Received: 20 May 2022 | Revised: 13 June 2023 | Accepted: 23 October 2023

DOI: 10.1002/pst.2348

WILEY

MAIN PAPER

An illness–death multistate model to implement delta adjustment and reference-based imputation with time-to-event endpoints

Alberto García-Hernandez[1] | **Teresa Pérez**[1] | **María del Carmen Pardo**[2] | **Dimitris Rizopoulos**[3]

[1]Facultad de Estudios Estadísticos, Univ. Complutense, Madrid, Spain

[2]Facultad de Ciencias Matemáticas, Univ. Complutense, Madrid, Spain

[3]Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

Correspondence
Alberto García-Hernandez, Facultad de Estudios Estadísticos, Ciudad Universitaria, Avenida Puerta de Hierro s/n, Madrid 28040, Spain.
Email: albega28@ucm.es

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-104681RBI00

Abstract

With a treatment policy strategy, therapies are evaluated regardless of the disturbance caused by intercurrent events (ICEs). Implementing this estimand is challenging if subjects are not followed up after the ICE. This circumstance can be dealt with using delta adjustment (DA) or reference-based (RB) imputation. In the survival field, DA and RB imputation have been researched so far using multiple imputation (MI). Here, we present a fully analytical solution. We use the illness–death multistate model with the following transitions: (a) from the initial state to the event of interest, (b) from the initial state to the ICE, and (c) from the ICE to the event. We estimate the intensity function of transitions (a) and (b) using flexible parametric survival models. Transition (c) is assumed unobserved but identifiable using DA or RB imputation assumptions. Various rules have been considered: no ICE effect, DA under proportional hazards (PH) or additive hazards (AH), jump to reference (J2R), and (either PH or AH) copy increment from reference. We obtain the marginal survival curve of interest by calculating, via numerical integration, the probability of transitioning from the initial state to the event of interest regardless of having passed or not by the ICE state. We use the delta method to obtain standard errors (SEs). Finally, we quantify the performance of the proposed estimator through simulations and compare it against MI. Our analytical solution is more efficient than MI and avoids SE misestimation—a known phenomenon associated with Rubin's variance equation.

1 | INTRODUCTION

In clinical trials, an estimand is a precise description of the treatment effect to be estimated, serving as the focal point of a clinical trial's design, conduct, and analysis. The study estimand is characterized by five attributes: (a) the treatments under comparison; (b) the target population; (c) the subject-level response variable; (d) the population-level summary measure; and (e) the handling of intercurrent events (ICEs). An ICE has been defined as an event that, occurring post-baseline, precludes the observation of the outcome variable or affects its measurement or interpretation.[1]

When a treatment policy strategy is used to handle a specific ICE, the response is evaluated regardless of whether or not participants experience the ICE.[1] Implementing the treatment policy strategy is problematic when subjects are not

followed up after the ICE. In this situation, the ICE has two simultaneous effects: (a) the response may be modified by the ICE; and (b) the response is no longer observed after the ICE. The consequence is a very particular form of informative censoring as the dual effect of the ICE induces an association between the unobserved responses and the censoring mechanism even if the ICE is a completely independent stochastic process. This manuscript focuses on the situation mentioned above.

This problem can be addressed using delta adjustment (DA) and reference-based (RB) imputation.[2–7] DA and RB imputation were first proposed for normally distributed longitudinal endpoints.[2–7] The most popular tool to implement DA and RB imputation—based on multiple imputation (MI)—has been referred to in the literature as the Carpenter, Roger, and Kenward (CRK) algorithm.[2–7] This paper's focus is on time-to-event variables. Table 1 contains an overview of recent research in which the CRK algorithm has been adapted to the survival analysis field.[8–16] These works have essentially followed the same steps proposed originally for longitudinal responses. An imputation model was first fitted on the observed data but not directly used to fill in the missing times after the ICE. Depending on the chosen imputation rule, a certain adjustment was applied to the model parameters. If J2R is used, for instance, the estimated risk function in the reference arm was used to impute times after the ICE in both treatment arms. Once the censored data have been replaced by particular event times, a standard analysis model was fitted on each of the imputed datasets. Finally, the resulting estimates and standard errors (SEs) were combined across imputations.

For longitudinal responses, a few papers have also provided fully analytical solutions to implement DA and RB imputation.[18–22] These methodologies make no use of MI and can be implemented using either a maximum likelihood (ML) approach or a Bayesian framework. However, as far as we are aware, no research has yet addressed DA or RB imputation analytically for time-to-event endpoints. This paper frames this problem as a particular case of the multistate illness–death model. Two of this model's transition intensities (from the initial state to the ICE and from the initial state to the event of interest) are assumed to be identifiable from the data. In addition, we use well-established DA and RB imputation assumptions to speculate the form of the intensity function for the unobserved transition from the ICE to the event of interest. We build estimators for the marginal survival function of interest, calculating the probabilities for the transition from the initial state to the event of interest regardless of the route that subjects followed (i.e., whether or not they passed by the ICE state along the way).

In Section 2, we introduce the proposed parameterization. In Section 3, we propose an estimator using flexible parametric survival models. In Section 4, we present the results of simulations to quantify the performance of the proposed models versus MI. In Section 5, we illustrate the use of the proposed methodology with a real dataset. In Section 6, we analyze the distinct viewpoints available in the literature on SE estimation. Finally, in Section 7, we end the work with some concluding remarks.

2 | THE ILLNESS–DEATH MODEL PARAMETERIZATION

In this section, we will frame the problem addressed in this paper as a specific instance of the multistate illness–death model, featuring the following transitions: (a) from the initial state to the event of interest, (b) from the initial state to the ICE, and (c) from the ICE to the event. For simplicity, we will present the proposed methodology using a specific ICE, namely the discontinuation of the randomized treatment. However, this parameterization can be applied to any other ICEs that we aim to address with the treatment policy strategy and for which no data is available post-ICE. However, we note that for ICEs such as death, implementing this strategy is not possible.[1]

In Section 2.1, we will introduce the proposed formulation. Section 2.2 will outline how we can employ well-established DA or RB imputation assumptions to identify the risk of the event of interest following the ICE. Subsequently, in Section 2.3, we will detail the process of constructing the statistics of interest from the functions used to characterize the illness–death model.

2.1 | Model formulation

We let T_E^* denote the time to the event of interest E regardless of drug discontinuation (event denoted as D). If study individuals were well followed up after D, we could identify directly T_E^* using standard survival models. However, here

TABLE 1 Delta-adjustment and reference-based imputation with time-to-event data.

Reference	Event type	Imputation model	Method to generate random times after the ICE	Imputation rules used	Analysis model or test	Effect size metric	Calculation of the SE
Keene et al. (2014)[8]	Recurrent events	Bayesian negative binomial	Conditional distribution (closed form equation)	DA, J2R, CR	Negative binomial	Rate ratio	Rubin's rule
Zhao et al. (2014)[9]	First event	Kaplan–Meier + bootstrap resampling	Uniform + linear interpolation	DA	Cox PH model, logrank, Wilcoxon test	Survival curves, HR	Rubin's rule
Jackson et al. (2014)	First event	Cox PH model + bootstrap resampling	Uniform + inverse method following Bender et al.[17]	User-defined function $\gamma(\cdot)$	Cox PH model	HR	Rubin's rule
Lu, Li and Koch (2015)[11]	First event	Bayesian PH model (piecewise constant and semi-parametric)	Uniform + inverse method	DA, TPA, J2R	Cox PH model	HR	Rubin's rule and bootstrapping
Akacha & Ogundimu (2016)[12]	Recurrent events	General model for recurrent event data + bootstrap resampling	Asymptotic & bootstrap imputation approach	DA, TPA, J2R	Negative binomial	Rate ratio	Rubin's rule
Lipkovich, Ratitch, and O'Kelly (2016)[13]	First event	Bayesian piecewise exponential PH model, frequentist Kaplan–Meier and Cox PH model (+ bootstrap resampling)	Uniform + inverse method	DA, TPA	Cox PH model and logrank	Survival curves, HR	Rubin's rule
Gao, Liu and Zeng (2017)[14]	Recurrent events	Bayesian piecewise exponential proportional intensity model with frailty	Fitted model (closed form equation)	CR, J2R	Negative binomial	Rate ratio (log-scale)	Rubin's rule
Atkinson et al. (2019)[15]	First event	Weibull PH model (ML)	Uniform + inverse method	J2R, LHCF, CIR, DA, TPA	Cox PH model	HR	Rubin's rule
Atkinson et al. (2020)[16]	First event	Tobit imputation model (ML)	Iterative method using the fitted model	J2R	t test	Difference in means	Rubin's rule

Abbreviations: CIR, copy increment from reference; CR, copy reference; DA, delta adjustment; HR, hazard ratio; ICE, intercurrent events; J2R, jump to reference; LHCF, last hazard carried forward; ML, maximum likelihood; PH, proportional hazards; SE, standard error; TPA, tipping point analysis.

we study the situation where the follow-up is stopped when the randomized therapy is prematurely discontinued (or soon after that).

We will frame our problem using multistate models, an extension of the competing risks class in survival analysis, wherein individuals can undergo various states or conditions over time.[23] In particular, we use the illness–death model, where only three health states are considered. In Figure 1, we illustrate graphically our particular implementation of this model using the following health states:

- A subject is said to be at time t at the *initial* state (I) if the subject has not yet suffered either the endpoint of interest or treatment discontinuation at time t. All subjects start at the initial time ($t = 0$) at state I and may transition to the other states during the study.
- Drug *discontinuation* (D) denotes the state occupied by a subject who has stopped the study treatment permanently but has not yet had the event of interest by time t. A subject who, at time t, is at state D should have transitioned from I at an earlier time.
- Finally, the *endpoint* state (E) is the absorbing state representing the event of interest. A subject who, at time t, is at state E should have transitioned there from either I or D at an earlier time.

We use k to represent a transition with $k \in \{I \to D, I \to E, D \to E\}$, and $\lambda_k(t|\text{arm})$ for the intensity function associated with transition k for a particular treatment arm. We note that the intensity function has close ties with the hazard function as used in standard survival analysis. The intensity function is used in the context of counting processes and multistate modelling and describes the instantaneous rate at which transitions occur.[23] For the sake of succinctness, throughout Section 2, intensities are defined conditioned on the treatment arm only. Later in Section 3, dependency on other baseline covariates will also be considered.

