation of the odds ratios (ORs) obtained from these different sampling designs is different.[1–3] With the concurrent sampling design, the OR estimator provides the incidence rate ratio (IRR) for an exposed group compared with an unexposed group.[1]

The concurrent design is more generally referred to as an incidence density sampling design,[1–3] which has the following defining characteristics: (i) controls are sampled at a rate proportional to the incidence rate of disease cases, (ii) the probability of any person being selected as a control is proportional to that person's amount of time at risk during the study period and (iii) it is possible for a subject to be sampled multiple times. Controls are most commonly chosen from among those at risk at the moment each case occurs, i.e. concurrent sampling. In this design, the ratio of exposed to unexposed person-time in the source population is estimated by the ratio of the number of exposed to unexposed controls.[1,3] It thus estimates the IRR. A graphical illustration of the concurrent design is included in Supplementary File 1, available as Supplementary data at *IJE* online.

Rodrigues and Kirkwood were explicit that the design can be used to study either first/terminal events such as death or first episode of a disease, or recurrent events such as episodes of diarrhoea.[1] In the conceptual framework of the concurrent design, a person may recover from the target disease and return to the at-risk population, becoming available again for sampling as a control or as a case. In contrast, epidemiology textbooks appear to implicitly limit the discussion of the incidence density sampling to the studies of first/terminal events, assuming that, once a person becomes a case, s/he permanently drops out from the at-risk population. A matched estimator of the OR has been recommended for the analysis of data obtained from the concurrent design unless the prevalence of exposure is constant over time in the source population.[1,4]

Recurrent events are common in clinical and epidemiological evaluations. Some examples include episodes of infectious diseases (where the infection does not confer lifetime immunity), injuries and hospital admissions. The concurrent design is an important extension of the case–control study approach. However, Rodrigues and Kirkwood did not shed light on how to perform statistical inference,[1] i.e. estimation of confidence interval and hypothesis testing, when a person can become a case multiple times. Standard statistical methods are not valid for such correlated data. Approaches to the analysis of recurrent events have been well discussed in the setting of prospective studies (e.g. [5–9]). There is a gap in the literature on case–control studies.

In the literature, case–control studies of 'disease recurrence' or 'recurrent event' often define the outcome as a binary variable that is not repeatable, such as 'having a disease the second time' (case) vs 'having a disease the first time' (control) (e.g. [10,11]). Despite using the phrase 'recurrence' or 'recurrent', they are studies of a single event. Examples of case–control studies that truly used the concurrent design for recurrent events include studies of diarrhoea episodes,[12] malaria episodes,[13] asthma attacks[14] and hospital readmissions.[15] Their statistical analyses treated the data as if all events were independent, mostly using conditional logistic regression (CLR).

In the context of some studies in ecology, a robust standard error for CLR analysis of clustered data has been proposed.[16,17] However, this method is not suitable for the data structure in the concurrent design. In these ecology studies, each animal makes choices. What an animal chooses and does not choose are the 'cases' and its matched 'controls', respectively. In this context, all observations—cases or controls—within a matched set are contributed by the same animal, making it appropriate for matched sets concerning the same animal to be treated as one cluster of

observations. In contrast, in the concurrent design, each time a person has a disease episode (case), it is matched with some other person(s) who are at risk (controls). The matched set here involves multiple subjects. Furthermore, this person who is the case (or control) may also contribute observations as a control (or case) in some other matched sets. The aforementioned robust standard error does not fit this complex data structure. Similarly, the bootstrapping method is not applicable because there is no resampling unit that is suitable for this data structure.

In this article, we propose an approach based on breaking the matches and applying unconditional logistic regression (ULR) for the statistical analysis of case–control studies that use the concurrent design with recurrent events. We evaluate this and other methods by extensive simulation. For illustration, we apply the methods to a study of childhood pneumonia.

Methods

Study design and analytic methods

In this section, we describe the statistical methods to be evaluated. Mathematical details of the statistical methods are available in Appendix 1.

In the concurrent design, when the i-th person in the at-risk population experiences his/her j-th outcome event at calendar time t_{ij}, the person is recruited as the index case and K ($K \geq 1$) persons who are at risk at time t_{ij} are sampled as the matched controls. That is, the matching variable is calendar time (t), which is a continuous variable. The i-th person remains eligible to be sampled as a case or control after t_{ij} as long as s/he remains in the source population during the study period.

A conventional view on the analysis of individually matched case–control studies is that the data should be analysed by CLR. However, in the analysis of case–control studies that uses the exclusive design, it has been demonstrated that ULR with adjustment for a binary matching factor gives an unbiased estimate of the OR.[18] For studies that individually match cases and controls for age, a continuous variable, it has been suggested that age can be grouped into broad intervals and adjusted for as indicator variables,[19] possibly with further adjustment for a linear residual term within each age interval to reduce the residual coarseness.[20,21] The adjustment for broad intervals assumed that the event rate is a step function that jumps at the boundaries of intervals but remains constant within each interval. This may not be sufficient for capturing the variation. The linear residual term within each interval removes the assumption of a constant event rate within the interval. It allows the event rate to decrease or increase linearly within the interval. Hence, the combined use of the intervals and linear residual term provides a very flexible, non-parametric form of statistical adjustment. Previous research has shown that directly adjusting for a continuous matching variable in an unconditional model is biased.[20,21] This is because, even if there is a smooth dose–response trend in the population, the case–control matching can distort the response curve to a discontinuous form in the study sample. Therefore, direct adjustment assuming a smooth trend is not justified. The more complex, discretized adjustment is needed.

A recent study on the analysis of case–control studies with incidence density sampling for a single event outcome shows that the use of ULR with adjustment for time in quintiles usually gives results highly comparable to CLR analysis.[22] The choice of quintiles was motivated by the research of Cox about grouping observations of a continuous variable into tertiles, quartiles or quintiles that would retain approximately 79%, 86% or 90% of the information, respectively.[23] The incremental information retention per additional group is limited beyond that; inclusion of too many groups may generate a sparse data bias.[22,24] Based on these recent findings, we propose to break the matches, and then use ULR with adjustment for calendar time in quintiles, with or without including a linear residual term within each quintile, and apply the robust standard-error estimator for clustered data,[25] defining each person as a cluster. For brevity, we call the methods that adjust for time in quintiles with and without linear residual terms ULR-QL and ULR-Q, respectively.

