ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)

A hybrid approach for the stratified mark-specific proportional hazards model with missing covariates and missing marks, with application to vaccine efficacy trials

Yanqing Sun,

University of North Carolina at Charlotte, USA

Li Qi,

Sanofi, Bridgewater, USA

Fei Heng

University of North Florida, Jacksonville, USA

and Peter B. Gilbert

University of Washington and Fred Hutchinson Cancer Research Center, Seattle, USA

[Received February 2019. Revised April 2020]

Summary. Deployment of the recently licensed tetravalent dengue vaccine based on a chimeric yellow fever virus, CYD-TDV, requires understanding of how the risk of dengue disease in vaccine recipients depends jointly on a host biomarker measured after vaccination (neutralization titre—neutralizing antibodies) and on a 'mark' feature of the dengue disease failure event (the amino acid sequence distance of the dengue virus to the dengue sequence represented in the vaccine). The CYD14 phase 3 trial of CYD-TDV measured neutralizing antibodies via case–cohort sampling and the mark in dengue disease failure events, with about a third missing marks. We addressed the question of interest by developing inferential procedures for the stratified mark-specific proportional hazards model with missing covariates and missing marks. Two hybrid approaches are investigated that leverage both augmented inverse probability weighting and nearest neighbourhood hot deck multiple imputation. The two approaches differ in how the imputed marks are pooled in estimation. Our investigation shows that nearest neighbourhood hot deck imputation can lead to biased estimation without properly selected neighbourhoods. Simulations show that the hybrid methods developed perform well with unbiased nearest neighbourhood hot deck imputations from proper neighbourhood selection. The new methods applied to CYD14 show that neutralizing antibody level is strongly inversely associated with the risk of dengue disease in vaccine recipients, more strongly against dengue viruses with shorter distances.

Keywords: Augmented inverse probability weighting; Competing risks failure time; Missing marks; Nearest neighbourhood hot deck multiple imputation; Semiparametric regression; Two-phase sampling

Address for correspondence: Yanqing Sun, Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, NC 28223, USA.
E-mail: yasun@uncc.edu

0035–9254/20/69000

1. Introduction

The CYD14 phase 3 trial randomized 2–14-year-old children within five countries of south-east Asia in 2:1 allocation to receive the dengue vaccine based on a chimeric yellow fever virus, CYD-TDV, or placebo in three injections at months 0, 6 and 12, where the CYD-TDV vaccine (Sanofi Pasteur) is a recombinant, live-attenuated, tetravalent vaccine containing one representative dengue strain from each of the four dengue serotypes (Capeding *et al.*, 2014). Participants underwent active surveillance for the primary study end point symptomatic virologically confirmed dengue (henceforth 'dengue disease') between month 13 and month 25 post first vaccination. Partly based on this trial that showed that the rate of dengue disease was an estimated 56% lower in the vaccine group than the placebo group ($p < 0.001$), this vaccine has been licensed in more than a dozen countries. The vaccine has been thought to work by inducing antidengue neutralizing antibodies (NAs).

We develop statistical methods to analyse the CYD14 efficacy trial data that are appropriate for interrogating how the association of antidengue NAs with dengue disease risk may differ depending on the amino acid sequence of the dengue virus causing the study end point, accounting for the fact that the expensive covariate of interest (NA titre) was measured through a classic case–cohort sampling design (measured from a Bernoulli simple random sample of 19.5% of participants at enrolment and from all disease cases) and that there is a substantial percentage (about a third) of missing dengue sequences among cases. This integrated assessment of how a host biomarker and a 'mark' feature of the failure event relate to failure risk has many applications, including general prospective studies that follow a cohort for acquisition of a genetically diverse infectious disease, encompassing many pathogens including human immunodeficiency virus (HIV) type 1, influenza and malaria.

To define the general statistical problem, let T be the time to a failure event of interest, and Z be a time-independent p-dimensional covariate. Under the competing risks model, a cause-of-failure mark V is observed when a failure event occurs. Let V be a continuous mark variable with bounded support $[0, 1]$. The mark-specific failure time data follow a competing risks model where the mark variable V plays the role of the cause of failure that is only observable on failure. In the motivating dengue vaccine study, the mark V measures the amino acid sequence distance of a dengue-disease-causing dengue sequence to the nearest dengue sequence inside the vaccine, which can only be observed in subjects experiencing the dengue disease end point and is not available or meaningfully defined in subjects without the end point.

The mark-specific hazard function, defined as

$$\lambda(t, v) = \lim_{\Delta_1 \to 0, \Delta_2 \to 0} P\{T \in [t, t + \Delta_1), V \in [v, v + \Delta_2) | T \geqslant t\} / (\Delta_1 \Delta_2),$$

was studied by Gilbert *et al.* (2004). It measures the instantaneous risk of failure by a mark in the presence of all marks, e.g. dengue sequences circulating in the efficacy trial region exposing trial participants through mosquito bites, and can be considered as an extension of the cause-specific hazard function to a continuous mark. Subsequent statistical methods have been developed to model the conditional mark-specific hazard function with applications to HIV vaccine efficacy studies; see Sun *et al.* (2009), Sun and Gilbert (2012), Juraska and Gilbert (2013, 2016) and Yang *et al.* (2017).

Suppose that the population of interest includes K subpopulations or strata, each with different baseline mark-specific hazard functions. Let $\lambda_k(t, v|z)$ be the conditional mark-specific hazard function at $(T, V) = (t, v)$ given covariate $Z = z$ for an individual in the kth stratum. The stratified mark-specific proportional hazards model postulates that

$$\lambda_k(t,v|z)=\lambda_{0k}(t,v)\exp\{\beta(v)^{\mathrm{T}}z\}, \qquad k=1,\ldots,K, \tag{1}$$

where $\lambda_{0k}(\cdot,v)=\lambda_k(t,v|z=0)$ is the unspecified baseline hazard function for the kth stratum, and $\beta(v)$ is the p-dimensional unknown regression coefficient function of v. Model (1) allows different baseline functions for different strata.

The mark-specific proportional hazards model (1) was first studied by Sun *et al.* (2009) under $K=1$ with the objective of evaluating mark-specific HIV vaccine efficacy, where the mark is an amino acid sequence distance of an infecting HIV strain to an HIV strain inside the vaccine. Model (1) was further studied by Sun and Gilbert (2012), Gilbert and Sun (2015) and Juraska and Gilbert (2016) for the situation where the marks are subject to missingness in subjects with observed failure times. Yang *et al.* (2017) investigated model (1) under two-phase sampling of components of Z allowing some participants to have missing covariates. However, the methods accounting for missing marks assumed complete measurements of all covariates, and the methods accounting for missing covariates assumed complete data on the marks of failures. Therefore, new methods are needed to account for both types of missing data.

In the motivating CYD14 efficacy trial, there are two types of missing data. The covariate NA titre is missing through a case–cohort sampling design and the mark V (dengue sequence distance) is missing for some cases. Multiple imputation has been widely used for handling missing data; see Rubin (1987). Two-phase sampling or case–cohort designs are common forms of studies with missing covariates, where covariates are divided into phase 1 or phase 2, with the former measured in all enrolled subjects and the latter measured only in a subset, typically because of expense of measurement. A 'case–cohort' design typically refers to randomly sampling subjects at enrolment into a subcohort for measuring the phase 2 covariates, which are also measured in all subjects outside the subcohort who experience the failure event and have the requisite samples available (White, 1982; Prentice, 1986; Breslow and Lumley, 2013). 'Two-phase sampling' typically refers to the generalization of outcome-dependent case–control sampling, where, within each cell of a 2 × K table defined by outcome status cross-classified with the K levels of a discrete phase 1 covariate, subjects are randomly sampled for measuring the phase 2 covariates (Breslow *et al.*, 2009). These designs can be implemented with Bernoulli or without replacement sampling, and our methods apply to any of the Bernoulli sampling versions. As is the usual case, application of the methods to the without-replacement sampling versions provides approximately correct results, with inferences tending to be slightly conservative. There is extensive literature on statistical methods for two-phase sampling or case–cohort designs, e.g. Prentice (1986), Robins *et al.* (1994), Borgan *et al.* (2000), Scheike and Martinussen (2004), Kulich and Lin (2004), Nan (2004), Breslow *et al.* (2009) and Breslow and Lumley (2013).

Nearest neighbourhood imputation is one of the hot deck imputation methods that are commonly used in survey sampling (Sedransk, 1985; Kovar *et al.*, 1988; Jonsson and Wohlin, 2004; Andridge and Little, 2010). The idea of nearest neighbourhood hot deck (NNHD) imputation is to replace each missing value with an observed response from a matching subject from the same data set. The hybrid approach proposed leverages both augmented inverse-probability-weighting (AIPW) complete-case estimation (Robins *et al.*, 1994) to handle the two-phase sampled covariates and NNHD imputation to fill in missing marks in failure cases (Chen and Shao, 2000; Beretta and Santaniello, 2016). AIPW estimation has a double-robust property, yielding consistent estimates if either the model for whether phase-2 covariates are missing or the model for the conditional expectations of phase 2 covariates is correctly specified (Robins *et al.*, 1994; Gao and Tsiatis, 2005). Most imputation methods assume a parametric model for the variable to be imputed. In contrast, as a non-parametric technique, NNHD imputation does not rely on model fitting for the variable to be imputed, and thus is potentially less sensitive to model misspecification than a parametric-model-based imputation method. However, our investigation shows that NNHD imputation can lead to biased estimation without proper neighbourhood selection.

We develop hybrid estimation and hypothesis testing procedures for model (1) that use both AIPW estimation and NNHD imputation. We investigate the neighbourhood selection for NNHD imputation for unbiased estimation. NNHD imputation is employed to impute the values of missing marks, followed by completed-marks two-phase sampling data analysis with an AIPW method similar to that of Yang *et al.* (2017) that did not account for missing marks. We investigate two hybrid estimation methods using the completed-marks two-phase sampling data that differ in the way in which the imputed marks are pooled in estimation. We develop hypothesis testing procedures to evaluate whether the mark-specific hazard ratios are unity and whether they change with the mark. The main contribution of this paper is the development of hybrid estimation and hypothesis testing methods for model (1) that relates the hazard of an outcome to both covariates and marks, accounting for missingness in both, including the investigation of neighbourhood selection for the NNHD imputation of marks to achieve valid inference on the association parameters. The procedures developed enable assessment of whether and how the hazard rate of an infectious disease with a pathogen genetically close to or far from a reference genetic sequence is modified by participant covariates. This application is exemplified by the dengue vaccine efficacy trial, with reference sequence the closest dengue strain in the vaccine construct and covariates age and immune response to the dengue vaccine strains.

In Section 2, we formulate the missing data problem, presenting notation and assumptions. The NNHD imputation technique is introduced in Section 2.1. The two hybrid estimation procedures are developed in Section 2.2. Techniques for estimation of the mark-specific cumulative incidence function rate (CIFR) are given in Section 2.3. Statistical procedures for hypothesis testing of the mark-specific hazard ratios are developed in Section 3. An extensive simulation study is conducted in Section 4 to examine the performances of the newly proposed methods, which are applied to the CYD14 data in Section 5. Some concluding remarks are given in Section 6. Additional discussions about the proposed hybrid methods along with more simulation results, analysis of the simulated data based on the CYD14 efficacy trial and additional analysis of the CYD14 efficacy trial are presented in the on-line supplementary materials.

2. Hybrid estimation using augmented inverse probability weighting and nearest neighbourhood hot deck multiple imputations

The AIPW estimation method was proposed by Robins *et al.* (1994) for missing data to improve robustness and efficiency over simple inverse probability weighted estimators. This important methodology has been widely used and has shown efficiency and the double-robust property in many studies; see Gao and Tsiatis (2005), Sun and Gilbert (2012), Yang *et al.* (2017) and Sun *et al.* (2018), among others. We investigate two hybrid methods of estimation of the mark-specific proportional hazards model that use both AIPW and NNHD imputation. We propose to employ the NNHD method to impute the values of missing marks, followed by completed-marks two-phase sampling data analysis with an AIPW method similar to that of Yang *et al.* (2017) that did not account for missing marks. The first approach follows the standard multiple-imputation scheme of Rubin (1987) whereas the second approach incorporates multiple imputations in estimating equations (MIEEs).