Let us focus first on the two former transitions. The interpretation of $\lambda_{I\to E}(t|\text{arm})$ and $\lambda_{I\to D}(t|\text{arm})$ is similar to that used for cause-specific hazards in survival problems with competing events. They could be used to characterize the distribution of the time-to-event variables that we would observe in a hypothetical world where competing transitions do not exist. When we target the "hypothetical strategy" estimand, we typically model transition $I \to E$ using standard survival models where subjects are censored at the time they transition to D.[24] Now, let us shift our focus to the $D \to E$ transition. The intensity function $\lambda_{D\to E,t_D}(t|\text{arm})$ represents the risk of the event E at time $t > t_D$ for a subject who has transitioned from I to D at time $t_D > 0$. We assume that transition $D \to E$ is not identifiable from the dataset and will characterize its intensity using DA and RB imputation. Throughout this manuscript, times will always have a "clock forward" interpretation, that is, they refer to the overall time from randomization (denoted as $t = 0$). This approach contrasts the "time reset" parameterization in which the intensities associated with transition $D \to E$ are expressed using relative times since entry into state D.[23]

Next, we will contextualize the proposed parameterization using Markov chain theory assumptions. The Markov assumption, in a strict sense, implies that the probability of transitioning to any particular state depends solely on the current state. For transitions $I \to E$ and $I \to D$, we slightly relax the Markov assumption, allowing the transition intensities to depend on duration the subjects have spent in state I. In other words, these two transitions are assumed to adhere to the semi-Markov property.[23] For transition $D \to E$, we generalize further the above assumption, allowing the

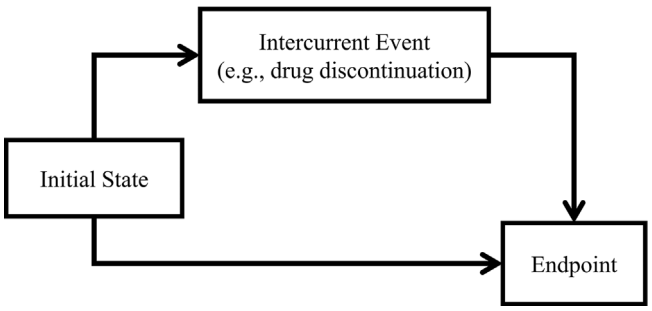

FIGURE 1 The illness–death model parameterization.

intensities to also depend on the time t_D at which the subject entered state D. This condition is said to fulfill a (state arrival) extended semi-Markov property.[23]

2.2 | Identifying transition $D \rightarrow E$ through DA and RB imputation

Throughout this article, we consider two study groups, namely the experimental (exp) and the reference (ref) arms. For the reference arm, either the "no ICE effect" or the DA assumption will be used. For the latter, we allow both a proportional hazards (PH) or additive hazards (AH) parameterization. For the experimental arm, in addition to the three rules above, J2R and copy increment from reference (CIR), using either PH or AH, will also be considered.[13,15,16]

Analytical expressions under the illness–death model for each of these assumptions are given in Equations (1a) and (1b).

$$\lambda_{D\rightarrow E,t_D}(t|\,\text{ref}) = \begin{cases} \lambda_{I\rightarrow E}(t|\,\text{ref}) & \text{if no ICE effect} \\ \lambda_{I\rightarrow E}(t|\,\text{ref})\cdot\delta_{\text{ref}} & \text{if DA}^{\text{PH}} \\ \lambda_{I\rightarrow E}(t|\,\text{ref})+\delta_{\text{ref}} & \text{if DA}^{\text{AH}} \end{cases} \tag{1a}$$

$$\lambda_{D\rightarrow E,t_D}(t|\,\text{exp}) = \begin{cases} \lambda_{I\rightarrow E}(t|\,\text{exp}) & \text{if no ICE effect} \\ \lambda_{I\rightarrow E}(t|\,\text{exp})\cdot\delta_{\text{exp}} & \text{if DA}^{\text{PH}} \\ \lambda_{I\rightarrow E}(t|\,\text{exp})+\delta_{\text{exp}} & \text{if DA}^{\text{AH}} \\ \lambda_{I\rightarrow E}(t|\,\text{ref}) & \text{if J2R} \\ \lambda_{I\rightarrow E}(t|\,\text{ref})\cdot \text{HR}_{I\rightarrow E}(t_D) & \text{if CIR}^{\text{PH}} \\ \lambda_{I\rightarrow E}(t|\,\text{ref})+\text{AHD}_{I\rightarrow E}(t_D) & \text{if CIR}^{\text{AH}} \end{cases} \tag{1b}$$

The "no ICE effect" assumption is defined above as the intensity for $D \rightarrow E$ coinciding with that estimated for $I \rightarrow E$. There is a connection between this rule and the censoring-at-random (CAR) assumption used in the literature when this problem is handled without the explicit modeling of transition $D \rightarrow E$. When transitions to D (i.e., the ICE) are simply regarded as shifts from the noncensored to the censored status, the CAR assumption allows us to use the risk estimated from the noncensored subjects to infer what has happen in the whole population—that is, including censored subjects. Using the proposed parameterization, we express the same concept differently. With "no ICE effect", we assume that the unobserved intensity of transition $D \rightarrow E$ can be characterized using the estimable intensity of transition $I \rightarrow E$.

Using DA, a fixed function $\delta_{\text{arm}}(t)$ or constant δ_{arm} is used to displace the intensity observed for transition $I \rightarrow E$ in a given arm.[8,11–13,15] Only constant shifts $\delta_{\text{arm}}(t) = \delta_{\text{arm}}$ are considered in this paper. We still study DA under both a PH and an AH approach, denoted as DA^{PH} and DA^{AH}, respectively. Coefficients δ_{arm} need to be fixed while guaranteeing that the resulting intensity function is positively defined. Using $\text{DA}^{\text{PH}}, \delta_{\text{arm}} > 0$ is to be interpreted in the ratio scale so $\delta_{\text{arm}} = 0.5$ would represent an assumption that the ICE reduces the intensity by half. To implement the DA method in the AH scale we could use, in principle, either negative or positive shifts. To avoid ending up with negative intensity functions, only positive additive shifts have been considered in this work.

A well-known challenge associated with the DA method is the choice and justification of the δ_{arm} values. This significant constraint has restricted the application of the DA method mostly to sensitivity analysis. One way to circumvent this limitation, the so-called tipping-point analysis (TPA), consists of repeating the DA method with multiple values of δ_{ref} and δ_{exp} to report, for instance, at which values of $(\delta_{\text{ref}}, \delta_{\text{exp}})$ the study results reverse from null hypothesis rejection to acceptance.

Two RB imputation rules have been considered in Equation (1b): J2R and CIR. With the J2R rule, we use the observable risk function in the reference arm, $\lambda_{I\rightarrow E}(t|\text{ref})$, to impute the risk of event in the experimental arm at times $t > t_D$. J2R may be the most widely used RB rule due to its conservative approach towards the null hypothesis, making it appealing for regulatory purposes. With CIR, the experimental drug benefit at t_D, represented by either the hazard ratio $\text{HR}_{I\rightarrow E}(t_D) = \lambda_{I\rightarrow E}(t_D|\text{exp})/\lambda_{I\rightarrow E}(t_D|\text{ref})$ or the AH difference $\text{AHD}_{I\rightarrow E}(t_D) = \lambda_{I\rightarrow E}(t_D|\text{exp}) - \lambda_{I\rightarrow E}(t_D|\text{ref})$, is assumed to be maintained after t_D.[15] The CIR rule has been said to be appropriate for drugs with a long duration of action, such as those studied for Alzheimer's disease.[6,7] Implementing CIR under an AH parameterization is possible only if the intensity resulting from the use of Equation (1b) is a positively defined function $\lambda_{D\rightarrow E,t_D}(t|\text{exp})$ for any times $t_D > 0$ and $t > t_D$. This is something we can check once the intensities of transitions $I \rightarrow D$ and $I \rightarrow E$ have been

characterized. If this condition does not hold, one could still cap to the resulting values to ensure that only positively defined intensities are used for transition D → E.

2.3 | Calculating the survival function of interest

Multistate modeling frequently requires two separate computational efforts. In an initial stage, the dataset is modeled to identify the transition intensities. This task is often insufficient to answer scientific questions of interest. A number of works in this field have focused on the *predictions* task and provided equations to calculate (conditional or not) probabilities of interest from the (estimated) transition hazards.[23–27]

Let us denote the time to enter the intermediate event D as T_D^* and the time of the clinical event of interest as T_E^*. We are interested in estimating the marginal survival function of T_E^*, that we will express as $S_{T_E^*}(t|arm)$, as this function characterizes the probability distribution of T_E^* regardless of whether or not the subject has stopped by state D along the way.

We can decompose

$$1 - S_{T_E^*}(t|arm) = P\left(T_E^* \le t|arm\right)$$

as

$$P\left(T_E^* \le t|arm\right) = \overbrace{P\left(T_E^* \le t, T_E^* < T_D^*|arm\right)}^{P_E^1(t)} + \overbrace{P\left(T_E^* \le t, T_E^* \ge T_D^*|arm\right)}^{P_E^2(t)} \tag{2}$$

where we use the short terms $P_E^1(t)$ and $P_E^2(t)$ to distinguish between having the event of interest E directly from state I and moving through state D respectively. The first element of (2) is the definition of the cumulative incidence function, also known as sub-distribution function, associated with transitions I → E, and is given by

$$P_E^1(t) = \int_0^t P\left(T_E^* > x, T_D^* > x|arm\right) \lambda_{I\to E}(x|arm)dx \tag{3}$$

where

$$P\left(T_E^* > t, T_D^* > t|arm\right) = \exp[-\{A_{I\to E}(t|arm) + A_{I\to D}(t|arm)\}] \tag{4}$$

and $A_k(t|arm)$ is used to denote the cumulative intensity function associated with transition k for arm.

From Putter, Fiocco, and Geskus, we also know that, using a "clock forward" approach, the second component of Equation (2) is given by

$$P_E^2(t) = \int_0^t P\left(T_E^* > x|T_D^* > x, arm\right) \lambda_{I\to D}(x|arm) \overbrace{P\left(T_E^* \le t, T_E^* > x, T_D^* = x|arm\right)}^{P_{E|x}^2(t)} dx \tag{5}$$

where we use $P_{E|x}^2(t)$ to denote the (conditional) probability of moving to state E before or at time t given that the subject has entered state D at time x. This probability is given by

$$P_{E|x}^2(t) = 1 - \exp\left\{-\int_x^t \lambda_{D\to E,x}(u|arm)du\right\}. \tag{6}$$

To calculate $P_{E|x}^2(t)$, Equation (6) considers the transition-specific intensity function for D → E as a valid hazard function as defined in survival analysis without competing events. We can do that because, when a subject is at state D, there are no future transitions competing with the transition to E.[27] We note that Equations (3), (5), and (6) are particular cases of eqs. (34), (31), and (29), respectively, from the work of Putter, Fiocco, and Geskus, as here we do not condition on results up to an interim data point s (with $0 \le s < t$) but set $s = 0$.[23]

We find it useful to sum $P^1_E(t)$ and $P^2_E(t)$ and, after some algebraic manipulation, obtain that

$$S_{T^*_E}(t|arm) = P\left(T^*_E > t | T^*_D > t, arm\right) + \int_0^t P\left(T^*_E > x | T^*_D > x, arm\right) \lambda_{I\to D}(x|arm)\, \exp\left\{ -\int_x^t \lambda_{D\to E,x}(u|arm)du \right\} dx \quad (7)$$

We have included a proof of Equation (7) in the Supplementary information. Resolving the outer integral in Equation (7) may or may not have a closed form depending on the form of the integrand components. In Section 4, we propose an estimator that resolves this integration numerically.

In Equations (1a) and (1b), we have seen that we may assume that the ICE has no effect. This is expressed within this framework as $\lambda_{D\to E,t_D}(t|arm) = \lambda_{I\to E}(t|arm)$ for any times $t_D > 0$ and $t > t_D$. Let us end this section with a little note. If, for a given arm, the ICE has no effect, then $S_{T^*_E}(t|arm) = \exp[-\{A_{I\to E}(t|arm)\}]$ at any $t > 0$. A proof of this assertion has also been included as Supplementary information. In summary, if we can assume no effect of the ICE, we could build estimates for T^*_E by using the standard survival models we typically use for the "hypothetical strategy" estimand in which we focus on transition $I \to E$ and censor otherwise.