Standard CLR with naïve standard error (CLR-N) assumes the outcome events to be independent. This tends to inflate the type 1 error rate when the assumption is false. However, if only a few people have more than one event during the study period, the impact of the violation of the assumption of independence may be small. On the other hand, whereas the aforementioned robust standard-error estimator for CLR analysis of clustered data proposed in the context of ecology research (CLR-C[16,17]) is in principle not suitable for the concurrent design, its performance in realistic epidemiological situations is unknown. It may happen that, if most people in a case–control study are recruited only once whereas some people are recruited multiple times as cases, the application of the CLR-C using the group of matched sets generated for the same person who is the case in these matched sets as a cluster of observations may be approximately correct. To understand whether there are situations in which the use of these methods may give approximately correct results, we included these two methods in the evaluations described below.

Note that CLR-N and CLR-C share the same point estimate of association. They only differ in the standard error

of the estimate. Therefore, we evaluate three methods in terms of bias in point estimates (CLR, ULR-Q and ULR-QL) but four methods in coverage probabilities of confidence intervals (CLR-N, CLR-C, ULR-Q and ULR-QL).

Simulation settings and procedures

This section outlines the key features of the simulation. Details of the procedures are available in Supplementary File 2, available as Supplementary data at *IJE* online.

Most case–control studies and realistic situations that epidemiologists encounter concern dynamic populations.[2,26] We therefore conducted simulation of dynamic populations over 144 scenarios defined by:

- high or low incidence;
- hazard is constant or changing over time;
- exposure is time-constant or time-varying;
- the effect of exposure is time-constant or time-varying (waning) or it has no effect;
- with or without confounding by and adjustment for age;
- number of controls per case is one, two or four.

Within the calendar time period, the study recruited all events in the population as cases and, for each case, randomly selected K controls ($K = 1$, 2 or 4) who were at risk of the disease at the same calendar time as the index case event occurred. The parameters that defined the incidence rate were chosen to resemble some previous reports of high incidence of malaria and acute otitis media[27,28] and low incidence of inpatient admission due to malaria and radiological pneumonia with consolidation in young children.[6,27] We used the Weibull distribution with shape parameter $\gamma = 1.0$ or 0.7 to generate two time patterns. Exposure is either time-constant (since birth) or time-varying (since a randomly generated age at initiation of exposure). We considered three patterns of exposure effect: no effect ($\beta = 0$), time-constant protective effect ($\beta = -0.3$) or waning effect [$\beta(t) = \beta \times (2 - t/365)$], where $\beta = -0.3$, such that the protective effect waned as time (in days) went by. We considered scenarios with or without confounding by age. Whenever there was confounding by age in the data-generation process, the analysis models adjusted for age but the study design did not match for age. This serves to reiterate a point that, in case–control studies, controlling for confounders does not necessitate matching for the confounders.

We set the source population size as 1000 or 4000 for the high- and low-incidence scenarios, respectively. This corresponds to realistic situations in which a larger population (or longer study duration) is needed for the study of a less common event. In a supplementary set of evaluations, we reduced the population size by 50%. Recursive algorithms were used to simulate the data.[29] They allowed the flexibility of introducing time-varying exposure and/or time-varying effects in the generation of recurrent-event data.

In each scenario, we conducted 1000 replicates of data simulation and data analysis by ULR-Q, ULR-QL, CLR-N and CLR-C. In scenarios with $\beta = 0$, for each analytic method, we calculated the absolute bias (mean of observed estimates for β − the true value of β|) and the coverage probability of the 95% confidence interval (CI). Since the absolute bias was small, we presented 100× absolute bias instead. In scenarios with $\beta \neq 0$, we calculated the relative bias as [(mean of observed estimates for β − true value of β)/true value of β] and the empirical coverage probability of the 95% CI. If the true coverage probability of the 95% CI is 0.95, the use of 1000 replicates in the simulation gives a standard error of $\sqrt{(0.95 \times 0.05)/1000} = 0.0069$.

Childhood pneumonia: a case study

For illustration, we conducted a case–control study nested within a population-based surveillance system for monitoring pneumonia and pneumococcal diseases in young children in the Gambia, Western Africa. Details of the surveillance system have been previously published.[30] We analysed clinical pneumonia as defined previously,[31] which was a high-incidence outcome, and radiological pneumonia with consolidation as defined by the World Health Organization,[32] which was a lower-incidence outcome, in a 2-year period among children below the age of 5 years in relation to time-varying (age) and time-constant (e.g. mother's education) exposure variables.

A successful case–control study should provide results similar to those in a cohort study.[3] We began with analysing the data as a cohort study using the Andersen–Gill model with calendar time as the timescale.[6] We applied the concurrent design to conduct a case–control study that included all clinical pneumonia episodes as cases and a case–control study of all radiological pneumonia episodes as cases. For each case, we randomly selected four controls, individually matched for calendar time and geographic areas defined by distance to the nearest health facilities (seven in total). We applied CLR-N, CLR-C, ULR-Q and ULR-QL to the case–control-study data. In addition to adjusting for calendar time, the ULR analyses also adjusted for the effects of different geographic areas, as this was one of the matching variables.[3]

Results

Simulation

Figure 1 shows the results for scenarios with time-decreasing hazard, time-constant exposure, time-constant protective effect of exposure and no confounding by age.

