Statistics in Biosciences
https://doi.org/10.1007/s12561-021-09317-3

# Estimation of Discrete Survival Function through Modeling Diagnostic Accuracy for Mismeasured Outcome Data

Hee-Koung Joeng[1] · Abidemi K. Adeniji[2] · Naitee Ting[3] · Ming-Hui Chen[1]

Received: 30 March 2020 / Revised: 21 June 2021 / Accepted: 27 June 2021

## Abstract

Standard survival methods are inappropriate for mismeasured outcomes. Previous research has shown that outcome misclassification can bias estimation of the survival function. We develop methods to accurately estimate the survival function when the diagnostic tool used to measure the outcome of disease is not perfectly sensitive and specific. Since the diagnostic tool used to measure disease outcome is not the gold standard, the true or error-free outcomes are latent, they cannot be observed. Our method uses the negative predictive value (NPV) and the positive predictive values (PPV) of the diagnostic tool to construct a bridge between the error-prone outcomes and the true outcomes. We formulate an exact relationship between the true (latent) survival function and the observed (error-prone) survival function as a formulation of time-varying NPV and PPV. We specify models for the NPV and PPV that depend only on parameters that can be easily estimated from a fraction of the observed data. Furthermore, we conduct an in-depth study to accurately estimate the latent survival function based on the assumption that the biology that underlies the disease process follows a gamma process. We examine the performance of our method by applying it to the Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (VIRAHEP-C) data. To show the broader relevance of our research, we apply our proposed methodology to a dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

**Keywords** Detection limit · Misclassification · Negative predictive value (NPV) · Positive predictive values (PPV) · Hepatitis C virus data

✉ Abidemi K. Adeniji
adeniji@m-estimator.com

[1] Department of Statistics, University of Connecticut, Storrs, CT, USA

[2] M-Estimator, LLC, Mansfield, TX, USA

[3] Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT, USA

Published online: 03 September 2021

## 1 Introduction

A centerpiece in the practice of medicine requires accuracy in the diagnosis of disease. However, in diagnostic testing patient outcomes could be misclassified due to false-negative or false-positive errors. In the diagnosis of Alzheimer's disease(AD), misclassification can occur between imaging results and autopsy [1–4]. In the assessment of viral negativity for hepatitis C virus (HCV), false-negative (or false-positive) results can be observed between assays with different limits of detection [5–8]. In oncology, discordance between the investigator and independent central review assessments can occur in the determination of progression free survival [9–12]. In microbiology, testing for helicobacter pylori infection can be performed with an error-prone urea breathe test or by the gold-standard biopsy examination; the urea breathe test can lead to incorrect conclusions [13–17].

Our work is motivated by two major and distinct clinical studies: (i) the VIRAHEP-C clinical trial and (ii) the ADNI. We first discuss the VIRAHEP-C clinical trial, this study had participants assessed for viral negativity at preset clinical visits. The gold-standard assay has a limit of detection of 50IU/ml while the error-prone assay has a limit of detection of 600 IU/ml. Viral negativity is defined as the point in time when the viral load falls below the limit of detection. The time to viral negativity is intrinsically discrete because it is observed only at predetermined clinical visits and there is no information on viral negativity between any two clinical visits. In the introductory section of Joeng et al. [18] a thorough discussion of "intrinsically discrete" survival time is discussed. Both assays are accurate in regard to their respective detection limit, however, the clinical outcomes of the gold-standard assay are expected to reflect the true underlying nature of disease. The discordance between the event times of the gold-standard (G-S) assay and the error-prone (E-P) assay is one of the two motivators of this research. The methodology we propose will construct a bridge between the G-S and E-P outcomes through time-varying PPV.

The other motivator of this research comes to us from the ADNI study. The ADNI study aims to improve clinical trials for the prevention and treatment of AD. The E-P examination for the detection of Alzheimer's is performed through clinical assessment. The amyloid beta ($A\beta$) protein biomarker from a cerebral spinal fluid (CSF $A\beta$42 or CSF for simplicity) assay has been shown to represent the pathological aspects of AD well and the abnormality of $A\beta$ can be used as a reliable (true, G-S) endpoint for studying time to pathological diagnosis of AD among living participants [19]. CSF assays were performed and $A\beta$ protein concentrations were measured. Participants with an $A\beta$ biomarker value greater than 192 pg/ml were classified as non-AD at baseline and those with an $A\beta$ value less than or equal to 192 pg/ml were classified as AD at baseline. The CSF biomarker assay involves a lumbar puncture, so it is often considered too invasive for many patients and therefore has limited availability. A subset of participants also had longitudinal (annual) CSF assays to measure $A\beta$ values, from which time to CSF diagnoses could be determined. The best single measure for discriminating between AD and control patients was CSF $A\beta42$ [20]. The discordance between the event times of the G-S CSF biomarker assay and the E-P clinical assessment is the second motivator of this research. Here, (and in contrast to the methodology we propose for the VIRAHEP-C data), we construct a bridge between the G-S and E-P outcomes through time-varying NPV.

In this paper, we develop two new estimators of the survival function that handle misclassified outcomes by incorporating the PPV(NPV) of the diagnostic procedure. In the presence of misclassified outcomes, the Kaplan–Meier estimator may lead to biased estimates of the survival rate of true outcomes [21–23]. The issue of estimating the survival function from misclassified outcomes has been studied, Racine-Poon and Hoel [22] derived a non-parametric estimation of the survival function for which the cause of death was uncertain. Richardson and Hughes [24] used the Expectation Maximization algorithm to obtain unbiased estimates of the conditional probability of disease. McKeown and Jewell [25] extended the non-parametric maximum likelihood estimator to allow for time-dependent misclassification rates. In contrast to other studies which used sensitivity and specificity as inputs in their models, we prefer the PPV because it quantifies the value of a test to clinicians.

In Sect. 2, we propose a model that is a bridge between the outcomes of the E-P assay and the G-S assay through a time-dependent PPV. The prevalence of disease can considerably influence PPV; hence, we derive an estimator of the survival function within a non-constant PPV framework. The gamma distribution is a common building block of stochastic epidemiologic models used to study the biology of disease stages (latent and infectious) of infectious diseases [26–34]. We therefore examine the properties of the proposed model (4) under stochastic processes in Sect. 3. First, we assume that viral load of HCV over time follows a gamma process, under which we established a novel iterative identity to derive a closed-form (deterministic) expression for the survival function to calculate the true PPV of the diagnostic assay. The gamma process is attractive since PPV is non-constant over time and it also allows us to examine the properties of the proposed model (4) analytically. Another contribution of the paper is the ability to conduct inferential statistical analyses which makes it possible to examine the effect of an unequal proportion of misclassified outcomes between treatment groups. The derivation of the variance of our proposed estimator (Sect. 2.2.2) required new techniques since the delta method may not be applicable due to the variation in the parameters of our model under small sample sizes. Furthermore, we model the course of HCV by a Wiener process. In contrast to the gamma process, the closed-form expression of the survival function is not available and our methods are evaluated empirically through simulation studies. We then examine the performance of our methodology by applying it to data from the VIRAHEP-C study.

The previous two paragraphs deal with the first estimator of the survival function which incorporates a time-varying PPV of the diagnostic procedure with application to the VIRAHEP-C study. However, we believe our methodology has broad applicability in the realm of misclassified outcomes such that by modeling the NPV (instead of the PPV) of the diagnostic tool, we obtain the other estimator of the survival function that can be applied to a very different therapeutic area, namely, Alzheimer's disease(AD). Our research is applicable to the lower limit and upper limit of detection framework. By modeling the time-varying NPV of the diagnostic tool in a similar fashion as done with the time-varying PPV, we construct the second survival rate estimator for application to the misclassified outcomes as observed in the ADNI data. We believe our two survival rate estimators broadly cover the problem of correctly estimating the survival function in the presence of disagreements between outcomes of the gold-standard and error-prone diagnostic procedures.

The rest of the paper is organized as follows. In Sect. 2, we provide a detailed development of the exact relationship between survival functions of G-S and E-P events, and their proposed models. We examine various properties of the proposed models, and provide inference procedure and implementation procedures. An extensive study under stochastic processes is carried out in Sect. 3. Particularly, we conduct in-depth simulation studies to evaluate the operating characteristics of our proposed procedures. Namely, Sect. 3.2 models the viral load course via a standard Brownian motion (Wiener) process, and evaluates the performance of the methods detailed in Sect. 2.2.2 (Inference for Unknown Model Parameters). Unlike the gamma process, the Wiener process does not have a closed-form formulation of the G-S survival function. As a result, the Wiener process provides a framework to evaluate the properties of our proposed estimator through simulation studies. Sect. 4 presents detailed analyses of clinical trial data, the analyses of the VIRAHEP-C data and ADNI datasets are given in Sects. 4.1 and 4.2, respectively. We conclude the paper with some discussion in Sect. 5. We provide proofs of theorems, lemmas, and propositions in Appendix A, as well as the extended results under the upper-limit detection problem paradigm process in Appendix B. The R codes to run our methodology are provided in the Supplementary Materials.

## 2 The Methods

### 2.1 The Hazards for Mismeasured and True Discrete Survival Times

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Wiener, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment(MCI) and early AD.

The design of this study builds on the notion that there exist three populations of events, (i) population of misclassified or E-P events, (ii) population of G-S events, and (iii) population of the underlying absolute state (true) of events, or the ground truth; this population is latent and G-S aims to represent the ground truth as closely as possible. As a result of the use of an E-P diagnostic test or procedure, the events in (i) are subject to be observed with error; while the events in (ii) are assumed to be error free, the use of an E-P diagnostic test renders the G-S status of these events to be unobservable. The rest of this section provides notation for the hazards of E-P and G-S events as well as their respective survival functions.

Define $T$ to be a discrete random variable representing the survival time for the E-P diagnostic test, it takes only positive values $t_0(=0) < t_1 < t_2 < \cdots$. The discrete-time hazard function for $T$ at time $t_j$ is defined as

$$h(j) = P(T = t_j | T \geq t_j), \text{ for } j = 1, 2, \cdots \tag{1}$$

Using (1), we have the probability of $T$ at time $t_j$ as

$$P(T = t_j) = h(j) \prod_{k=1}^{j-1} \{1 - h(k)\},$$

and the survival function of E-P events as

$$S(j) = P(T > t_j) = \prod_{k=1}^{j} \{1 - h(k)\},$$

for $j = 1, 2, \ldots$. Next, we discuss the basic formulation of the hazard and survival functions of G-S failure time.

Let $T^*$ be a G-S discrete random variable taking positive values $t_0(=0) < t_1 < \ldots$. The discrete hazard function of G-S events at time $t_j$ for $T^*$ is defined as

$$h^*(j) = P(T^* = t_j | T^* \geq t_j),$$

for $j = 1, 2, \ldots$. In a similar fashion to survival function of E-P events ($S(j)$), the probability of $T^*$ at time $t_j$ is

$$P(T^* = t_j) = h^*(j) \prod_{k=1}^{j-1} \{1 - h^*(k)\},$$

therefore, the survival function of G-S events is

$$S^*(j) = P(T^* > t_j) = \prod_{k=1}^{j} \{1 - h^*(k)\}.$$

$E_j = I(T = t_j)$ for $j = 1, 2, \ldots$ denote the error-prone population and $E_j^* = I(T^* = t_j)$ for $j = 1, 2, \ldots$, denote the true population. Under this notation, the probability of E-P events at time $t_j$ is $P(T = t_j) = P(E_j = 1)$ while the probability of G-S events at time $t_j$ is $P(T^* = t_j) = P(E_j^* = 1)$. In our framework, once an individual is observed to have an event the individual is no longer followed, therefore, if $E_j = 0$ then $E_k = 0$ for $k < j$. At time $t_j$, define $\gamma_j$ as the negative predicted value (NPV); $\tau_j$ as the positive predicted value (PPV). Their expressions are as follows:

$$\gamma_j = P(T^* > t_j | T > t_j); \quad \text{and} \quad \tau_j = P(T^* \leq t_j | T \leq t_j), \tag{2}$$

for $j = 1, 2, \ldots$.

***Remark 2.1*** Under our notation we have the following relationships at time $t_j$,

(i) $S(j) = P(E_j = 0)$; and $S^*(j) = P(E_j^* = 0)$ are the probabilities of no event (survival);

(ii) $P(T = t_j) = P(E_j = 1)$; and $P(T^* = t_j) = P(E_j^* = 1)$ are the probabilities of an event;

(iii) $\gamma_j = P(E_j^* = 0|E_j = 0)$ is the NPV; and $\tau_j = 1 - \frac{\sum_{k=1}^{j} P(E_j^*=0|E_k=1)P(E_k=1)}{1-P(E_j=0)}$ is the PPV.

### 2.2 Assumptions and Proposed Methods

The main goal of our paper is to develop a link between the G-S and E-P populations of events which will allow the accurate estimation of the survival function of G-S events. We formulate an exact relationship between the survival functions of G-S and E-P events by using the NPV and PPV of the diagnostic tool, Lemma 2.1 provides the details.

**Lemma 2.1** *The survival function of G-S events at time $t_j$ can be expressed as*

$$S^*(j) = (1 - \tau_j)\{1 - S(j)\} + \gamma_j S(j). \tag{3}$$

If the G-S events are observable, the above formula provides the exact relationship between the survival functions of E-P ($S(j)$) and G-S ($S^*(j)$) events. But the G-S outcomes are unobserved, thus, we develop a new method to estimate the survival function of G-S outcomes by using Lemma 2.1 and modeling NPV and PPV. A method to estimate the survival function of G-S events using a constant PPV was proposed in Adeniji et al. [21]. We will extend their method by allowing a time-varying PPV and the following assumption.

**Assumption 1** A G-S event does not happen before an E-P event. That is,

$$P(T^* \geq T) = 1.$$

Assumption 1 is especially reasonable for mismeasurement caused by lower detection limit, we will further elaborate with the application of our methods to the VIRAHEP-C data.