To conclude this section, we observe that the methodology proposed in this paper essentially constructs estimators for the treatment policy by merging parameters derived from a model targeting the hypothetical strategy estimand (i.e., using data up to the ICE) with those related to the time to the ICE process. The unobserved post-ICE evolution is integrated into the estimator using DA and RB imputation. An approach quite similar to this was recently proposed by Garcia-Hernandez et al. for normally distributed longitudinal variables.[22]

2.4 | On the dependence between transitions $I \to E$ and $I \to D$

In this section, we will focus on an important assumption behind the parameterization proposed above. The transition-specific intensities obtained for $I \to E$ and $I \to D$ are known to be valid estimates of the true intensities even if these two processes are correlated with each other. We also know that the transition-specific sub-distribution function, $P\left(T^*_E \le t, T^*_E < T^*_D | arm\right)$, can be obtained from those intensities via Equations (3) and (4) under either independence or dependence.[23] However, these intensities should not be interpreted as standard hazard functions unless we can assume independence. The expression $\exp[-\{A_{I\to E}(t|arm)\}]$ represents the marginal survival function of T^*_E in a hypothetical world without transitions to D only if the two competing transitions $I \to E$ and $I \to D$ are independent.[23,28]

In Equation (6), we have calculated $P\left(T^*_E > t | T^*_E > x, T^*_D = x, arm\right)$ from the transition-specific intensities associated with $D \to E$ using an equation that considers this intensity function as a competing-events-free hazard function. We can do that because, when the subject is in state D, there are no future transitions competing with the transition to E. The problem is that we do not observe $\lambda_{D\to E,x}(\cdot)$. In Equations (1a) and (1b) we take the intensities for transition $I \to E$, obtained in a competing environment, and use them to identify $\lambda_{D\to E,x}(\cdot)$ that we later interpret as a hazard function. To accomplish this, the proposed parameterization needs to assume independence between the event of interest and the occurrence of the ICE (at least conditional on the baseline covariates).

We note, though, that this important assumption is not specific to this solution but also applies to the MI-based CRK algorithm. For longitudinal responses, Seaman et al. have used the term "noninformative deviation" for a similar assumption of this estimator.[29] Recently, Garcia-Hernandez et al. have shown, for longitudinal variables, that DA and RB imputation can be implemented combining parameters coming from a model targeting the hypothetical strategy estimand (i.e., using data up to the ICE) with those associated with the time to the ICE process. They have shown that the "noninformative deviation" is essentially the MAR assumption behind the former model.[22] This behavior is mirrored here as the set of models typically used to characterize transition $I \to E$ are also the models we would use in a study that targets the hypothetical strategy estimand (i.e., censoring times at the onset of the ICE). In Section 4, we will illustrate empirically this point via simulations.

2.5 | On the treatment effect size under a "treatment policy strategy" estimand

In comparative clinical trials, it is useful to build a clinically interpretable measurement of the treatment effect size. In the survival analysis field, we observe two major tendencies. In some therapeutic areas (e.g., oncology), the absolute

difference between arms of the median time is frequently used. In other diseases, a constant HR of the experimental vs. the reference arm is typically reported. Another quantity, increasingly used, is the mean time on its version for right-censored data—the restricted mean survival time (RMST).[30–33]

In Section 3, we propose an estimator that can be used to obtain any of these statistics, but one important clarification is required with respect to the HR. By definition, the "treatment policy strategy" estimand looks at the drug's response beyond the ICE. DA and RB imputation, in one way or the other, modifies the risk function once subjects suffer the ICE. This alteration is generally incompatible with the constant HR assumption. Using J2R, for example, the treatment effect is assumed to fully disappear for those with ICE after its occurrence (i.e., the HR after the transition to D is assumed to be 1). This feature has an evident consequence. Even if the PH assumption holds for transition $I \to E$, this property will no longer hold for the marginal survival curve $S_{T_E^*}(t)$ that looks at the risk of event beyond t_D. This issue has been acknowledged in previous manuscripts in this field.[9,11,13,15] These works still used Cox regression (i.e., reporting one single HR) but recognized that this is just a simple way to report the average hazard ratio (AHR) over the study period.

The illness–death parametrization proposed in this paper does not lead to a constant HR for the process of interest, T_E^*, even if we use PH models to model transitions $I \to E$ and $I \to D$. Under the proposed illness–death model parameterization, we are still able to report the AHR over time. For this, we will use Kalbfleisch and Prentice's AHR function.[34,35]

3 | A FLEXIBLE PARAMETRIC ESTIMATOR

Section 2 introduced an illness–death model formulation for implementing DA and RB imputation with time-to-event endpoints. In Section 3.1, we will employ flexible parametric survival models to fit the data, enabling us to characterize the intensity functions of the transitions. Section 3.2 will present an algorithm to calculate the statistics of interest from the estimated model parameters. Finally, Section 3.3 will present alternative modeling options, distinct from the one utilized in this paper.

3.1 | Modeling transitions $I \to E$ and $I \to D$

Multistate models are an extension of the competing risks models' class. In the problem covered by this paper, we have no data to characterize the intensity of transition $D \to E$ but identify it using DA and RB imputation. Leaving aside this transition, the data modeling effort we encounter is essentially a standard survival analysis with two competing events (i.e., our transitions $I \to D$ and $I \to E$). To identify these two transitions, we could use either nonparametric, semiparametric, or fully parametric methods.[23,27,36,44] In this paper, we will use flexible (possibly stratified) parametric models. We have chosen this approach as it allows us to adjust for baseline covariates using either the PH assumption or stratification. A parametric framework, with close equations for the transition-specific intensities $I \to D$ and $I \to E$, is useful to eventually estimate the time-to-event of interest T_E^*.

We model the transition-specific intensity of transition k for a subject with covariate vector Z that belongs to the j^{th} stratum as

$$\lambda_{k,j}(t|Z;\beta) = \lambda_{0,k,j}\left(t|\gamma_{k,j}\right)\exp\left(\beta_k^\top Z\right) \tag{8}$$

where $\lambda_{0,k,j}\left(t|\gamma_{k,j}\right)$ is the baseline intensity for transition k on stratum j that depends on the vector of parameters $\gamma_{k,j}$. The vector β_k represents the proportional-hazard effects for transition k associated with covariate vector Z. We have not included the treatment arm identifier in Equation (8) to allow this item to be part of either Z (PH effect) or the strata definition (non-PH effect).

There are many options in the literature to model parametrically the baseline risk function. The most popular parametric model is arguably the Weibull model. This model can be formulated in terms of the log-transformed baseline cumulative intensity function as an increasing linear equation $\log\{A_0(t)\} = \gamma_0 + \gamma_1 \cdot x$ with $x = \log(t)$ and $\gamma_1 > 0$. The Royston–Parmar (RP) model is a flexible generalization of this model, with $\log\{A_0(t)\} = s(x;\gamma)$ where $s(x;\gamma)$ is a smooth function of x and γ is the vector of coefficients associated with the smooth function parameters.[37] Using the RP

model, we can express the cumulative baseline intensity associated with transition k for a subject with covariate vector Z that belongs to the j^{th} stratum as

$$\log\{A_{k,j}(t)|Z,\gamma\} = \overbrace{\sum_{q=1}^{Q}\gamma_{k,j,q}B_{k,j,q}(x)}^{\log\{A_{0,k,j}(t)|\gamma\}} + \beta_k^{\top} Z \tag{9}$$

where coefficient $\gamma_{k,j,q}$ is associated with the basis function $B_{k,j,q}(x)$ with $q \in \{1,...,Q\}$. Several options are available for the parameterization of the smooth function $s(x;\gamma)$. We have chosen natural cubic splines (NCS) as they fit straight lines (following a Weibull pattern) at the boundary knots. With NCS, we need as many basis functions (and corresponding parameters) as the total number of knots (including boundary and internal knots). A particular case with no internal knots would reduce this smooth function to the two-parameter Weibull model. Further guidance on how to build the NCS basis functions $B_{k,j,q}(x)$ for a given t and set of knots $\upsilon_{k,j} = \{\upsilon_{k,j,1},...,\upsilon_{k,j,Q}\}$ is provided as Supplementary information.[38] We note that, above, we have assumed that a separate set of knots $\upsilon_{k,j}$ is used for each transition k and stratum j. In case we go for a common set of knots υ across transitions/strata, the basis functions can be expressed simply as $B_q(x)$.

Since $\lambda(t) = dA(t)/dt$, the log-intensity function of the RP model can be formulated in terms of the basis functions as follows:

$$\log\{\lambda_{k,j}(t)|Z;\beta,\gamma\} = -\log(t) + \sum_{q=1}^{Q}\gamma_{k,j,q}B_{k,j,q}(x) + \beta_k^{\top} Z + \log\left\{\sum_{q=1}^{Q}\gamma_{k,j,q}B'_{k,j,q}(x)\right\}. \tag{10}$$

A known challenge that comes with the RP model is the decision on the number and placement of internal knots. For the boundary knots, it is customary to use maximum and minimum uncensored log times.[37,39] For internal knots, Royston and Parmar, following Durrleman and Simon's recommendation, favored using up to three knots at equally distributed quantiles of the log-transformed uncensored event times.[37,40] Other simulation studies have confirmed four internal knots are likely to fit well even the more complex shapes of the baseline hazard function.[41,42] However, to obtain stable and accurate estimates, we must also have a minimum number of events within each pair of consecutive knots.

The model above can be estimated via ML. In this work, we perform this task using the SAS® NLMIXED procedure. A SAS® script is included in the Supplementary information.

To end this section, we note that the above model allows having censored observations apart from those driven by the ICE. Censoring unrelated to the ICE is handled as usual in the survival field. Two frequent events triggering censoring in randomized clinical trials are (a) the administrative censoring caused by the moment when the database is cut off for analysis, and (b) the censoring caused by a subject being loss to follow up during the study. These censoring mechanisms are also assumed to be at random, at least, once we condition on the baseline covariates used in the model.

3.2 | Building drug's effect size quantities for T_E^*

Following the results of Section 2, we propose estimating the survival function of interest for a given covariate vector Z, stratum j, and time t by plugging in the estimates of the above model into Equation (7). The second term of Equation (7) may not have a tractable form but can be approximated using numerical integration. In the following, we detail the algorithm we propose to calculate all metrics of interest numerically.

To obtain the survival value at a given time horizon t^* (e.g., end of the planned treatment period) for a given pair $\{j,Z\}$ we first create a partition of $(0,t^*]$ into n (e.g., 100) intervals of identical length $\Delta = t^*/n$. We calculate $S_{T_E^*}(t^*|j,Z;\hat{\theta})$ using Equation (7) where the second summand is approximated using the trapezoidal rule, except for the first interval where we use a right Riemann sum approach to avoid evaluation at time $t = 0$.[43]

1. We also approximate the survival function $S_{T_E^*}$ at the interim datapoints u_l (viz., the upper limit of subinterval l) using the same methodology on the subpartition of $(0,u_l]$ formed by the first l intervals of the above partition.

2. The RMST function at each timepoint u_l, defined as $\int_0^{u_l} S_{T_E^*}(t|j,Z)dt$, is estimated using the trapezoidal rule and thus re-uses the $S_{T_E^*}(u_l|j,Z;\hat{\theta})$ elements approximated before. We will use $h_{T_E^*}(t)$ to denote the hazard function associated with the time-to-event endpoint of interest T_E^*. For a given covariate vector Z, stratum j, and $l \in \{1,..,n\}$, $h_{T_E^*}(m_l|j,Z;\hat{\theta})$ is approximated as the increment in the estimated $-\log\{S_{T_E^*}(t|j,Z;\hat{\theta})\}$ function from $t = u_{l-1}$ to $t = u_l$ divided by Δ where u_0 here is set to 0 and $m_l = (u_{l-1} + u_l/2)$.
3. Finally, we estimate the median by first observing which item $l^* \in \{1,...n\}$ meets that $S_{T_E^*}(u_{l^*-1}|j,Z;\hat{\theta}) > 0.5$ and $S_{T_E^*}(u_{l^*}|j,Z;\hat{\theta}) \leq 0.5$. Afterward, we obtain the median by using linear interpolation from u_{l^*-1} and u_{l^*} and their corresponding survival function evaluations.