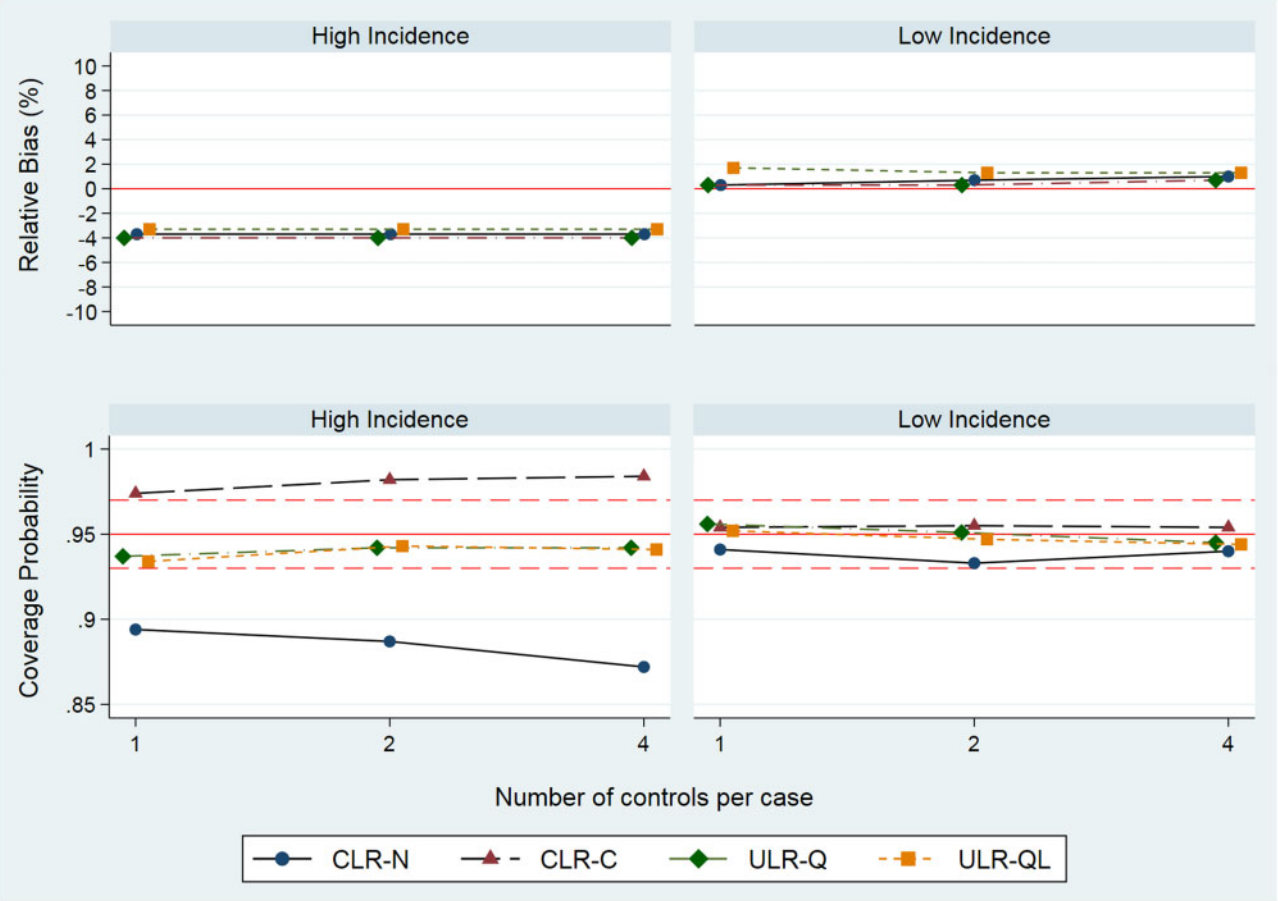

Figure 1. Relative bias and coverage probability in relation to levels of disease incidence, control-to-case ratios and analytic methods for a scenario in which the hazard declines over time, exposure is time-constant, the protective effect of the exposure is time-constant and there is no age effect. Source population size = 1000 and 4000 for high- and low-incidence scenarios, respectively. Solid reference lines: relative bias 0% and coverage probability 95%; dashed reference lines: coverage probability 93% and 97%.

Regardless of the level of disease incidence and the number of controls per case, the two unconditional logistic-regression methods, ULR-Q and ULR-QL, gave point estimates that closely followed the CLR estimates. In the high-incidence scenarios, all methods had a relative bias of about −4%. Since the true β was −0.3 (IRR 0.741), the relative bias implied $\beta \approx -0.288$ (IRR 0.750). The magnitude of bias was negligible.

The lower panel shows the coverage probability of the 95% CI. CLR-N gave confidence intervals that were too narrow. In the high-incidence scenarios, the 95% CI covered the true β <90% of the times. CLR-C had coverage probability that was higher than the 95% target level in the high-incidence scenarios. The coverage probabilities obtained from ULR-Q and ULR-QL were approximately within a ±2% range of the 95% target.

Figure 2 shows the results in scenarios with time-varying exposure. Otherwise, the parameters are the same as in Figure 1. The under-coverage of the 95% CIs from CLR-N was less serious than in Figure 1. Other findings are similar to those shown in Figure 1.

Figure 3 shows scenarios with the same setting as in Figure 2, except that the exposure has no effect, i.e. $\beta = 0$, and there was adjustment for age as a confounder. In both the high- and low-incidence scenarios, all the regression models show a slight degree of absolute bias. There was no big difference in bias between the methods. This is similar to the findings in Figure 2. ULR-QL and ULR-Q had similar coverage probability that varied within the range from about 93% to 95%. CLR-N and CLR-C had more deviation from the target level of 95% coverage probability than ULR-Q and ULR-QL in scenarios with high incidence.

Further simulation results are included in Supplementary File 3, available as Supplementary data at *IJE* online. They showed patterns similar to Figures 1–3.