**Proposition 2.1** *Under Assumption 1, we have*

(i) *the NPV at time $t_j$ is $\gamma_j = 1$ for all $j = 1, 2, \ldots$; and*

(ii) *the PPV at time $t_j$ is $\tau_j = \frac{1-P(E_j^*=0)}{1-P(E_j=0)}, j = 1, 2, \ldots$.*

Lemma 2.1 shows the importance of obtaining accurate measures of $\tau_j$ and $\gamma_j$ in order to obtain an unbiased estimate of the survival function of G-S events. We derive the model of time-varying PPV by modeling the conditional probabilities of G-S events given E-P events. For a known $\tau_0$ and for $t_k \leq t_j$, we propose that the probability of the occurrence of a G-S event at time $t_j$, given an E-P event at a specified time $t_k$ is

$$P(T^* \leq t_j | T = t_k) = 1 - \left\{1 - \tau_0\right\}^{(t_j - t_1)\omega_1 + (t_j - t_k)\omega_2 + 1}, \tag{4}$$

where $\omega_1 \geq 0$, $\omega_2 \geq 0$, and $0 < \tau_0 \leq 1$. The proposed model implies that given the prior occurrence of an E-P event, the probability of correctly classifying the G-S event increases with time. The following proposition presents properties of our proposed model (4).

**Proposition 2.2** *Under our proposed model* (4) *and a known* $\tau_0$ *we have the following:*

(i) *if* $P(E_1 = 1) > 0$, $\tau_0$ (i.e., $P(E_1^* = 1 | E_1 = 1)$) *is the conditional probability of the G-S failure time at* $t_1$ *given E-P failure time at* $t_1$;

(ii) *for a fixed* $\tau_0$ *with* $0 < \tau_0 < 1$, *the logarithm of the G-S rate at* $t_j$, *given observed event at* $t_j$, *is proportional to* $\omega_1(t_j - t_1)$ *as*

$$\log\{P(T^* > t_j | T = t_j)\} = \omega_1(t_j - t_1)\log(1 - \tau_0);$$

(iii) *for a fixed* $\tau_0$ *with* $0 < \tau_0 < 1$, *the difference between the logarithm of the G-S rate at* $t_j$, *given observed event at* $t_j$ *and given observed event at* $t_k$ *is proportional to* $\omega_2(t_j - t_k)$ *as*

$$\log\left\{\frac{P(T^* > t_j | T = t_j)}{P(T^* > t_j | T = t_k)}\right\} = \omega_2(t_j - t_k)\log(1 - \tau_0);$$

(iv) *the PPV at time* $t_j$ *is written as* $1 - \frac{\sum_{k=1}^{j} P(E_k=1)(1-\tau_0)^{(t_j-t_1)\omega_1+(t_j-t_k)w_2+1}}{1-P(E_j=0)}$;

(v) *if* $\omega_2 = 0$, *the survival rate of G-S event at* $t_j$ *given observed event at* $t_k$ *is constant for all observed events at* $t_k$ ($t_k \leq t_j$) *as* $\left\{1 - \tau_0\right\}^{(t_j - t_1)\omega_1 + 1}$; *and*

(vi) *if* $\omega_1 = 0$ *and* $\omega_2 = 0$, *the G-S survival rate at* $t_j$, *given observed event at* $t_k$ *is constant as* $1 - \tau_0$ *for all* $t_k$ ($t_k \leq t_j$) *and for any* $t_j$, *which shows that the G-S survival function is a cure rate function.*

The first three points of Proposition 2.2 explain and interpret the parameters of our model given in (4) in terms of probability of events and PPV, while (iv) through (vi) of Proposition 2.2 provide derivations of PPV under our proposed model in relation to the parameters from (i)-(iii).

We now advance toward the goal of obtaining the survival function of G-S (unobserved) outcomes as formulations of the survival function of E-P (error-prone) outcomes, NPV and PPV. Under Assumption 1, the formula below provides a way to obtain the survival function of G-S outcomes ($S^*(j)$) from the survival function of E-P outcomes ($S(j)$).

**Theorem 2.1** *Under Assumption* 1 *and the proposed model* (4), *the survival function of the G-S event time is given by*

$$S^*(j) = S(j) + \sum_{k=1}^{j} P(E_k = 1)\{1 - \tau_0\}^{(t_j - t_1)\omega_1 + (t_j - t_k)\omega_2 + 1}, \; j = 1, 2, \dots, \tag{5}$$

*where* $\omega_1 \geq 0, \omega_2 \geq 0$, and $0 \leq \tau_0 \leq 1$.

The proof of Theorem 2.1 directly follows from Lemma 2.1, (i) of Proposition 2.1, and (ii) of Proposition 2.2. The $S^*(j)$ in Theorem 2.1 is non-increasing in $j$ as $S^*(j-1) - S^*(j) = P(E_j = 1)\left\{1 - (1 - \tau_0)^{(t_j - t_1)\omega_1 + 1}\right\}(\geq 0)$. Within the survival framework, we wish to eventually express Theorem 2.1 in terms of survival rates.

The probability of E-P events at time $t_j$, $P(E_j = 1)$, can be expressed with probabilities of non-events as

$$P(E_j = 1) = P(E_{j-1} = 0) - P(E_j = 0), \text{ for } j = 1, 2, \dots.$$

The next step is to express Theorem 2.1 in regard to PPV and the survival vector of E-P events. Let $S^*(j)(S(j))$ denote the survival rate of the G-S (E-P) events at time $t_j$, for $j \in \{1, 2, \dots, J\}$. Also, define the error-prone survival vector, $\mathbb{S}_0 = (S(1), S(2), \dots, S(J))^T$, for time points 1 to $J$. The following Eq. (6) obtains our goal, it expresses Theorem 2.1 as a function of PPV and the error-prone survival vector $\mathbb{S}_0$ as follows:

$$S^*(j) = f_j(\mathbb{P}) + \left\{g_j(\mathbb{P})\right\}^T \mathbb{S}_0, \tag{6}$$

where $\mathbb{P} = (\omega_1, \omega_2, \tau_0)$, $f_j(\mathbb{P}) = (1 - \tau_0)^{(t_j - t_1)\omega_1 + (t_j - t_1)\omega_2 + 1}$, and

$$g_j(\mathbb{P}) = (g_{j1}(\mathbb{P}), g_{j2}(\mathbb{P}), \dots, g_{jJ}(\mathbb{P}))^T.$$

Each $g_{jk}(\mathbb{P})$ in $g_j(\mathbb{P})$ is defined as

$$\begin{cases} \{1 - (1 - \tau_0)^{(t_{k+1} - t_k)\omega_2}\}(1 - \tau_0)^{(t_j - t_1)\omega_1 + (t_j - t_{k+1})\omega_2 + 1} & \text{for } k = 1, \dots, j - 1 \\ 1 - (1 - \tau_0)^{(t_j - t_1)\omega_1 + 1} & \text{for } k = j \\ 0 & \text{for } k > j, \end{cases}$$

for $\omega_1 \geq 0, \omega_2 \geq 0$, and $0 \leq \tau_0 \leq 1$.

Situations exist for which inaccurate assessment of disease status is not the result of an ambiguous diagnosis or the lack of medical expertise, but rather, due to a lower-limit detection of the diagnostic procedure. For instance, if the analytical lower limit is below the detection of the assay, the outcomes may be misclassified. Although we develop methodology for the broad problem of estimating the survival function of G-S events from mismeasured outcomes, we focus our data analysis and simulation studies for mismeasured outcomes that originate from lower-limit detection. Note that within the framework of lower-limit detection, and as discussed in Proposition 2.1, the NPV ($\gamma_j$) at time $t_j$ equals 1 for all time points.

Thus far, we have worked under the framework that the timing of the observance of the G-S event does not occur prior to the observance of the E-P event. However, in the ADNI study, the G-S event almost exclusively occurs prior to, or by the time of the E-P event. As such, we develop a second assumption and proposition based on the second assumption.

**Assumption 2** A G-S event does not happen after an E-P event. That is,

$$P(T^* \leq T) = 1.$$

**Proposition 2.3** *Under Assumption 2, we have*

(i) *the PPV at time $t_j$ is $\tau_j = 1$ for all $j = 1, 2, \ldots$; and*
(ii) *the NPV at time $t_j$ is $\gamma_j = \frac{P(E_j^*=0)}{P(E_j=0)}, j = 1, 2, \ldots$.*

Under Assumption 2, a proposed model for NPV ($\gamma_j$) at time $t_j$ is

$$P(T^* > t_j | T > t_j) = \gamma_0^{(t_j-t_1)^2\psi_1+(t_j-t_1)\psi_2+1}, \tag{7}$$

where $0 < \gamma_0 \leq 1$ and $\psi_1, \psi_2 \geq 0$.

**Theorem 2.2** *Under Assumption 2 and the proposed model (7), the survival function of G-S outcomes is obtained as*

$$S^*(j) = \gamma_0^{(t_j-t_1)^2\psi_1+(t_j-t_1)\psi_2+1} S(j) \text{ for } j = 1,\ 2, \ldots, \tag{8}$$

*where* $0 < \gamma_0 \leq 1$ and $\psi_1, \psi_2 \geq 0$.

By setting $\mathbb{P} = (\psi_1, \psi_2, \gamma_0)$, $f_j(\mathbb{P}) = 0$ and $g_j(\mathbb{P}) = (g_{j1}(\mathbb{P}), g_{j2}(\mathbb{P}), \ldots, g_{jJ}(\mathbb{P}))^T$ with each $g_{jk}(\mathbb{P}) = \gamma_0^{(t_j-t_1)^2\psi_1+(t_j-t_1)\psi_2+1}$ for $k = j$ and $g_{jk}(\mathbb{P}) = 0$, the model in (8) is expressed in the similar fashion to (6)—expressed as a function of NPV and the error-prone survival vector $\mathbb{S}_0$.

Thus far, we have presented two main results, Theorems 2.1 and 2.2, which pertain to Assumptions 1 and 2, respectively. Specifically, Theorem 2.1 deals with the case where the G-S event does not happen before an E-P event (Assumption 1). Theorem 2.2 addresses the scenario for which the G-S event does not happen after an E-P event (Assumption 2). In Sect. 2.2.1, we present methodology for implementing Theorems 2.1 and 2.2 when the model parameters are known. Likewise, Sect. 2.2.2 provides inference under a framework where the model parameters are unknown and need to be estimated.

### 2.2.1 Inference for Known Model Parameters

As mentioned in the preceding section, there are two options for data analysis, we discuss the case where model parameters are known. We do not estimate any parameters from the clinical study. The three parameters, $\mathbb{P} = (\omega_1, \omega_2, \tau_0)$ in (6) under Theorem 2.1 ($\mathbb{P} = (\psi_1, \psi_2, \gamma_0)$ under Theorem 2.2) are acquired from medical experts. The data analyst in collaboration with medical personnel may obtain three parameters from previous clinical studies or literature. At first, we discuss parameters, $\mathbb{P} = (\omega_1, \omega_2, \tau_0)$. These estimates are assumed to be known with confidence prior to the conduct of the clinical study, for example, Adeniji et al. [21] assumed $\tau_0$ was known before the beginning of the clinical trial. The measure for which the probability of misclassification changes over time is $\omega_1$, while $\omega_2$ is the measure for which the probability of misclassification changes over time after the occurrence of an E-P event.

Since we assume that $\omega_1$, $\omega_2$ and $\tau_0$ are known and fixed prior to the start of the clinical study, the variance–covariance formula for the variance of (6) is not very complex, this is because the variability of $\omega_1$, $\omega_2$ and $\tau_0$ will be excluded from the variance–covariance matrix of $S^*$ in (6). The elements of $\mathbb{S}_0$ can be estimated by the product limit estimator in Kaplan and Meier [35], which is also called the Kaplan–Meier (KM) estimator. We let $S^*(j)(S(j))$ denote the survival rate of the G-S (E-P) events at time $t_j$, for $j \in \{1, 2, \dots, J\}$. Also, we define the error-prone survival vector, $\mathbb{S}_0 = (S(1), S(2), \dots, S(J))^T$ for time points 1 to $J$. This error-prone survival vector, $\mathbb{S}_0$, does not take into account potentially misclassified events, as such it will be estimated by the KM estimator. The elements of $\mathbb{S}_0$ can be estimated by the KM estimator as follows:

$$\hat{S}(j) = \prod_{k=1}^{j} \left( 1 - \frac{\sum_{i=1}^{n_k} (1 - \delta_{ik}) I(E_{ik} = 1)}{n_k} \right), \; j = 1, 2, \dots, J,$$

where $E_{ij}$ is the event indicator for the $i$-th subject at time $t_j$ (i.e., $E_{ij} = I(T_i = t_j)$), the indicator $\delta_{ij} = 1$ represents the $i$-th subject censored at time $t_j$, and 0 otherwise, and $n_j$ is the number of subjects known to have survived (have not yet had an event or been censored) up to time $t_{j-1}$ (i.e., $n_j = \sum_{i=1}^{n_{j-1}} (1 - \delta_{i(j-1)}) I(E_{i(j-1)} = 0)$). Let $\hat{\mathbb{S}}_0 = (\hat{S}(1), \hat{S}(2), \dots, \hat{S}(J))^T$. An expression for the estimator of the survival function of G-S events at time $t_j$ is given by

$$\hat{S}^*(j) = f_j(\mathbb{P}) + \left\{ g_j(\mathbb{P}) \right\}^T \hat{\mathbb{S}}_0, \; j = 1, 2, \dots, J. \tag{9}$$

Breslow and Crowley [36] showed that as $n \to \infty$, $\sqrt{n}(\hat{S}(j) - S(j))$ converges in distribution to a Gaussian process with expectation 0 and a variance–covariance function that could be approximated using Greenwoods formula in Greenwood [37]. By adapting their techniques, we derive the asymptotic variance of the KM estimates and thus obtain the asymptotic covariance matrix of our proposed estimator in the presence of right censoring and mismeasured events. The estimated variance of the estimated survival rate of G-S events at time $t_j$ is given by

$$\widehat{\text{Var}}\,(\hat{S}^*(j)) = \left\{ g_j(\mathbb{P}) \right\}^T \widehat{\text{Var}}\,(\hat{\mathbb{S}}_0) \left\{ g_j(\mathbb{P}) \right\}, \tag{10}$$

where

$$\widehat{\mathrm{Var}}\,(\hat{\mathbb{S}}_0) = \begin{pmatrix} \widehat{\mathrm{Var}}\,(\hat{S}(1)) & \widehat{\mathrm{Cov}}\,(\hat{S}(1), \hat{S}(2)) & \cdots & \widehat{\mathrm{Cov}}\,(\hat{S}(1), \hat{S}(J)) \\ \widehat{\mathrm{Cov}}\,(\hat{S}(2), \hat{S}(1)) & \widehat{\mathrm{Var}}\,(\hat{S}(2)) & \cdots & \widehat{\mathrm{Cov}}\,(\hat{S}(2), \hat{S}(J)) \\ \cdots & \cdots & \cdots & \cdots \\ \widehat{\mathrm{Cov}}\,(\hat{S}(J), \hat{S}(1)) & \widehat{\mathrm{Cov}}\,(\hat{S}(J), \hat{S}(2)) & \cdots & \widehat{\mathrm{Var}}\,(\hat{S}(J)) \end{pmatrix},$$

$\widehat{\mathrm{Var}}\,(\cdot)$ is obtained from Greenwood [37] with

$$\widehat{\mathrm{Var}}\,(\hat{S}(j) = \hat{S}^2(j) \sum_{k=1}^{j} \left( \frac{\sum_{i=1}^{n_k} I(E_{ik} = 1)}{n_k \left[ n_k - \sum_{i=1}^{n_k} I(E_{ik} = 1) \right]} \right), j = 1, 2, \dots, J,$$

and $\widehat{\mathrm{Cov}}\,(\cdot)$ is obtained from Breslow and Crowley [36] for $j < k$ and $j, k = 1, \dots, J$ with $\widehat{\mathrm{Cov}}\,(\hat{S}(j), \hat{S}(k)) = \frac{\hat{S}(k)}{\hat{S}(j)} \widehat{\mathrm{Var}}\,(\hat{S}(j))$. Log–log transformed $(1 - \alpha)$ CI at time $t_j$, suggested by Borgan and Liestøl [38], is given by

$$\left( [\hat{S}^*(j)]^{\frac{1}{\theta}},\ [\hat{S}^*(j)]^{\theta} \right), \tag{11}$$

where $\theta = \exp\left\{ \frac{Z_{\alpha/2} \hat{\sigma}_{S^*(j)}}{\log[\hat{S}^*(j)]} \right\}$ and $\hat{\sigma}^2_{S^*(j)} = \frac{\widehat{\mathrm{Var}}\,\{\hat{S}^*(j)\}}{\{\hat{S}^*(j)\}^2}$.