In this section, we have made no reference yet to the treatment arm to allow this important attribute to be modelled by using either the covariates vector Z or the strata identifier j. Treatment effect can be obtained by first calculating the above statistics for each treatment group and, afterward, calculating the appropriate comparison (i.e., subtraction or ratio). Regardless of how we model treatment, we can also estimate the AHR (experimental vs. placebo) associated with probability distribution of interest (T_E^*) over the period $(0,t^*]$ using, for instance, the AHR function of Kalbfleisch and Prentice.[34] In this paper, we will use the parameterization of this function provided below.[35]

$$AHR(Z,\hat{\theta}) = \frac{\sum_{l=1}^{n} \frac{h_{T_E^*}(m_l|exp,Z;\hat{\theta})}{h_{T_E^*}(m_l|exp,Z;\hat{\theta}) + h_{T_E^*}(m_l|ref,Z;\hat{\theta})} f_{T_E^*}(m_l|Z;\hat{\theta})\omega(m_l)}{\sum_{l=1}^{n} \frac{h_{T_E^*}(m_l|ref,Z;\hat{\theta})}{h_{T_E^*}(m_l|exp,Z;\hat{\theta}) + h_{T_E^*}(m_l|ref,Z;\hat{\theta})} f_{T_E^*}(m_l|Z;\hat{\theta})\omega(m_l)} \tag{11}$$

where

$$f_{T_E^*}(m_l|Z;\hat{\theta}) = h_{T_E^*}(m_l|exp,Z;\hat{\theta}) S_{T_E^*}(m_l|exp,Z;\hat{\theta}) +$$

$$h_{T_E^*}(m_l|ref,Z;\hat{\theta}) S_{T_E^*}(m_l|ref,Z;\hat{\theta}), \tag{12}$$

$S_{T_E^*}(m_l|exp,Z;\hat{\theta})$ is approximated as the average of the survival at the subinterval limits (u_{l-1} and u_l) and the weight function $\omega(t)$ is chosen to reflect the importance of the different time periods. Frequent choices for $\omega(t)$ are a constant function or the estimated survival function at t. With the former option, all times are treated as equally important while the latter considers the importance of HRs at different times as being proportional to the numbers of individuals at risk at these times.[35] We note that Equation (11) assumes that treatment has been modeled as strata and no other strata variables are modeled.

SEs for the between-arm difference in medians and RMST, and the log-transformed AHR are built in this work using the delta method (the method used by the SAS® NLMIXED procedure to estimate variances for functions of model parameters).

It is worth mentioning that the effect size metrics obtained above represent conditional treatment effects (unless the model considers no baseline variables other than the treatment arm). In a particular trial, these amounts can be reported for a theoretical average subject on the baseline covariates $(Z = \overline{Z})$. It is also possible to build a fully marginalized effect size using, for example, the plug-in estimator proposed by Freedman.[44] We will illustrate how to implement this technique in Section 5.

3.3 | Other modeling options

The estimator proposed in Sections 3.1 and 3.2 is one among the many options available in the literature to fit an illness–death model.[23,27,36,44,45] In this section, we will consider a few alternatives.

We have chosen to use the RP model to identify transitions $I \rightarrow E$ and $I \rightarrow D$. A parametric alternative to model the baseline hazard would be to use the piecewise exponential model and thus assume the intensities to be discontinuous step functions. Depending on the form of the assumed risk after the ICE, the integrand in Equation (7) may have a tractable form making it possible to calculate $S_{T_E^*}(t|j,Z;\hat{\theta})$ analytically. In this paper, we use the RP model (followed by numerical integration) as the added computational burden is manageable. The algorithm proposed above to obtain

$S_{T_E^*}(t|j,Z;\hat{\theta})$ can be used straight-away with any possible assumptions for the post-ICE risk (i.e., beyond the ones considered in this paper).

Another option we have left out of the scope of this paper is using a nonparametric approach to model our data. Under this framework, we could estimate the marginal survival function $S_{T_E^*}(t|arm)$ using the Aalen–Johansen estimator with the caveat that we do not have data to estimate one component, $P\left(T_E^* \leq t | T_E^* > x, T_D^* = x, arm\right)$, but need to impute it using DA or RB rules.[27] An evident limitation of using a nonparametric approach is that we cannot adjust for covariates as we do in the model proposed above. We also note that nonparametric methods have been discouraged for the calculation of the RMST as survival estimators can be unstable at times for which the number of subjects at risk is small.[30,31]

To end this section, one more modeling variation will be contemplated. Back in Section 2, we saw that we need to assume independence between transitions $I \rightarrow E$ and $I \rightarrow D$ (at least conditioned on the baseline variables used in the model). It could be the case that baseline covariates are not enough to explain the association, but we need to include postbaseline intermediate outcomes (e.g., biomarkers). A postbaseline variable may be affected by the studied treatment and could then be a mediator. It is generally inadequate to adjust a clinical trial analysis for a postbaseline response. Using the CRK algorithm, it is possible to adjust the imputation model for this postbaseline (i.e., time dependent) variable while taking the variable out of the analysis model. Using a fully analytical solution, this is more difficult as imputation and analysis models are unified in a single effort. One possible solution is to use inverse probability of censoring weighting (IPCW).[44] Using this methodology, the analysis of a certain transition (e.g., $I \rightarrow E$) could use weights built using the probabilities of not having transitioned to a competing transition at that time. To calculate these probabilities (and the subsequent weights), we can use a survival model with time-varying covariates.[44,46] This possible extension has been left out of the scope of this work.

4 | SIMULATION STUDY

4.1 | Simulation study design

A two-year treatment period clinical trial with an overall sample size of 500 subjects allocated to either placebo (arm = 0) or experimental (arm = 1) using a 1:1 ratio was simulated 5000 times. Only one continuous baseline covariate, denoted as bsln, was simulated using a normal distribution with mean = 0 and standard deviation = 1. Two competing event times, $T_{i,I\rightarrow E}^*$ and $T_{i,I\rightarrow D}^*$ were simulated using the following hazard functions:

$$\lambda_{I\rightarrow E}(t|bsln, arm) = \sqrt{t} \times \exp(0.5bsln - 0.3arm) \tag{13}$$

and

$$\lambda_{I\rightarrow D}(t|bsln, arm) = 0.2. \tag{14}$$

The event of interest was associated with the simulated times $T_{i,I\rightarrow E}^*$. The ICE was represented by the simulated times $T_{i,I\rightarrow D}^*$. These two events competed with each other and with an administrative censoring mechanism ($C_i = 2$). The targeted estimand follows a treatment policy strategy. That means that we wish to ignore the occurrence of the ICE (times $T_{i,I\rightarrow D}^*$). However, we assumed that we did not follow up the subjects following the occurrence of the first among these three events ($T_{i,I\rightarrow E}^*$, $T_{i,I\rightarrow D}^*$, & C_i) so we observed only $T_i = \min\left(T_{i,I\rightarrow E}^*, T_{i,I\rightarrow D}^*, C_i\right)$ and two indicators $\delta_{i,I\rightarrow E} = I\left(T_i = T_{I\rightarrow E}^*\right)$ and $\delta_{i,I\rightarrow D} = I\left(T_i = T_{I\rightarrow D}^*\right)$. We did not allow for ties, so it is not possible that, for a particular subject, both $\delta_{i,I\rightarrow E}$ and $\delta_{i,I\rightarrow D}$ are 1. Traditional survival methods treat $T_{i,I\rightarrow D}^*$ purely as a censoring mechanism, assuming that the instantaneous risk of event E following $T_{i,I\rightarrow D}^*$ can be accurately represented using the data observed up to $T_{i,I\rightarrow D}^*$. As we do not anticipate this assumption to be valid, we employ DA or RB imputation.

For each simulated dataset, transitions $I \rightarrow E$ and $I \rightarrow D$ were modeled using the following models:

$$\lambda_{I\rightarrow E}(t|bsln, arm) = \lambda_{0,I\rightarrow E,arm}(t)\exp(\beta_{I\rightarrow E}bsln), \tag{15}$$

and

$$\lambda_{I\to D}(t|\text{bsln},\text{arm}) = \lambda_{0,I\to D,\text{arm}}(t)\exp(\beta_{I\to D}\text{bsln}) \tag{16}$$

where the baseline intensity functions, $\lambda_{0,k,\text{arm}}(t)$, were modeled using the RP model with a five-knot NCS smooth function. This model uses 5 parameters $\gamma_{k,\text{arm},1}$ to $\gamma_{k,\text{arm},5}$ per arm and transition. For each transition $k \in \{I\to D, I\to E\}$, the three internal knots were fixed at the quartiles of the noncensored event times for each trial. Both the minimum and the maximum transition-specific event times were used as lower and upper boundary knots, respectively.

From Equation (1a) and (1b), we have up to $3\times 6 = 18$ possible combinations on the assumed risk after D. For the sake of succinctness, the intensity of transition $D\to E$ was always assumed to coincide with that estimated for transition $I\to E$ (no ICE effect) for the placebo arm. For the experimental arm, the following DA and RB rules were used: no ICE effect, DA^{PH} with $\delta_{exp} = 1.5$, DA^{AH} with $\delta_{exp} = 0.3$, J2R, CIR^{PH} and CIR^{AH}.

The following amounts were obtained for each sample: log-transformed AHR, difference of medians, and difference of RMSTs. The estimator proposed in Section 3 was implemented using a partition of the $(0,2]$ years interval formed by 100 equally sized subintervals. For the calculation of the AHR, we used Equation (11) with a weighting function set to the average of the estimated survival functions across arms. For each trial, these amounts were calculated for a theoretical subject with $\text{bsln} = \overline{\text{bsln}}$. Figure 2 shows the true treatment effect size versus the baseline value (bsln).