Childhood pneumonia: a case study

Table 1 shows the number of clinical and radiological pneumonia episodes per child in the cohort. Most children had none, 2256 children had one and 448 children had two or more episodes of clinical pneumonia. Only eight

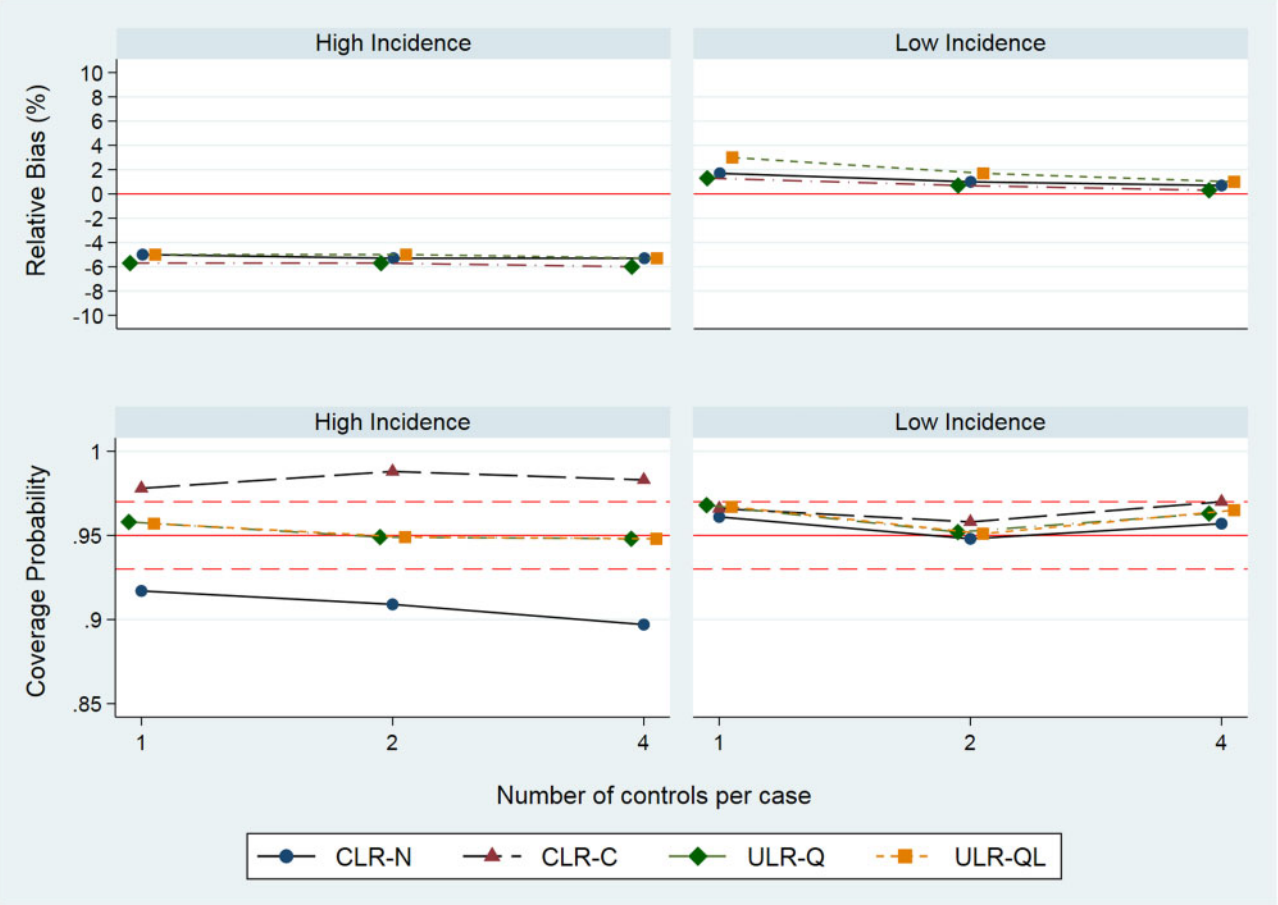

Figure 2. Relative bias and coverage probability in relation to levels of disease incidence, control-to-case ratio and analytic methods for a scenario in which the hazard declines over time, exposure is time-varying, the protective effect of the exposure is time-constant and there is no age effect. Source population size = 1000 and 4000 for high- and low-incidence scenarios, respectively. Solid reference lines: relative bias 0% and coverage probability 95%; dashed reference lines: coverage probability 93% and 97%.

children had two or more episodes of radiological pneumonia.

Table 2 shows the analytic results for clinical pneumonia in the cohort and in the nested case–control study. The cohort analysis and the four ways of analysing the nested case–control studies mostly gave similar results, in terms of both the IRR estimates and their level of statistical significance. ULR-Q and ULR-QL gave IRR estimates that were mostly identical to CLR estimates (up to two decimal places). A difference was that CLR-N showed a statistically significant difference between levels of mother's education—Madrassa/Quranic Education vs No Education (IRR 0.87; 95% CI 0.77 to 0.98)—whereas the cohort and the other analyses of the case–control study did not. For time-constant covariates such as gender, there was a tendency that the width of the CI was in the direction of CLR-N < ULR-Q/ULR-QL < CLR-C, which agreed with the simulation results. For the time-varying covariate, age, the width of the CI was similar for all the analyses of the case–control-study data. This was also supported by the simulation findings, reflecting the fact that a child who had multiple disease episodes might not have multiple episodes within the same age interval. So the recurrence might be less impactful on the analysis of time-varying covariates.

Analyses of cohort and case–control studies of radiologically confirmed pneumonia are included in Supplementary File 4, available as Supplementary data at *IJE* online. The findings were similar between the cohort analysis and the four sets of case–control-study analyses.