Suppose that the failure time T is subject to right censoring and is partially observed through observation of $X=\min\{T,C\}$ and $\delta=I(T\leqslant C^*)$, where $I(\cdot)$ is the indicator function and $C^*=\min(C,\tau)$ is the right censoring time with τ the end of follow-up and C the right censoring random variable. Let Z be a time-independent covariate vector. We assume independent censoring—that C is independent of (T, V) conditionally on Z. Suppose that $Z=(Z_1^{\mathrm{T}}, Z_2^{\mathrm{T}})^{\mathrm{T}}$ consists of two parts—Z_1 are observed in all subjects (phase 1) and Z_2 are measured in only a subset (phase 2 sample). In addition, the mark variable V is subject to missingness. Let $\xi=(\xi_z, \xi_v)$ be the vector of missing data indicators, where ξ_z is the indicator for whether a subject has complete covariate information, and ξ_v is the indicator for whether the mark variable V is observed. We set $\xi_v=1$ if $\delta=0$ since the mark V is inherently not available and is not considered as missing. We also set $\xi_v=1$ if $\delta=1$ and V is observed; otherwise $\xi_v=0$. Let $A=(A_z, A_v)$ be auxiliary variables, with A_z the auxiliary variable predictive of phase 2 covariates and A_v the auxiliary variable predictive of missing marks. For convenience, we denote $\Omega=(X, Z_1, A)$ and represent the observed data by $\tilde{\Omega}_{\mathrm{o}}=(\Omega,\ \xi_z Z_2,\ \xi_v \delta V, \delta)$.

We assume that Z_2 and V are missing at random (Rubin, 1976), satisfying missingness at random (MAR):

(a) $P(\xi_z=1|X, Z_1, Z_2, A, \delta V, \delta)=P(\xi_z=1|X, Z_1, A_z, \delta)$,
(b) $P(\xi_v=1|X, Z_1, Z_2, A, \delta V, \delta=1)=P(\xi_v=1|X, Z_1, Z_2, A_v, \delta=1)$ and
(c) $P(\xi_z=0, \xi_v=0|X, Z_1, Z_2, A, \delta V, \delta)=0$.

MAR (a) assumes that the missingness of Z_2 does not depend on the value of Z_2 and δV, MAR (b) assumes that the missingness of V does not depend on the value of V and MAR (c) implies that Z_2 and V do not have missing values on the same subjects, which is always satisfied under Prentice's (1986) original case–cohort sampling design for which no cases have missing Z_2-values. It is approximately satisfied for implemented case–cohort sampling designs (including our example) that intend to measure Z_2 in all cases but end up with a small number of happenstance missing values.

Suppose that there are K strata. Let n_k be the number of subjects in the kth stratum and $n=\Sigma_{k=1}^{K} n_k$. We label the ith subject in the kth stratum with a pair of subscripts $\{ki\}$. Let $Z_{1,ki}$ and $Z_{2,ki}$ be copies of covariates Z_1 and Z_2 for subject i in stratum k respectively. Similarly, $\xi_{z,ki}$ and $\xi_{v,ki}$ are copies of ξ_z and ξ_v respectively. Let $Z_{ki}=(Z_{1,ki}^{\mathrm{T}}, Z_{2,ki}^{\mathrm{T}})^{\mathrm{T}}$, $\xi_{ki}=(\xi_{z,ki}, \xi_{v,ki})$ and $\Omega_{ki}=(X_{ki}, Z_{1,ki}, A_{ki})$. The observed data are $\tilde{\Omega}_{\mathrm{o},ki}=(\Omega_{ki},\ \xi_{z,ki}Z_{2,ki},\ \xi_{v,ki}\delta_{ki}V_{ki}, \delta_{ki})$, for $i=1,\ldots,n_k$, $k=1,\ldots,K$. We assume that $\{T_{ki}, C_{ki}, V_{ki}, Z_{ki}, \xi_{ki}, A_{ki}; i=1,\ldots,n_k\}$ are independent identically distributed replicates of (T, C, V, Z, ξ, A) from stratum k, $k=1,\ldots,K$.

2.1. Nearest neighbourhood hot deck imputation of missing marks

In the competing risks setting, V_{ki} is observable if a failure is observed, i.e. $\delta_{ki}=1$. If the mark value V_{ki} is not available for $\delta_{ki}=1$, then we have a missing mark indicated by $\xi_{v,ki}=0$. The standard imputation approach involves first drawing the parameters of the posterior distribution of the missing variables given the observed data, and then drawing M sets of imputed values for the missing data from their posterior distribution given the observed data; see Rubin (1987). However, parametric multiple imputation can be sensitive to misspecification of the imputation model (Carroll *et al.*, 1984).

NNHD imputation, as a hot deck imputation method, replaces each missing value with an observed response from a matching subject from the same data set. Hot deck imputation methods have been studied by Little (1988), Reilly (1993), Chen and Shao (2000) and Beretta and Santaniello (2016), among others. Using the hot deck method, we impute a missing value V of a subject by choosing at random from observed V-values among matching donors. Donors are matched for their similarity in regard to some metric. This approach does not rely on model fitting for the variable to be imputed and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model.

We describe the NNHD imputation procedure as follows. Suppose that V_{ki} is missing, in which case $\xi_{v,ki}=0$. We impute missing values V_{ki} by using hot deck imputation from donors with similar $\mathcal{H}_{ki}=(T_{ki}, Z_{ki})$ or $\mathcal{H}_{ki}=(T_{ki}, Z_{ki}, A_{v,ki})$ in the case that a relevant $A_{v,ki}$ is available. Let $d(\mathcal{H}_{ki},\mathcal{H}_{kj})$ be a measure of similarity between $\mathcal{H}_{ki}$ and $\mathcal{H}_{kj}$. Each hot deck imputation of V_{ki} is obtained by randomly selecting a donor's mark from the L nearest neighbourhood $\mathcal{L}_{ki}$ matched on the basis of the similarity measure $d(\mathcal{H}_{ki},\mathcal{H}_{kj})$, where L is a number less than the number of non-missing marks for observed failures. Let $R_{ki,L}$ be the Lth order statistic of $d(\mathcal{H}_{ki},\mathcal{H}_{kj})$ for subjects with $\delta_{kj}=1, \xi_{v,kj}=1, j=1,\ldots,n_k$. An L nearest neighbourhood of V_{ki} is defined as $\mathcal{L}_{ki}=\{V_{kj}: d(\mathcal{H}_{ki},\mathcal{H}_{kj})\leqslant R_{ki,L}, \delta_{kj}=1, \xi_{v,kj}=1, j=1,\ldots,n_k\}$. The implementation of the nearest neighbourhood hot deck depends on the choice of metric and the variables that are included for the neighbourhood selection. If some components of Z_{ki} are discrete, then the L nearest neighbourhood imputations are carried out on the basis of the remaining variables $\mathcal{H}^{\mathrm{s}}_{ki}$ in $\mathcal{H}_{ki}$ stratified by the values of the discrete components of Z_{ki}. Further, the similarity measure $d(\mathcal{H}^{\mathrm{s}}_{ki},\mathcal{H}^{\mathrm{s}}_{kj})$ can be calculated on the basis of the z-scores or the ranks of variables, which eliminates the effects of scales or units of the variables on the nearest neighbour selections. Let $V^{(m)}_{ki}$, $m=1,\ldots,M$, be M random selections from $\mathcal{L}_{ki}$ with replacement. If V_{ki} is not missing, in which case $\xi_{v,ki}=1$, then we let $V^{(m)}_{ki}=V_{ki}$.

NNHD imputation is related to variable bandwidth L nearest neighbours kernel smoothing that is widely used in non-parametric density estimation and regression; see Stone (1977), Li (1984) and Altman (1992). Every case with missing V has the same number of marks imputed from the L nearest neighbours. The NNHD approach with a fixed number L of neighbours is similar to defining neighbourhoods by a metric with varying bandwidth such as $R_{ki,L}$, whereas an alternative approach with a fixed bandwidth B is similar to allowing variable L. A fixed B-bandwidth neighbourhood of V_{ki} is defined as $\mathcal{B}_{ki}=\{V_{kj}: d(\mathcal{H}_{ki},\mathcal{H}_{kj})\leqslant B, \delta_{kj}=1, \xi_{v,kj}=1, j=1,\ldots,n_k\}$. In this case, the number L of neighbours belonging to $\mathcal{B}_{ki}$ varies between subjects with missing marks. An advantage of using fixed L is that the bandwidth is allowed to be larger when data are sparse, which is a common non-parametric smoothing approach to guard against incorporating too few points that could occur by using a fixed bandwidth. Although we study NNHD with a fixed L, the method could also be implemented with a fixed bandwidth or variable L.

Choosing the set $\mathcal{H}_{ki}$ of variables for neighbourhood selection is very important. Our investigation shows that NNHD imputation can lead to biased estimation without proper selection of the neighbourhood. Let $W=(T, Z, A_v)$ and $\rho_k(v, W)=P(V\leqslant v|\delta=1, W)$ be the conditional distribution of V given W for cases. For an observed value $w=(t,z,a)$ of W of an individual in the kth stratum, $\rho_k(v,w)=P(V\leqslant v|\delta=1, W=w)$. Let $g_k(a|t,v,z)=P(A_{v,ki}=a|T_{ki}=t, V_{ki}=v, Z_{ki}=z, \delta_{ki}=1)$ be the probability density of a possible auxiliary variable for V. By Sun and Gilbert (2012),

$$\rho_k(v,w)=\int_0^v \lambda_k(t,u|z)g_k(a|t,u,z)\,\mathrm{d}u\Big/\int_0^1 \lambda_k(t,u|z)g_k(a|t,u,z)\,\mathrm{d}u. \tag{2}$$

If $A_{v,ki}$ is not available or independent of V_{ki} given $(T_{ki}, Z_{ki}, \delta_{ki})$, then

$$\rho_k(v,w)=\int_0^v \lambda_k(t,u|z)\,\mathrm{d}u\Big/\int_0^1 \lambda_k(t,u|z)\,\mathrm{d}u.$$

Equation (2) shows that the conditional distribution of V_{ki} depends on $(T_{ki}, Z_{ki}, A_{v,ki})$ in general. Unbiased imputation of V_{ki} should be selected from a neighbourhood defined on the basis of $\mathcal{H}_{ki}=(T_{ki}, Z_{ki}, A_{v,ki})$ except for certain special situations where $\beta(v)$ in model (1) does not change with v and $A_{v,ki}$ is conditionally independent of Z_{ki} given $(T_{ki}, V_{ki}, \delta_{ki})$. In this case, z cancels out from equation (2) under model (1). A simulation example in Section 4 shows that the NNHD imputation leads to biased estimation without including $Z_{2,ki}$ in $\mathcal{H}_{ki}$

2.2. Hybrid estimation procedures

We propose two hybrid approaches for estimation of model (1). The first approach follows the standard multiple-imputation scheme of Rubin (1987) such that the NNHD estimator is the average of the AIPW estimates of Yang *et al.* (2017) for two-phase sampling of covariates for completed marks under each imputation and the variance estimator is adjusted by using Rubin's formula. The second approach utilizes multiple imputations in a single AIPW estimating equation.

Let $Y_{ki}(t)=I(X_{ki}\geqslant t)$ be the at-risk process. The sampling probabilities of the phase 2 covariates are given by $\pi_{z,ki}(t)=P_k\{\xi_{z,ki}=1|\Omega_{ki},\delta_{ki},Y_{ki}(t)=1\}$. Suppose that $\hat{\pi}_{z,ki}(t)$ is an estimator of $\pi_{z,ki}(t)$ based on parametric models as discussed in Yang *et al.* (2017). Let $W_{ki}(t)=\xi_{z,ki}\{\pi_{z,ki}(t)\}^{-1}$ and $\hat{W}_{ki}(t)=\xi_{z,ki}\{\hat{\pi}_{z,ki}(t)\}^{-1}$. We define the marked counting processes for the completed marks by $N_{ki}^{(m)}(t,v)=I(X_{ki}\leqslant t,\delta_{ki}=1,V_{ki}^{(m)}\leqslant v)$ for $m=1,\ldots,M$. If V_{ki} is not missing, $V_{ki}^{(m)}=V_{ki}$ and $N_{ki}^{(m)}(t,v)=N_{ki}(t,v)\equiv I(X_{ki}\leqslant t,\delta_{ki}=1,V_{ki}\leqslant v)$.

2.2.1. *Hybrid estimation using standard multiple imputation*

Standard multiple-imputation estimation of $\beta(v)$ uses the average of estimates obtained for each imputation. Following Yang *et al.* (2017), for the mth imputation, $m=1,\ldots,M$, let $\hat{\beta}_{\mathrm{R}}^{(m)}(v)$ be the solution to the estimating equation for $\beta=\beta(v)$ for $v\in(0,1)$:

$$U^{(m)}(v,\beta)=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^1\int_0^{\tau}K_h(u-v)[\hat{W}_{ki}(t)\{Z_{ki}-\hat{Z}_k(t,\beta)\}\\+\{1-\hat{W}_{ki}(t)\}\{\hat{E}_k(Z_{ki}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)})-\hat{Z}_k(t,\beta)\}]N_{ki}^{(m)}(\mathrm{d}t,\,\mathrm{d}u),\tag{3}$$

where $K_h(x)=K(x/h)/h$, $K(\cdot)$ is a kernel function and h the bandwidth, and $\hat{E}_k(Z_{ki}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)})$ and $\hat{Z}_k(t,\beta)=\hat{S}_k^{(1)}(t,\beta)/\hat{S}_k^{(0)}(t,\beta)$ are the estimates that were described in Yang *et al.* (2017).