The fully analytical approach proposed in this paper was compared against the CRK algorithm. The latter neither uses multistate modeling nor distinguishes explicitly transition $I\to E$ from transition $I\to D$, as the latter is simply regarded as a censoring mechanism. In the following, we briefly explain how we implemented the CRK algorithm in this paper. A Bayesian version of the model in Equation (9) was fitted using Markov chain Monte Carlo (MCMC). Only transition $I\to E$ was modeled. Flat independent priors were used for parameter β and the baseline-hazard parameters $\gamma_{\text{arm},1}$ to $\gamma_{\text{arm},5}$. Parameters $\gamma_{\text{arm},2} > 0$ so its log-transformation was included as a model parameter. Subjects who discontinued the treatment during the study required imputing their time from the ICE onset up to the event of interest or the end of follow up time (whatever came first). We drew $L = 20$ approximately independent samples from the posterior distribution of the model parameters. To minimize autocorrelation, a 1/200 thinning rate was used. Using these values and one particular rule (e.g., J2R), we imputed the unobserved post-ICE times. Let us use T_i to denote the observed time at which subject i discontinues treatment and T_i^s the imputed time of the event of interest given a sample s of model parameters. To simulate the T_i^s values, we used the probabilities $P(T_i^s \le t | T_i^s > T_i) = \{-\int_{T_i}^{t} h(u)du\}$ for $t > T_i$ where $h(u)$

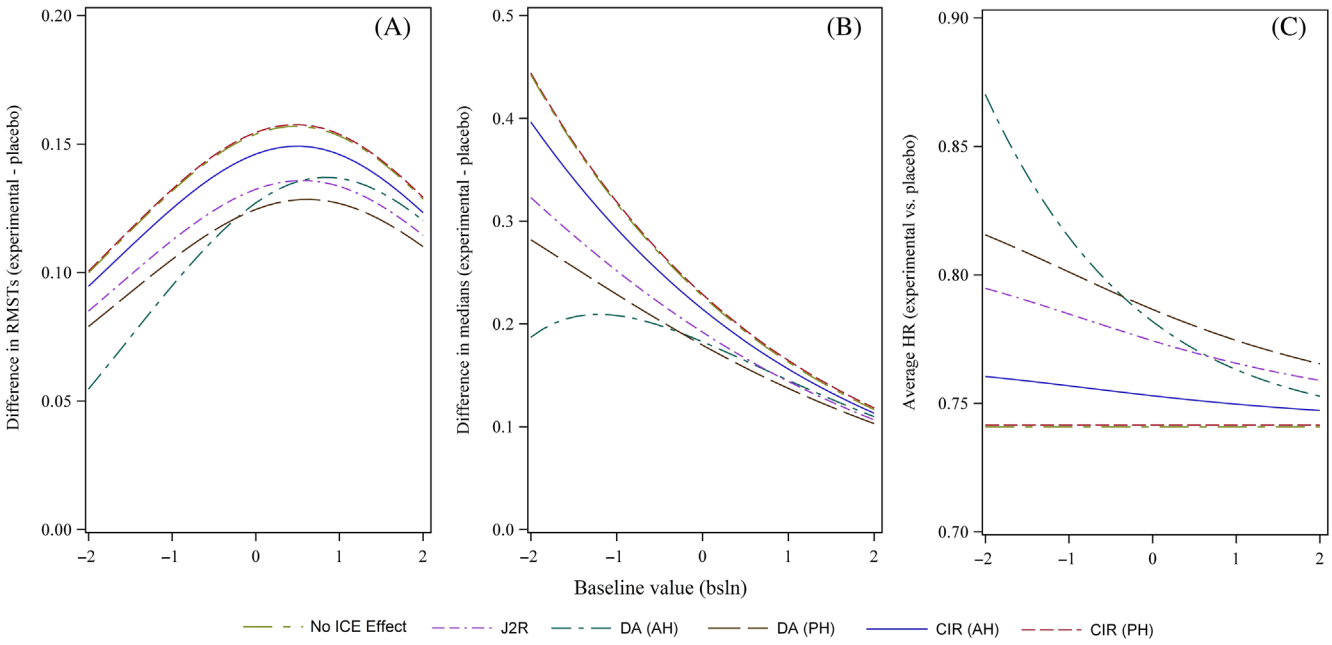

FIGURE 2 Simulation study: True treatment effect by rule and baseline value. DA (AH) with $\delta = 1.5$; DA (PH) with $\delta = 0.3$. AH, additive hazard; CIR, copy increment from reference; DA, delta adjustment; HR, hazard ratio; ICE, intercurrent events; J2R, jump to reference; PH, PH, proportional hazard; RMST, restricted mean survival time.

is the hazard function estimated initially from the data and afterward altered using the chosen DA or RB rule. As inverting the RP model survival function is not straightforward, we divided the time interval from T_i to $C_i = 2$ into 100 equally sized subintervals of length $\Delta = (C_i - T_i)/100$ with $C_i = 2$. For each subinterval l, with $l \in \{1,...100\}$, an exponentially distributed time was simulated with parameter λ set to the geometric mean of the hazards at the boundaries of the subinterval. This loop was repeated as long as the simulated time for subinterval l, that we will denote as $T_i^{s,l}$, was greater than Δ (as this represents a situation in which the event has not happened in subinterval l). If at some subinterval l^* $T_i^{s,l^*} < \Delta$, then the event was concluded to happen at time $T_i^s = T_i + \Delta(l^* - 1) + T_i^{s,l^*}$. Cases where $T_i^{s,l} \geq \Delta$ for all $l \in \{1,...100\}$ were regarded censored observations at time $C_i = 2$. Afterward, all 20 imputed datasets were analyzed with an ML version of the same survival model, and the effect size metrics were calculated using the same methodology followed for the analytical solution. The resulting treatment effects and corresponding SEs were finally combined using the standard Rubin's equation (via the SAS® MIANALYZE procedure).

The following performance metrics were obtained for each model: median execution time (MET), mean bias, mean (estimated) SE, empirical standard error (ESE), 95% confidence interval (CI) coverage rate, and test power. In addition, we calculated the relative increase in precision of the analytical solution with respect to the CRK-algorithm. For a detailed definition of these performance measurements, we refer to the work of Morris et al.[47]

A additional set of simulations was conducted for sensitivity purposes: (i) two alternative baseline intensities were used for the transition $I \rightarrow E$: $\lambda_{0,I\rightarrow E}(t) = t^{-0.1}$ and $\lambda_{0,I\rightarrow E}(t) = 1 + t$; (ii) two other sample sizes were studied: $n = 150$ and 2400; and (iii) the study follow-up period was considered variable among subjects using an administrative censoring mechanism with times $C_i = 2 - (i-1)/(n-1)$ years with $n = 600$ subjects. That is, the first randomized subject ($i = 1$) was on the study for 2 years while the last randomized subject ($i = n$) was followed up for 1 year.

To end this experiment, two additional questions were studied:

1. Performance of the analysis solution versus the CRK algorithm depending on the value L (number of imputed datasets) used to implement the latter. In particular, we studied the CRK algorithm using $L = 5$, 10, 20, 40, and 80.
2. Impact of the association between transitions $I \rightarrow E$ and $I \rightarrow D$ to the results. For this, subjects were assumed to be split at random into two equally sized latent (unobserved) classes. We will use SEV (1 = Yes, 0 = No) to identify the class of severe disease subjects with a worse prognosis. We assumed that the severe disease subjects all had (a) a higher event rate; (b) a higher discontinuation rate; and (c) a stronger treatment effect. To simulate this situation, we used the following intensity functions:

$$\lambda_{I\rightarrow E}(t|\alpha, arm) = \sqrt{t} \times \exp(-1 + SEV - 0.25 \times arm - 0.75 \times arm \times SEV),$$

$$\lambda_{I\rightarrow D}(t|\alpha, arm) = 0.2 \times \exp(-1 + 3 \times SEV).$$

For the sake of succinctness, the latter analyses (sensitivity and additional) were studied only assuming "no ICE effect" for the placebo group and J2R for the experimental arm.

4.2 | Simulation study results

Tables 2a and 2b summarize the performance of the studied tools. With any of the models and rules studied, mean bias was practically zero and explainable by the sampling variation—that is, within $\pm 2 \times ESE/\sqrt{5{,}000}$. We studied the estimated sampling variance by comparing the average estimated SE against the ESE and, in addition, by observing the coverage of the 95% CI. The proposed analytical solution provided, for all studied rules, average estimated SEs equal or close to the ESEs and coverage rates of 95%. The MI-based CRK algorithm provided adequate SE estimates when used with the "no ICE effect" and DA imputation. With J2R, though, the average estimated SE was fairly higher than the ESE with coverage rates ranging from 97% to 98% across the three studied statistics. This expected finding, explained by the known incompatibility between imputation and analysis models, is discussed further in Section 6. With CIR, the CRK overestimated slightly the SE when an AH approach was used. On the other hand, an underestimated SE was observed when CIR was implemented using PH. The bias observed with CIR was always smaller than that observed using J2R with coverage rates remaining within the 94% to 96% interval across the three studied statistics.

TABLE 2a Simulation study results: performance of the illness–death model.

Rule placebo arm	Rule experimental arm	MET	Difference of RMSTs (years) (experimental minus placebo) over the (0, 2] interval						Difference of medians (years) (experimental minus placebo)						Log-AHR (experimental vs. placebo) over the (0, 2] years interval					
			True mean	Mean bias	Mean SE	ESE	Coverage	Power	True Value	Mean bias	Mean SE	ESE	Coverage	Power	True Value	Mean bias	Mean SE	ESE	Coverage	Power
No ICE effect	No ICE effect	2.7	0.154	0.000	0.058	0.058	95.3%	74.8%	0.227	0.000	0.100	0.101	95.2%	61.3%	−0.300	−0.001	0.114	0.114	95.4%	75.0%
No ICE effect	DA^{PH} $\delta_{exp} = 1.5$	54.8	0.125	0.000	0.059	0.059	95.2%	56.3%	0.179	0.000	0.097	0.097	95.2%	45.1%	−0.242	−0.001	0.114	0.114	95.5%	56.3%
No ICE effect	DA^{AH} $\delta_{exp} = 0.3$	52.3	0.127	0.000	0.057	0.057	95.3%	59.4%	0.183	0.000	0.095	0.095	95.2%	47.9%	−0.247	−0.001	0.113	0.112	95.5%	59.7%
No ICE effect	J2R	62.6	0.132	0.000	0.051	0.051	95.3%	73.5%	0.192	−0.001	0.083	0.083	95.4%	63.2%	−0.258	0.000	0.100	0.100	95.4%	73.7%
No ICE effect	CIR^{PH}	82.6	0.154	0.000	0.061	0.061	94.9%	71.9%	0.227	−0.001	0.100	0.101	95.1%	61.1%	−0.300	0.000	0.119	0.119	94.9%	71.7%
No ICE effect	CIR^{AH}	102.1	0.146	0.001	0.057	0.057	95.1%	72.9%	0.214	0.000	0.094	0.094	95.4%	62.4%	−0.285	−0.001	0.111	0.111	95.2%	73.3%

Abbreviations: AH, additive hazard; CIR, copy increment from reference; DA, delta adjustment; ESE: empirical standard error; ICE, intercurrent events; J2R, jump to reference; MET, median execution time (seconds); PH, proportional hazard; RMST, restricted mean survival time; SE, standard error.

TABLE 2b Simulation study results: performance of the Carpenter, Roger, and Kenward (CRK) algorithm.

Rule placebo arm	Rule experimental arm	MET	Difference of RMSTs (years) (experimental minus placebo) over the (0, 2] interval						Difference of medians (years) (experimental minus placebo)						Log-AHR (experimental vs. placebo) over the (0, 2] years interval					
			True value	Mean bias	Mean SE	ESE	Coverage	Power	True Value	Mean bias	Mean SE	ESE	Coverage	Power	True Value	Mean bias	Mean SE	ESE	Coverage	Power
No ICE effect	No ICE effect	63.2	0.154	0.000	0.059	0.059	95.0%	73.4%	0.227	0.000	0.102	0.101	95.2%	60.1%	−0.300	0.002	0.115	0.115	95.2%	73.8%
No ICE effect	DA^{PH} $\delta_{exp} = 1.5$	65.5	0.125	0.000	0.059	0.059	95.0%	55.6%	0.179	0.000	0.098	0.098	95.0%	44.9%	−0.242	0.002	0.114	0.115	95.1%	55.4%
No ICE effect	DA^{AH} $\delta_{exp} = 0.3$	57.4	0.127	0.000	0.059	0.058	95.6%	57.3%	0.183	−0.001	0.098	0.096	95.7%	45.4%	−0.247	0.002	0.114	0.112	95.6%	57.4%
No ICE effect	J2R	65.7	0.132	−0.001	0.058	0.052	97.3%	62.7%	0.192	−0.002	0.097	0.084	97.6%	49.8%	−0.258	0.002	0.113	0.100	97.5%	62.8%
No ICE effect	CIR^{PH}	64.2	0.154	0.000	0.059	0.062	93.8%	72.3%	0.227	−0.001	0.103	0.102	95.2%	59.1%	−0.300	0.002	0.116	0.120	94.0%	72.4%
No ICE effect	CIR^{AH}	56.6	0.146	0.001	0.059	0.057	95.4%	70.8%	0.214	0.000	0.100	0.095	96.2%	57.1%	−0.285	0.000	0.114	0.112	95.4%	70.9%

Note: Number of imputed datasets (L) is 25.