Discussion

In clinical trials of infectious diseases with outcomes that can recur, until recently, it was a common practice to limit analyses to the first episode of the disease.[33] In 2008, a WHO advisory committee highlighted the importance of including all disease episodes in the analysis of randomized clinical trials.[33] Recently, improved statistical methods and consensus on how best to analyse recurrent events in the context of randomized clinical trials have emerged (e.g. [5,6]). There is also an awareness that analyses of first and all

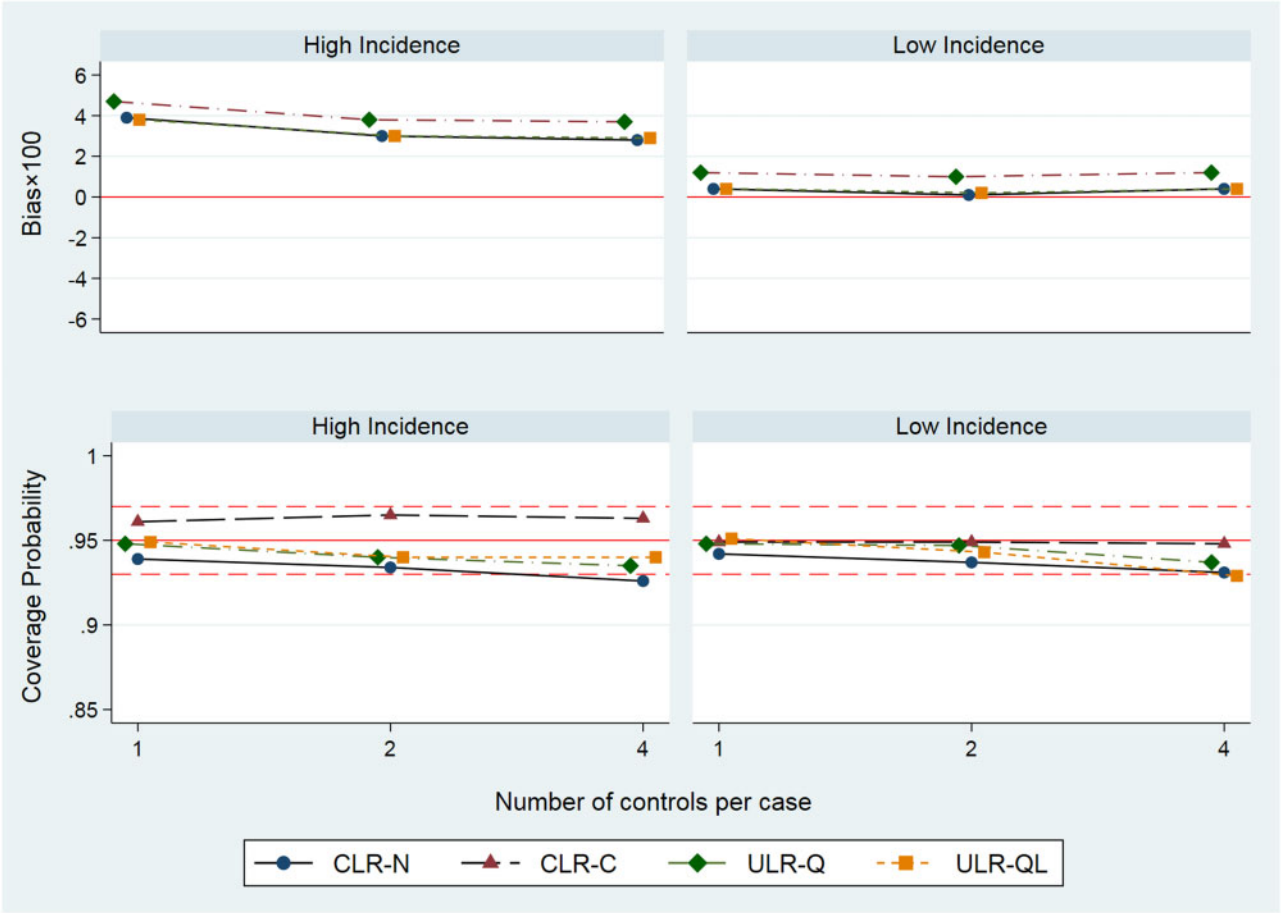

Figure 3. Absolute bias × 100 and coverage probability in relation to levels of disease incidence, control-to-case ratios and analytic methods for a scenario in which the hazard declines over time, exposure is time-varying, the exposure has no effect ($\beta = 0$) and there is an adjustment for age effect. Source population size = 1000 and 4000 for high- and low-incidence scenarios, respectively. Solid reference lines: bias 0 and coverage probability 95%; dashed reference lines: coverage probability 93% and 97%.

Table 1. Number of episodes of clinical pneumonia and radiological pneumonia per child in the Gambian cohort

Episodes	Clinical pneumonia	Radiological pneumonia
0	43 477	45 857
1	2256	316
2	316	7
3	81	0
4	31	1
5	12	0
6	3	0
7	2	0
8	2	0
9	0	0
≥10	1	0

episodes do not necessarily estimate the same effect. For example, analysis of the first episode only would likely emphasize the short-term effect of the exposure and downplay the long-term effect, and vice versa. Nevertheless, the approaches for analysis of recurrent events in case–control studies have remained unclear.

Capitalizing on the emerging consensus on the validity of the use of ULR with adjustment for matching variables for individually matched case–control studies,[18,20,22] we propose a strategy for analysis of case–control studies that use the concurrent design, which involves matching for time, a continuous variable. After breaking the matches, it becomes straightforward to apply the robust standard error for cluster data, with each person being a cluster. Our extensive simulation evaluation has shown that this approach gave estimates that approximately followed the CLR estimates. They gave accurate coverage probability of the CI. In contrast, the use of CLR-N and CLR-C tended to give CIs that were too narrow or too wide, respectively, although they had acceptable performance in scenarios with low incidence. Some researchers had advocated that, after breaking the matches, a residual term within each interval of continuous matching variable should be included.[20,21] In some simulation scenarios, we did see that this approach performed

Table 2. Analysis of data from cohort and nested case–control study of clinical pneumonia and four controls per case