In particular, $\hat{E}_k(Z_{ki}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)})$ is the estimate of $E_k(Z_{ki}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)})$, and

$$\hat{S}_k^{(j)}(t,\beta)=n_k^{-1}\sum_{i=1}^{n_k}[\hat{W}_{ki}(t)Y_{ki}(t)\exp(\beta^{\mathrm{T}}Z_{ki})Z_{ki}^{\otimes j}\\+\{1-\hat{W}_{ki}(t)\}Y_{ki}(t)\hat{E}_k\{\exp(\beta^{\mathrm{T}}Z_{ki})Z_{ki}^{\otimes j}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}\}],\tag{4}$$

for $j=0,1,2$, where $\hat{E}_k\{\exp(\beta^{\mathrm{T}}Z_{ki})Z_{ki}^{\otimes j}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}\}$ is the estimate of the conditional expectation $E_k\{\exp(\beta^{\mathrm{T}}Z_{ki})Z_{ki}^{\otimes j}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}\}$ for $j=0,1,2$. Write $\beta=(\beta_1^{\mathrm{T}},\beta_2^{\mathrm{T}})^{\mathrm{T}}$, where β_1 and β_2 are the coefficients for $Z_{1,ki}(t)$ and $Z_{2,ki}$ respectively. Note that $Z_{1,ki}(t)$ is a part of Ω_{ki}. For given β, the first part of $E_k(Z_{ki}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)})$ is $Z_{1,ki}(t)$ and the second part is $E_k(Z_{2,ki}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)})$. Similarly, $E_k\{\exp(\beta^{\mathrm{T}}Z_{ki})Z_{ki}^{\otimes j}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}\}$, for $j=0,1,2$, depend on the observed data and are functions of the conditional expectations $E_k\{\exp(\beta_2^{\mathrm{T}}Z_{2,ki})Z_{2,ki}^{\otimes r}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}\}$, $r=0,1,2$. Yang *et al.* (2017) considered using parametric models for $E_k\{g(Z_{2,ki})|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}\}$ to obtain the estimate $\hat{E}_k\{g(Z_{2,ki})|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}\}$, where $g(Z_{2,ki})$ is a specified function of $Z_{2,ki}$ such as $Z_{2,ki}$, $\exp(\beta_2Z_{2,ki})$ or $Z_{2,ki}\exp(\beta_2Z_{2,ki})$.

By the standard multiple-imputation scheme of Rubin (1987), the hybrid Rubin estimator is defined by $\hat\beta_{\mathrm{R}}(v)=M^{-1}\Sigma_{m=1}^{M}\hat\beta_{\mathrm{R}}^{(m)}(v)$. The variance estimate of $\hat\beta_{\mathrm{R}}(v)$ adjusting for multiple imputation by using Rubin's (Rubin (1987), page 76) rule equals

$$\widehat{\mathrm{var}}\{\hat\beta_{\mathrm{R}}(v)\}=M^{-1}\sum_{m=1}^{M}\widehat{\mathrm{var}}\{\hat\beta_{\mathrm{R}}^{(m)}(v)\}+(1+M^{-1})(M-1)^{-1}\sum_{m=1}^{M}\{\hat\beta_{\mathrm{R}}^{(m)}(v)-\hat\beta_{\mathrm{R}}(v)\}^2, \quad (5)$$

where $\widehat{\mathrm{var}}\{\hat\beta_{\mathrm{R}}^{(m)}(v)\}$ is the variance estimator of Yang *et al.* (2017) based on the mth imputation. The first part accounts for within-imputation variability, and the second part $(M-1)^{-1}\Sigma_{m=1}^{M}\{\hat\beta_{\mathrm{R}}^{(m)}(v)-\hat\beta_{\mathrm{R}}(v)\}^2$ for between-imputation variability. The term $1+M^{-1}$ corrects for bias due to the finite number of multiply imputed data sets.

2.2.2. *Hybrid estimation via the estimating equations approach*

This subsection proposes another hybrid approach that incorporates multiple imputations into a single estimating equation. A subject with a missing mark receives M imputed marks, which are associated with a particular subject and are dependent. The M imputed marks can be considered as a cluster. We consider the following hybrid estimating equation for $\beta=\beta(v)$ for $v\in(0,1)$:

$$U(v,\beta)=M^{-1}\sum_{m=1}^{M}U^{(m)}(v,\beta)=0, \quad (6)$$

where $U^{(m)}(v,\beta)$ is defined in equation (3). The estimating equation (6) for the hybrid MIEE resembles the generalized estimation equation approach for repeated measures analysis that assumes working independence (Liang and Zeger, 1986). A subject with observed mark, in which case $V_{ki}^{(m)}=V_{ki}$, receives weight 1, whereas the weight for a subject with missing mark is $1/M$ for each imputation. The estimator $\hat\beta(v)$ that solves $U(v,\beta)=0$ is termed the hybrid MIEE estimator.

The estimator $\hat\beta(v)$ can be implemented by using the Newton–Raphson iterative algorithm. Starting with an initial value $\beta^{(0)}(v)$, let $\beta^{(l)}(v)$ be the estimate of $\beta(v)$ at step l. The estimator $\hat\beta(v)$ is obtained by iterating steps (a) and (b) as follows until convergence:

(a) estimate the conditional expectations $E_k[\exp\{(\beta_2^{(l)}(v))^{\mathrm{T}}Z_{2,ki}\}Z_{2,ki}^{\otimes j}|\Omega_{ki},\delta_{ki}V_{ki}]$ for $j=0,1,2$, and calculate $\hat Z_k\{t,\beta^{(l)}(v)\}$;

(b) update the estimate $\beta^{(l+1)}(v)$ at step $l+1$ by $\beta^{(l+1)}(v)=\beta^{(l)}(v)-[\partial U\{v,\beta^{(l)}(v)\}/\partial\beta]^{-1}\times U\{v,\beta^{(l)}(v)\}$.

Estimation of the stratified mark-specific proportional hazards model (1) also involves estimation of the baseline mark-specific hazard function $\lambda_{0k}(t,v)$. The MIEE approach treats the multiple imputations for a given subject as a cluster. As such, the Nelson–Aalen-type estimator

$$\hat\Lambda_{0k}(t,v)=M^{-1}\sum_{m=1}^{M}\sum_{i=1}^{n_k}\int_0^t\int_0^v[n_k\hat S_k^{(0)}\{s,\hat\beta(u)\}]^{-1}N_{ki}^{(m)}(\mathrm{d}s,\mathrm{d}u)$$

is a natural estimator of the doubly cumulative baseline function $\Lambda_{0k}(t,v)=\int_0^t\int_0^v\lambda_{0k}(s,u)\,\mathrm{d}s\,\mathrm{d}u$. The baseline function $\lambda_{0k}(t,v)$ can be estimated by $\hat\lambda_{0k}(t,v)$ obtained by smoothing the increments of the estimator $\hat\Lambda_{0k}(t,v)$. For example, one can use kernel smoothing

$$\hat\lambda_{0k}(t,v)=\int_0^\tau\int_0^1 K_{h_1}^{(1)}(t-s)K_{h_2}^{(2)}(v-u)\,\hat\Lambda_{0k}(\mathrm{d}s,\mathrm{d}u),$$

where $K_{h_1}^{(1)}(x)=K^{(1)}(x/h_1)/h_1$ and $K_{h_2}^{(2)}(x)=K^{(2)}(x/h_2)/h_2$, with $K^{(1)}(\cdot)$ and $K^{(2)}(\cdot)$ kernel functions and h_1 and h_2 bandwidths. Other model parameters of interest that are discussed in Section 2.3 such as the overall conditional survival function, the conditional cumulative incidence function (CIF) and the mark-specific CIFR can be estimated under the same framework.

Next, we propose an estimator of the variance of $\hat{\beta}(v)$. Let $\hat{J}_k(t,\beta)=\hat{S}_k^{(2)}(t,\beta)/\hat{S}_k^{(0)}(t,\beta)-\{\hat{Z}_k(t,\beta)\}^{\otimes 2}$. The derivative of $U(v,\beta)$ with respect to β equals

$$U'(v,\beta)=-M^{-1}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^1\int_0^\tau K_h(u-v)\hat{J}_k(t,\beta)\,N_{ki}^{(m)}(\mathrm{d}t,\,\mathrm{d}u).$$

Following the proof of theorem 2 of Yang *et al.* (2017), we have the approximation

$$(nh)^{1/2}\{\hat{\beta}(v)-\beta(v)\}\approx\hat{\Sigma}(v)^{-1}n^{-1/2}h^{1/2}M^{-1}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{n_k}Q_{ki}^{(m)}(v), \tag{7}$$

where $\hat{\Sigma}(v)=-n^{-1}U'\{v,\hat{\beta}(v)\}$, and, similarly to Yang *et al.* (2017), $Q_{ki}^{(m)}(v)$ is approximated by

$$\begin{aligned}\hat{Q}_{ki}^{(m)}(v)=&\int_0^1\int_0^\tau K_h(u-v)\hat{W}_{ki}(t)[Z_{ki}-\hat{Z}_k\{t,\hat{\beta}(u)\}]\hat{M}_{ki}^{(m)}(\mathrm{d}t,\,\mathrm{d}u)\\&+\int_0^1\int_0^\tau K_h(u-v)\{1-\hat{W}_{ki}(t)\}\hat{M}_{ki,\bar{z}}^{\mathrm{o}\,(m)}(\mathrm{d}t,\,\mathrm{d}u).\end{aligned} \tag{8}$$

Here $\hat{M}_{ki}^{(m)}(\mathrm{d}t,\,\mathrm{d}u)=N_{ki}^{(m)}(\mathrm{d}t,\,\mathrm{d}u)-Y_{ki}(t)\exp\{\hat{\beta}(u)^{\mathrm{T}}Z_{ki}\}\hat{\Lambda}_{0k}(\mathrm{d}t,\,\mathrm{d}u)$ and $\hat{M}_{ki,\bar{z}}^{\mathrm{o}\,(m)}(\mathrm{d}t,\,\mathrm{d}u)$ is the estimator of $M_{ki,\bar{z}}^{\mathrm{o}\,(m)}(\mathrm{d}t,\,\mathrm{d}u)$ given by

$$\begin{aligned}&[E(Z_{ki}|\Omega_{ki})-\bar{z}_k\{t,\beta_0(u)\}]N_{ki}^{(m)}(\mathrm{d}t,\mathrm{d}u)-(E[Z_{ki}\exp\{\beta_0^{\mathrm{T}}(v)Z_{ki}\}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}]\\&\qquad-\bar{z}_k\{t,\beta_0(u)\}E[\exp\{\beta_0^{\mathrm{T}}(v)Z_{ki}\}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}])Y_{ki}(t)\lambda_{0k}(t,u)\,\mathrm{d}t\,\mathrm{d}u.\end{aligned}$$

Hence, $\hat{M}_{ki,\bar{z}}^{\mathrm{o}\,(m)}(\mathrm{d}t,\,\mathrm{d}u)$ is obtained by replacing $\bar{z}_k\{t,\beta_0(u)\}$ with $\hat{Z}_k\{t,\hat{\beta}(u)\}$ and $\lambda_{0k}(t,u)\,\mathrm{d}t\,\mathrm{d}u$ with $\hat{\Lambda}_{0k}(\mathrm{d}t,\,\mathrm{d}u)$, and by replacing $E(Z_{ki}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)})$, $E[\exp\{\beta_0^{\mathrm{T}}(v)Z_{ki}\}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}]$ and $E[Z_{ki}\exp\{\beta_0^{\mathrm{T}}(v)Z_{ki}\}|\Omega_{ki},\delta_{ki}V_{ki}^{(m)}]$ with their estimates.

Using Rubin's idea to account for the between-imputation variability, we propose to estimate the variance of $\hat{\beta}(v)$ by $\hat{\Sigma}_{\hat{\beta}}(v)=\hat{\Sigma}(v)^{-1}\hat{\Sigma}_{\mathrm{R}}^{*}(v)\hat{\Sigma}(v)^{-1}/(nh)$, where

$$\hat{\Sigma}_{\mathrm{R}}^{*}(v)=\frac{h}{n}\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\{\hat{Q}_{ki}^{(m)}(v)\}^{\otimes 2}+\frac{M+1}{M}\frac{1}{M-1}\sum_{m=1}^{M}\{U^{(m)}(v,\hat{\beta})-U(v,\hat{\beta})\}^{\otimes 2}\right]. \tag{9}$$

In the web appendix A, we present heuristic arguments to show that the proposed hybrid MIEE and hybrid Rubin estimators are unbiased for large samples by using NNHD imputation and under the model assumptions that were given by Yang *et al.* (2017). The hybrid MIEE and hybrid Rubin estimators also enjoy the double-robustness properties similarly to the AIPW estimators of Yang *et al.* (2017).

Parametric multiple imputation often uses between two and 10 imputations (Rubin (1987), page 15). Reilly (1993) recommended that hot deck estimation be performed with three, five and 10 imputations.