Abbreviations: AH, additive hazard; CIR, copy increment from reference; DA, delta adjustment; ESE, empirical standard error; ICE, intercurrent events; J2R, jump to reference; MET, median execution time (seconds); PH, proportional hazard; SE, standard error.

Sensitivity analyses (i), (ii), and (iii) provided results consistent with the results of Tables 2a and 2b. The results are available in the Supplementary information.

Afterward, we studied the efficiency of these tools by contrasting the ESE obtained using the illness–death model against the values obtained with the CRK algorithm. As the latter depends heavily on the number of imputed datasets, the CRK algorithm was implemented using L = 5, 10, 20, 40, and 80. Figure 3 shows both the increase in precision and the decrease in execution time of the illness–death model as compared to the CRK algorithm depending on the number imputed datasets (L) used to implement the latter. The MI solution was computationally faster than the analytical solution if a reduced number of imputed datasets (L = 5 or L = 10) were used at the expense of a dismissed precision. Figure 3 also shows that, using the CRK algorithm, it is possible to obtain an efficiency that is practically identical to that obtained with the analytical solution. The price to pay to achieve this is a considerably high execution time.

Finally, we used the simulations to illustrate empirically the impact of the association between transitions I → E and I → D on the model estimates. We used the intensity functions mentioned in Section 4.1. This point was researched assuming "no ICE effect" for the placebo group and J2R for the experimental arm. Table 3 shows that point estimates presented bias that was identical with both models. We have discussed this topic in detail in Section 2.4.

5 | A REAL EXAMPLE

We will use this section to illustrate that the proposed framework has applications beyond the handling of treatment discontinuation through a post-hoc analysis of a study that compared two antiretroviral drugs.[48] From December 1990 through September 1991, 467 patients with CD4 ≤ 300 cells per cubic millimeter were randomized to either zalcitabine (experimental) or didanosine (active control). In this analysis, we aimed to quantify the net benefit that subjects in the experimental arm had on the overall survival by calculating the between-groups difference in RMSTs over the (0, 21 months] time interval. This statistic gives us an amount with a straightforward clinical interpretation as the lifetime that an average subject in the experimental arm has gained during the study period.[32,33]

Treatment discontinuation did not require using DA or RB imputation because, in general, subjects were well observed for overall survival during the entire follow-up period regardless of their adherence to the randomized

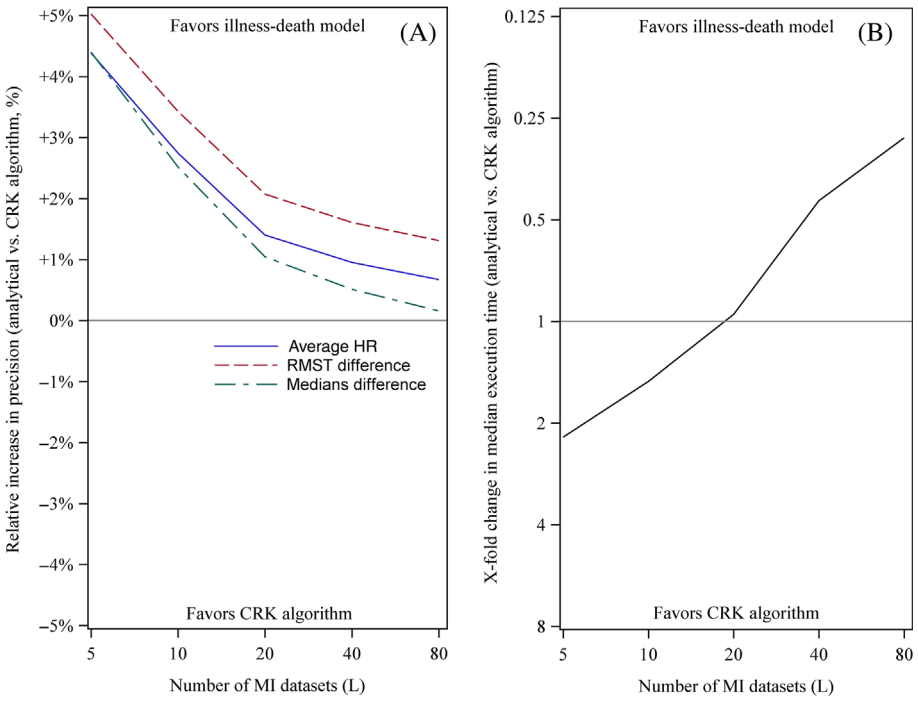

FIGURE 3 Relative increase in precision and execution time (illness–death model vs. CRK algorithm). Relative increase in precision = 100 × (ESE [CRK algorithm]^2/ESE (illness–death model)^2 – 1). X-fold change in execution time = MET (illness–death model)/MET (CRK algorithm). CRK, Carpenter, Roger, and Kenward; ESE, empirical standard error; HR, hazard ratio; MET, median execution time; MI, multiple imputation; RMST, restricted mean survival time.

TABLE 3 Simulation study results: impact of the association between transitions I → E and I → D.

	Model	True value	Mean bias	Mean SE	ESE	Coverage	Power
Difference of RMSTs (years) (experimental minus placebo) over the (0, 2] interval	Illness–death model	0.140	0.027	0.048	0.047	92%	94%
	CRK algorithm		0.027	0.066	0.048	98%	78%
Difference of medians (years) (experimental minus placebo)	Illness–death model	0.230	0.069	0.097	0.095	90%	89%
	CRK algorithm		0.068	0.143	0.096	99%	57%
Log-AHR (experimental vs. placebo) over the (0, 2] years interval	Illness–death model	0.140	−0.014	0.090	0.008	95%	93%
	CRK algorithm		−0.014	0.129	0.091	99%	77%

Note: The "no ICE effect" rule was used for the placebo rule while the J2R rule was used for the experimental arm. This table assumes that the modeled RB rule adequately represents the true unobserved risk.
Abbreviations: CRK, Carpenter, Roger, and Kenward; ESE: empirical standard error; MET: median execution time (seconds); RMST, restricted mean survival time; SE: standard error.

treatment. Actually, only 4 of 467 subjects were lost (and thus censored) for survival status during the study. However, in this dataset, we encountered another ICE influencing notably the interpretation of the RMST. As planned in the protocol, follow-up was abruptly stopped due to study termination in September of 1991. At that moment, follow-up periods varied notably across subjects. While the first subject recruited in the study had been followed up for 21 months, the last subject had been studied for only 12 months. A standard survival analysis, censoring at study termination, inherently provides expected lifetimes in an imaginary world where all subjects have been exposed to the studied therapies beyond study closure. This feature may be regarded acceptable, as clinical trial results are commonly extrapolated beyond the constraints of the study design. Some stakeholders, though, may find it more informative to consider an estimand that represents better what has truly happened to the study participants during the study. This option could be addressed using a treatment policy strategy for the ICE defined as the end of the follow-up due study closure. Subjects in this trial predominantly returned to didanosine and zidovudine, a standard of care therapy that we assumed to have an efficacy comparable to that seen during the study in the control arm. We have thus imputed events possibly occurring after this ICE using the "no ICE effect" rule for the control group and J2R for zalcitabine. This imputation strategy essentially assumes no between-arms risk difference in the period of time after study closure.

The model presented in Section 3 was fitted using the two strata, j = zalcitabine or j = didanosine, and three proportional-hazard baseline covariates: reason for inclusion in the study (intolerance or failure to zidovudine), previous infection (yes or no) and baseline CD4 level (after fourth root transformation). Natural cubic splines were used to model the baseline intensities of I → E and I → D. The estimated model parameters have been included as Supplementary information. Using these estimates and the algorithm proposed in Section 3, the survival and RMST curve over time was calculated for each combination of baseline values in the dataset. RMST estimates were marginalized using the plug-in estimator proposed by Freedman.[44] Accordingly, the RMST per arm (and its difference) was first estimated for each vector of baseline values in the database and, subsequently, all these amounts were averaged. The variance of the marginalized RMST was approximated as the average of the estimated (conditional) RMST variances plus the variance of the (conditional) point estimates.

Results can be found in Table 4 and Figures 4 and 5. Using the hypothetical strategy, the (marginalized) model estimates of the survival function (Figure 4B) are similar to the standard Kaplan–Meier estimates reported in the original paper (Figure 4A).[48] However, a treatment policy strategy with J2R imputation for discontinuation of the follow-up due to study closure (Figure 4C) leads to survival probabilities for the experimental arm that are higher than those obtained using the hypothetical strategy. This result is consistent with the between-group difference in RMSTs up to month 21 (Figure 5A and Table 4). While the hypothetical strategy provided a 0.74-month difference in favor of the experimental arm, with the treatment policy strategy and J2R imputation, we obtained a 0.86-month difference.

Differences between groups are not statistically significant so might be simply explained by sampling variability. However, it is still of interest to understand the phenomenon by which the effect size obtained using a treatment policy strategy and J2R imputation is greater than that obtained through standard survival model (i.e., using a hypothetical strategy). J2R is normally expected to imply a penalization on the treatment effect, but here the correction goes in the other direction. The answer to this dilemma can be found in Figure 4B. In this study, the experimental arm does better than the control arm (i.e., HR <1) during most of the study follow-up, but the HR reverses and becomes well >1 in the

TABLE 4 Real example: Estimated marginal RMST up to month 21.

Strategy to handle end of follow-up due to study closure	Statistic	Didanosine (N = 230)	Zalcitabine (N = 237)	Zalcitabine minus Didanosine
Hypothetical	Mean	15.20	15.93	0.74
	SE	3.92	3.58	0.71
	95% confidence interval	7.49 to 22.91	8.89 to 22.98	−0.67 to 2.14
	p-value			0.3030
Treatment policy (J2R)	Mean	15.20	16.06	0.86
	SE	3.92	3.55	0.71
	95% confidence interval	7.49 to 22.91	9.07 to 23.04	−0.54 to 2.25
	p-value			0.2265

Abbreviations: J2R, jump to reference; SE: standard error.