Exposure variables	Cohort study		Case–control study							
			CLR-N		CLR-C		ULR-Q		ULR-QL	
	IRR	95% CI	IRR	95% CI	IRR	95% CI	IRR	95%CI	IRR	95%
Age (years)										
0–1	1.00		1.00		1.00		1.00		1.00	
1–2	0.77	(0.70–0.84)	0.74	(0.66–0.82)	0.74	(0.66–0.82)	0.74	(0.67–0.82)	0.74	(0.66–0.82)
≥3	0.23	(0.21–0.25)	0.23	(0.21–0.26)	0.23	(0.21–0.26)	0.24	(0.21–0.26)	0.23	(0.21–0.26)
Gender										
Female	1.00		1.00		1.00		1.00		1.00	
Male	1.18	(1.09–1.28)	1.17	(1.08–1.27)	1.17	(1.06–1.29)	1.16	(1.06–1.27)	1.16	(1.06–1.27)
Ethnicity										
Fula	1.00		1.00		1.00		1.00		1.00	
Mandinka	1.58	(1.41–1.78)	1.60	(1.42–1.80)	1.60	(1.40–1.83)	1.62	(1.42–1.85)	1.62	(1.42–1.85)
Sarahule	1.19	(1.06–1.32)	1.23	(1.11–1.36)	1.28	(1.08–1.40)	1.23	(1.09–1.39)	1.23	(1.09–1.39)
Other	1.20	(0.98–1.48)	1.26	(0.99–1.60)	1.26	(0.97–1.62)	1.25	(0.97–1.60)	1.25	(0.97–1.61)
Mother's education										
None	1.00		1.00		1.00		1.00		1.00	
Basic	0.88	(0.71–1.09)	0.84	(0.69–1.03)	0.84	(0.66–1.08)	0.84	(0.66–1.06)	0.84	(0.66–1.06)
Secondary	1.08	(0.83–1.41)	1.20	(0.90–1.60)	1.20	(0.85–1.69)	1.20	(0.87–1.64)	1.20	(0.88. 1.64)
College/University	0.80	(0.67–0.97)	0.74	(0.62–0.89)	0.74	(0.59–0.93)	0.73	(0.59–0.90)	0.73	(0.59. 0.90)
Madrassa/Quranic	0.93	(0.80–1.07)	0.87	(0.77–0.98)	0.87	(0.74–1.03)	0.87	(0.75–1.02)	0.87	(0.75–1.01)
Other	0.82	(0.68–0.99)	0.80	(0.66–0.96)	0.80	(0.64–0.99)	0.78	(0.63–0.96)	0.78	(0.63–0.96)
Distance to health centre (per km)	0.95	(0.94–0.95)	0.94	(0.93–0.95)	0.94	(0.93–0.95)	0.94	(0.93–0.95)	0.94	(0.93–0.95)

better than its counterpart without the residual terms, although the difference was quite small.

These patterns were also found in the study of childhood pneumonia in the Gambia. Nevertheless, in the study of radiological pneumonia, the number of persons with multiple episodes was small and there was no discernible difference in the results between any of the analyses. Although previous case–control studies that used the concurrent design might have used naïve methods to perform statistical inference,[12–15] we speculate that their conclusions were approximately correct unless the degree of disease recurrence was as common as that of clinical pneumonia we analysed here.

In conclusion, the strategy of breaking the matches followed by ULR with adjustment for time in quintiles and residual time within quintile and robust standard error for clustered observations provides correct statistical inference for case–control studies of recurrent events.

Supplementary data

Supplementary data are available at *IJE* online.

Funding

The funding agencies played no role in the study design, analysis and interpretation of data, or writing and submission of the article.

Acknowledgements

The methodological work was supported by the National Medical Research Council, Singapore (NMRC/ CIRG/1475/ 2017). The population-based surveillance system in the Gambia was funded by the Gavi Alliance's PneumoADIP—Bloomberg School of Public Health, Johns Hopkins University; the Bill & Melinda Gates Foundation (OPP 1020372); and MRC (UK).

Conflict of interest: None declared.

Author Contributions

Y.B.C. conceived the study, designed the study, interpreted the findings and wrote the first and final version of the article. X.M. implemented the simulation and data analysis, interpreted the findings and critically reviewed and revised the draft article. G.M. collected the data used in the case study. K.F.L., J.L., C.F.Y., G.M. and P.M. participated in the design of the study, interpreted the findings and critically reviewed and revised the draft article. All authors approved the final version and agreed to be accountable for all aspects of the work.

References

1. Rodrigues L, Kirkwood BR. Case-control designs in the study of common diseases. *Int J Epidemiol* 1990;**19**:205–13.
2. Knol MJ, Vandenbroucke JP, Scott P, Egger M. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol* 2008;**168**: 1073–81.

3. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*, 3rd edn. Philadelphia, PA: Lippincott, Williams & Wilkins, 2008.
4. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 1982;**116**: 547–53.
5. Cairns M, Cheung YB, Xu T *et al*. Analysis of malaria cohort studies: exploring partial and complete protection, and total and primary intervention effects. *Am J Epidemiol* 2015;**181**:1008–17.
6. Cheung YB, Xu Y, Tan SH, Cutts F, Milligan P. Estimation of intervention effects using first or multiple episodes in clinical trials: the Andersen-Gill model re-examined. *Stat Med* 2010;**29**:328–36.
7. Jahn-Eimermacher A, Du Prel JB, Schmitt HJ. Assessing vaccine efficacy for the prevention of acute otitis media by pneumococcal vaccination in children: a methodological overview of statistical practice in randomized controlled clinical trials. *Vaccine* 2007; **25**:6237–44.
8. Kelly PJ, Lim LL. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Stat Med* 2000;**19**: 13–33.
9. Metcalfe C, Thompson SG. Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome? *Stat Methods Med Res* 2007;**16**:103–22.
10. Mohammad R, Halboub E, Mashlah A, Abou-Hamed H. Levels of salivary IgA in patients with minor recurrent aphthous stomatitis: a matched case–control study. *Clin Oral Invest* 2013;**17**:975–80.
11. Muñoz P, Vena A, Valerio M *et al*. Risk factors for late recurrent candidaemia: a retrospective matched case–control study. *Clin Microbiol Infect* 2016;**22**:277.e11–20.
12. Mertens TE, Fernando MA, Cousens SN, Kirkwood BR, Marshall TF, Feachem RG. Childhood diarrhoea in Sri Lanka: a case-control study of the impact of improved water sources. *Trop Med Parasitol* 1990;**41**:98–104.
13. DE LA Hoz F, Cruz J, Hall AJ *et al*. Case-control study of mosquito nets against malaria in the Amazon region of Colombia. *Am J Trop Med Hygiene* 2005;**73**:140–48.
14. Loyo-Berríos NI, Irizarry R, Hennessey JG, Tao XG, Matanoski G. Air pollution sources and childhood asthma attacks in Catano, Puerto Rico. *Am J Epidemiol* 2007;**165**:927–35.
15. Silva NC, Bassani DG, Palazzo LS. A case-control study of factors associated with multiple psychiatric readmissions. *Psychiatr Serv* 2009;**60**:786–91.
16. Craiu RV, Duchesne T, Fortin D. Inference methods for the conditional logistic regression model with longitudinal data. *Biom J* 2008;**50**:97–109.
17. Prima MC, Duchesne T, Fortin D. Robust inference from conditional logistic regression applied to movement and habitat selection analysis. *PLoS One* 2017;**12**:e0169779.
18. Pearce N. Analysis of matched case-control studies. *BMJ* 2016; **352**:i969.
19. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford: Oxford University Press, 1993.
20. Mansournia MA, Jewell NP, Greenland S. Case-control matching: effects, misconceptions, and recommendations. *Eur J Epidemiol* 2018;**33**:5–14.
21. Greenland S, Partial and marginal matching in case-control studies. In Moolgavkar SH, Prentice RL (eds). *Modern Statistical Methods in Chronic Disease Epidemiology*. New York: Wiley, 1986, pp. 35–49.
22. Cheung YB, Ma X, Lam KF, Li J, Milligan P. Bias control in the analysis of case-control studies with incidence density sampling. *Int J Epidemiol* 2019;**48**:1981–91.
23. Cox DR. Note on grouping. *J Am Stat Assoc* 1957;**52**:543–47.
24. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *J R Stat Soc: Ser B* 1991;**53**:629–43.
25. Gould W, Sribney W. *Maximum Likelihood Estimation with Stata*. College Station, TX: Stata Press, 1999.
26. Vandenbroucke JP, Pearce N. Case-control studies: basic concepts. *Int J Epidemiol* 2012;**41**:1480–89.
27. Chandramohan D, Owusu-Agyei S, Carneiro I *et al*. Cluster randomised trial of intermittent preventive treatment for malaria in infants in area of high, seasonal transmission in Ghana. *Br Med J* 2005;**331**:727–33.
28. Eskola J, Kilpi T, Palmu A *et al*. Efficacy of a pneumococcal conjugate vaccine against acute otitis media. *N Engl J Med* 2001; **344**:403–09.
29. Xu J, Lam KF, Chen F, Milligan P, Cheung YB. Semiparametric estimation of time-varying intervention effects using recurrent event data. *Stat Med* 2017;**36**:2082–96.
30. Mackenzie GA, Plumb ID, Sambou S *et al*. Monitoring the introduction of pneumococcal conjugate vaccines into West Africa: design and implementation of a population-based surveillance system. *PLoS Med* 2012;**9**:e1001161.
31. Mackenzie GA, Hill PC, Sahito SM *et al*. Impact of the introduction of pneumococcal conjugate vaccination on pneumonia in the Gambia: population-based surveillance and case-control studies. *Lancet Infect Dis* 2017;**17**:965–73.
32. World Health Organization, Department of Vaccines and Biologicals. *Standardization of Interpretation of Chest Radiographs for the Diagnosis of Pneumonia in Children*. WHO/V&B/01.35. Geneva: WHO, 2001.
33. Moorthy VS, Reed Z, Smith PG. WHO Malaria Vaccine Advisory Committee. MALVAC 2008: measures of efficacy of malaria vaccines in phase 2b and phase 3 trials. *Vaccine* 2009; **27**:624–28.

Appendix 1. Statistical models

Suppose there are N individuals in the source population. Without loss of generality, we assume that the first M individuals had at least one episode during the study duration. Suppose individual i $(i = 1, 2 \ldots, M)$ with exposure status X_i had m_i episodes during follow-up time (where $m_i > 0$). At each event time t_{ij} $(i = 1, \ldots M; j = 1, \ldots, m_i)$, we randomly select K controls from the at-risk population without replacement using the concurrent design and matching criteria as described in the text. In total, we have $\sum_{i=1}^{M} m_i = S$ matched case–control sets in the data set. The total number of observations in the case–control study is $S(1 + K)$. In each matched set, there is one case with indicator $ij0$ $(Y_{ij0} = 1)$ and K controls with indicators ijk $(Y_{ijk} = 0,\ k = 1, \ldots,\ K)$ for the selected control subjects associated with episode j of subject i.

1. Conditional logistic regression

Conditional logistic-regression (CLR) estimates β by maximizing the log conditional likelihood $l_{CLR}(\beta\ ) = \sum_{i=1}^{M} \ln L_{CLR,i}(\beta\ )$, where $L_{CLR,i}(\beta)$ is:

$$L_{CLR,\ i}(\beta_E) = \prod_{j=1}^{m_i} \frac{e^{\beta X_{ij0}}}{e^{\beta X_{ij0}} + \sum_{k=1}^{K} e^{\beta X_{ijk}}}.$$

Let $U_{CLR}(\beta)$ be the score function,

$$U_{CLR}(\beta) = \frac{\partial l_{CLR}(\beta\ )}{\partial \beta} = \sum_{i=1}^{M} U_{CLR,\ i}(\beta) = \sum_{i=1}^{M} \left[m_i X_i - \sum\nolimits_{j=1}^{m_i} \frac{X_i e^{\beta X_i} + \sum_{k=1}^{K} X_{ijk} e^{\beta X_{ijk}}}{e^{\beta X_i} + \sum_{k=1}^{K} e^{\beta X_{ijk}}} \right],$$

since $X_{ij0} = X_i$ for $j = 1, \ldots,\ m_i$. Let $D_{CLR} = \frac{\partial U_{CLR}(\beta)}{\partial \beta}$ be the second derivatives of the log likelihood and $\hat{\beta}_{CLR}$ is the solution of the score equation $U_{CLR}(\beta) = 0$. Let $\hat{D}_{CLR}$ be D_{CLR} with β replaced by $\hat{\beta}_{CLR}$. The naïve variance estimator for the conditional logistic regression (CLR-N) is $\hat{\text{var}}_{CLR}(\beta) = -\hat{D}_{CLR}^{-1}$.