The next step is to prove consistency and asymptotic normality of our estimator of the survival distribution of G-S events.

**Theorem 2.3** (Consistency) *In model* (5) *under Assumption* 1, *the estimators defined in* (9) *are consistent.*

The result follows from the fact that the KM estimator $\hat{S}(j)$ of $S(j)$ is consistent in Gill [39] and the estimator $\hat{S}^*(j)$ is a linear combination of $\hat{S}(j)$.

**Theorem 2.4** (Asymptotic normality) *In model* (5) *under Assumption* 1, *the estimators defined in* (9) *are asymptotically normal with mean* $S^*(j)$ *and variance*

$$\mathrm{Var}\,(\hat{S}^*(j)) = \left\{ g_j(\mathbb{P}) \right\}^T \mathrm{Var}\,(\hat{\mathbb{S}}_0) \left\{ g_j(\mathbb{P}) \right\}^T, \tag{12}$$

*where* $\mathrm{Var}\,(\hat{\mathbb{S}}_0)$ *is a variance–covariance matrix, with* $\mathrm{Var}\,(\hat{S}^*(j))$ *along the diagonal and* $\mathrm{Cov}\,(\hat{S}^*(j), \hat{S}(k))$, $j < k; j, k = 1, \dots, J$, on *the off-diagonal.*

The above approaches can be similarly applied to the model in (8) under Assumption 2, which is for the case where the G-S event does not happen after an E-P event.

### 2.2.2 Inference for Unknown Model Parameters

There are situations for which $\omega_1$, $\omega_2$ and $\tau_0$ in the proposed model (4), and $\psi_1$, $\psi_2$ and $\gamma_0$ in the model (7) are not known and the aforementioned parameters, $\mathbb{P} = (\omega_1, \omega_2, \tau_0)$ in (6) under Theorem 2.1 ($\mathbb{P} = (\psi_1, \psi_2, \gamma_0)$ under Theorem 2.2), will need to be estimated from data. In this framework, we estimate $\mathbb{P}$ directly from the on-going clinical study. We first need to obtain the "pilot data" (complete data), which is data on a randomly selected small portion of the entire clinical study subjects with E-P and G-S outcomes and is used to estimate $\mathbb{P}$. The remaining (unselected) participants in the clinical study would only have the E-P outcomes, this set of observations we call the "analysis data." Under this setting, the pilot data and the analysis data are independent.

First, we discuss parameters, $\mathbb{P} = (\omega_1, \omega_2, \tau_0)$, using sustained virologic response (SVR), defined as lack of detectable serum HCV RNA in serum after 24 weeks of completing treatment was the primary endpoint in the VIRAHEP-C study. There were two assays used to test viral load, the quantitative PCR-based assay (E-P) and the qualitative PCR-based assay (G-S). Serum samples were tested for HCV RNA levels using the quantitative PCR-based assay which had a lower limit of sensitivity of 600 IU/ml, while the qualitative PCR-based had a lower limit of sensitivity of 50 IU/ml. Viral negativity was assessed by the more sensitive qualitative assay. If the qualitative assay (G-S) was not available due to costs or other reasons, it is reasonable to deduce that the outcomes from the less sensitive quantitative (E-P) assay are prone to error. Our research specifically addresses this issue, and we shall illustrate in Sect. 4.1 that the survival function of events from the G-S assay can be accurately obtained from a small pilot dataset.

We apply our methods to the study of the G-S event time to viral negativity from the VIRAHEP-C clinical trial. As discussed in Sect. 2.1, there are the G-S population (unobserved) and the potentially misclassified (E-P) population. In this view, the derivation of the survival function of G-S events is intractable. Using the pilot dataset, the estimates $\hat{\mathbb{P}} = (\hat{\omega}_1, \hat{\omega}_2, \hat{\tau}_0)$ are obtained by minimizing the weighted sum of squared distances between the estimated KM survival rates of G-S events($S_P^*(j)$) and the estimated approximated survival rates of G-S events ($\hat{S}^*(j)$) based on (4). Under Assumption 1, the estimates, $(\hat{\omega}_1, \hat{\omega}_2, \hat{\tau}_0)$, are obtained as follows:

$$(\hat{\omega}_1, \hat{\omega}_2, \hat{\tau}_0) = \underset{0 \le \tau_0 \le 1,\ \omega_1, \omega_2 \ge 0}{\operatorname{argmin}} \left\{ \sum_{j=1}^{J} w(j) \left( S_P^*(j) - \hat{S}^*(j) \right)^2 \right\}, \tag{13}$$

where the weight $w(j)$ is $\{S_P^*(j)\}^{\rho_1}\{1 - S_P^*(j)\}^{\rho_2}$ for $0 \le \rho_1, \rho_2 \le 1$ and $j = 1, 2, \ldots, J$. Let $\mathbb{D} = (\mathbb{S}_0^T, \mathbb{P}^T)^T$. Then, the $\hat{\mathbb{S}}_0$ are the KM estimates using the analysis dataset and $\hat{\mathbb{P}}$ can be obtained by (13) using the pilot dataset. The extended expression for the estimator of the survival distribution of G-S events is given by

$$\hat{S}^*(j) = f_j(\hat{\mathbb{P}}) + \left\{ g_j(\hat{\mathbb{P}}) \right\}^T \hat{\mathbb{S}}_0,\ j = 1, 2, \ldots, J. \tag{14}$$

Since $\hat{\mathbb{P}}$ and $\hat{\mathbb{S}}_0$ are correspondingly obtained from the pilot dataset and analysis dataset, they are independent.

Computing the standard error of $\hat{S}^*(j)$ in (14) is quite challenging since the delta method may not be applicable due to the small size of the pilot data. Here, we develop a new approach to estimate the variance of $\hat{S}^*(j)$. Using the standard variance decomposition formula, we have

$$\text{Var}\left[\hat{S}^*(j)\right] = E\left[\text{ Var }(\hat{S}^*(j)|\hat{\mathbb{P}})\right] + \text{ Var }\left[E\left(\hat{S}^*(j)|\hat{\mathbb{P}}\right)\right], j = 1, 2, \ldots, J.$$

Since Var $\left[\hat{S}^*(j)|\hat{\mathbb{P}}\right]$ and $E\left[\hat{S}^*(j)|\hat{\mathbb{P}}\right]$ are functions of $\mathbb{S}_0$ and $\hat{\mathbb{P}}$, we write

$$\sigma_j^2(\mathbb{S}_0,\ \hat{\mathbb{P}}) = \ \text{Var}\left[\hat{S}^*(j)|\hat{\mathbb{P}}\right],$$

and

$$\mu_j(\mathbb{S}_0,\ \hat{\mathbb{P}}) = E\left[\hat{S}^*(j)\middle|\hat{\mathbb{P}}\right], j = 1, 2, \ldots, J.$$

Using (12), we have

$$\sigma_j^2(\mathbb{S}_0, \hat{\mathbb{P}}) = \{g_j(\hat{\mathbb{P}})\}^T \ \text{Var}\ (\hat{\mathbb{S}}_0)\{g_j(\hat{\mathbb{P}})\}, j = 1, 2, \ldots, J.$$

Since the size of the analysis dataset is relatively large and KM estimates, $\hat{\mathbb{D}}$, are consistent, $\mu_j(\mathbb{S}_0,\ \hat{\mathbb{P}})$ can be approximated by

$$\tilde{\mu}_j(\mathbb{S}_0,\ \hat{\mathbb{P}}) = f_j(\hat{\mathbb{P}}) + \left\{g_j(\hat{\mathbb{P}})\right\}^T \mathbb{S}_0, j = 1, 2, \ldots, J.$$

To estimate $E\left[\sigma_j^2(\mathbb{S}_0,\ \hat{\mathbb{P}})\right]$ and Var $\left[\tilde{\mu}_j(\mathbb{S}_0,\ \hat{\mathbb{P}})\right]$ at time $t_j$, we use the bootstrapping method. Let $\{\hat{\mathbb{P}}^{(b)} = (\hat{\omega}_1^{(b)},\ \hat{\omega}_2^{(b)},\ \hat{\tau}_0^{(b)}),\ b = 1, 2, \ldots, B\}$ denote a bootstrap sample of size $B$ using the pilot dataset. For given $\mathbb{S}_0$, we compute

$$\hat{E}\left[\sigma_j^2(\mathbb{S}_0,\ \hat{\mathbb{P}})\right] = \frac{1}{B}\sum_{b=1}^{B}\{g_j(\hat{\mathbb{P}}^{(b)})\}^T \ \text{Var}\ (\mathbb{S}_0), \{g_j(\hat{\mathbb{P}}^{(b)})\} \tag{15}$$

and

$$\widehat{\text{Var}}\ (\tilde{\mu}_j(\mathbb{S}_0,\ \hat{\mathbb{P}})) = \frac{1}{B-1}\sum_{b=1}^{B}(\tilde{\mu}_j(\mathbb{S}_0,\ \hat{\mathbb{P}}^{(b)}) - \bar{\mu}_j(\mathbb{S}_0,\ \hat{\mathbb{P}}))^2, \tag{16}$$

where $\bar{\mu}_j(\mathbb{S}_0,\ \hat{\mathbb{P}}) = \frac{1}{B}\sum_{b=1}^{B}\tilde{\mu}_j(\mathbb{S}_0,\ \hat{\mathbb{P}}^{(b)})$. Finally, letting $\mathbb{S}_0 = \hat{\mathbb{S}}_0$ in (15) and (16), we obtain an approximate standard error (se) of $\hat{S}^*(j)$ as follows:

$$se(\hat{S}^*(j)) = \left\{\hat{E}\left[\sigma_j^2(\hat{\mathbb{S}}_0,\ \hat{\mathbb{P}})\right] + \widehat{\text{Var}}\ (\tilde{\mu}_j(\hat{\mathbb{S}}_0,\ \hat{\mathbb{P}}))\right\}^{1/2}, j = 1, 2, \ldots, J. \tag{17}$$

Note that to compute $\hat{E}\left[\sigma_j^2(\hat{\mathbb{S}}_0,\ \hat{\mathbb{P}})\right]$ in (17), we use $\{g_j(\hat{\mathbb{P}}^{(b)})\}^T\ \widehat{\text{Var}}\ (\hat{\mathbb{S}}_0)\{g_j(\hat{\mathbb{P}}^{(b)})\}$ in (15), where $\widehat{\text{Var}}\ (\hat{\mathbb{S}}_0)$ is given by (10).

In a similar fashion, the above approaches can be applied to the parameters, $\psi_1,\ \psi_2$ and $\gamma_0$ in (8)—the scenario for which the G-S event does not happen after an E-P event.

## 3 Stochastic Process-Based Discrete Survival Times

The latent course of a foreign body (e.g., viral load, bacteria count) which underlies disease progression may have a random probability distribution or pattern. In oncology, tumor growth could be studied as a stochastic process. Even the spread of a fatal disease within a closed community could be modeled through a random probability distribution. The opportunities of real-life applications of time-to-event analysis derived from stochastic processes are exciting. Thus, we examine the gamma process and Wiener process within the discrete time-to-event setting.

### 3.1 Gamma Process-Based Discrete Survival Times

We infer that the course of the viral load can be modeled via a gamma process. If this conjecture is approximately correct, it will mean that the closed-form expression of the true survival function can be derived analytically, therefore simulation studies are unnecessary. It will mean that the gamma process represents the underlying absolute state of information (truth), thus allowing for inference as discussed in Sect. 2.1. This is a very favorable quality because the survival estimates from the gold-standard diagnostic tool can be directly calculated and $\tau_0$ will depend on the diagnostic test. From the gamma process, we generate discrete-time survival data using a specified detection limit. For a gamma distribution denoted as Gamma $(a, b)$ $(a, b > 0)$ with mean $ab$ and variance $ab^2$, suppose $\alpha(t)$ is an increasing and right continuous function on $[0, \infty)$ with $\alpha(0) = 0$. Furthermore, let $W = \{W_t,\ t \geq 0\}$ be a gamma process with the following properties: (i) $W_0 = 0$, (ii) $W$ has independent increments in disjoint intervals, and (iii) for $t > s$, $W_t - W_s \sim$ Gamma $(\alpha(t) - \alpha(s),\ b)$, where $b > 0$ is a constant. Then $W$ is called a gamma process (GP), denoted by $W \sim$ GP $(\alpha(t),\ b)$. Let $W_j^* = W_j - E[W_j]$ and assume that we only observe $W_j$ at integer times, i.e., $j = 1, 2, 3, \ldots$.