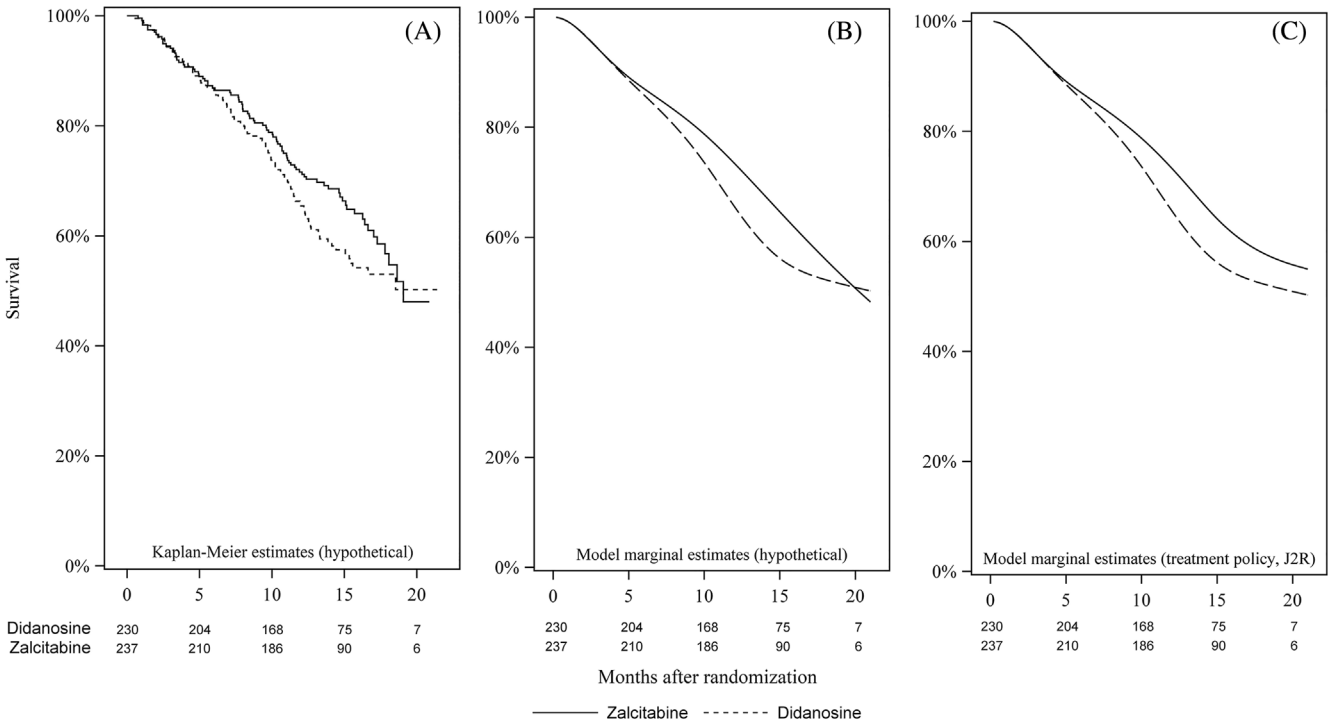

FIGURE 4 Real example: Survival function estimates. The strategy used to handle discontinuation of follow-up due to study termination is specified between brackets. J2R, jump to reference.

ending period. The hypothetical strategy targets an imaginary world where all subjects have continued treatment with the randomization therapies for 21 months. Using treatment policy strategy with J2R imputation we assume that, following study closure, subjects move to a sort of "HR = 1" status. Using J2R, the benefit we assume for the experimental arm after study closure, none, is still higher than the negative benefit we observe in the database at later times.

6 | ON THE SE ESTIMATION

The simulation study of Section 4 has evidenced that the analytical solution based on the illness–death model estimates adequately the SE. On the other hand, biased SE estimates have been found when J2R imputation was implemented using the CRK algorithm. This is a well-known feature of the CRK algorithm due to a lack of congeniality between the

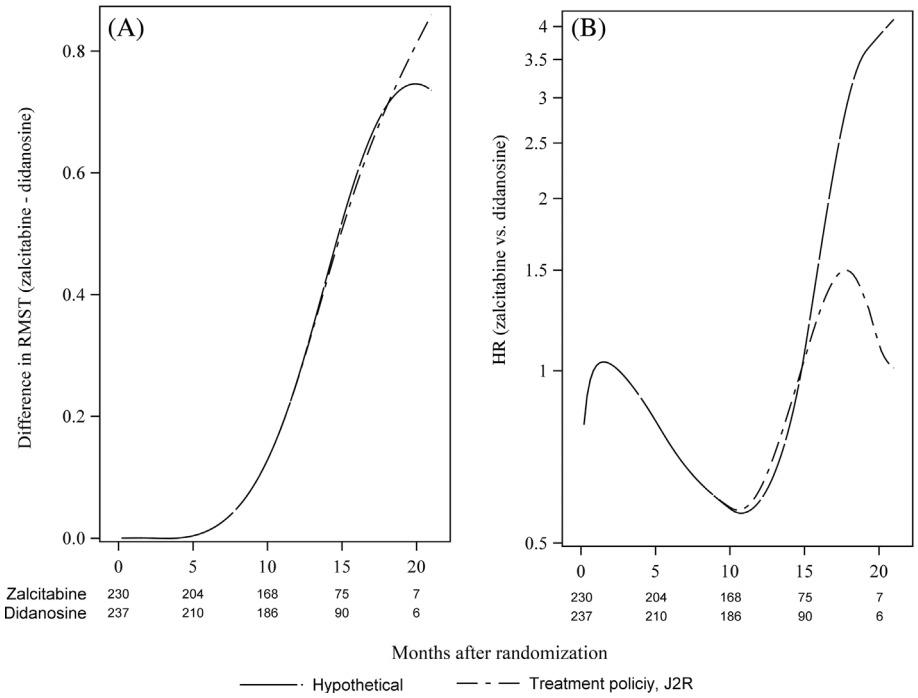

FIGURE 5 Real example: Estimated difference in RMST and HR. HR, hazard ratio; J2R, jump to reference; RMST, restricted mean survival time.

imputation and the analysis model.[18–20] We know that it is possible to avoid this bias by using bootstrapping.[49] This option has been proven to provide consistent confidence intervals with nominal coverage, even if the imputation and analysis models are uncongenial. The price to pay for this route, though, is its computational burden.[49]

We acknowledge that this topic remains controversial as sampling variability is constrained artificially when we use RB imputation. Some authors find it counter-intuitive that the more subjects transition to the ICE the smaller the SE is.[15,16,50] Following this line of thinking, Cro et al. have introduced a concept by which a certain model is said to be information-anchored (IA) if it holds the same proportion of information lost due to the ICE as that observed in the model used as reference (transition $I \rightarrow E$, in our case).[50] These authors have proven that the CRK algorithm, using Rubin's variance equation, provides approximately IA confidence intervals.[16,50]

Other authors, however, have provided arguments to use SEs that represent the empirical variability.[22,49] Using RB imputation, we introduce an external assumption to our estimator. Using J2R, for instance, the treatment effect disappears as soon as the treatment is stopped. The validity of our J2R estimate depends entirely on the correctness of the corresponding (unverifiable) assumption. If we implement an estimator following the IA concept, we essentially act as if the deterministic component of our estimator experiences sampling variability when, in reality, it does not. On this topic, Bartlett has reviewed in detail the arguments made by both sides of this debate and concluded that the SE estimators that target the sampling variance are more adequate.[49]

In this paper, we have primarily followed the latter line of thinking and proposed a variance estimator that represents the sampling variability. However, we note that it is possible to use the proposed analytical solution to build IA confidence intervals. To do this, we can use the algorithm proposed by Lu.[51] This methodology essentially makes use of the two following facts: (a) using DA, we obtain IA variance estimates[19,50]; and (b) it is possible to implement the DA method iteratively until we find a value $\delta_{exp} = \delta^*$ for which the estimated between-arm difference coincides with that obtained using RB imputation. Consequently, Lu proposes using DA with this very particular $\delta_{exp} = \delta^*$ to build IA confidence intervals for RB imputation rules and provides theoretical and empirical justifications for this approach.[51]

7 | DISCUSSION

The ICH E9 (R1) addendum, focused on estimands, came into effect in 2019.[1] This guideline has arguably become the most life-changing document released by regulatory authorities in decades. This document introduced a new concept,

the ICEs, and encouraged trialists to pre-define an approach to handle them. The guideline defined five possible strategies but did not firmly favor some over the others.[1] Since this document was first released for consultation in 2017, a number of therapeutic area (TA) related guidelines and publications have provided more precise recommendations on the strategies to be used to handle ICEs for some indications. For treatment discontinuation, an overall trend to favor the treatment policy and composite strategies over the hypothetical strategy is observed.[52–55]

The focus of this paper is the treatment policy strategy. Confronting this situation appears simple at first sight because we ought to ignore the ICE. A key point here is whether we have adequately followed up the study participants after the ICE. There is no unique answer to this question since it depends on the indication and factors like the study duration and how invasive the observation of the event of interest is. For example, it may be simple to follow up subjects after treatment discontinuation (e.g., via a phone call) to assess overall survival, but it may be more problematic for other endpoints if an invasive test is required. In this paper, we have focused on a situation whereby subjects are not (or rarely) followed up after the ICE. In this very particular setting, the ICE has two potential effects. The ICE may modify the response we intend to observe, and it is also the cause of not observing this (modified) response. This dual effect induces an association between the unobserved responses and the missingness mechanism.

DA and RB imputation allow us to speculate the form of the unobserved risk function after the ICE. DA uses fixed coefficients to modify the risk function after the ICE. The biggest challenge with applying DA is justifying the δ values. By using a TPA, we conduct a stress test on the drug's efficacy by repeating the DA method with several values of δ. J2R assumes that, after the ICE, both arms share the same hazard function (i.e., the observed risk in the reference arm). With CIR, the treatment effect achieved at the ICE onset is assumed to be maintained after the ICE.

In the survival field, DA and RB imputation have been studied so far using the MI-based CRK algorithm.[8–16] In this paper, we have proposed a fully analytical framework based on the illness–death multistate model. We have also proposed a specific parametric model to implement the proposed parameterization. The proposed model allows us to estimate (and test) statistics such as the between-arm difference of medians, RMSTs or log-transformed AHRs. With this model, we can also report the nonconstant HR function over time. This plot is helpful to understand the drug's effect as DA and RB imputation imply alterations of the risk function that can be incompatible with the constant HR assumption.

In what follows, we will focus on practicability and address the challenges end users encounter when implementing DA and RB imputation with time-to-event endpoints. Let us first focus on the solutions based on MI. This set of methodologies adheres to a divide-and-conquer principle, breaking down a complex problem into simpler steps. However, the imputation phase still presents complexities that warrant attention. Imputations should consider both population variability and uncertainty around model estimates. The latter can be accounted for by employing a Bayesian imputation model, which allows sampling from the posterior distribution of the parameters. This involves decisions related to the parameters' prior distribution, chain sampler, and thinning rate. These choices require careful consideration, as relying solely on software's default parameters may not always be optimal.[56] Some researchers have circumvented the need for fitting fully Bayesian models by utilizing the asymptotic equivalence between Bayesian and ML estimates, while others have captured parameter uncertainty through bootstrapping.[9,10,12,13,15,16] Another challenge for the imputation task is drawing post-ICE event time using the survival function conditional to the time the subject has been monitored without event. This often necessitates inverting the survival function, which might not be straightforward depending on the chosen imputation model. Once the missing data has been imputed, the remaining task is simple, as it merely involves analyzing the datasets using standard methods and pooling the estimates using Rubin's equation. We will now turn to the analytical solution proposed in this manuscript, which comes with its own set of challenges. The data modeling step (refer to Sections 2.2 and 3.1) is relatively straightforward and can be implemented using standard software. However, calculating the survival function of interest and subsequent quantities such as median, RMST, and AHR, requires significant computational effort as described in Sections 2.3 and 3.2. Given all these considerations, choosing between these two solutions based on practicality alone is challenging. As such, the choice might come down to a matter of preference. Like most of the works listed in Table 1, which are supplemented with either SAS® or R scripts for end-user implementation, this manuscript also provides a SAS® script in the Supplementary information.

We have studied empirically the performance of both the proposed model and the CRK algorithm through simulations. Both tools provided unbiased estimates of the effect size point estimates. It is important to emphasize that, using DA and RB imputation, unbiased estimates are obtained only if the imputation rule assumed by the modeler coincides with the true evolution of the response after the ICE. That assumption is, unfortunately, unverifiable if subjects have not been followed up after the ICE. The simulation study has also shown that the proposed ML estimator estimates adequately both point estimates and SEs. In contrast, we have observed biased SE estimates when RB imputation was

implemented using MI. Two more results from the simulations deserve attention. In the studied conditions, the illness–death model was slightly more efficient than the CRK algorithm. The efficiency of the MI-based model could be improved by using a large number of imputed datasets (L) at the expense of a significantly larger execution time.