2.3. Estimation of the mark-specific cumulative incidence function rate

By definition of the conditional mark-specific hazard function, $\lambda_k(t,v|z)\,\mathrm{d}v$ measures the instantaneous rate of failure at time t with failure type or mark (e.g. dengue sequence distance) $V\in[v,v+\mathrm{d}v)$ in the presence of all other possible failure types (dengue viruses with different sequence distances) for a very small $\mathrm{d}v$. In this section, we introduce the mark-specific CIFR that provides interpretable results (e.g. through visual display of estimates) and is useful for prediction. The conditional CIF for stratum k is defined by $\mathcal{F}_k(t,v|z)=P(T_k \leqslant t, V_k \leqslant v|Z_k=z)$, which has the interpretation of the classical CIF as the conditional probability of failure by time t with failure cause $V_k \leqslant v$. The conditional mark-specific CIFR $f_{k,v}(t,v|z)$ is the derivative of $\mathcal{F}_k(t,v|z)$ with respect to v. The quantity $f_{k,v}(t,v|z)\,\mathrm{d}v$ is the conditional probability that failure with mark $V\in[v,v+\mathrm{d}v)$ occurs by time t. The cumulative incidence of failure with mark V in an interval $(v_1,v_2]\subset(0,1)$ is given by $P(T_k\leqslant t, v_1<V_k\leqslant v_2|Z_k=z)=\int_{v_1}^{v_2} f_{k,v}(t,v|z)\,\mathrm{d}v$ for $(v_1,v_2]\subset(0,1)$.

Whereas $\lambda_k(t,v|z)$ is useful for measuring the instantaneous rate of failure occurrence at time t for those at risk, the mark-specific CIFR $f_{k,v}(t,v|z)$ is useful to estimate or predict the probability of failure by time t with $V\in[v,v+\mathrm{d}v)$. As with the classical competing risks model, the mark-specific hazard function is related to the CIFR through the simple formula

$$f_{k,v}(t,v|z)=\exp\{\beta(v)^{\mathrm{T}}z\}\int_0^t S_k(s|z)\,A_{0k}(\mathrm{d}s,v),$$

where $A_{0k}(t,v)=\int_0^t \lambda_{0k}(s,v)\,\mathrm{d}s$ and $S_k(t|z)$ is the conditional overall survival function of T_k given $Z_k=z$ that is given by

$$S_k(t|z)=\exp\left\{-\int_0^1 A_{0k}(t,v)\exp\{\beta(v)^{\mathrm{T}}z\}\,\mathrm{d}v\right\}$$

under model (1). The CIF and CIFR can be estimated by plugging in the estimates of $\beta(v)$ and $\lambda_{0k}(t,v)$. The details of estimation are given in the web appendix B. The relationships between the estimated conditional mark-specific hazard function, CIF and CIFR are the same as for their population quantities by using the hybrid MIEE approach with multiple imputations, but this is not so for the hybrid Rubin approach.

3. Statistical inferences for $\beta(v)$

We develop procedures for testing two sets of hypotheses regarding $\beta(v)$. Let $\beta_r(v)$ be the rth component of $\beta(v)$, $1\leqslant r\leqslant p$. We first test the null hypothesis $H_{10}:\beta_r(v)=0$ for $v\in[a,b]\subset(0,1)$ against the general alternative $H_{1\mathrm{a}}:\beta_r(v)\neq 0$ for at least some $v\in[a,b]$, and against the monotone alternative $H_{1\mathrm{m}}:\beta_r(v)\leqslant 0$ with $\beta_r(v)<0$ for some $v\in[a,b]$. The testing procedure can be used to test $\beta_r(v)\geqslant 0$ with simple modifications. The second hypothesis H_{20} concerns whether $\beta_r(v)$ does not depend on v for $v\in[a,b]$. We test H_{20} against the general alternative $H_{2\mathrm{a}}$ that $\beta_r(v)$ depends on v for $v\in[a,b]$ and the monotone alternative $H_{2\mathrm{m}}$ that $\beta_r(v)$ is a monotone increasing function. The test can be modified to test the monotone alternative that $\beta_r(v)$ is a monotone decreasing function. The tests of H_{10} are helpful for identifying covariates that are correlated with risk for at least some failure types or marks. The tests of H_{20} evaluate whether the strength of association of a covariate with risk varies with values of the failure type or mark V.

We construct the following test procedures based on the hybrid MIEE estimator of $\beta(v)$. Let $0<v_1<\ldots<v_G<1$ be a grid of G points in the range of the marks $(0,1)$. By Aalen and Johansen (1978) and Sun *et al.* (2009), it can be shown that $\hat{\beta}(v_1),\ldots,\hat{\beta}(v_G)$ are asymptotically independent and approximately normal. The estimated variance of $\hat{\beta}_r(v)$, $\widehat{\mathrm{var}}\{\hat{\beta}_r(v)\}$, is the rth element on the diagonal of $\hat{\Sigma}_{\hat{\beta}}(v)$. Let $\hat{\beta}_r=(\hat{\beta}_r(v_1),\hat{\beta}_r(v_2),\ldots,\hat{\beta}_r(v_G))^{\mathrm{T}}$. We propose the following test statistic to test $H_{10}:\beta_r(v)=0$ against $H_{1\mathrm{a}}:\beta_r(v)\neq 0$:

$$T_{1\mathrm{a}}=\sum_{g=1}^{G}\hat{\beta}_r(v_g)^2/\widehat{\mathrm{var}}\{\beta(v_g)\}.$$

The following test statistic is used to test H_{10} against $H_{1\mathrm{m}}: \beta_r(v) \leqslant 0$:

$$T_{1\mathrm{m}} = \sum_{g=1}^{G} \hat{\beta}_r(v_g) / \widehat{\mathrm{var}}\{\beta(v_g)\}^{1/2}.$$

Under the null hypothesis H_{10}, $T_{1\mathrm{a}}$ has an approximately χ^2-distribution χ^2_G with G degrees of freedom, and $T_{1\mathrm{m}}$ has an approximate normal distribution with mean 0 and variance G. A larger value of $T_{1\mathrm{a}}$ indicates departures from H_{10}, rejecting H_{10} in favour of $H_{1\mathrm{a}}$ at significance level α if $T_{1\mathrm{a}}$ is greater than the $(1-\alpha)$-percentile of χ^2_G. A smaller value of $T_{1\mathrm{m}}$ shows evidence in favour of $H_{1\mathrm{m}}$, rejecting H_{10} at significance level α if $T_{1\mathrm{m}}$ is less than the α-percentile of $N(0, G)$.

To test the null hypothesis H_{20} that the rth component $\beta_r(v)$ does not depend on v, we let $Q_{\hat{\beta}} = (\hat{\beta}(v_2) - \hat{\beta}(v_1), \ldots, \hat{\beta}(v_G) - \hat{\beta}(v_{G-1}))^{\mathrm{T}}$. Then $Q_{\hat{\beta}} = A\hat{\beta}_r$, where A is the $(G-1) \times G$ matrix with -1 as the (i, i)th element, 1 as the $(i, i+1)$th element for $i = 1, \ldots, G-1$ and the rest of the elements 0. Thus the covariance matrix of $Q_{\hat{\beta}}$ is $\mathrm{cov}(Q_{\hat{\beta}}) = A\,\mathrm{cov}(\hat{\beta}_r)\,A^{\mathrm{T}}$, where $\mathrm{cov}(\hat{\beta}_r)$ is the diagonal matrix with $\mathrm{var}\{\hat{\beta}(v_g)\}, g = 1, \ldots, G$, on the diagonals. The following expressions are the two test statistics for testing H_{20}:

$$T_{2\mathrm{a}} = Q^{\mathrm{T}}_{\hat{\beta}} \widehat{\mathrm{cov}}(Q_{\hat{\beta}})^{-1} Q_{\hat{\beta}},$$

and

$$T_{2\mathrm{m}} = J^{\mathrm{T}} \widehat{\mathrm{cov}}(Q_{\hat{\beta}})^{-1/2} Q_{\hat{\beta}},$$

where $\widehat{\mathrm{cov}}(Q_{\hat{\beta}})$ is the diagonal matrix with $\widehat{\mathrm{var}}\{\hat{\beta}(v_g)\}$, $g = 1, \ldots, G$, on the diagonals, and J is a $(G-1)$-dimensional vector of 1s. Under H_{20}, $T_{2\mathrm{a}}$ has an approximately χ^2-distribution χ^2_{G-1} with $G-1$ degrees of freedom, and $T_{2\mathrm{m}}$ has an approximate normal distribution with mean 0 and variance $G-1$. We reject H_{20} in favour of $H_{2\mathrm{a}}$ at level of significance α if $T_{2\mathrm{a}}$ is greater than the $(1-\alpha)$-percentile of χ^2_{G-1} and we reject H_{20} in favour of $H_{2\mathrm{m}}$ if $T_{2\mathrm{m}}$ is greater than the $(1-\alpha)$-percentile of $N(0, G-1)$.

In practice, we recommend that G takes a value from 3 to 5 with approximately evenly spaced grid points with spacing greater than the size of the bandwidth for better approximations of the null distributions of the test statistics.

4. Simulation study

We conducted a simulation study to evaluate the finite sample performance of the estimation and hypothesis testing procedures proposed. Let U_1, U_2 and U_3 be independent uniformly distributed random variables on (0, 1). Let $Z_1 = U_1 + 2U_3$ be a phase 1 covariate, and $Z_2 = -U_1 + 2U_2$ a phase 2 covariate, with resulting correlation coefficient -0.2. We study the scenario of one stratum $K = 1$. The (T_{1i}, V_{1i}) are generated from the following mark-specific proportional hazards model:

$$\lambda(t, v|Z) = \lambda_0(t, v) \exp\{\beta_1(v) Z_1 + \beta_2(v) Z_2\}, \qquad t \geqslant 0, \quad 0 \leqslant v \leqslant 1, \tag{10}$$

where the mark-specific baseline function is $\lambda_0(t, v) = \exp(-0.3v)$, $\beta_1(v) = -0.2v$ and $\beta_2(v) = \alpha + \theta v$. We study the performance of the hypothesis tests of H_{10} and H_{20} for $\beta_2(v)$. The parameters α and θ are chosen to examine the sizes and powers of the tests proposed. All failure times that are greater than $\tau = 2.0$ are right censored at τ. Censoring times are generated from an exponential distribution, independent of (T, V), with parameter adjusted so that the overall censoring rate during follow-up is approximately 40%.

For two-phase sampling, we consider a simple Bernoulli random sample taken separately for cases and controls, with selection probability $\pi_{z,1i} = 1$ for cases $(\delta_{1i} = 1)$ and 0.5 for controls

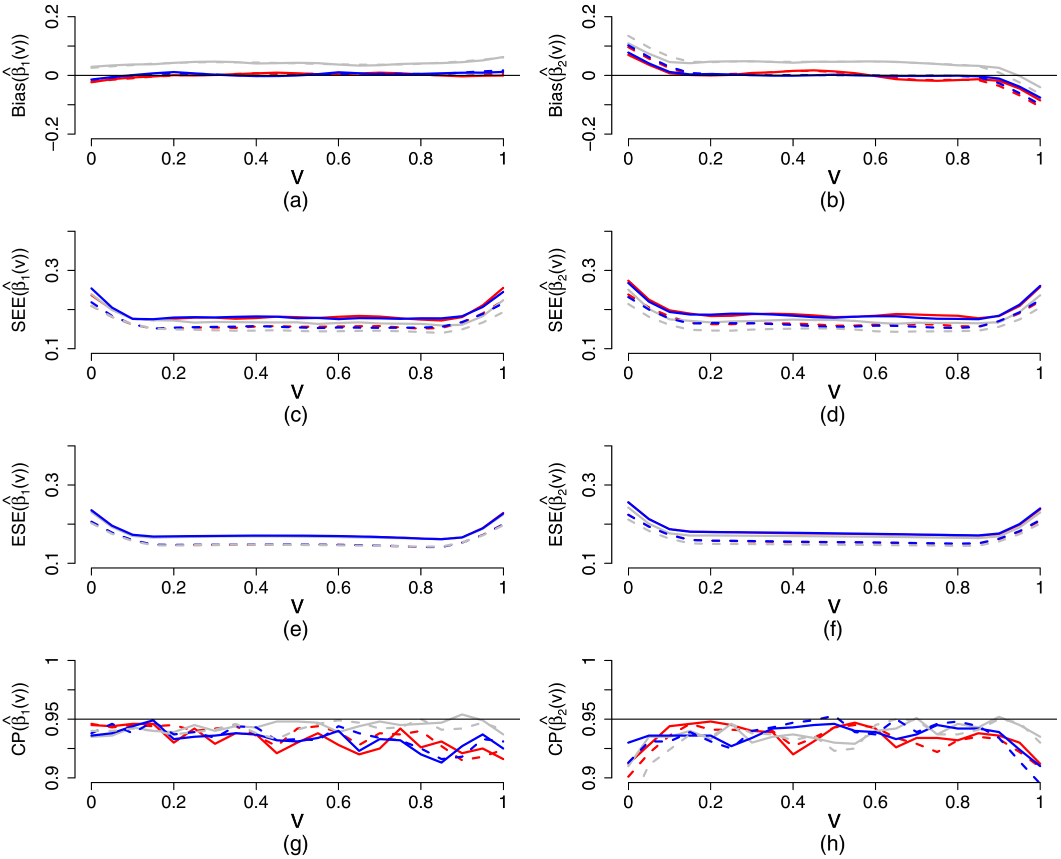

Fig. 1. Bias, SEE, ESE and CP for $\hat{\beta}_1(v)$ and $\hat{\beta}_2(v)$ for $n = 800$ under the setting P4 for model (10) with $M = 5$ imputations from the five nearest neighbourhoods $\mathcal{L}_{1i}$ of cases i with missing marks based on 1000 simulations (the $\mathcal{L}_{1i}$ are calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j} = (T_{1j}, Z_{1j})$ for cases (with $\delta_{1j} = 1$): ———, MIEE($h = 0.13$); – – – –, MIEE($h = 0.17$); ———, Rubin($h = 0.13$); – – – –, Rubin($h = 0.17$); ———, CC($h = 0.13$); – – – –, CC($h = 0.17$)

($\delta_{1i} = 0$). Suppose that there is an auxiliary variable A_z correlating with Z_2, $A_z = Z_2 + \epsilon$, where ϵ is normally distributed with mean 0 and standard deviation 0.5, which corresponds to a Pearson correlation coefficient between Z_2 and A_z of $\rho = 0.75$. The conditional expectations involving the phase 2 covariate Z_2 are estimated by using linear models with $(1, \delta, Z_1, A_z, \delta Z_1, \delta A_Z)$ as predictors based on the subjects with observed Z_2. Covariates $(1, Z_1, A_z)$ are used for estimating the logit linear model for $\pi_{z,1i}$ for subjects with $\delta_{1i} = 0$.