Let $W = \{W_j, j \geq 0\}$ be a GP $(j, 1)$, where $W_j = X_1 + \cdots + X_j$ and the $X_j$ are i.i.d. from Gamma $(1, 1)$ for $j = 1, \ldots$. The survival function at time $t_j = j$ with detection level $c$ is defined as $S_c(j) = P(X_1 \geq 1 + c, X_1 + X_2 \geq 2 + c, \ldots, X_1 + \cdots + X_j \geq j + c)$. The survival function with a lower detection limit level $c$ is expressed as

$$\begin{aligned} S_c(j) = &\ P(X_1 \geq 1 + c, X_1 + X_2 \geq 2 + c, \ldots, X_1 + \cdots + X_j \geq j + c) \\ = &\int_{c+j}^{\infty} \exp(-w_j) B_j(c, w_j) dw_j \end{aligned} \tag{18}$$

where $B_j(c, w_j) = \int_{c+j-1}^{w_j} \cdots \int_{c+1}^{w_2} dw_1 \cdots dw_{j-1}$ for $j > 1$ and $B_1(c, w_1) = 1$. The following lemma provides the closed-form expression of $B_j(c, w_j)$.

**Lemma 3.1** *The $B_j(c, w_j)$ in (18) is written as*

$$B_j(c, w_j) = \frac{(w_j - c)^{j-1}}{(j-1)!} - \frac{(w_j - c)^{j-2}}{(j-2)!}$$

*for $j = 2, 3, 4, \ldots$.*

Using Lemma 3.1, we obtain the closed-form expression of the survival function, which is given in the next theorem.

**Theorem 3.1** *Suppose $W = \{W_j, j = 1, 2, \ldots\}$ follows $GP(j,\ 1)$. The survival function at time $t_j$ with a lower detection limit level $c$ is given by*

$$S_c(j) = \frac{j^{(j-1)}}{(j-1)!}\exp\{-(c+j)\}.$$

Under the lower-limit detection framework, if $c^* \leq c$ then $P(T^* \geq T) = 1$, where $^*$ denotes the gold standard, we examine our proposed model under this framework. We consider two detection limits for the G-S and E-P events as $c^* = -0.8$, and $c = -0.4$. In this case, the G-S survival function is $S_{c^*}(j) = \frac{j^{(j-1)}}{(j-1)!}\exp\{-(c^*+j)\}$ and the E-P survival function is $S_c(j) = \frac{j^{(j-1)}}{(j-1)!}\exp\{-(c+j)\}$. Let $\hat{S}^*_c(j)$ be the approximate survival function based on (6), where $\mathbb{S}_0$ is computed using $S_c(j)$. It does not appear that there exist $(\omega_1, \omega_2, \tau_0)$ such that $\hat{S}^*_c(j)$ is exactly equal to $S_{c^*}(j)$. Therefore, we use (13) with $S^*_P(j)$ and $\hat{S}^*(j)$ replaced by $S_{c^*}(j)$ and $\hat{S}^*_c(j)$, respectively, to find the optimal values of $(\omega_1, \omega_2, \tau_0)$. The optimal values of $(\omega_1, \omega_2, \tau_0)$ based on the pilot data are (0.00, 0.59, 0.48) for $\rho_1 = 1$ and $\rho_2 = 0$, (0.31, 0.00, 0.53) for $\rho_1 = \rho_2 = 0.5$, and (0.22, 0.000, 0.59) for $\rho_1 = 0$ and $\rho_2 = 1$. By considering $S_{c^*}(j)$ as $P(E^*_j = 0)$ and $S_{c*}(j)$ as $P(E_j = 0)$ in (ii) of the Proposition 2.1, we obtain the vector of PPV, $\tau = (\tau_1 = 0.402,\ \tau_2 = 0.667,\ \tau_3 = 0.753,\ \tau_4 = 0.798,\ \tau_5 = 0.826,\ \tau_6 = 0.845,\ \tau_7 = 0.859,\ \tau_8 = 0.871)^T$. The vector of PPV, $\tau$, shows that PPV is not constant over time and therefore supports the approach of developing our methodology from a time-varying PPV standpoint. In addition, Table 1 shows G-S survival rates, E-P survival rates, and approximated survival for $\rho_1 = 1,\ \rho_2 = 0$; $\rho_1 = \rho_2 = 0.5$; and for $\rho_1 = 0,\ \rho_2 = 1$ based on Theorem 2.1.

The approximated survival rates with $\rho_1 = 0.5$ and $\rho_2 = 0.5$ are robust. The approximation with $\rho_1 = 1$ and $\rho_2 = 0$ is best at $t_1$, but worse at $t_8$, whereas the approximation with $\rho_1 = 0$ and $\rho_2 = 1$ is best at $t_8$, but worse at $t_1$. The approximated survival functions ($\rho_1 = \rho_2 = 0.5$; $\rho_1 = 0,\ \rho_2 = 1$) in Table 1 are illustrated in Fig. 1. The difference between the survival rates of E-P and G-S are due to mismeasured outcomes.

From Fig. 1, we observe that the approximated survival function is very close to the survival function of G-S outcomes. This shows that the model in Theorem 2.1 works well under the gamma process. We extend Theorem 3.1 with $X_j \sim$ Gamma $(1, \lambda)$. Since $\frac{X_j}{\lambda} \sim$ Gamma $(1, 1)$, we have

$$\begin{aligned} S_{(c,\lambda)}(j) =&\ P(X_1 \geq \lambda + c, X_1 + X_2 \geq 2\lambda + c, \dots, X_1 + \cdots + X_j \geq j\lambda + c) \\ =&\ P(\frac{X_1}{\lambda} \geq 1 + \frac{c}{\lambda}, \frac{X_1 + X_2}{\lambda} \geq 2 + \frac{c}{\lambda}, \dots, \frac{X_1 + \cdots + X_j}{\lambda} \geq j + \frac{c}{\lambda}) \\ =&\ \int_{\frac{c}{\lambda}+j}^{\infty} \exp(-y_j) B_j\Big(\frac{c}{\lambda}, y_j\Big) dy_j = S_{\frac{c}{\lambda}}(j), \end{aligned} \tag{19}$$

where $Y_j = \frac{X_1 + \cdots + X_j}{\lambda}$. Using (19), the formula of $S_{(c,\lambda)}(j)$ with $X_j \sim$ Gamma $(1, \lambda)$ is given in Corollary 3.1.

**Table 1** Results of the approximation of survival probabilities for time to viral negativity at selected time points for $c^* = -0.8$ and $c = -0.4$

| Time | $c^* = -0.8,\ c = -0.4$ | | | $\rho_1 = 0.5,\ \rho_2 = 0.5$ | | $\rho_1 = 1,\ \rho_2 = 0$ | | $\rho_1 = 0,\ \rho_2 = 1$ | |
|---|---|---|---|---|---|---|---|---|---|
| | G-S | E-P | PPV | Approximated | PPV-Approx | Approximated | PPV-Approx | Approximated | PPV-Approx |
| $t_1$ | 0.819 | 0.549 | 0.402 | 0.759 | 0.478 | 0.784 | 0.534 | 0.735 | 0.587 |
| $t_2$ | 0.602 | 0.404 | 0.667 | 0.623 | 0.604 | 0.640 | 0.632 | 0.606 | 0.661 |
| $t_3$ | 0.499 | 0.334 | 0.753 | 0.528 | 0.704 | 0.531 | 0.709 | 0.519 | 0.722 |
| $t_4$ | 0.435 | 0.291 | 0.798 | 0.455 | 0.779 | 0.448 | 0.769 | 0.453 | 0.772 |
| $t_5$ | 0.391 | 0.262 | 0.826 | 0.396 | 0.835 | 0.384 | 0.818 | 0.400 | 0.812 |
| $t_6$ | 0.357 | 0.240 | 0.845 | 0.349 | 0.875 | 0.334 | 0.856 | 0.357 | 0.846 |
| $t_7$ | 0.332 | 0.222 | 0.859 | 0.311 | 0.905 | 0.296 | 0.886 | 0.321 | 0.873 |
| $t_8$ | 0.311 | 0.208 | 0.871 | 0.280 | 0.927 | 0.266 | 0.910 | 0.291 | 0.896 |

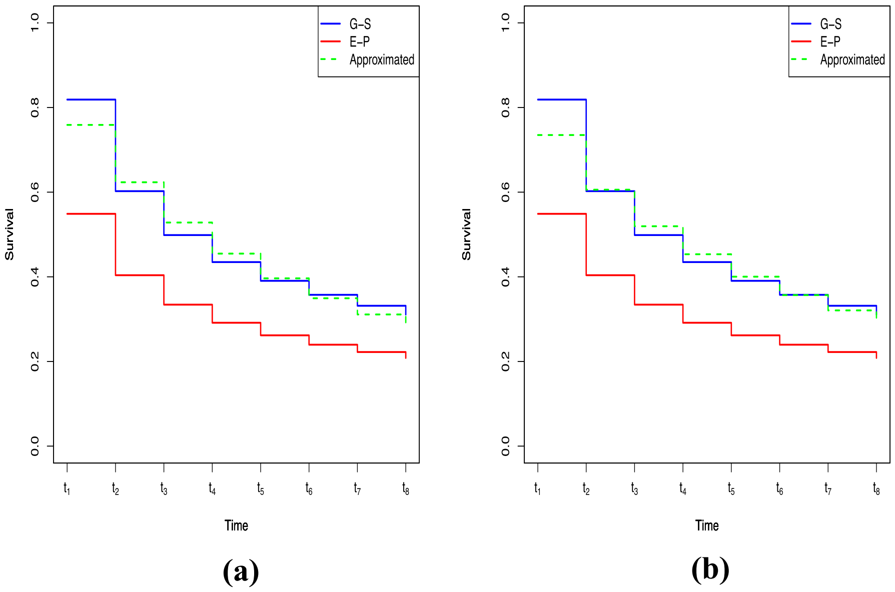


**Fig. 1** The survival functions of G-S and E-P events and approximated survival functions with the lower detection limit levels as $c^* = -0.8$ and $c = -0.4$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (**a**), and for $\rho_1 = 0$ and $\rho_2 = 1$ (**b**)

**Corollary 3.1** *Suppose that the* $X_j$ *are i.i.d. from* Gamma $(1, \lambda)$ *for* $j = 1, 2, \ldots$. *Then, the survival function at time* $t_j$ *with a lower detection limit level as c is given by*

$$S_{(c,\lambda)}(j) = P(X_1 \geq \lambda + c, X_1 + X_2 \geq 2\lambda + c, \ldots, X_1 + \cdots + X_j \geq j\lambda + c) = S_{\frac{c}{\lambda}}(j).$$

There is no information to adequately model the course of viral load to any specific stochastic process. However, under the scenario that the observed viral load originates from a gamma process, we have demonstrated an application of our method to the lower-limit detection problem and shown the favorable quality of a closed-form formulation of the G-S survival function.

### 3.2 Wiener Process

We now model the viral load course via a standard Brownian motion process(Wiener process). Unlike the gamma process, this approach does not have the favorable quality of a closed-form formulation of the G-S survival function. As a result, we assess the properties of our estimator through simulation studies. The BM process on the interval [0, $K$] is a random variable, $W(t)$, which depends continuously on $t \in [0, K]$ and satisfies the following: $W(0) = 0$ and for $0 \leq t_s < t_j \leq t_J$, $W(t_j) - W(t_s) \sim \sqrt{t_j - t_s} * N(0, \frac{1}{18000})$, where $t_J = 18000$ and $J = 8$, which is the maximum predetermined number of clinical visits.

To simulate, we discretize the BM with unevenly spaced time step, $t_1 = 20, t_2 = 30, t_3 = 40, t_4 = 50, t_5 = 80, t_6 = 200, t_7 = 1000, t_8 = 18000$. We conduct the simulation study as follows. For each simulated dataset of size $n = 400$, we generate $B_i = (B_{ik})'$ as $B_{ik} \sim N(0, \frac{1}{18000})$ and obtain $W_{it_j} = \sum_{k=1}^{t_j} B_{ik}$, for $i = 1, \ldots, n$, $j = 1, 2, \ldots, 8$ and $k = 1, \ldots, 18000$. By setting $W_{i0} = 0$, $t_8 = 18000$, we have $W_{it_j} - W_{i0} \sim N(0, \frac{t_j}{18000})$. We consider only 8 time points of $W_{it_j}$ and the 8 time points are selected to generate similar survival rates compared to HCV data described in Sect. 4.1.

Since we are comparing the E-P diagnostic test to the G-S, we therefore specify two detection levels, $c = -0.04$ ($c^* = -0.056$) for E-P (G-S) survival time. The E-P ($T_i$) and G-S ($T_i^*$) survival times are generated as $T_i = \min\{j : W_{it_j} \leq -0.04\}$ and $T_i^* = \min\{j : W_{it_j} \leq -0.056\}$. We then generate 500 datasets with $n = 400$. For the $\ell$ th analysis dataset, a pilot dataset is randomly sampled with $n_0$ subjects for $n_0 = 40$ and $n_0 = 80$, which correspond to the 10%, and 20% of $n$. Using the pilot data, we obtain parameter estimates, $\hat{\mathbb{P}}_{\ell 1} = (\hat{\omega}_{\ell 1}, \hat{\omega}_{\ell 2}, \hat{\tau}_{\ell 0})$ and $B = 200$ sets of bootstrapping estimates, $\hat{\mathbb{P}}_{\ell 1}^{(b)} = (\hat{\omega}_{\ell 1}^{(b)}, \hat{\omega}_{\ell 2}^{(b)}, \hat{\tau}_{\ell 0}^{(b)})$, for $b = 1, \ldots, B$. For the $\ell$ th analysis dataset with $n - n_0$ subjects, the approximated survival function and estimated variance of the approximated survival function are obtained using $\hat{\mathbb{P}}_{\ell 1}$ and $\hat{\mathbb{P}}_{\ell 1}^{(b)}$ for $b = 1, \ldots, B$. For each simulated dataset of size $n = 400$, the running times in minutes with the bootstrap sample of size $B = 200$ are 2.05 ($\rho_1 = 1$ and $\rho_2 = 0$), 2.74 ($\rho_1 = 0.5$ and $\rho_2 = 0.5$), and 2.65 ($\rho_1 = 0$ and $\rho_2 = 1$) for the pilot data with $n_0 = 40$. Likewise, for the pilot data with $n_0 = 80$ the times (in minutes) are 0.90 ($\rho_1 = 1$ and $\rho_2 = 0$), 1.29 ($\rho_1 = 0.5$ and $\rho_2 = 0.5$), and 1.93 ($\rho_1 = 0$ and $\rho_2 = 1$). Simulations were performed on an intel core i7 processor machine with 16 GB of RAM memory using a Windows 10 operating system for computing.