Finally, our simulations study has also illustrated empirically that both the analytical and the MI solution assume independence between the event of interest and the occurrence of the ICE conditional on the included baseline covariates. Some authors have used the term "noninformative deviation" for this condition.[29] This prerequisite can also be seen as the CAR assumption behind the subprocess targeting the hypothetical strategy that we use as a base to build estimates for the treatment policy strategy estimand.[22] This important assumption may not have been emphasized sufficiently in the literature. DA and RB imputation have been said to address informative censoring, but this family of models actually deals only with the "response modification" effect of the ICE. DA and RB imputation still assume the ICE triggering censoring to be noninformative.

To end this manuscript, we recommend further research on the use of multistate models to analyze time-to-event endpoints that compete with ICEs since they can innately factor the distribution of the ICE process in our estimator. This work has primarily concentrated on situations in which a single ICE such as treatment discontinuation is to be addressed using the treatment policy strategy. Alternative parameterizations may be useful to manage other strategies or a hybrid combination thereof. In the survival field, multistate modeling appears to be a natural route for statisticians that tend to prefer fully analytical solutions over those based on MI.

ACKNOWLEDGMENTS

The authors would like to sincerely thank the editor and the anonymous referees for their meticulous review and valuable suggestions.

FUNDING INFORMATION

This work was partially supported by grant PID2019-104681RBI00 of the Spanish Ministry of Science and Innovation.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to report.

DATA AVAILABILITY STATEMENT

The data produced for the simulations task are available from the corresponding author upon request. The dataset analyzed in Section 5 is part of the JM R package.[57]

ORCID

Alberto García-Hernandez https://orcid.org/0000-0001-5265-8160

REFERENCES

1. International Council for Harmonisation (ICH) of technical requirements for pharmaceuticals for human use. ICH E9 (R1) Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Adopted on 20 November 2019 https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf [Last accessed 4-Dec-2021]
2. Carpenter JR, Roger JH. Kenward MG (2013) analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *J Biopharm Stat*. 2013;23(6):1352-1371. doi:10.1080/10543406.2013.834911
3. Mallinckrodt C, Linand Q, Molenberghs M. A structured framework for assessing sensitivity to missing data assumptions in longitudinal clinical trials. *Pharm Stat*. 2013;12:121-126. doi:10.1002/pst.1547
4. Mallinckrodt C, Molenberghs G, Rathmann S. Choosing estimands in clinical trials with missing data. *Pharm Stat*. 2017;16:29-36. doi:10.1002/pst.1765
5. Mallinckrodt C, Roger J, Chuang-stein C, et al. Missing data: turning guidance into action. *Statistics in Biopharmaceutical Research*. 2013;5(4):369-382. doi:10.1080/19466315.2013.848822
6. Kenward M. Controlled multiple imputation methods for sensitivity analyses in longitudinal clinical trials with dropout and protocol deviation. *Clin Invest (Lond)*. 2015;5(3):311-320.
7. Cro S, Morris T, Kenward M, Carpenter J. Reference-based sensitivity analysis via multiple imputation for longitudinal trials with protocol deviation. *Stata J*. 2016;16(2):443-463.
8. Keene ON, Roger JH, Hartley BF, Kenward MG. Missing data sensitivity analysis for recurrent event data using controlled imputation. *Pharm Stat*. 2014;13:258-264.
9. Zhao Y, Herring AH, Zhou H, Mirza AW, Koch GG. A multiple imputation method for sensitivity analysis of time-to-event data with possibly informative censoring. *J Biopharm Stat*. 2014;24(2):229-253.

10. Jackson D, White IR, Seaman S, Evans H, Baisley K, Carpenter J. Relaxing the independent censoring assumption in the cox proportional hazards model using multiple imputation. *Stat Med*. 2014;33(27):4681-4694. doi:10.1002/sim.6274
11. Lu K, Li D, Koch GG. Comparison between two controlled multiple imputation methods for sensitivity analyses of time-to-event data with possibly informative censoring. *Stat Biopharm Res*. 2015;7:199-213.
12. Akachaa M, Ogundimub EO. Sensitivity analyses for partially observed recurrent event data. *Pharm Stat*. 2016;15:4-14.
13. Lipkovich I, Ratitch B, O'Kelly M. Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints. *Pharm Stat*. 2016;15:216-229.
14. Gao F, Liu GF, Zeng D, et al. Control-based imputation for sensitivity analyses in informative censoring for recurrent event data. *Pharm Stat*. 2017;16(6):424-432.
15. Atkinson A, Kenward MG, Clayton T, Carpenter JR. Reference-based sensitivity analysis for time-to-event data. *Pharm Stat*. 2019;18:645-658.
16. Atkinson A, Cro S, Carpenter JR, Kenward MG. Information anchored reference-based sensitivity analysis for truncated normal data with application to survival analysis. *Stat Neerl*. 2021;1–24:500-523.
17. Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazards models. *Stat Med*. 2005;24(11):1713-1723. doi:10.1002/sim.2059
18. Lu K. An analytic method for the placebo-based pattern-mixture model. *Stat Med*. 2014;33:1134-1145. doi:10.1002/sim.6008
19. Tang Y. Short notes on maximum likelihood inference for control-based pattern-mixture models. *Pharm Stat*. 2015;14:395-399.
20. Liu GF, Pang L. On analysis of longitudinal clinical trials with missing data using reference-based imputation. *J Biopharm Stat*. 2016;26:924-936.
21. Lu K. An extension of the placebo-based pattern-mixture model. *Pharm Stat*. 2014;13:103-109. doi:10.1002/pst.1605
22. García-Hernandez A, Pérez T, Pardo MC, Rizopoulos D. A flexible analytical framework for reference-based imputation, delta adjustment and tipping-point stress-testing. *Statistics in Biopharmaceutical Research*. 2023;1-13. doi:10.1080/19466315.2022.2151506
23. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statist Med*. 2007;26:2389-2430.
24. Flyer P, Hirman J. Missing data in confirmatory clinical trials. *J Biopharm Stat*. 2009;19(6):969-979. doi:10.1080/10543400903242746 PMID: 20183459.
25. Klein JP, Keiding N, Copelan EA. Plotting summary predictions in multistate survival models—probabilities of relapse and death in remission for bone-marrow transplantation patients. *Stat Med*. 1994;12:2315-2332.
26. Arjas E, Eerola M. On predictive causality in longitudinal studies. *J Stat Plan Infer*. 1993;34:361-384.
27. Borgan Ø. Aalen–Johansen Estimator. *Stat Ref*. 2014. doi:10.1002/9781118445112.stat06001.pub2
28. Tsiatis AA. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci USA*. 1975;72:20-22.
29. Seaman SR, White IR, Leacy FP. Comment on 'analysis of longitudinal trials with protocol deviations: a framework for relevant, accessible assumptions, and inference via multiple imputation', by Carpenter, Roger, and Kenward. *J Biopharm Stat*. 2014;24:1358-1362.
30. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30:2409-2421.
31. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13:152. doi:10.1186/1471-2288-13-152
32. Kloecker DE, Davies MJ, Khunti K, Zaccardi F. Uses and limitations of the restricted mean survival time: illustrative examples from cardiovascular outcomes and mortality trials in type 2 diabetes. *Ann Intern Med*. 2020;172(8):541-552. doi:10.7326/M19-3286 Epub 2020 Mar 24. PMID: 32203984.
33. Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. *JAMA Cardiol*. 2017;2(11):1179-1180. doi:10.1001/jamacardio.2017.2922 PMID: 28877311; PMCID: PMC6359932.
34. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. *Biometrika*. 1981;68(1):105-112.
35. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted cox regression. *Stat Med*. 2009;28(19):2473-2489. doi:10.1002/sim.3623 PMID: 19472308.
36. Crowther MJ, Lambert PC. Parametric multi-state survival models: flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Stat Med*. 2017;36:1-24.
37. Royston P, Parmar MKB. Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21:2175-2197.
38. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: Data mining, inference and prediction*. 2nd ed. Springer-Verlag, New York; 2009. doi:10.1007/978-0-387-84858-7
39. Ng R, Kornas K, Sutradhar R, Wodchis WP, Rosella LC. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. *Diagn Progn Res*. 2018;2:4. doi:10.1186/s41512-018-0026-5
40. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statist Med*. 1989;8:551-561. doi:10.1002/sim.4780080504
41. Rutherford M, Crowther M, Lambert P. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J Stat Comput Sim*. 2015;85(4):777-793. doi:10.1080/00949655.2013.845890
42. Bower H, Crowther M, Rutherford M, et al. Capturing simple and complex time-dependent effects using flexible parametric survival models: a simulation study. *Commun Stat-Simul C*. 2021;50(11):3777-3793. doi:10.1080/03610918.2019.1634201
43. Davis P, Rabinowitz P. Automatic Integration. *Methods of Numerical Integration*. Dover Publications; 1984.
44. Freedman DA. Randomization does not justify logistic regression. *Statist Sci*. 2008;23(2):237-249.

45. Hinchliffe SR, Scott DA, Lambert PC. Flexible parametric illness-death models. *Stat J*. 2013;13(4):759-775.
46. Matsuyama Y, Yamaguchi T. Estimation of the marginal survival time in the presence of dependent competing risks using inverse probability of censoring weighted (IPCW) methods. *Pharmaceut Statist*. 2008;7:202-214. doi:10.1002/pst.290
47. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074-2102. doi:10.1002/sim.8086
48. Abrams DI, Goldman AI, Launer C, et al. A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human-immunodeficiency-virus infection. *N Engl J Med*. 1994;330:657-662. doi:10.1056/nejm199403103301001
49. Bartlett JW. Reference-based multiple imputation—what is the right variance and how to estimate it. *Stat Biopharm*. 2021;15:178-186. doi:10.1080/19466315.2021.1983455
50. Cro S, Carpenter JR, Kenward MG. Information anchored sensitivity analysis. *J R Stat Soc Ser A*. 2018;82(2):623-645.
51. Lu K. Alternative implementation of reference-based controlled imputation procedures. *Stat Biopharm*. 2021;13(4):483-491. doi:10.1080/19466315.2020.1796781
52. Committee for Medicinal Products for Human Use (CHMP). Guideline on the clinical investigation of medicines for the treatment of Alzheimer's disease. 2018 CPMP/EWP/553/95 Rev.2.
53. Committee for Medicinal Products for Human Use (CHMP). Guideline on the development of new medicinal products for the treatment of Ulcerative Colitis. 2018 CHMP/EWP/18463/2006 Rev.1.
54. Committee for Medicinal Products for human use (CHMP). Guideline on the development of new medicinal products for the treatment of Crohn's disease. 2018 CPMP/EWP/2284/99 Rev 2.
55. Committee for Medicinal Products for human use (CHMP). Draft guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus. 2018 CPMP/EWP/1080/00 Rev 2.
56. Wang J, Johnson DE. An examination of discrepancies in multiple imputation procedures between SAS® and SPSS®. *Am Stat*. 2018;73:80-88. doi:10.1080/00031305.2018.1437078
57. Rizopoulos DJM. An R package for the joint modelling of longitudinal and time-to-event data. *J Stat Softw*. 2010;35(9):1-33. doi:10.18637/jss.v035.i09

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: García-Hernandez A, Pérez T, del Carmen Pardo M, Rizopoulos D. An illness–death multistate model to implement delta adjustment and reference-based imputation with time-to-event endpoints. *Pharmaceutical Statistics*. 2023;1-23. doi:10.1002/pst.2348