The mark V_{1i} is missing following the conditional probability $\text{logit}(\pi_{v,1i}) = \text{logit}\{P(\xi_{v,1i} = 1|\Omega_{1i})\} = 0.3Z_{1,1i} + 0.8$ for $\delta_{1i} = 1$, yielding about 22% missing marks. The hot deck imputation of a missing V_{ki} is obtained from donors from the same stratum k with $\delta_{ki} = 1$ and with similar $\mathcal{H}_{ki}$ defined by Euclidean distance and z-scores of (T_{ki}, Z_{ki}); for our simulations we study only one stratum $k = 1$. The L nearest neighbourhood imputations are carried out on the basis of $\mathcal{H}_{1i}$ with the Euclidean metric. By considering the z-scores of variables, we eliminate the effects of scales or units of the variables on the nearest neighbour selections. We consider $M = 3$ and $M = 5$ imputations from the M nearest neighbourhoods for the cases with missing marks.

The performances of the test procedures proposed are evaluated through simulations under model (10) for the parameter settings P1–P5 that are defined as follows: $P1, (\alpha, \theta) = (0, 0)$;

$P2, (\alpha, \theta) = (-0.4, 0)$; $P3, (\alpha, \theta) = (-0.5, 0)$; $P4, (\alpha, \theta) = (-1, 1.5)$; $P5, (\alpha, \theta) = (-1, 2)$. P1 and P3 are models under the null hypothesis H_{10} and H_{20} respectively; P2 and P3 are $H_{1\mathrm{m}}$ alternatives to H_{10}, and P4 and P5 are $H_{2\mathrm{m}}$ alternatives to H_{20}.

The Epanechnikov kernel $K(x) = 0.75(1 - x^2)I(|x| \leqslant 1)$ is used for kernel smoothing. The bandwidth is selected by using the formula $h = C\hat{\sigma}_v n^{-1/3}$, where $\hat{\sigma}_v$ is the estimated standard error of the observed marks for uncensored failure times and C is a constant ranging from 2 to 5. Sun *et al.* (2016) and Yang *et al.* (2017) showed that this formula works well in simulations. A larger C can be used if the distribution of the observed marks is skewed or marks are sparse in some areas. Alternatively, the formula $h = C\hat{\sigma}_v n_\mathrm{o}^{-1/3}$ has also been used in situations with a very large phase 1 sample and low event rate (Yang *et al.*, 2017), where n_o is the observed number of events. The values of $\hat{\sigma}_v$ under model (10) for settings P1–P5 are approximately 0.29, yielding $h = 4\hat{\sigma}_v n^{-1/3} = 0.15$ for $n = 500$ and $h = 0.13$ for $n = 800$. We also studied the effect of using larger bandwidths: $h = 0.20$ for $n = 500$ and $h = 0.17$ for $n = 800$.

We estimate $\beta(v)$ over 21 evenly spaced grid points in $[0, 1]$ with spacing 0.05 such that $v_1 = 0, v_2 = 0.05, \ldots, v_{21} = 1$. The initial value for estimating $\beta(v_1)$ is set to 0. The estimate $\hat{\beta}(v_1)$ is used as the initial value for estimating $\beta(v_2)$ such that $\hat{\beta}(v_{i-1})$ is used as the initial value for estimating $\beta(v_i)$ for $i = 2, \ldots, N$. The Newton–Raphson iterative algorithm that was proposed in Section 2.2.2 is not overly sensitive to the choice of initial values.

Fig. 1 shows the simulation results for estimating $\beta(v) = (\beta_1(v), \beta_2(v))^\mathrm{T}$ under setting P4 for model (10) without auxiliary A_z with $M = 5$ imputations from five nearest neighbourhoods of cases with missing marks based on 1000 simulations by using bandwidths $h = 0.13$ and $h = 0.15$ for five nearest neighbourhoods $\mathcal{L}_{1i}$ calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j} = (T_{1j}, Z_{1j})$ for cases (with $\delta_{1j} = 1$), where Bias is the bias, SEE is the sample standard error of the estimator, ESE is the sample mean of the estimated standard errors and CP is the 95% empirical coverage probability. Fig. 2 compares the estimates by using different neighbourhood selections under setting P4 for model (10).

Additional simulation studies are presented in the web appendix C, which includes simulation results under setting P3 of model (10), and for a different mark-specific proportional hazards model with $K = 2$ strata. Web appendix C also includes a real data simulation study that applies the proposed methods to a data set generated on the basis of the CYD14 trial data.

The simulation study shows that the biases of both the hybrid MIEE estimator and the hybrid Rubin estimator are very small except in the left-hand and right-hand tails for $\beta_2(v)$ (which are the expected boundary effects in non-parametric estimation) by using the five nearest neighbourhoods $\mathcal{L}_{1i}$ of cases with missing marks calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{kj} = (T_{kj}, Z_{kj})$ for cases. The pointwise coverage probabilities are slightly below but very close to 95% for $v \in (0, 1)$ except in the left-hand and right-hand tails for $\beta_2(v)$, indicating adequate performance of the variance estimators proposed. For larger bandwidths, SEE and ESE of the estimator are smaller. The study also shows that estimation based on the L nearest neighbourhoods calculated by using only a subset of $\mathcal{H}_{kj} = (T_{kj}, Z_{kj})$ can yield much larger biases unless $\beta(v)$ does not depend on v. In particular, under setting P4 for model (10), Fig. 2(b) shows that using $\mathcal{H}_{kj} = (T_{kj})$ yields much larger biases than using $\mathcal{H}_{kj} = (T_{kj}, Z_{kj})$. We also note from Fig. 3 in the web appendix C that the biases are small for both selections of $\mathcal{H}_{kj}$ under setting P3 since $\beta(v)$ does not vary with v in this setting.

Using the five nearest neighbourhoods $\mathcal{L}_{1i}$ calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j} = (T_{1j}, Z_{1j})$ for all cases (with $\delta_{1j} = 1$), Fig. 3 shows the simulation results for estimating the conditional mark-specific CIFR $f_{1,v}(t, v|z)$ at $Z_1 = 1.5$ and at the 10th, 50th and 90th percentiles of Z_2 for $t = 1$ and $n = 800$ under setting P4 for model (10) with $M = 5$ based

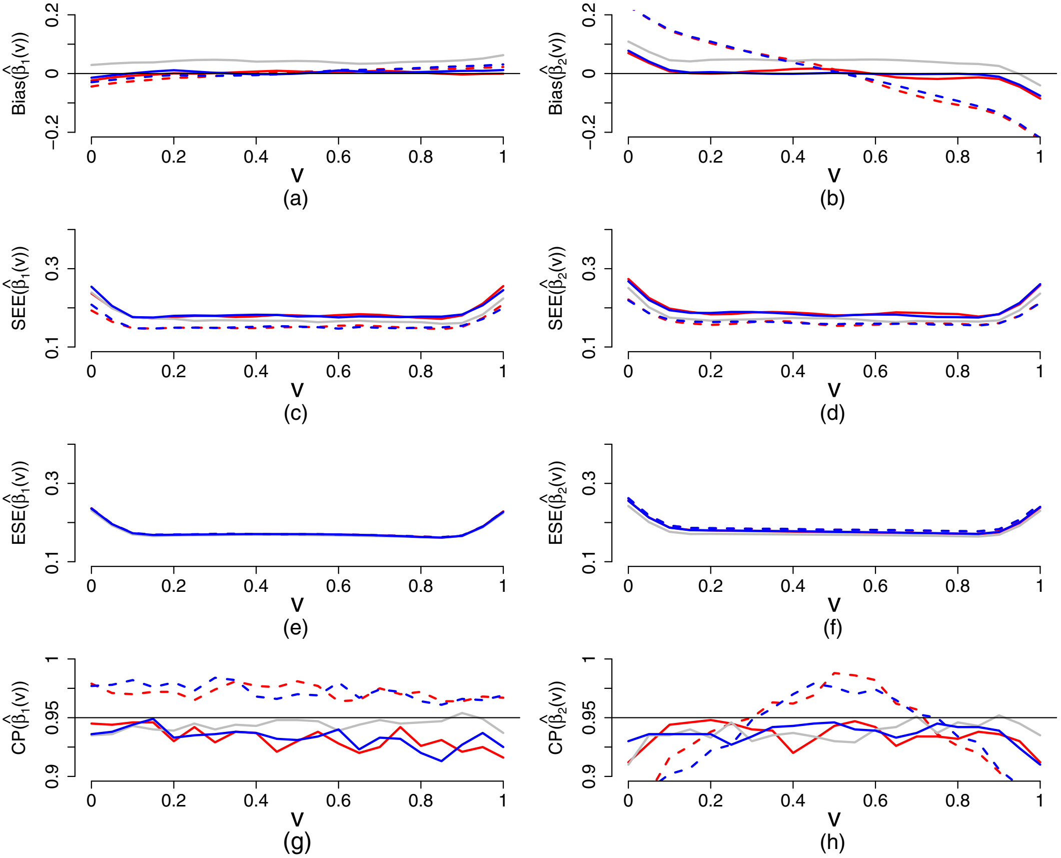

Fig. 2. Bias, SEE, ESE and CP for $\hat{\beta}_1(v)$ and $\hat{\beta}_2(v)$ for $n = 800$ under the setting P4 for model (10) with $M = 5$ imputations from the four nearest neighbourhoods of cases with missing marks based on 1000 simulations (MIEE-NN(T, Z) is for the hybrid MIEE estimator with the five nearest neighbourhoods $\mathcal{L}_{1i}$ calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j} = (T_{1j}, Z_{1j})$ for cases (with $\delta_{1j} = 1$), whereas MIEE-NN(T) is the same except z-scores of the $\mathcal{H}_{1j} = (T_{1j})$ are used; Rubin-NN(T, Z) and Rubin-NN(T) are defined similarly for the hybrid Rubin estimator): ———, MIEE-NN(T, Z); - - - -, MIEE-NN(T); ———, Rubin-NN(T, Z); - - - -, Rubin-NN(T); ———, CC

on 1000 simulations using bandwidth $h = 0.13$. Fig. 3 shows that the average of the estimated $f_{1,v}(t, v|z)$ are close to the true values $f_{1,v}(t, v|z)$.