The sandwich robust estimator of variance for clustered data with the matched sets that have the same individual as the case considered a cluster of observations is obtained by (CLR-C):

$$\hat{\text{var}}_{CLR}(\beta)^{cluster} = \hat{D}_{CLR}^{-1} \sum_{i=1}^{M} \left[\hat{U}_{CLR,\ i}\left(\hat{\beta}_{CLR}\right) \right]^2 \hat{D}_{CLR}^{-1} = \hat{D}_{CLR}^{-1} \sum_{i=1}^{M} \left[m_i X_i - \sum\nolimits_{j=1}^{m_i} \frac{X_i e^{\hat{\beta} X_i} + \sum_{k=1}^{K} X_{ijk} e^{\hat{\beta} X_{ijk}}}{e^{\hat{\beta} X_i} + \sum_{k=1}^{K} e^{\hat{\beta} X_{ijk}}} \right]^2 \hat{D}_{CLR}^{-1}.$$

2. Unconditional Logistic Regression

An unconditional logistic regression (ULR) has the following likelihood function:

$$L_{ULR}(\beta) = \prod_{i=1}^{M} \prod_{j=1}^{m_i} \prod_{k=0}^{K} \frac{e^{Y_{ijk}(\beta_0 + \beta X_{ijk})}}{1 + e^{\beta_0 + \beta X_{ijk}}} = \prod_{i'=1}^{n} \prod_{j'=1}^{n_{i'}} \frac{e^{Y_{i'j'}(\beta_0 + \beta X_{i'j'})}}{1 + e^{\beta_0 + \beta X_{i'j'}}},$$

where n is the total number of individuals recruited in the case–control study and $n_{i'}$ is the total number of times an individual is recruited as either a case or a control. $Y_{i'j'} = 1$ if individual i' is recruited as a case at j'-th time, otherwise $Y_{i'j'} = 0$.

Let $U_{ULR}(\beta)$ be the corresponding score function,

$$U_{ULR}(\beta) = \frac{\partial \ln(L_{ULR}(\beta\ ))}{\partial \beta} = \sum_{i'=1}^{n} U_{ULR,\ i'}(\beta) = \sum_{i'=1}^{n} \left[\sum\nolimits_{j'=1}^{n_{i'}} \left(Y_{i'j'} X_{i'j'} - \frac{X_{i'j'} e^{\beta_0 + \beta X_{i'j'}}}{1 + e^{\beta_0 + \beta X_{i'j'}}} \right) \right].$$

Let $D_{ULR} = \frac{\partial U_{ULR}(\beta)}{\partial \beta}$ be the second derivatives of the log likelihood and $\hat{\beta}_{ULR}$ is the solution of the score equation $U_{ULR}(\beta) = 0$. Let $\hat{D}_{ULR}$ be D_{ULR} with β replaced by $\hat{\beta}_{ULR}$. The naïve variance estimator for the ULR is $\hat{\text{var}}_{ULR}(\beta) = -\hat{D}_{ULR}^{-1}$.

The sandwich robust estimator of variance for clustered data with all observations from an individual considered a cluster of observations is obtained by:

$$\hat{\text{var}}_{ULR}(\beta)^{cluster} = \hat{D}_{ULR}^{-1} \sum_{i'=1}^{n} \left[\hat{U}_{ULR,\ i'}\left(\hat{\beta}_{ULR}\right) \right]^2 \hat{D}_{ULR}^{-1}$$

The unconditional logistic regression with adjustment for time in quintiles (ULR-Q) estimates β by adding to the simple ULR four indicator variables that contrast the second to fifth quintiles of event time against the first quintile as the reference:

$$Prob(Y = 1 | X, Q_2, Q_3, Q_4, Q_5) = \frac{e^{\beta_0 + \beta X + \sum_{q=2}^{5} \beta_{Q_q} Q_q}}{1 + e^{\beta_0 + \beta X + \sum_{q=2}^{5} \beta_{Q_q} Q_q}},$$

where Q_q $(q = 2,\ 3,\ 4,\ 5)$ is an indicator variable for the q-th event-time quintile, $Q_q = 0$ or 1. Due to the matching for calendar time, a case and its control(s) must be in the same event-time quintile.

The unconditional logistic regression with adjustment for time in quintiles and a linear trend for residual time within quintile (ULR-QL) estimates β by adding to ULR-Q a linear trend to account for residual changes over time within each time quintile:

$$Prob(Y=1|X,Q_2,\ldots,Q_5,R_1,\ldots,R_5) = \frac{e^{\beta_0+\beta X+\sum_{q=2}^{5}\beta_{Q_q}Q_q+\sum_{l=1}^{5}\beta_{R_l}R_l}}{1+e^{\beta_0+\beta X+\sum_{q=2}^{5}\beta_{Q_q}Q_q+\sum_{l=1}^{5}\beta_{R_l}R_l}},$$

where $R_l = (t-\bar{t}_{Q_l})Q_l,\ (l=1,\ 2,\ldots,5)$, t is event time, $\bar{t}_{Q_l}$ is the mean time within the l-th quintile and Q_l is an indicator variable for the l-th event-time quintile.

The likelihood functions of ULR-Q and ULR-QL are straightforward modifications of the $L_{ULR}(\beta)$ aforementioned by including the appropriate covariates. The sandwich robust-variance estimator for cluster data, $\hat{var}_{ULR}(\beta)^{cluster}$, described above is then applied.

Stata codes for illustration of the ULR-Q and ULR-QL are included in Supplementary File 5, available as Supplementary data at *IJE* online.