We evaluate the performance of the methods detailed in Sect. 2.2.2 by simulating 500 datasets. In Table 2, we present the average of the approximated survival rates (Approximated), along with the the average of standard errors (ASE), the Monte Carlo standard error (MCSE) and the coverage probability (CP) are also presented. Figure 2 shows the means of survival rates of G-S and E-P events, and the means of approximated survival rates from the analysis datasets using the estimated parameters (from pilot datasets) with $n_0 = 40$ and $n_0 = 80$, respectively. The approximated survival function is very close to the G-S survival function across all time points.

These results from $t_1$ to $t_8$ suggest that (i) the differences between survival rates of G-S events and approximated survival rates are less than 0.012 (ii) the differences between ASE and MCSE are less than 0.005 (iii) except at $t_1$ with small number of events, CPs for $n_0 = 40$ and $n_0 = 80$ are from 0.912 to 0.954 and from 0.924 to 0.964, respectively.

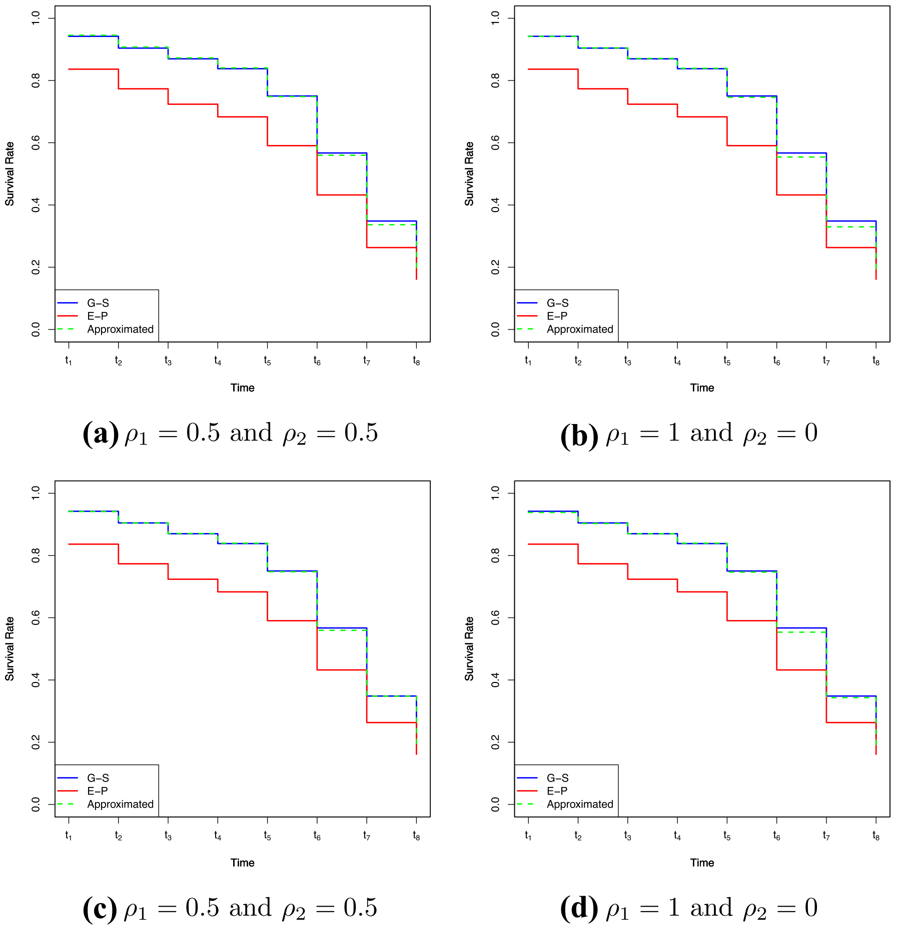
Most importantly, the results of our simulation study validate the mathematical results from Sect. 2.2.2. When the course of the viral load does not follow a gamma process, we have shown that the parameters, $\omega_1$, $\omega_2$ , and $\tau_0$ can be estimated through a small pilot study. This is a useful development for clinical trial studies for which the parameters are unknown and the latent stochastic process of disease cannot be confirmed.

**Table 2** The Estimates under the Brownian motions process with $c^* = -0.056$ and $c = -0.04$

| Time | $n_0 = 40$ | | | | | | $n_0 = 80$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G-S | E-P | Approximated | ASE | MCSE | CP | G-S | E-P | Approximated | ASE | MCSE | CP |
| $\rho_1 = 1,\ \rho_2 = 0$ | | | | | | | | | | | | |
| $t_1$ | 0.958 | 0.891 | 0.957 | 0.020 | 0.022 | 0.928 | 0.958 | 0.891 | 0.954 | 0.018 | 0.020 | 0.876 |
| $t_2$ | 0.905 | 0.809 | 0.907 | 0.028 | 0.030 | 0.942 | 0.905 | 0.808 | 0.905 | 0.023 | 0.024 | 0.944 |
| $t_3$ | 0.854 | 0.743 | 0.858 | 0.036 | 0.038 | 0.940 | 0.854 | 0.743 | 0.857 | 0.029 | 0.029 | 0.952 |
| $t_4$ | 0.808 | 0.690 | 0.814 | 0.044 | 0.047 | 0.940 | 0.808 | 0.69 | 0.813 | 0.035 | 0.036 | 0.946 |
| $t_5$ | 0.725 | 0.606 | 0.720 | 0.053 | 0.056 | 0.912 | 0.726 | 0.606 | 0.718 | 0.042 | 0.045 | 0.920 |
| $t_6$ | 0.585 | 0.479 | 0.573 | 0.054 | 0.051 | 0.942 | 0.586 | 0.479 | 0.571 | 0.042 | 0.039 | 0.952 |
| $t_7$ | 0.408 | 0.330 | 0.396 | 0.048 | 0.045 | 0.904 | 0.408 | 0.331 | 0.401 | 0.040 | 0.038 | 0.960 |
| $t_8$ | 0.241 | 0.195 | 0.232 | 0.037 | 0.040 | 0.912 | 0.241 | 0.195 | 0.235 | 0.034 | 0.034 | 0.954 |
| $\rho_1 = 0.5,\ \rho_2 = 0.5$ | | | | | | | | | | | | |
| $t_1$ | 0.958 | 0.891 | 0.957 | 0.021 | 0.023 | 0.934 | 0.958 | 0.891 | 0.956 | 0.020 | 0.022 | 0.892 |
| $t_2$ | 0.905 | 0.809 | 0.907 | 0.028 | 0.031 | 0.946 | 0.905 | 0.808 | 0.906 | 0.024 | 0.025 | 0.946 |
| $t_3$ | 0.854 | 0.743 | 0.857 | 0.036 | 0.039 | 0.936 | 0.854 | 0.743 | 0.857 | 0.029 | 0.030 | 0.958 |
| $t_4$ | 0.808 | 0.690 | 0.813 | 0.044 | 0.047 | 0.938 | 0.808 | 0.69 | 0.812 | 0.035 | 0.036 | 0.952 |
| $t_5$ | 0.725 | 0.606 | 0.719 | 0.052 | 0.055 | 0.908 | 0.726 | 0.606 | 0.717 | 0.042 | 0.044 | 0.926 |
| $t_6$ | 0.585 | 0.479 | 0.576 | 0.053 | 0.049 | 0.946 | 0.586 | 0.479 | 0.574 | 0.040 | 0.039 | 0.956 |
| $t_7$ | 0.408 | 0.330 | 0.400 | 0.047 | 0.044 | 0.918 | 0.408 | 0.331 | 0.405 | 0.039 | 0.039 | 0.964 |
| $t_8$ | 0.241 | 0.195 | 0.234 | 0.037 | 0.039 | 0.916 | 0.241 | 0.195 | 0.237 | 0.034 | 0.034 | 0.956 |
| $\rho_1 = 0,\ \rho_2 = 1$ | | | | | | | | | | | | |
| $t_1$ | 0.958 | 0.891 | 0.958 | 0.022 | 0.024 | 0.936 | 0.958 | 0.891 | 0.959 | 0.021 | 0.023 | 0.910 |
| $t_2$ | 0.905 | 0.809 | 0.907 | 0.030 | 0.032 | 0.948 | 0.905 | 0.808 | 0.909 | 0.026 | 0.027 | 0.952 |
| $t_3$ | 0.854 | 0.743 | 0.858 | 0.038 | 0.040 | 0.942 | 0.854 | 0.743 | 0.859 | 0.031 | 0.032 | 0.970 |
| $t_4$ | 0.808 | 0.690 | 0.813 | 0.045 | 0.049 | 0.938 | 0.808 | 0.69 | 0.813 | 0.036 | 0.038 | 0.962 |
| $t_5$ | 0.725 | 0.606 | 0.718 | 0.053 | 0.056 | 0.902 | 0.726 | 0.606 | 0.715 | 0.042 | 0.044 | 0.924 |

**Table 2** (continued)

| Time | $n_0 = 40$ | | | | | | $n_0 = 80$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G-S | E-P | Approximated | ASE | MCSE | CP | G-S | E-P | Approximated | ASE | MCSE | CP |
| $t_6$ | 0.585 | 0.479 | 0.577 | 0.052 | 0.050 | 0.954 | 0.586 | 0.479 | 0.575 | 0.040 | 0.039 | 0.960 |
| $t_7$ | 0.408 | 0.330 | 0.403 | 0.046 | 0.044 | 0.924 | 0.408 | 0.331 | 0.409 | 0.039 | 0.040 | 0.964 |
| $t_8$ | 0.241 | 0.195 | 0.237 | 0.037 | 0.039 | 0.918 | 0.241 | 0.195 | 0.238 | 0.035 | 0.035 | 0.944 |



**(a)** $\rho_1 = 0.5$ and $\rho_2 = 0.5$

**(b)** $\rho_1 = 1$ and $\rho_2 = 0$

**(c)** $\rho_1 = 0.5$ and $\rho_2 = 0.5$

**(d)** $\rho_1 = 1$ and $\rho_2 = 0$

**Fig. 2** The means of G-S and E-P, and approximated G-S survival rates using $n_0 = 40$ (**a** and **b**) and $n_0 = 80$ (**c** and **d**)

# 4 Data Analyses

## 4.1 Analysis of VIRAHEP-C Data

The VIRAHEP-C study is an international clinical trial sponsored by the NIDDK-NIH that was designed to test the hypothesis that African Americans respond less well to antiviral therapy than Caucasian Americans. A total of 401 chronically infected participants with Hepatitis C virus (HCV) of genotype 1 were enrolled. We study up to the 24-week timepoint as this was the time of the primary endpoint in the VIRAHEP-C study, the visit times are as follows: $t_1 = 1$, $t_2 = 2$, $t_3 = 7$, $t_4 = 14$, $t_5 = 28$, $t_6 = 56$, $t_7 = 84$, and $t_8 = 168$ day. G-S and E-P events at time $t_j$ are defined as $E^*(j) = I\{\text{viral levels } \leq 50 \text{ IU/mlat } t_j\}$ and $E(j) = I\{\text{viral levels } \leq 600 \text{ IU/mlat } t_j\}$, respectively.

As described in Sect. 2.2.2, we assume parameters, $\omega_1$, $\omega_2$, and $\tau_0$ are unknown. Two pilot datasets with $n_0 = 40$ and $n_0 = 80$ are randomly sampled from the VIRAHEP-C data ($n = 401$). Estimates $\hat{\mathbb{P}}$ and $\hat{\mathbb{P}}^{(b)}$ for $b = 1, \dots, 200$ in Sect. 2.2.2 are obtained using G-S and E-P survival functions from each pilot dataset. These parameter estimates optimize the distance metric,

$$\sum_{j=1}^{8} w(j)\{S_P^*(j) - \hat{S}^*(j)\}^2,$$

where the weight $w(j)$ is $\{S_P^*(j)\}^{\rho_1}\{1 - S_P^*(j)\}^{\rho_2}$.

The estimates of parameters for $\rho_1 = \rho_2 = 0.5$ are $\hat{\mathbb{P}} = (\omega_1 = 0.018,\ \omega_2 = 21.425,\ \tau_0 = 0.169)$ for the pilot data of 40 subjects ($n_0 = 40$), and $\hat{\mathbb{P}} = (\omega_1 = 0.007,\ \omega_2 = 0.14,\ \tau_0 = 0.355)$ for the pilot data with 80 subjects ($n_0 = 80$). However, the estimated parameters for $\rho_1 = 0$ and $\rho_2 = 1$ are $\hat{\mathbb{P}} = (\omega_1 = 0.036,\ \omega_2 = 21.155,\ \tau_0 = 0.111)$ and $\hat{\mathbb{P}} = (\omega_1 = 0.015,\ \omega_2 = 0.168,\ \tau_0 = 0.291)$ for pilot datasets of $n_0 = 40$ and $n_0 = 80$, respectively.

Using the aforementioned parameter estimates, Table 3 provides approximated G-S survival function (Approximated), the standard errors (SE) by the methods in Eq. (17), 95% confidence interval (95% CI) by the method in Eq. (11), and 95% confidence interval (95% CI-Boots) based on Bootstrapping method. Table 3 shows that (i) for both ($\rho_1 = \rho_2 = 0.5$) and ($\rho_1 = 0,\ \rho_2 = 1$), the difference between the G-S and approximated G-S survival rates are less than 0.015 except at $t_6$ for $n_0 = 80$; and (ii) the approximated G-S survival rates for ($\rho_1 = \rho_2 = 0.5$) and ($\rho_1 = 0,\ \rho_2 = 1$) are robust, especially even for the small sample of $n_0 = 40$.

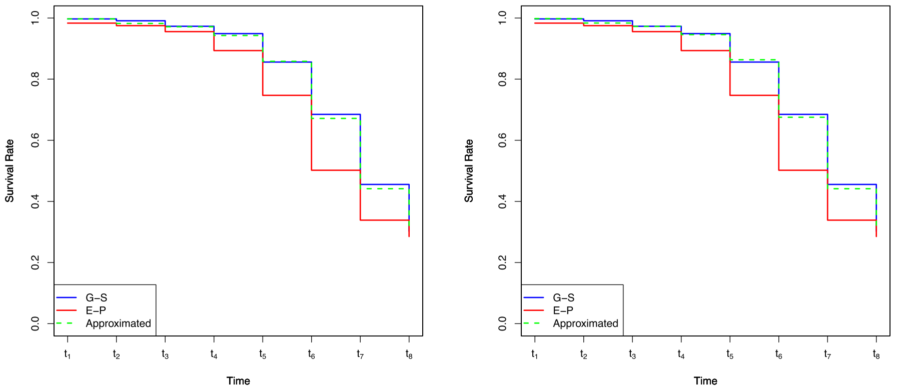
Figure 3 shows G-S, E-P, and approximated G-S survival functions of the analysis dataset using parameter estimates from the pilot dataset ($n_0 = 40$). The distances between the E-P and G-S survival rates at each time point are due to mismeasured outcomes. The G-S survival function, approximated G-S survival function, as well as the 95% confidence bands for each time point are displayed in Fig. 4. We observe from Figs. 3 and 4 that our proposed method performs very well using the pilot dataset with $n_0 = 40$.