The simulations are carried out to examine the performances of the proposed tests with the nearest neighbourhoods $\mathcal{L}_{1i}$ calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j} = (T_{1i}, Z_{1i})$ for subjects with $\delta_{1j} = 1$. Table 1 presents the empirical sizes and powers of tests $T_{1\mathrm{a}}$ and $T_{1\mathrm{m}}$ for testing H_{10} and tests $T_{2\mathrm{a}}$ and $T_{2\mathrm{m}}$ for testing H_{20} at nominal level 0.05 by using $M = 3$ and $M = 5$ imputations from the M nearest neighbourhoods based on 1000 simulations. The test statistics are calculated by using bandwidth $h = 0.15$ for $n = 500$ and $h = 0.13$ for $n = 800$ and using $G = 3$ grid points with $v_1 = 0.2$, $v_2 = 0.5$ and $v_3 = 0.8$. The empirical sizes for testing H_{10} under setting P1 and for testing H_{20} under P3 are slightly higher but very close to the nominal level 0.05, indicating adequate performance of the tests proposed. The powers of the tests for testing H_{10} increase as the model moves from P1 to P3, whereas the powers of the tests for testing H_{20} increase as the model moves from P3 to P5, representing increasing departures from the null hypotheses H_{10} and H_{20}. Powers of the tests with auxiliary variable A_z are slightly

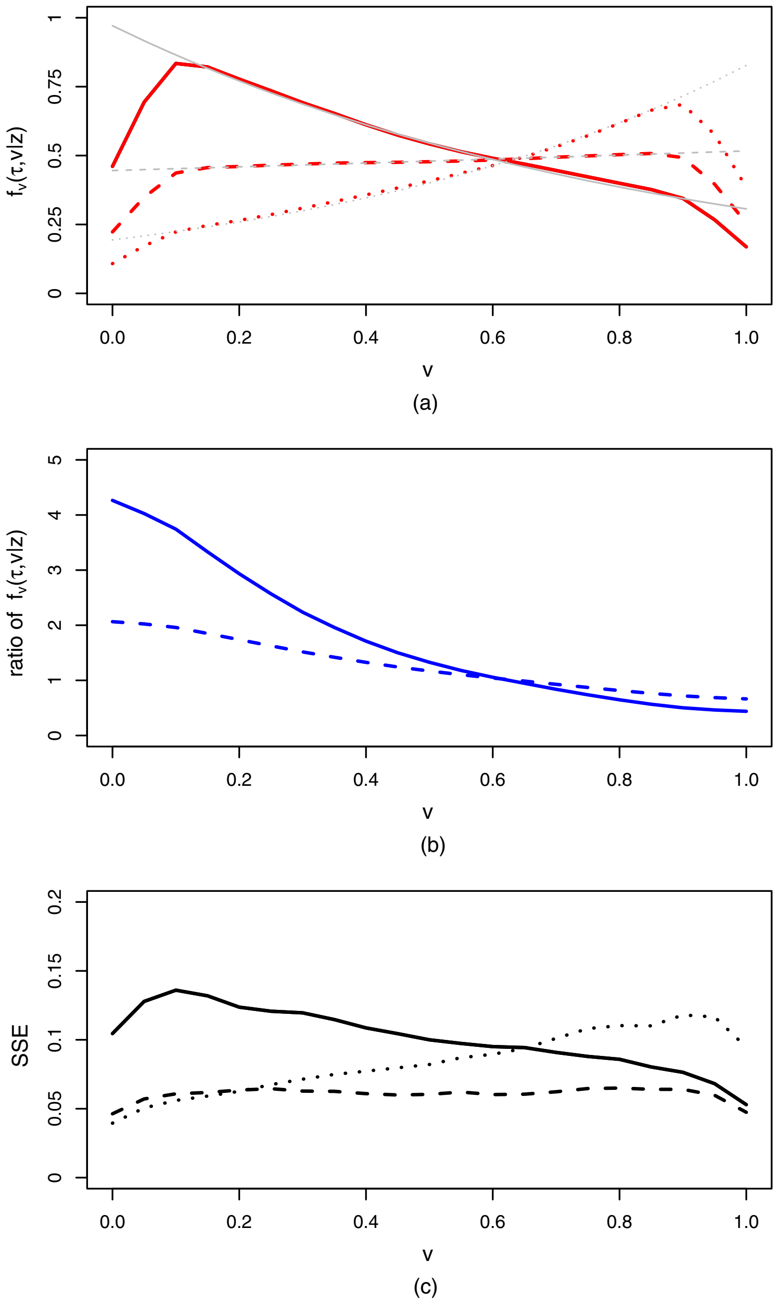

Fig. 3. Estimation of the conditional mark-specific CIFR $f_{1,v}(\tau, v|z)$ at $Z_1 = 1.5$ and at the 10th, 50th and 90th percentiles of Z_2 for $\tau = 1$ and $n = 800$ under the setting P4 for model (10) with $M = 5$ imputations from the five nearest neighbours of the missing marks using bandwidth $h = 0.13$ based on 1000 simulations (——, true values); the five nearest neighbourhoods are calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j} = (T_{1j}, Z_{1j})$ for cases (with $\delta_{1j} = 1$): (a) averages of the estimated $f_{1,v}(\tau, v|z)$ (——, MIEE; ——, true; ——, 10th percentile Z_2; – – – –, 50th percentile Z_2; · · · · , 90th percentile Z_2); (b) estimated ratios of the $f_{1,v}(\tau, v|z)$ at the 10th (– – – –) and 50th (– – – –) percentiles of Z_2 divided by $f_{1,v}(\tau, v|z)$ at the 90th percentiles of Z_2 respectively, for $Z_1 = 1.5$ and $\tau = 1$; (c) SSEs of the estimated $f_{1,v}(\tau, v|z)$ (——, 10th percentile Z_2; – – – –, 50th percentile Z_2; · · · · , 90th percentile Z_2)

Table 1. Empirical sizes and powers of the test statistics T_{1a} and T_{1m} for testing H_{10} and the test statistics T_{2a} and T_{2m} for testing H_{20} under model (10) with $M=3$ and $M=5$ imputations from the M nearest neighbourhoods of the missing marks at nominal level 0.05 based on 1000 simulations†

Model	*n*	*M*	*Without* A_z				*With* A_z			
			BAND1		*BAND2*		*BAND1*		*BAND2*	
			T_{1a}	T_{1m}	T_{1a}	T_{1m}	T_{1a}	T_{1m}	T_{1a}	T_{1m}
Testing H_{10}										
P1	500	3	0.064	0.062	0.067	0.065	0.070	0.051	0.063	0.054
		5	0.067	0.066	0.071	0.072	0.079	0.061	0.081	0.061
	800	3	0.069	0.060	0.062	0.058	0.072	0.051	0.070	0.045
		5	0.073	0.060	0.060	0.058	0.066	0.045	0.063	0.043
P2	500	3	0.771	0.934	0.876	0.968	0.814	0.959	0.915	0.983
		5	0.746	0.921	0.843	0.960	0.800	0.941	0.891	0.974
	800	3	0.897	0.974	0.959	0.996	0.933	0.991	0.980	0.999
		5	0.886	0.982	0.946	0.990	0.911	0.991	0.964	0.997
P3	500	3	0.912	0.982	0.966	0.995	0.945	0.992	0.984	0.999
		5	0.907	0.977	0.955	0.989	0.943	0.990	0.976	0.997
	800	3	0.974	0.998	0.992	1.000	0.986	0.998	0.999	1.000
		5	0.983	0.999	0.998	1.000	0.994	0.999	1.000	1.000
Testing H_{20}										
P3	500	3	0.070	0.049	0.055	0.047	0.079	0.055	0.062	0.053
		5	0.056	0.063	0.050	0.059	0.070	0.069	0.062	0.069
	800	3	0.066	0.060	0.069	0.054	0.075	0.062	0.083	0.058
		5	0.066	0.061	0.056	0.059	0.073	0.066	0.069	0.063
P4	500	3	0.757	0.883	0.839	0.942	0.779	0.891	0.867	0.953
		5	0.755	0.909	0.868	0.964	0.772	0.918	0.896	0.970
	800	3	0.894	0.973	0.959	0.990	0.904	0.980	0.967	0.993
		5	0.871	0.970	0.955	0.991	0.891	0.975	0.966	0.992
P5	500	3	0.953	0.990	0.989	0.999	0.996	0.998	0.993	0.999
		5	0.946	0.988	0.987	0.998	0.999	1.000	0.993	0.999
	800	3	0.991	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		5	0.995	0.999	0.999	1.000	1.000	1.000	0.999	1.000

†The test statistics are constructed by using $G=3$, $v_1=0.2, v_2=0.5$ and $v_3=0.8$. 'Without A_z' refers to the scenario where there is no auxiliary A_z, and 'with A_z' refers to the scenario where the auxiliary A_z is used. BAND1 is the bandwidth setting of $h=0.15$ for $n=500$ and $h=0.13$ for $n=800$ whereas BAND2 is the bandwidth setting of $h=0.20$ for $n=500$ and $h=0.17$ for $n=800$.

higher than those without using A_z. The powers of the tests are not overly sensitive to the number of imputations M but seem to increase slightly for larger bandwidth.

5. Dengue vaccine efficacy trial analysis

The CYD14 cohort for data analysis is all participants attending the month 13 study visit without previously experiencing the dengue disease primary end point, comprising 6639 vaccine recipients and 3220 placebo recipients. Of these, 116 vaccine recipients and 129 placebo recipients experienced the dengue end point by month 25, constituting an estimated 56.5% vaccine reduction in the hazard of dengue disease between month 13 and 25 (Capeding *et al.*, 2014). The percentage of right censoring by month 25 was 98.3%. An important scientific question is how does natural and vaccine immunity work in preventing dengue disease? NAs are generally believed to be important for both natural and vaccine-induced protection, which are present in many placebo recipients (caused by prior dengue exposures and infections), and are boosted or increased in many vaccine recipients (caused by dengue vaccination) (Moodie *et al.*, 2018). In this section, we apply the developed methods to analyse the CYD14 data with the objective of understanding, for each of the placebo and vaccine groups, the association of month 13 NA levels (‘NA titre’) with subsequent occurrence of the dengue disease primary end point through month 25, and whether and how the associations depend on dengue amino acid sequence. The NA titre marker is the average of an individual’s log-base-10 50% neutralization titre to each of the four dengue strains in the vaccine (one strain for each dengue serotype), where the 50% neutralization titre quantifies the ability of antibodies in an individual’s blood sample to kill a given dengue strain (defined in detail in Moodie *et al.* (2018)). The analyses by treatment group can be interpreted as assessing NA titre as a marker of different kinds of acquired protection or disease resistance—for placebo recipients’ naturally acquired resistance and for vaccine recipients a combination of naturally and vaccine-acquired resistance. This integrated analysis of host and pathogen data types would increase knowledge of NA titre as a correlate of risk of dengue disease, with many applications including aiding refinement of models for bridging vaccine efficacy to new settings that were not studied in CYD14.

As summarized in Section 1, NA titre was measured from month 13 blood samples from a subset of participants who were selected through a case–cohort sampling design. With controls defined as participants reaching the month 25 visit never experiencing the dengue disease study end point, the NA titre marker was measured from $n = 1879$ controls (1275 vaccine; 604 placebo), and from all $n = 245$ cases (116 vaccine; 129 placebo).

From blood samples drawn at dengue disease failure event times, dengue virus nucleotide sequences of the complete antigen coding region of the dengue genome represented in the CYD-TDV vaccine (prM/E) were measured by using 454-sequencing (Rabaa *et al.*, 2017). The prM/E dengue genome (1985 base pairs for serotypes 1, 2 and 4 and 1979 base pairs for serotype 3) was sequenced and translated to 661 amino acid positions (659 for serotype 3). The amino acid sequences were multiply aligned with the four vaccine strain sequences. A subset of 65 of the prM/E amino acid positions have been documented to be ‘NA contact sites’, defined as positions on the outer surface of dengue that have been documented to interact with antidengue NAs. Because sequence variation in these contact sites was hypothesized to be especially relevant for potential protection against dengue disease, we studied the mark V defined as the Hamming distance based on these NA contact sites. The distance V ‘Hamming distances: NA contact sites’ was calculated, which is the percentage amino acid mismatch in the 65 NA contact sites between the dengue sequence from a given disease case and the closest dengue sequence among the four vaccine strain sequences. The mark V was measured from 76 (66%) of the 116 vaccine recipient cases and from 84 (65%) of the 129 placebo recipient cases.

Vaccine recipients who were exposed to dengue sequences with short distances to the vaccine may be more likely to be protected by antibodies than vaccine recipients who were exposed to dengue sequences with large distances. Therefore, if NA titre is important for protection, its inverse correlation with dengue disease risk would be expected to be strongest against dengue viruses with small distances and to be weakest or non-existent against dengue sequences with large distances. The results on these hypotheses may provide insights into how the vaccine partially worked and thereby guide next steps of vaccine research. In addition, the same analysis in placebo recipients aids understanding of how naturally acquired NA titres associate with sequence-specific dengue risk.

Let T be the time between the month 13 visit until diagnosis of dengue disease to month 25. We consider the following mark-specific proportional hazards model:

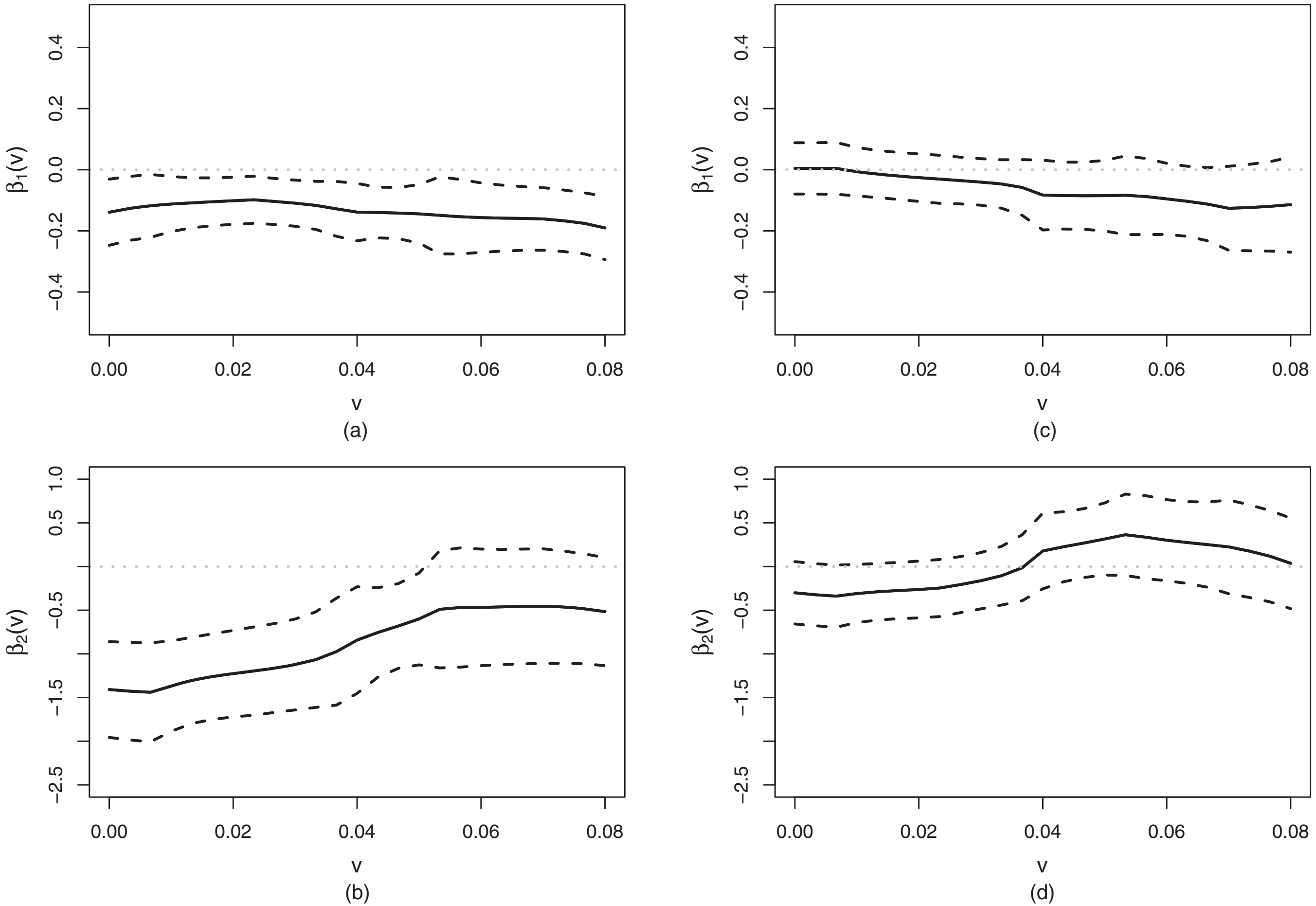

Fig. 4. Estimation of the associations of age and NA titre with the mark-specific hazard of the dengue disease end point with mark 'Hamming distances: NA contact sites' using $M=5$ imputations from the five nearest neighbourhoods of missing marks, with bandwidth $h=0.023$ for the vaccine group and $h=0.024$ for the placebo group (the five nearest neighbourhoods $\mathcal{L}_{1i}$ are calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j}=(T_{1j}, \mathrm{Age}_{1j}, \mathrm{NAb}_{1j})$ values of cases (with $\delta_{1j}=1$)): estimated log-mark-specific hazard ratios $\beta_1(v)$ for Age and $\beta_2(v)$ for NA titre are given for (a), (b) the vaccine group and (c), (d) the placebo group

Table 2. Results of hypothesis tests of H_{10} and H_{20} for the CYD14 trial

	NA titre				*Age*			
	Testing H_{10}		*Testing* H_{20}		*Testing* H_{10}		*Testing* H_{20}	
	T_{1a}	T_{1m}	T_{2a}	T_{2m}	T_{1a}	T_{1m}	T_{2a}	T_{2m}
Vaccine	<0.001	<0.001	0.069	0.005	<0.001	<0.001	0.857	0.762
Placebo	0.116	0.393	0.070	0.012	0.170	0.016	0.383	0.961

$$\lambda_k\{t, v|z(t)\} = \lambda_{k0}(t, v) \exp\{\beta_1(v)\text{Age} + \beta_2(v)\text{NAb}\}, \tag{11}$$

with $K=1$ baseline stratum, where NAb is the month 13 NA titre and Age is age at enrolment. Age is a phase 1 variable whereas NAb is a phase 2 variable. For both the vaccine and the placebo groups, NA titre is observed for all the cases but missing for 80.5% of the non-cases.

We implement the proposed estimation and testing procedures that were described in Sections 2 and 3. We estimated the probability of observing the NA titre marker with a logistic regression model, with $\text{logit}\{P(\xi_z = 1|\Omega)\}$ a linear function of (1, Age, Sex). To implement the AIPW method, we use linear models for $E(\text{NAb}|\Omega)$ and $E[\exp\{\beta_2(v)\text{NAb}^{\otimes j}\}|\Omega]$ for $j=0,1,2$, with predictors (1, Age, Sex). For each case i with missing mark V_i, we use $M=5$ imputed marks from the five nearest neighbourhoods $\mathcal{L}_{1i}$ calculated by using z-scores of $\mathcal{H}_{1j} = (T_{1j}, \text{Age}_{1j}, \text{NAb}_{1j})$ from all cases j (with $\delta_{1j}=1$).

Because of the very large phase 1 sample and the low event rate, we used the bandwidth $h = 5\hat{\sigma}_v n_o^{-1/3}$, where $\hat{\sigma}_v$ is the estimated standard error of the observed marks and n_o is the number of cases. The standard deviation of the observed mark 'Hamming distances: NA contact sites' is 0.0225 for the vaccine group and 0.0243 for the placebo group, resulting in bandwidth $h=0.023$ and $h=0.024$ respectively.

Fig. 4 shows point and 95% confidence interval estimates of $\beta_1(v)$ for Age and $\beta_2(v)$ for NA titre, by treatment arm. Greater age is associated with a lower risk of dengue disease for the vaccine group and apparently not for the placebo group, and the associations do not appear to depend on the mark. NA titre is strongly inversely associated with risk of dengue disease in the vaccine group, with stronger association for dengue viruses that are closest to the vaccine strains. In the placebo group the results suggest a weak inverse association of NA titre with dengue disease, only for dengue viruses that are close to the vaccine strains.

Augmenting results from Fig. 4, Fig. 5 shows the estimated conditional mark-specific CIFR $f_v(\tau, v|z)$ at month $\tau=25$ for the 10th, 50th and 90th percentiles of the NA titre marker and at the average age 8.35 years old, by treatment arm. Fig. 5 shows that $\hat{f}_{1,v}(\tau=25, v|z)$ is highest at the 10th percentile of NA titre and lowest at the 90th percentile. Fig. 5 also shows the ratios of $\hat{f}_v(\tau=25, v|z)$ for the 10th *versus* 90th percentiles and 50th *versus* 90th percentiles of NA titre at the average age. For the vaccine group, this ratio for the 10th *versus* 90th percentile is almost twice that of the ratio for the 50th *versus* 90th percentile for mark values $v<0.021$. Such differences in estimated $f_v(\tau=25, v|z)$ are not observed for the placebo group.

Table 2 presents the results of the hypothesis testing for $\beta_1(v)$ and $\beta_2(v)$ under H_{10} and H_{20} for the vaccine group and placebo group. The p-values are calculated by using $G=4$ grid points with

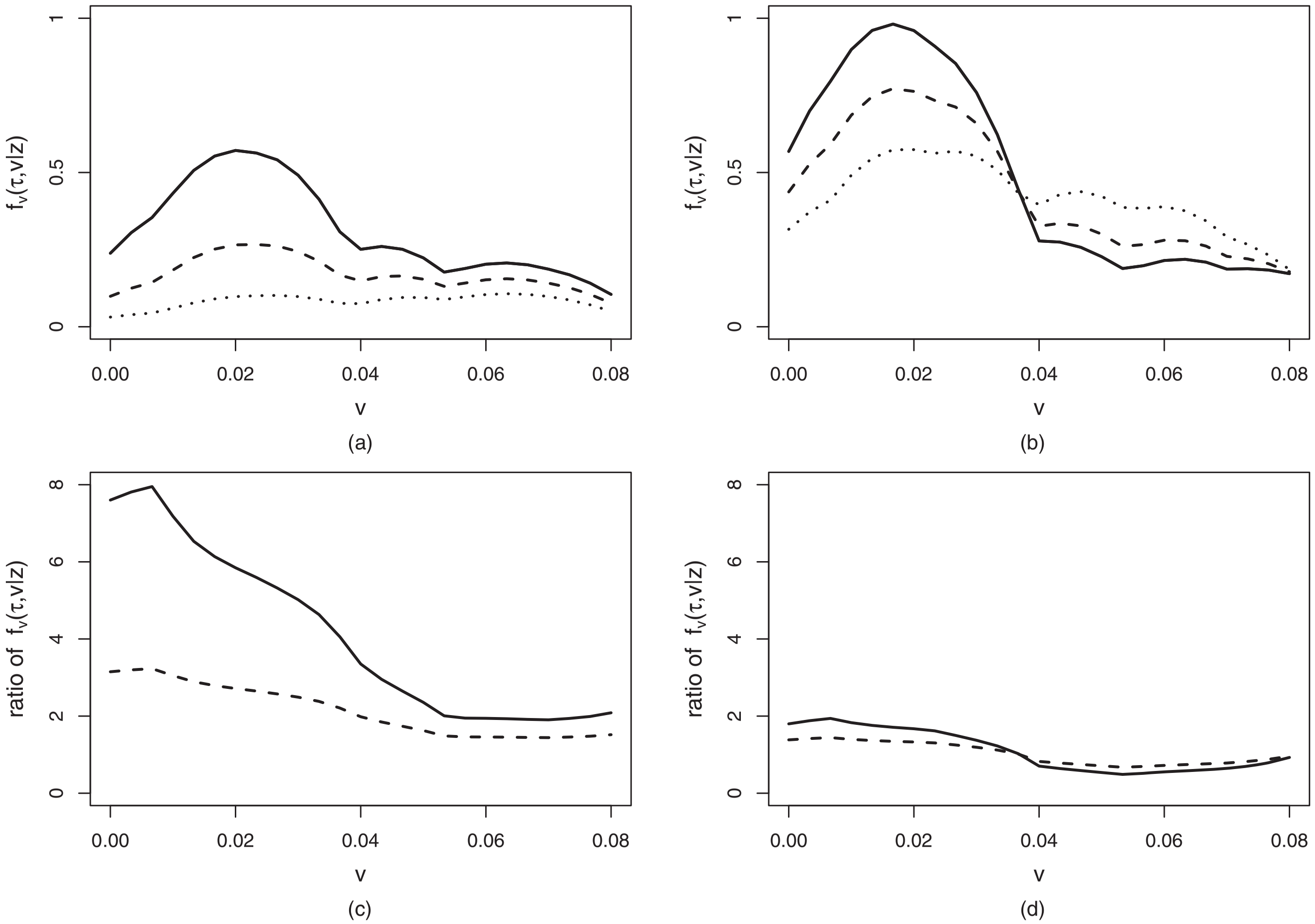

Fig. 5. Estimation of the conditional mark-specific CIFR $f_V(\tau, v|z)$ at month $\tau = 25$ and the ratios with the mark calculated as 'Hamming distances: NA contact sites' using $M = 5$ imputations from the five nearest neighbourhoods of missing marks, with bandwidth $h = 0.023$ for the vaccine group and $h = 0.024$ for the placebo group (the five nearest neighbourhoods $\mathcal{L}_{1i}$ are calculated by using Euclidean distance and z-scores of the $\mathcal{H}_{1j} = (T_{1j}, \mathrm{Age}_{1j}, \mathrm{NAb}_{1j})$ values of cases (with $\delta_{1j} = 1$): (a), (b) estimated $f_V(\tau = 25, v|z)$ at the 10th (———), 50th (- - - -) and 90th (. . . .) percentiles of NA titre at the average age 8.35 years old by treatment group; (c), (d) ratios at the 10th *versus* 90th (———) and the 50th *versus* 90th (- - - -) percentiles

$v_1 = 0.01, v_2 = 0.03, v_3 = 0.05$ and $v_4 = 0.07$. The results support that the risk of dengue disease decreases as the NA titre increases for the vaccine group, but not for the placebo group. Older children are at lower risk of dengue disease for both treatment groups but more significantly for the vaccine group. There are statistically significant results that the magnitude of the mark-specific association parameter $\beta_2(v)$ for NA titre decreases with increasing mark values for both treatment groups.