## 4.2 Analysis of ADNI Data

We further evaluate the new methods using data from the ADNI-1 and ADNI-GO segments of the ADNI study (Wiener, 2012 [20]). For the analyses, we consider annual outcomes of the clinical (as E-P) and biomarker (as G-S) diagnoses. At the time of the data extract (December 2, 2016), there were 755 non-AD subjects and 193 AD subjects according to clinical diagnosis. Of the 755 subjects, 185 subjects have both G-S and E-P outcomes (pilot data), 565 subjects have only the E-P outcome (analysis data), and 5 are excluded due to Assumption 2 (G-S occurred prior to E-P). Table 4 shows the KM estimates of G-S (G-S) and E-P (E-P), and the standard

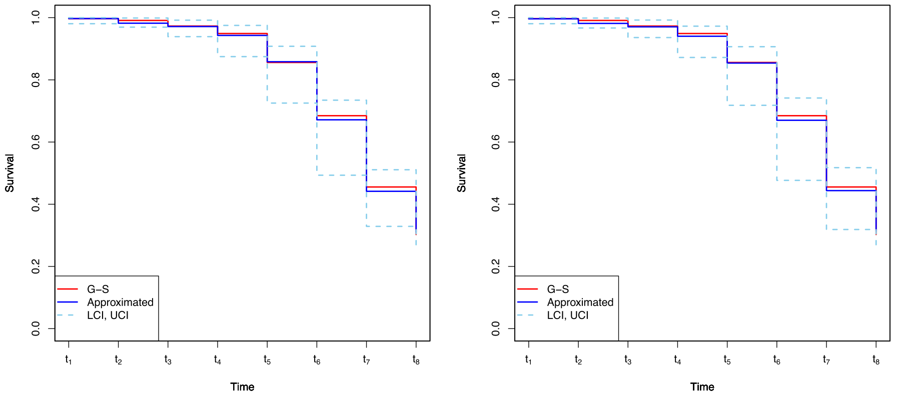
**Table 3** Data analyses results using analysis data and estimated parameters from pilot data with $n_0 = 40$ and $n_0 = 80$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$, and for $\rho_1 = 0$ and $\rho_2 = 1$

| $n_0 = 40$ | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time | G-S | E-P | $(\rho_1 = 0.5,\ \rho_2 = 0.5)$ | | | | | | $(\rho_1 = 0,\ \rho_2 = 1)$ | | | | | |
| | | | Approx. | SE | 95%CI | | 95 %CI-Boots | | Approx. | SE | 95%CI | | 95 %CI-Boots | |
| $t_1$ | 0.997 | 0.983 | 0.997 | 0.005 | 0.891 | 1 | 0.981 | 1 | 0.998 | 0.005 | 0.625 | 1 | 0.983 | 1 |
| $t_2$ | 0.991 | 0.975 | 0.982 | 0.008 | 0.956 | 0.993 | 0.970 | 0.999 | 0.984 | 0.008 | 0.956 | 0.994 | 0.970 | 0.999 |
| $t_3$ | 0.973 | 0.956 | 0.971 | 0.012 | 0.937 | 0.987 | 0.939 | 0.992 | 0.972 | 0.012 | 0.937 | 0.988 | 0.950 | 0.992 |
| $t_4$ | 0.949 | 0.894 | 0.943 | 0.022 | 0.879 | 0.973 | 0.875 | 0.975 | 0.946 | 0.022 | 0.881 | 0.976 | 0.888 | 0.974 |
| $t_5$ | 0.856 | 0.747 | 0.858 | 0.042 | 0.752 | 0.921 | 0.725 | 0.908 | 0.863 | 0.041 | 0.759 | 0.925 | 0.738 | 0.910 |
| $t_6$ | 0.685 | 0.502 | 0.672 | 0.060 | 0.538 | 0.774 | 0.493 | 0.735 | 0.675 | 0.059 | 0.544 | 0.776 | 0.498 | 0.737 |
| $t_7$ | 0.455 | 0.339 | 0.442 | 0.049 | 0.344 | 0.534 | 0.329 | 0.511 | 0.441 | 0.049 | 0.344 | 0.535 | 0.329 | 0.508 |
| $t_8$ | 0.303 | 0.285 | 0.311 | 0.030 | 0.253 | 0.370 | 0.252 | 0.359 | 0.309 | 0.031 | 0.249 | 0.371 | 0.255 | 0.364 |
| $n_0 = 80$ | | | | | | | | | | | | | | |
| Time | G-S | E-P | $(\rho_1 = 0.5,\ \rho_2 = 0.5)$ | | | | | | $(\rho_1 = 0,\ \rho_2 = 1)$ | | | | | |
| | | | Approx. | SE | 95%CI | | 95 %CI-Boots | | Approx | SE | 95%CI | | 95 %CI-Boots | |
| $t_1$ | 0.997 | 0.981 | 0.993 | 0.004 | 0.977 | 0.998 | 0.983 | 1 | 0.995 | 0.004 | 0.976 | 0.999 | 0.983 | 1 |
| $t_2$ | 0.990 | 0.972 | 0.989 | 0.007 | 0.959 | 0.997 | 0.969 | 0.998 | 0.991 | 0.007 | 0.959 | 0.998 | 0.972 | 0.999 |
| $t_3$ | 0.970 | 0.950 | 0.976 | 0.010 | 0.944 | 0.990 | 0.950 | 0.991 | 0.979 | 0.010 | 0.945 | 0.992 | 0.955 | 0.992 |
| $t_4$ | 0.946 | 0.893 | 0.945 | 0.016 | 0.905 | 0.969 | 0.910 | 0.971 | 0.950 | 0.016 | 0.908 | 0.973 | 0.915 | 0.973 |
| $t_5$ | 0.859 | 0.748 | 0.855 | 0.025 | 0.799 | 0.897 | 0.807 | 0.895 | 0.862 | 0.024 | 0.807 | 0.902 | 0.811 | 0.898 |
| $t_6$ | 0.690 | 0.504 | 0.654 | 0.036 | 0.578 | 0.720 | 0.575 | 0.715 | 0.656 | 0.035 | 0.582 | 0.719 | 0.574 | 0.713 |
| $t_7$ | 0.460 | 0.345 | 0.449 | 0.037 | 0.375 | 0.520 | 0.358 | 0.513 | 0.446 | 0.037 | 0.372 | 0.518 | 0.367 | 0.515 |
| $t_8$ | 0.309 | 0.286 | 0.309 | 0.030 | 0.252 | 0.367 | 0.253 | 0.362 | 0.305 | 0.030 | 0.247 | 0.364 | 0.248 | 0.365 |



**(a)** $\rho_1 = 0.5$ and $\rho_2 = 0.5$ **(b)** $\rho_1 = 0$ and $\rho_2 = 1$

**Fig. 3** The survival functions of analysis dataset using pilot dataset with $n_0 = 37$



**(a)** $\rho_1 = 0.5$ and $\rho_2 = 0.5$ **(b)** $\rho_1 = 0$ and $\rho_2 = 1$

**Fig. 4** G-S and approximated G-S survival functions of analysis dataset, and 95 % CI's using pilot data with $n_0 = 37$

errors of G-S (SE of G-S) and E-P (SE of E-P) using the pilot data. These estimates optimize the distance metric

$$\left\{ \sum_{j=1}^{8} w(j)\left(S_P(j) - \hat{S}(j)\right)^2 \right\},$$

where the weights $w(j)$ are defined as $\{S_p(j)\}^{\rho_1}\{1 - S_p(j)\}^{\rho_2}$ for $0 \leq \rho_1,\ \rho_2 \leq 1$ and $j = 2, \ldots, 8$, and $w(1) = w(2)$. The estimated parameters for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ are $\hat{\mathbb{P}} = (\psi_1 = 0.007,\ \psi_2 = 0.000,\ \gamma_0 = 0.438)$, and for $\rho_1 = 0$ and $\rho_2 = 1$ are $\hat{\mathbb{P}} = (\psi_1 = 0.008,\ \psi_2 = 0.000,\ \gamma_0 = 0.440)$. As no event is observed at $t_1$ based on E-P test, constant weights at the early visit are used. The approximated G-S survival

**Table 4** KM Survival Estimates based on ADNI Pilot dataset

| Ttime | E-P | SE of E-P | G-S | SE of G-S | $(\rho_1 = 0.5,\ \rho_2 = 0.5)$ Approximated | $(\rho_1 = 0,\ \rho_2 = 1)$ Approximated |
|---|---|---|---|---|---|---|
| $t_1$ | 1 | 0 | 0.405 | 0.036 | 0.438 | 0.440 |
| $t_2$ | 0.924 | 0.019 | 0.400 | 0.036 | 0.402 | 0.404 |
| $t_3$ | 0.843 | 0.027 | 0.389 | 0.036 | 0.360 | 0.361 |
| $t_4$ | 0.786 | 0.030 | 0.332 | 0.037 | 0.326 | 0.326 |
| $t_5$ | 0.759 | 0.032 | 0.318 | 0.038 | 0.302 | 0.300 |
| $t_6$ | 0.751 | 0.033 | 0.301 | 0.040 | 0.283 | 0.279 |
| $t_7$ | 0.705 | 0.038 | 0.215 | 0.059 | 0.249 | 0.244 |
| $t_8$ | 0.662 | 0.043 | 0.215 | 0.059 | 0.216 | 0.210 |

Pilot data consist of 185 subjects (of a total of 755) that had both G-S and E-P outcomes

functions of the pilot dataset for $\rho_1 = 0.5$ and $\rho_2 = 0.5$, and for $\rho_1 = 0$ and $\rho_2 = 1$ are obtained using the proposed model in (8). Our proposed model works well. The approximated G-S survival rates show that (i) the difference between the G-S and approximated G-S survival rates are less than 0.035; and (ii) the approximated G-S survival rates for $(\rho_1 = 0.5,\ \rho_2 = 0.5)$ and $(\rho_1 = 0.5,\ \rho_2 = 0.5)$ are robust (difference of approximated G-S survival rates < 0.007). For the analysis dataset, the approximated G-S survival rates are obtained using the $\hat{\mathbb{P}}$ based on pilot data and the estimated E-P survival rates from the analysis dataset. The estimated standard errors of G-S survival rates are estimated by the bootstrap approach in a similar fashion to Eq. (11). In Table 5, we present the KM estimates of E-P (E-P), the approximated G-S (Approximated), standard errors of G-S (SE), and the confidence intervals of G-S (LCI, and UCI). We now evaluate the performance of our proposed estimator in estimating the survival distribution of the G-S outcomes by using 3 parameters $(\psi_1,\ \psi_2,\ \gamma_0)$ and along with only the error-prone (E-P) outcomes. In Table 5, we see that the estimated E-P survival rates from the analysis dataset are consistent to

**Table 5** Survival Estimates based on ADNI Analysis dataset

| Time | E-P | $(\rho_1 = 0.5,\ \rho_2 = 0.5)$ | | | | $(\rho_1 = 0,\ \rho_2 = 1)$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Approximated | SE | LCI | UCI | Approximated | SE | LCI | UCI |
| $t_1$ | 1 | 0.438 | 0.019 | 0.401 | 0.474 | 0.440 | 0.020 | 0.401 | 0.478 |
| $t_2$ | 0.899 | 0.391 | 0.017 | 0.358 | 0.423 | 0.393 | 0.017 | 0.359 | 0.427 |
| $t_3$ | 0.780 | 0.333 | 0.013 | 0.307 | 0.360 | 0.334 | 0.014 | 0.308 | 0.361 |
| $t_4$ | 0.716 | 0.297 | 0.011 | 0.275 | 0.319 | 0.297 | 0.012 | 0.274 | 0.320 |
| $t_5$ | 0.671 | 0.267 | 0.012 | 0.243 | 0.291 | 0.265 | 0.015 | 0.237 | 0.294 |
| $t_6$ | 0.650 | 0.245 | 0.016 | 0.213 | 0.278 | 0.242 | 0.021 | 0.203 | 0.283 |
| $t_7$ | 0.617 | 0.218 | 0.022 | 0.176 | 0.262 | 0.214 | 0.028 | 0.162 | 0.270 |
| $t_8$ | 0.617 | 0.201 | 0.029 | 0.148 | 0.261 | 0.196 | 0.035 | 0.132 | 0.269 |

Analysis data consist of 565 subjects (of a total of 755) that had only E-P outcomes the estimated E-P survival rates from the pilot dataset (Table 4). It is reasonable to assume the distribution of G-S outcomes from the pilot data (randomly selected) ought to be consistent with the unobserved distribution of G-S outcomes from the analysis data. Our proposed survival rate estimator of the unobserved G-S outcomes performs well, as the approximated G-S survival rates between the pilot data and the analysis data are similar. The results are robust to the choice of weights.

## 5 Discussion

In this paper, we have developed a framework to correctly estimate the discrete-time survival function to handle mismeasured outcomes through the modeling of time-dependent NPV and PPV. Our discrete-time survival estimator (i) allows for the probability of correctly classifying a G-S event to increase with time in a specific manner given the occurrence of an E-P event; and (ii) allows for the inclusion of estimated model parameters $\hat{\mathbb{P}} = (\hat{\omega}_1, \hat{\omega}_2, \hat{\tau}_0)$ through a validation subsample ("pilot dataset"). This is very useful in scenarios for which $\hat{\mathbb{P}}$ is not known with confidence and must be estimated from the imminent clinical trial.

The prevalence does impact the PPV and NPV. For mismeasured outcomes caused by lower detection limit, we have that the G-S event cannot happen prior to the E-P event, this idea is formulated in Assumption 1 and given as $P(T^* \geq T) = 1$. Under this assumption, Proposition 2.1 provides the behavior of NPV and PPV over time, namely that (i) the NPV($\gamma_j$) at time $t_j$ is 1 for all $j = 1, 2, \ldots$; and (ii) the PPV($\tau_j$) at time $t_j$ is $\frac{1-P(E_j^*=0)}{1-P(E_j=0)}$ for $j = 1, 2, \ldots$. Therefore under Assumption 1, we see that the PPV, $\tau_j$, is dynamic across time. We applied (ii) to the G-S and E-P survival rates of Table 1 which has both low prevalence and high prevalence and observed that the PPV changes for every visit. In Table 1, we presented the true (or population) PPV. The true PPV can be calculated since we have G-S and E-P outcomes for all. Here, we observed the true PPV is dynamic, with a value of 0.402 at $t_1$, and 0.871 at $t_8$. In addition, we provide approximated PPV which are based on our approximated G-S survival rates for different choices of $\rho_1$ and $\rho_2$. Additionally, by similar arguments as stated above, the PPV can be calculated under the Brownian motion process of Table 2 and the VIRAHEP-C data of Table 3. These results are a motivating reason for conducting our research with non-constant or dynamic PPV, it is in fact to adjust for low prevalence and high prevalence. Similar arguments can be made for dynamic NPV under Assumption 2 and Proposition 2.3.

With the assumption of a constant PPV using VIRAHEP-C data, Table S1 of the Supplementary Materials provides the approximated G-S survival function (Approximated), the standard errors (SE) by the method in (17), the 95% confidence interval (95% CI) by the method in (11), and the 95% confidence interval (95% CI-Boots) based on the Bootstrapping method. The results of Table S1 show that for both ($\rho_1 = 0.5,\ \rho_2 = 0.5$) and ($\rho_1 = 0,\ \rho_2 = 1$), the approximation of the G-S survival rates under the assumption of a constant PPV is not as good as the one under the time-varying PPV. Note that under the framework of a time-varying PPV (Table 3), we observe that the differences between the G-S survival rates and approximated G-S survival rates are less than 0.015 (except at $t_6$ for $n_0 = 80$).

To further illustrate the need for a time-varying PPV, we conduct a simulation study under the Brownian motion process similar to Sect. 3.2, except that now we fix PPV, the results are shown in Table S2 of the Supplementary Materials. In Table S2, we report the average of the approximated survival rates (Approximated), along with ASE, MCSE, and CP. We see that under the framework of a constant PPV, the approximated survival function does not perform well in estimating the G-S survival function as compared to the framework of the time-varying PPV. Therefore, the PPV is indeed time-varying, which further supports our proposed research.

To assess the merits of our estimator, we model the course of viral load through two stochastic processes, the gamma process and the Wiener process. For the gamma process as shown in Sect. 3.1, one obtains elegant closed-form expressions of the survival function. Such a result supports our approach of developing our methodology from a time-varying PPV perspective because it allows, through Proposition 2.1, the exact calculation of the PPV vector, the result of which is consistent with a time-dependent PPV model. In Sect. 3.2, we take a different viewpoint on viral load and model its course through a Wiener process. Since an analytical expression of the survival function under the Wiener process is unavailable, we examine the properties of our estimator through simulation studies. The findings of the simulation study under the Wiener process are important because it validates the mathematical results of Sect. 2.2.2. Namely, that our estimator performs very well even when $\mathbb{P}_1$ is estimated through a very small pilot dataset.

In the real data example of Sect. 4.1, we illustrate the performance of our estimator under the notion that the parameters of our model, $\mathbb{P} = (\omega_1, \omega_2, \tau_0)$ are unknown (as described in Sect. 2.2.2). To estimate $\mathbb{P}$, we constructed two sets of pilot data which are a random sample of 10% and 20% of the VIRAHEP-C data. The results of the data analyses showed that our approximated G-S survival function is a good fit to the G-S survival function. Our method which produces an approximated G-S survival distribution is consistent to G-S survival distribution of G-S outcomes, hence, demonstrating that our method works well with a pilot data size of 10% and 20% from the VIRAHEP-C study. However, to assess the robustness of the result, we used data from the ADNI study which is different in trial design, objectives, disease indication, and primary endpoint. The VIRAHEP-C study was a randomized, interventional clinical trial, while ADNI is an observational and non-interventional study. In the ADNI data, indeed of the 755 subjects, only 185 subjects have both G-S and E-P outcomes. The G-S outcome in ADNI represents the result of the cerebral spinal fluid (CSF) biomarker assay which involves a lumbar puncture, so it is often considered too invasive for many patients and therefore has limited availability. Given that the available sample (pilot data) constituted about 25% of the total data, there was no need for a smaller random sample. There is an inherent random selection in this pilot data since the likelihood of not getting a lumbar puncture does not depend on any study design, but on personal characteristics and circumstances. Therefore, we believe that the pilot data is a reasonable reflection of the targeted population. We found that pilot data of size 10% to about 20% performed well across both datasets. However, the size of the pilot dataset depends on the study in question, it is contingent on the number of events at each timepoint, and the study's total sample size.

We developed a new approach to derive the standard errors of $\hat{S}^*(j)$ in (14) due to the behavior of the variability of $\hat{\mathbb{P}} = (\hat{\omega}_1, \hat{\omega}_2, \hat{\tau}_0)$ under a small pilot dataset. We initially examined the utility of the delta method in estimating the standard errors of $\hat{S}^*(j)$, however, the results were unimpressive. Hence, the motivation to develop a new method to estimate the standard errors of $\hat{S}^*(j)$ by way of the standard variance decomposition approach (provided in (17)). The results of Table 2 in Sect. 3 show the impressive empirical performance of our variance estimate, as the differences between ASEs and MCSEs are less than 0.005.

Regarding the selection of weights, we recommend setting $\rho_1$ and $\rho_2 = 0.5$ as the survival estimates with $\rho_1$ and $\rho_2 = 0.5$ are robust. Table 1 applies our proposed method to a course of viral load modeled after a gamma process. The table gives the approximated G-S survival rates under 3 different weights, namely, (1) $\rho_1 = 1$ and $\rho_2 = 0$, (2) $\rho_1 = 0.5$ and $\rho_2 = 0.5$ , and (3) $\rho_1 = 0$ and $\rho_2 = 1$. If early in the study there are very few events as compared to later in the study, then placing a lower weight early on by setting $\rho_1$ close to 0 and $\rho_2$ closer to 1 could lead to a better approximation of the G-S survival rates for the earlier timepoints. However, if it is vice versa, where there are more events early and few later in the study, then setting $\rho_1$ close to 1 and $\rho_2$ close to 0 is reasonable. In Table 1, we see that the approximation is best at $t_1$ when $\rho_1 = 1$ and $\rho_2 = 0$, but worse at $t_8$. Whereas the approximation is best at $t_8$ when $\rho_1 = 0$ and $\rho_2 = 1$, but worse at $t_1$. We recommend setting $\rho_1$ and $\rho_2 = 0.5$ for a consistent performance over time.

A limitation of our work is that the estimation of the 3 unknown parameters requires a fraction (albeit small) of the data from the on-going study. An interesting further investigation would be to directly model the underlying stochastic process of the course of disease to serve as a bridge to between E-P and G-S survival functions. Such a development would be important and useful for the lower-limit detection problem. It would have important clinical implications in the design of clinical trials because it could be possible to directly calculate the probability of having an outcome at the design stage of a clinical trial. Thus, the timing and spacing of clinical trial visits could be more strategic. The work of Huang et al. [40] contains further insights, they directly modeled the course of viral load using latent variable and stochastic processes to capture the viral load between the predetermined discrete time points and the dependency of binary response over time. Under investigation is the extension of our work to develop Bayesian methods via the logistic regression model for the G-S hazard at each time point $t_j$.

We have conducted extensive research of the performance of our methods using stochastic processes, namely a gamma process and a Wiener process. Our methodology performed well. Furthermore, we evaluated the performance of our methods using observed data from a multi-national HCV clinical trial and from the ADNI studies. Our proposed method with only 3 unknown parameters works well in approximating the G-S survival function. We demonstrated that our proposed methodology works well under the lower-limit detection framework. Early detection for serious diseases is important, our method offers a way to conduct time-to-event analyses that can be generalized to the broader population based on a validation subsample.

## Appendix A: Proofs of Propositions, Lemmas, and Theorems

**Proof of Lemma 2.1**

$$\begin{aligned} S^*(j) &= Pr(T^* > t_j | T \le t_j)P(T \le t_j) + Pr(T^* > t_j | T > t_j)P(T > t_j) \\ &= \{1 - Pr(T^* \le t_j | T \le t_j)\}\{1 - P(T > t_j)\} + Pr(T^* > t_j | T > t_j)P(T > t_j). \end{aligned}$$

Using the definition of PPV and NPV in (2), we obtain Lemma 2.1. □

**Proof of Proposition 2.1** For (i), it is trivial by considering j=1, k=1 in 4. For (ii), Using iv) in Remark 1, it can be easily obtained. For (iii), it is obvious for the first part of (iii) and for the last part, since $P(T^* \le t_j | T = t_k) = 1 - \left\{1 - \tau_0\right\}^{(t_j-1)\omega_1+1}$ for all $t_k \le t_j$, which is constant for $t_k$ and $\sum_{k=1}^{j} P(T = t_k) = P(T \le t_j)$, the proof is done. For (iv), it is trivial by considering $P(T^* \le t_j | T = t_k) = \tau_0$ for the proof of (iii). □

**Proof of Proposition 2.3** (i) We can rewrite $\gamma_j(\boldsymbol{x}) = P(T^* > t_j | T > t_j)$ as $\gamma_j = 1 - P(T^* \le t_j | T > t_j)$. Under Assumption 1, $P(T^* \le t_j | T > t_j) = 0$. (ii) $\tau_j = P(T^* \le t_j | T \le t_j) = \frac{P(T^* \le t_j, T \le t_j)}{P(T \le t_j)}$. Under Assumption 1, we have $P(T^* \le t_j, T \le t_j) = P(T^* \le t_j)$, which completes the proof. □

**Proof of Theorem 2.3** By Theorem 5 of [36], let $t_J < \infty$ satisfy $1 - S(j|\boldsymbol{x}) < 1$. Then the random variable $\sqrt{n}(\hat{S}(j) - S(j))$, for $0 < j < J$, converges weakly to a mean zero normal random variable $Z_j$. Moreover, $\text{Cov}\,(Z_j, Z_k) = S(j)S(k)\sum_{m=0}^{j}(S(m))^{-2}(1 - H(t_j))^{-1}P(E_m = 1), j \le k$ where $1 - H(t_j)$ is the right censoring distribution function. Since $\hat{S}^*(j)$ is a linear combination of the KM estimator, it therefore follows an asymptotically normal distribution. □

**Proof of Lemma 3.1** For n=2, $B_2(c, x) = \int_{c+1}^{x} B_1(c, u)du = x - (c + 1) = \frac{(x-c)}{1!} - 1$, which is G-S. Suppose it is G-S for $n = k$, then $B_{c,\,k+1}(x)$ is obtained as follows:

$$\begin{aligned} B_{k+1}(c, x) = \int_{c+k}^{x} B_k(c, u)du &= \int_{c+k}^{x} \frac{(u-c)^{k-1}}{(k-1)!} - \frac{(u-c)^{k-2}}{(k-2)!}du \\ &= \frac{(x-c)^k}{k!} - \frac{k^k}{k!} - \frac{(x-c)^{k-1}}{(k-1)!} + \frac{k^{k-1}}{(k-1)!}. \end{aligned}$$

Since $-\frac{k^k}{k!} + \frac{k^{k-1}}{(k-1)!} = \frac{k^k - k^k}{k!} = 0$, we obtain that $B_{k+1}(c, x) = \frac{(x-c)^k}{k!} - \frac{(x-c)^{k-1}}{(k-1)!}$, which is G-S for $n = k + 1$. By induction, we complete the proof. □

**Proof of Theorem 3.1** We have $S_c(1) = \exp\{-(c + j)\}$, which implies it is G-S for $j = 1$. To prove for $j \ge 2$, let $G_n(a) = \int_a^\infty u^n \exp(-u)du$. Then, it is easily obtained that $G_n(a) = a^n \exp(-a) + nG_{n-1}(a) = \sum_{m=0}^{n} \frac{n!}{(n-m)!} a^{(n-m)} \exp(-a)$. Using that fact and the Lemma 3.1, we have

$$S_c(j) = \int_{c+j}^{\infty} \exp(-w_j) B_j(c, w_j) dw_j = \int_{c+j}^{\infty} \exp(-w_j) \Big[ \frac{(w_j - c)^{j-1}}{(j-1)!} - \frac{(w_j - c)^{j-2}}{(j-2)!} \Big] dw_j$$
$$= \sum_{m=0}^{j-1} \frac{j^m}{m!} \exp\{-(c+j)\} - \sum_{m=0}^{j-2} \frac{j^m}{m!} \exp\{-(c+j)\} = \frac{j^{(j-1)}}{(j-1)!} \exp\{-(c+j)\}.$$

## Appendix B: Upper-Limit Detection Problem Under Gamma Process

Now, we discuss the upper detection limit problem focused on gamma process and derive lemmas, theorem, and corollary for that. Under the gamma process discussed in Sect. 3.1, if we consider the upper limit of detection level as $c$, then survival function at time $t_j$ is

$$S^c(j) = P(X_1 \le 1 + c, X_1 + X_2 \le 2 + c, \dots, X_1 + \cdots + X_j \le j + c).$$

Using the same transformation for the low limit detection problem from $(X_1, \dots, X_j)$ to $(W_1, \dots, W_j)$, where $W_j = X_1 + \cdots + X_j$ for $j = 1, 2, \dots$, we can rewrite $S^c(j)$ as

$$S^c(j) = \int_0^{c+1} \int_0^{c+2} \cdots \int_0^{c+j} \exp(-w_j) 1(w_1 \le \cdots \le w_{j-1} \le w_j) dw_j \cdots dw_2 dw_1$$
$$= \int_0^{c+1} \int_{w_1}^{c+j-(j-2)} \cdots \int_{w_{(j-1)}}^{c+j} \exp(-w_j) dw_j \cdots dw_2 dw_1.$$

Define a new sequence of random variables as $Y_1 = W_j, \dots, Y_j = W_1$. Then, $S^c(j)$ is given by

$$S^c(j) = \int_0^{c+1} \int_{y_j}^{c+j-(j-2)} \cdots \int_{y_2}^{c+j} \exp(-y_1) dy_1 \cdots dy_{(j-1)} dy_j.$$

To obtain a general expression of $S^c(j)$, define a new sequential function $U_n(y, b)$ as

$$U_n(y, b) = \int_y^{b-(n-2)} U_{(n-1)}(z, b) dz, \tag{B.1}$$

for $n = 2, 2, \dots$, where $U_1(y, b) = \exp(-y)$. We derive an iterative expression for $U_n(y, b)$ in Lemma B.1.