The analysis that is presented in this paper imputes missing dengue sequence distances from subjects with similar event times, ages and NA titre in a neighbourhood. In the web appendix C, we present the results of the data analysis using the alternative hot deck imputations that were implemented by Juraska *et al.* (2018), which were obtained by using information on study site and local clinic, as well as on dengue genotype and serotype. Similar results are obtained but with slightly weaker evidence that the magnitude of $\beta_2(v)$ decreases with increasing mark values in the vaccine group.

Because the CYD-TDV vaccine is licensed for children 9 years of age or older, we repeated the analyses restricting to 9–14-year-olds (web appendix D). For the vaccine group, the results for inference on $\beta_2(v)$ with covariate NA titre are similar to those for all ages 6–14 years (web appendix D Table 3, Fig. 8 and Fig. 9). However, for the placebo group, the analysis restricting to 9–14-year-olds supports an inverse correlation of NA titre with dengue disease for low dengue mark values, whereas the analysis of 6–14-year-olds did not suggest a correlation for any mark values.

6. Concluding remarks

Motivated by the CYD14 dengue vaccine efficacy trial, this paper has developed estimation and hypothesis testing procedures for $\beta(v)$ in model (1) under two-phase sampling of some covariates and with missing marks for some individuals with the failure event. We investigated two hybrid approaches that utilize non-parametric NNHD multiple imputations to impute missing marks of observed failures, followed by application of the AIPW technique to the completed-marks case–cohort sampled data sets. The two hybrid methods differ in how the imputed marks are pooled. Our simulations show that the hybrid Rubin and the hybrid MIEE estimators have similar performances in estimation.

We consider hot deck imputations of missing marks from donors with similar characteristics $\mathcal{H}_{kj} = (T_{kj}, Z_{kj}, A_{v,kj})$ among the observed failures. The implementation of the NNHD depends on the choice of metric and the variables that are included for the neighbourhood selection. The imputation based on a subset of (T_{kj}, Z_{kj}) can lead to biased estimation. Our L nearest neighbourhoods imputations are carried out on the basis of the z-scores of the $\mathcal{H}_{kj}$ for cases and with the Euclidean metric. By considering the z-scores of variables, we eliminate the effects of scales or units of the variables on the nearest neighbour selections. Hsu and Yu (2019) recently studied a Cox model with missing covariates by using the non-parametric multiple-imputation approach with the neighbourhood selected on the basis of the predictive scores of two working regression models. We conducted a limited simulation study and found no advantages of the predictive score approach for neighbourhood selection.

Achieving consistent variance estimation in the presence of imputed data remains a challenge. Rubin's (1987) rule of adjusting for multiple imputation has been widely used in practice. Other methods for estimating variances have been investigated, but few are rigorously justified; see, for example, Kovar and Chen (1994), Lee *et al.* (1994, 1995), Rancourt *et al.* (1994) and Montaquila and Jernigan (1997). Chen and Shao (2000) investigated the theoretical properties of the NNHD imputation method and showed that the NNHD method provides asymptotically unbiased estimators for population means, quantiles and univariate distributions. They also derived consistent variance estimators of the NNHD estimators. The proposed hybrid Rubin and hybrid MIEE estimators for the mark-specific proportional hazards model (1) work very well with small biases in the many different models that we examined. However, finding consistent variance estimators is very challenging for the NNHD imputation of missing mark under two-phase sampling of covariates. We adopted Rubin's rule for the variance estimators, which seems to underestimate the true variances slightly under some situations. The underestimated variances also lead to slightly inflated observed sizes for the tests proposed. Further investigation of variance estimation is needed.

For the analysis of the CYD14 efficacy trial, model (11) assumes that the mark-specific log-hazard ratio for Age is the same for every unit increase in Age and similarly for the mark-specific log-hazard ratio for NAs. However, the model assumptions may fail and thus model checking is an important problem. Sun *et al.* (2016) proposed a goodness-of-fit test procedure for the stratified mark-specific proportional hazards model (1) when covariates are observed and there are no missing marks. Developing the goodness-of-fit test procedure for model (1) with missing data is a project meriting future research.

The paper presents the analysis of model (11) for children of all ages. However, the mark-specific effects $\beta(v)$ may be different for different age groups. In the web appendix D of the supplementary material, we conducted separate analyses for children in two different age groups: 2–8- and 9–14-year-olds. The additional analyses provide some insights on whether the effects of Age and NA titre on the mark-specific risk of the dengue disease are different for different age groups.

The web appendix E of the supplementary material also includes the analyses using the hot deck imputations that were implemented in Juraska *et al.* (2018) that defined the neighbourhood on the basis of biological and geographic information, i.e. dengue genotype, serotype, study site and local clinic, and Juraska *et al.* (2018) validated that these hot deck imputations were highly accurate. These hot deck imputations are scientifically based and more robust to model misspecifications, whereas the other hot deck imputations approach that we studied in this paper exploits the link between the failure time data and observed marks specified by the mark-specific proportional hazards model, which can improve power but at the expense of being less robust to model misspecifications. Further research is warranted to investigate the neighbourhood selections and their effects.

7. Supplementary materials

The web appendices A, B, C, D and E that are referenced in this paper are given in the supplementary material that is available from the journal's website. The MATLAB code and instructions for doing the analysis for a simulated data set that is presented in section 3.3 of the web-based supplementary material is available also from

https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets.

Acknowledgements

The authors thank the participants and investigators of the CYD14 trial. This research was partially supported by National Institute of Allergy and Infectious Diseases National Institutes of Health award R37AI054165, and by a contract from the CYD14 study sponsor Sanofi Pasteur. Dr Sun's research was partially supported by National Science Foundation grants DMS1513072 and DMS1915829. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Aalen, O. O. and Johansen, S. (1978) An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scand. J. Statist.*, **5**, 141–150.

Altman, N. S. (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Statistn*, **46**, 175–185.

Andridge, R. R. and Little, R. J. A. (2010) A review of hot deck imputation for survey non-response. *Int. Statist. Rev.*, **78**, 40–64.

Beretta, L. and Santaniello, A. (2016) Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med. Informat. Decsn Makng*, **16**, 197–208.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000) Exposure stratified case-cohort designs. *Liftim. Data Anal.*, **6**, 39–58.

Breslow, N. E. and Lumley, T. (2013) *Semiparametric Models and Two-phase Samples: Applications to Cox Regression*, pp. 65–77. Beachwood: Institute of Mathematical Statistics.

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. and Kulich, M. (2009) Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statist. Biosci.*, **1**, 32–49.

Capeding, M., Tran, N., Hadinegoro, S., Ismail, H., Chotpitayasunondh, T., Chua, M., Luong, C., Rusmil, K., Wirawan, D., Nallusamy, R., Pitisuttithum, P., Thisyakorn, U., Yoon, I., van der Vliet, D., Langevin, E., Laot, T., Hutagalung, Y., Frago, C., Boaz, M., Wartel, T., Tornieporth, N., Saville, M., Bouckenooghe, A. and CYD14 Study Group (2014) Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *Lancet*, **384**, 1358–1365.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. (1984) On errors-in-variables for binary regression models. *Biometrika*, **71**, 19–25.

Chen, J. and Shao, J. (2000) Nearest neighbor imputation for survey data. *J. Off. Statist.*, **16**, 113–141.

Gao, G. and Tsiatis, A. A. (2005) Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*, **92**, 875–891.

Gilbert, P., McKeague, I. and Sun, Y. (2004) Tests for comparing mark-specific hazards and cumulative incidence functions. *Liftim. Data Anal.*, **10**, 5–28.

Gilbert, P. B. and Sun, Y. (2015) Inferences on relative failure rates in stratified markspecific proportional hazards models with missing marks, with application to human immunodeficiency virus vaccine efficacy trials. *Appl. Statist.*, **64**, 49–73.

Hsu, C.-H. and Yu, M. (2019) Cox regression analysis with missing covariates via nonparametric multiple imputation. *Statist. Meth. Med. Res.*, **28**, 1676–1688.

Jonsson, P. and Wohlin, C. (2004) An evaluation of k-nearest neighbour imputation using Likert data. In *Proc. 10th Int. Symp. Software Metrics*, pp. 108–118. New York: Institute of Electrical and Electronics Engineers.

Juraska, M. and Gilbert, P. (2013) Mark-specific hazard ratio model with multivariate continuous marks: an application to vaccine efficacy. *Biometrics*, **69**, 328–337.

Juraska, M. and Gilbert, P. (2016) Mark-specific hazard ratio model with missing multivariate marks. *Liftim. Data Anal.*, **22**, 606–625.

Juraska, M., Magaret, C., Shao, J., Carpp, L., Fiore-Gartland, A., Benkeser, D., Girerd-Chambaz, Y., Langevin, E., Frago, C., Guy, B., Jackson, N., Duong, T., Simmons, C., Edlefsen, P. and Gilbert, P. (2018) Viral genetic diversity and protective efficacy of a tetravalent dengue vaccine in two phase 3 trials. *Proc. Natn. Acad. Sci. USA*, **115**, E8378–E8387.

Kovar, J., Whitridge, P. and MacMillan, J. (1988) Generalized edit and imputation system for economic surveys at Statistics Canada. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 627–630.

Kovar, J. G. and Chen, E. J. (1994) Jackknife variance estimation of imputed survey data. *Surv. Methodol.*, **20**, 45–52.

Kulich, M. and Lin, D. (2004) Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Am. Statist. Ass.*, **99**, 832–844.

Lee, H., Rancourt, E. and Särndal, C. (1995) Variance estimation in the presence of imputed data for the generalized estimation system. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 384–389.

Lee, H., Rancourt, E. and Särndal, C. E. (1994) Experiments with variance estimation from survey data with imputed values. *J. Off. Statist.*, **10**, 231–243.

Li, K.-C. (1984) Consistency for cross-validated nearest neighbor estimates in nonparametric regression. *Ann. Statist.*, **12**, 230–240.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Little, R. J. A. (1988) Missing-data adjustments in large surveys. *J. Bus. Econ. Statist.*, **6**, 287–296.

Montaquila, J. and Jernigan, R. (1997) Variance estimation in the presence of imputed data. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 273–277.

Moodie, Z., Juraska, M., Huang, Y., Zhuang, Y., Fong, Y., Carpp, L., Self, S., Chambonneau, L., Small, R., Jackson, N., Noriega, F. and Gilbert, P. (2018) Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. *J. Infect. Dis.*, **217**, 742–753.

Nan, B. (2004) Efficient estimation for case-cohort studies. *Can. J. Statist.*, **32**, 403–419.

Prentice, R. L. (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **73**, 1–11.

Rabaa, M. A., Girerd-Chambaz, Y., Duong Thi Hue, K., Vu Tuan, T., Wills, B., Bonaparte, M., van der Vliet, D., Langevin, E., Cortes, M., Zambrano, B., Dunod, C., Wartel-Tram, A., Jackson, N. and Simmons, C. P. (2017) Genetic epidemiology of dengue viruses in phase iii trials of the CYD tetravalent dengue vaccine and implications for efficacy. *eLife*, **6**, article e24196.

Rancourt, E., Särndal, C. and Lee, H. (1994) Estimation of the variance in the presence of nearest neighbor imputation. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 888–893.

Reilly, M. (1993) Data analysis using hot deck multiple imputation. *Statistician*, **42**, 307–313.

Robins, J., Rotnitzky, A. and Zhao, L. (1994) Estimation of regression-coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Scheike, T. H. and Martinussen, T. (2004) Maximum likelihood estimation for Cox's regression model under case–cohort sampling. *Scand. J. Statist.*, **31**, 283–293.

Sedransk, J. (1985) The objective and practice of imputation. In *Proc. 1st A. Res. Conf. US Bureau of the Census, Washington DC*, pp. 445–452.

Stone, C. J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–620.

Sun, Y. and Gilbert, P. (2012) Estimation of stratified mark-specific proportional hazards models with missing marks. *Scand. J. Statist.*, **39**, 34–52.

Sun, Y., Gilbert, P. and McKeague, I. (2009) Proportional hazards models with continuous marks. *Ann. Statist.*, **37**, 394–426.

Sun, Y., Li, M. and Gilbert, P. (2016) Goodness-of-fit test of the stratified mark-specific proportional hazards model with continuous mark. *Computnl Statist. Data Anal.*, **93**, 348–358.

Sun, Y., Qi, L., Yang, G. and Gilbert, P. (2018) Hypothesis tests for stratified mark-specific proportional hazards models with missing covariates, with application to HIV vaccine efficacy trials. *Biometr. J.*, **60**, 516–536.

White, J. E. (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. Epidem.*, **115**, 119–128.

Yang, G., Sun, Y., Qi, L. and Gilbert, P. (2017) Estimation of stratified mark-specific proportional hazards models under two-phase sampling with application to HIV vaccine efficacy trials. *Statist. Biosci.*, **9**, 259–283.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supplementary materials for "A hybrid approach for the stratified mark-specific proportional hazards model with missing covariates and missing marks, with application to vaccine efficacy trials"'.