**Lemma B.1** *The sequence $U_n(y, c+n)$ can be expressed as*

$$U_n(y, b) = U_{n-1}(y, b-1) + \exp(-b) \Big[ \frac{(b-y)^{(n-3)}}{(n-3)!} I(n \ge 3) - \frac{(b-y)^{(n-2)}}{(n-2)!} \Big],$$

*for* $n = 2, 3, \ldots,$ *where* $U_1(y, b) = \exp(-y)$ *and* $I(a)$ *is the indicator function that takes a value of 1 if a is G-S and 0 otherwise.*

From (B.1), we have

$$\begin{aligned} U_n(y_n, b) = & \int_{y_n}^{b-(n-2)} U_{n-1}(y_{n-1}, b) dy_{n-1} \\ = & \int_{y_n}^{b-(n-2)} \cdots \int_{y_2}^{b} \exp(-y_1) dy_1 \cdots dy_{n-1}, \end{aligned}$$

for $n = 2, \ldots$ where $U_1(y_1, b) = \exp(-y_1)$ and $S^c(j) = \int_0^{c+1} U_j(y, c+j) dy$ for $j = 1, \ldots$. Using Lemma B.1 , we discuss a sequential relationship of the survival function, $S^c(j)$ in the next Lemma.

**Lemma B.2** *Suppose that the* $X_j$ *are i.i.d. from* Gamma $(1, 1)$ for $j = 1, 2, 3 \ldots$. *Then, the survival function at time* $t_j$ *with an upper detection limit c has the relationship as follows:*

$$\begin{aligned} S^c(j) = & P(X_1 \le 1 + c, X_1 + X_2 \le 2 + c, \ldots, X_1 + \cdots + X_j \le j + c) \\ = & S^c(j-1) - \frac{\exp\{-(c+j)\}}{(j-1)!} \left[ (c+j)^{(j-2)} \{(c+j) - (j-1) I(j \ge 3)\} \right] \\ & + \frac{\exp\{-(c+j)\}}{(j-1)!} \left[ (j-1)^{(j-1)} I(j = 2) \right], \end{aligned}$$

*for* $j = 1, 2, \ldots$ *where* $S^c(0) = 1$, *an explicit expression of the survival function is given in the following theorem.*

**Theorem B.2** *Suppose that* $X_j$ *is i.i.d. from* Gamma $(1, 1)$ *for* $j = 1, \ldots$. *Then, the survival function at time* $t_j$ *with an upper detection limit c is given by*

$$S^c(j) = 1 - \sum_{k=1}^{j} \left[ \frac{\exp\{-(c+k)\}}{(k-1)!} \left\{ (c+1)(c+k)^{(k-2)} \right\}^{I(k \ge 2)} \right].$$

The proof of Theorem B.2 directly follows from $S^c(0) = 1$ and basic algebra.

Similar to the low limit detection problem, we can extend Theorem B.2 with $X_j \sim$ Gamma $(1, \lambda)$. Since $\frac{X_j}{\lambda} \sim$ Gamma $(1, 1)$, we have

$$\begin{aligned} S^{(c,\lambda)}(j) = & P(X_1 \le \lambda + c, X_1 + X_2 \le 2\lambda + c, \ldots, X_1 + \cdots + X_j \le j\lambda + c) \\ = & P\left( \frac{X_1}{\lambda} \le 1 + \frac{c}{\lambda}, \frac{X_1 + X_2}{\lambda} \le 2 + \frac{c}{\lambda}, \ldots, \frac{X_1 + \cdots + X_j}{\lambda} \le j + \frac{c}{\lambda} \right). \end{aligned} \tag{B.2}$$

Using (B.2), $S^{(c,\lambda)}(j)$ with $X_j \sim$ Gamma $(1, \lambda)$ can be obtained in the next corollary.

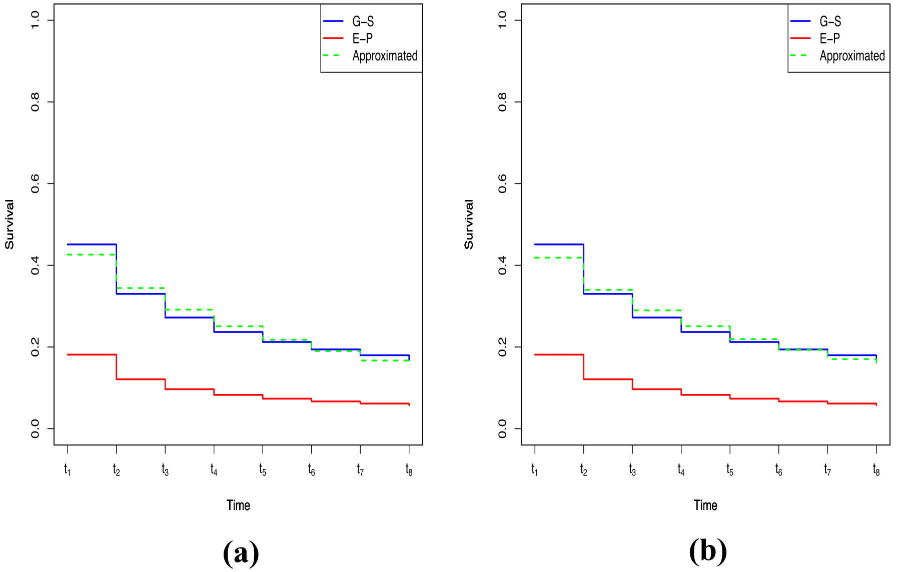


**Fig. 5** G-S, E-P, and approximated G-S survival functions with upper detection limits with $c^* = -0.4$ and $c = -0.8$ for $\rho_1 = 0.5$ and $\rho_2 = 0.5$ (**a**), and for $\rho_1 = 0$ and $\rho_2 = 1$ (**b**)

**Corollary B.1** *Suppose that the $X_j$ are i.i.d. from* Gamma $(1, \lambda)$ *for $j = 1, \dots$. Then, the survival function at time $t_j$ with an upper detection limit c is given by*

$$S^{(c,\lambda)}(j) = P(X_1 \le \lambda + c, X_1 + X_2 \le 2\lambda + c, \dots, X_1 + \cdots + X_j \le j\lambda + c) = S^{\frac{c}{\lambda}}(j).$$

Within the upper-limit detection framework, if the upper limit of the G-S test is above that of the E-P ftest, then $P(T^* \ge T)$ in Assumption 1 is reasonable, therefore $P(T^* \ge T) = 1$ if $c^* \ge c$.

To examine our proposed model in (4) for the upper-limit detection problem, for which $P(T^* \ge T) = 1$, we consider two different detection limits for G-S and E-P events as $c^* = -0.4$, and $c = -0.8$. After which, we obtain the approximated survival function using the optimal values of $(\omega_1, \omega_2, \tau_0)$, which minimize (13) for $\rho_1 = 0.5$ and $\rho_2 = 0.5$, and for $\rho_1 = 0$ and $\rho_2 = 1$ for $w(k)$. The pairs of estimates are (0.135, 0.000, 0.701) and (0.124, 0.000, 0.710) for $\rho_1 = 0.5$ and $\rho_2 = 0.5$, and for $\rho_1 = 0$ and $\rho_2 = 1$, respectively. Figure 5 shows G-S (blue solid line), E-P (red solid line), and approximated (green dashed line) G-S survival function. Similar to Fig. 1, the approximated G-S survival function is much closer to the G-S survival function than the E-P survival function. The observation from Fig. 5 confirms that the proposed models in (4) work well for the upper detection limit problem under the gamma process.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s12561-021-09317-3.

**Acknowledgements** Dr. M.-H. Chen's research was partially supported by NIH grants #GM70335 and #P01CA142538. Data collection and sharing for this project was funded by the ADNI(National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. A special thanks to Danielle J Harvey for her support in helping us navigate the ADNI database.

## References

1. Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K et al (2014) Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol 13:614–629
2. Raman MR, Preboske GM, Przybelski SA, Gunter JL, Senjem ML, Vemuri P et al (2014) Antemortem MRI findings associated with microinfarcts at autopsy. Neurology 82:1951–1958
3. Johnson KA, Fox NC, Sperling RA, Klunk WE (2012) Brain imaging in Alzheimer disease. Cold Spring Harb Perspect Med 2:a006213
4. Fearing MA, Bigler ED, Norton M, Tschanz JA, Hulette C, Leslie C, Welsh-Bohmer K, Investigators Cache County (2007) Autopsy-confirmed Alzheimer's disease versus clinically diagnosed Alzheimer's disease in the Cache County Study on Memory and Aging: a comparison of quantitative MRI and neuropsychological findings. J Clin Exp Neuropsychol 23:553–560
5. Khuroo MS, Khuroo NS, Khuroo MS (2014) Accuracy of rapid point-of-care diagnostic tests for hepatitis B surface antigen: a systematic review and meta-analysis. J Clin Exp Hepatol 4:226–240
6. Franzeck FC, Ngwale R, Msongole B, Hamisi M, Abdul O, Henning L et al (2013) Viral hepatitis and rapid diagnostic test based screening for HBsAg in HIV-infected patients in rural Tanzania. PLoS ONE 8:e58468
7. Kamili S, Drobeniuc J, Araujo AC, Hayden TM (2012) Laboratory diagnostics for hepatitis C virus infection. Clin Infect Dis 55:S43-48
8. Conjeevaram HS, Fried MW, Jeffers LJ, Terrault NA, Wiley-Lucas TE, Afdhal N et al (2006) Peginterferon and ribavirin treatment in African American and Caucasian American patients with hepatitis C genotype 1. Gastroenterology 131:470–477
9. Floquet A, Vergote I, Colombo N, Fiane B, Monk BJ, Reinthaller A et al (2015) Progression-free survival by local investigator versus independent central review: comparative analysis of the AGO-OVAR16 trial. Gynecol Oncol 136:37–42
10. Dodd LE, Korn EL, Freidlin B, Gray R, Bhattacharya S (2011) An audit strategy for progression-free survival. Biometrics 67:1092–1099
11. Amit O, Bushnell W, Dodd L, Roach N, Sargent D (2010) Blinded independent central review of the progression-free survival endpoint. Oncologist 1:492–495
12. U.S. Department of Health and Human Services (2007) FDA guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics, May
13. Kalali B, Formichella L, Gerhard M (2015) Diagnosis of Helicobacter pylori: changes towards the Future. Diseases 3:122–135

14. Queiroz DM, Harris PR, Sanderson IR, Windle HJ, Walker MM, Rocha AMC (2013) Iron status and Helicobacter pylori infection in symptomatic children: an international multi-centered study. PLoS ONE 8:e68833
15. Choi J, Kim CH, Kim D, Chung SJ, Song JH, Kang JM et al (2011) Prospective evaluation of a new stool antigen test for the detection of Helicobacter pylori, in comparison with histology, rapid urease test, (13)C-urea breath test, and serology. J Gastroenterol Hepatol 26:1053–1059
16. Gatta L, Ricci C, Tampieri A, Osborn J, Perna F, Bernabucci V et al (2006) Accuracy of breath tests using low doses of 13C-urea to diagnose Helicobacter pylori infection: a randomised controlled trial. Gut 55:457–462
17. Gisbert JP, de la Morena F, Abraira V (2006) Accuracy of monoclonal stool antigen test for the diagnosis of H. pylori infection: a systematic review and meta-analysis. Am J Gastroenterol 101:1921–1930
18. Joeng H-K, Chen M-H, Kang S (2016) Proportional exponentiated link transformed hazards (ELTH) models for discrete time survival data with application. Lifetime Data Anal 22:38–62
19. Shaw LM, Vanderstichele H, Knapik-Czajka M et al (2009) Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Ann Neurol 65:403–413
20. Wiener MW, Veitch DP, Aisen PS et al (2012) The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimer's Dement 8:S1–S68
21. Adeniji AK, Belle AH, Wahed AS (2014) Incorporating diagnostic accuracy into the estimation of discrete survival function. J Appl Stat 41:60–72
22. Racine-Poon AH, Hoel DG (1984) Nonparametric estimation of the survival function when cause of death is uncertain. Biometrics 40:1151–1158
23. Snapinn SM (1999) Survival analysis with uncertain endpoints. Biometrics 54:209–218
24. Richardson BA, Hughes J (2000) Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. Biostatistics 1:341–354
25. McKeown K, Jewell NP (2010) Misclassification of current status data. Lifetime Data Anal 16:215–230
26. Cappaso V (1993) Mathematical structures of epidemic systems, 2nd edn. Lectures notes in biomathematics. Springer, Berlin
27. Lloyd AL (2001) Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. Theor Popul Biol 60:59–71
28. Anderson D, Watson R (1980) On the spread of a disease with gamma distributed latent and infectious periods. Biometrika 67:191198
29. Hethcote HW, Tudor DW (1980) Integral equation models for endemic infectious diseases. J Math Biol 9:37–47
30. Wearing HJ, Rohani P, Keeling MJ (2005) Appropriate models for the management of infectious diseases. PLoS Med 2:e174
31. Feng Z, Xu D, Zhao H (2007) Epidemiological models with non-exponentially distributed disease stages and applications to disease control. Bull Math Biol 69:1511–1536
32. Hernandez-Ceron N, Feng Z, Castillo-Chavez C (2013) Discrete epidemic models with arbitrary stage distributions and applications to disease control. Bull Math Biol 75:1716–1746
33. Krylova O, Earn DJD (2013) Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. J R Soc Interface 10:20130098
34. Vergu E, Busson H, Ezanno P (2010) Impact of the infection period distribution on the epidemic spread in a metapopulation model. PLoS ONE 5:e9371
35. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
36. Breslow N, Crowley J (1974) A large sample study of the life table and product limit estimates under random censorship. Ann Stat 2:437–453
37. Greenwood M (1926) The natural duration of cancer (Reports on Public Health and Medical Subjects 33). Her Majesty's Stationery Office, , LondonLondon, pp 1–26
38. Borgan Ø, Liestøl K (1990) A note on confidence intervals and bands for the survival curve based on transformations. Scand J Stat 17:35–41
39. Gill RD (1983) Large sample behavior of the product-limit estimator on the whole line. Ann Stat 11:49–58
40. Huang P, Chen M-H, Sinha D (2009) A latent model approach to define event onset time in the presence of measurement error. Stat Interface 2:425–435