RESEARCH ARTICLE

Model-Assisted Estimators for Time-to-Event Data from Complex Surveys

Benjamin M. Reist*[1] | Richard Valliant[2]

[1]Office of the CIO, National Aeronautics
 and Space Administration, Washington DC,
 USA

[2]Survey Research Center, University of
 Michigan, Ann Arbor MI, USA

Correspondence
*Benjamin M. Reist, Data Scientist, NASA,
Washington, DC 20546. Email:
Benjamin.M.Reist@nasa.gov

Summary

We develop model-assisted estimators for complex survey data for the proportion
of a population that experienced some event by a specified time t. Theory for the
new estimators uses time-to-event models as the underlying framework but have both
good model-based and design-based properties. The estimators are compared in a
simulation to traditional survey estimation methods and are also applied to a study
of nurses' health. The new estimators take advantage of covariates predictive of the
event and reduce standard errors compared to conventional alternatives.

KEYWORDS:
doubly robust, general difference estimator, model calibrated estimator, time-to-failure model

1 | INTRODUCTION

We use time-to-event models to develop model-assisted estimators for complex survey data that can be used to estimate the
proportion of the population that have experienced an event by some time t. Such models are staples of survival analysis
in medical studies. Many complex surveys also collect the time at which a sampled unit experiences a given event. As an
example, consider the National Longitudinal Study of 1972 conducted by the US National Center for Education Statistics
(https://nces.ed.gov/surveys/nls72/), which surveyed a nationally representative sample of high school 12th graders. One item
collected during follow-up interviews was the date after graduation at which each respondent was hired for his or her first full-
time job. From this, we can estimate the proportion of people who were 12th graders in 1972 and were hired within five years
of graduation. Another example is the Survey of Income and Program Participation (SIPP), which measures how long individu-
als participate in various government assistance programs like Medicaid, Supplemental Nutrition Assistance Program (SNAP),
Housing Assistance, Supplemental Security Income (SSI), and Temporary Assistance for Needy Families (TANF)[1]. The Panel
Study of Income Dynamics (PSID, https://psidonline.isr.umich.edu/), conducted by the University of Michigan, is another longi-
tudinal survey that has collected data since 1968 on health, employment, income, wealth, expenditures, marriage, childbearing,
child development, philanthropy, and education. The Health and Retirement Study (HRS, http://hrsonline.isr.umich.edu/) is
also a large longitudinal, panel survey done by the University of Michigan to collect data on health status, aging, income, and
biomarkers. Many different endpoints can be derived from both PSID and HRS that can be used in time-to-event modeling.

We assume that a single-stage probability sample has been selected. The size of the finite population is N, s is the set of units
sampled from the population, π_i is the probability of selection for unit i, and the basic weight assigned to unit i is $d_i = \pi_i^{-1}$.
The proportion of a given population that has experienced an event by time t is $p_N(t) = N^{-1} \sum_{i=1}^{N} I_{\{T_i \leq t\}}$ where T_i is the time
at which the event happened for unit i, and I is the indicator function. The population proportion, p_N, can also be thought
of as a type of cumulative distribution function. This proportion can be estimated using a π-estimator as follows:

$$\hat{p}_\pi(t) = N^{-1} \sum_{i \in s} d_i I_{\{T_i \leq t\}}. \tag{1}$$

We assume that the survey closes out before all units have experienced a given event, i.e., T_i is only observed for $T_i \leq t_o$ where t_o is the time at which the survey ends. This means that T_i is right censored for units for which $T_i > t_o$, in which case the π-estimator will not correctly estimate $p_N(t)$ for $t > t_o$. Additionally this estimator cannot be used if any units are censored before time t.

Although research on survival analysis is abundant, literature for estimating survival models from complex survey data is much more limited. Binder[2] studied point and design-based variance estimation of the regression parameter in a Cox proportional hazards model using data from a complex survey. Lin[3] formalized the theory for some of Binder's heuristic results. Boudreau and Lawless[4] extended work on the Cox model to stratified, cluster sampling. There has been work in the case control literature for accelerated failure time models by Kong et al.[5], Kong and Cai[6], and Chiou et al.[7,8] for simple random samples, stratified simple random samples, and stratified simple random cluster samples. These methods use superpopulation models to develop theory and not the design-based approach we take here. Finally, we note that Heeringa, et al.[9] reviewed some of the software options available for survival analysis from survey data, including Kaplan-Meier estimation.

We propose estimators of the cumulative distribution of event times while accounting for the right-censoring and for complex survey designs that are used in data collection. In the remainder of this section, we briefly review some options for model-assisted estimation in finite populations for cross-sectional data (i.e., non-time-to-event data) and time-to-event estimation for non-survey data. The ideas of finite population model-assistance and time-to-event modeling will be combined to produce the estimators studied here.

1.1 | Model-Assisted Estimation in Finite Populations

There are a variety of model-assisted estimators in the survey literature, including generalized regression estimators (GREG), calibration estimators, general difference estimators (GDE), and model calibrated estimators (MCE). Model-assisted estimators of finite population means and totals are motivated by the model

$$E[y_i \mid \mathbf{x}_i] = \mu(\mathbf{x}_i, \boldsymbol{\theta}), \qquad V[y_i \mid \mathbf{x}_i] = v_i \sigma^2 \tag{2}$$

where $\mathbf{x}_i$ is a p-vector of covariates for unit i, $\boldsymbol{\theta}$ is a parameter vector, $\mu(\mathbf{x}_i, \boldsymbol{\theta})$ is the mean of y_i given the covariates, σ^2 is a variance parameter, and v_i is a scaling factor that can vary among the units. The form of the function μ is assumed to be known.

Wu and Sitter[10] define the GDE for the population mean of variable y_i as follows:

$$\hat{\bar{Y}}_{GD} = N^{-1} \left(\sum_{i=1}^{N} \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) + \sum_{i \in s} \frac{1}{\pi_i} \left[y_i - \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right] \right), \tag{3}$$

where N is the size of the population and $\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ is the model prediction for y_i based on $\mathbf{x}_i$ found by inserting an estimator of $\boldsymbol{\theta}$. For use below, define $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$ to be the vector of population totals of the covariates. If $\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \mathbf{x}_i^T \hat{\boldsymbol{\theta}}$, then $\hat{\bar{Y}}_{GD}$ is the classic GREG of Särndal, Swensson, and Wretman[11]. The GDE has the advantage in being more flexible than the GREG in accommodating more realistic models for binary and ordinal categorical variables. Its computational disadvantage for nonlinear models is that the individual values of covariates must be known for all units in the population to evaluate each $\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$.

Another estimator proposed by Wu and Sitter[10] is the MCE. This estimator is based on the traditional calibration estimator of Deville and Särndal[12]. The general form of a calibration estimator is $\hat{\bar{Y}}_{Cal} = N^{-1} \sum_{i \in s} w_i y_i$ where w_i satisfies the constraint $\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{X}$ while minimizing the average deviation of the calibration weights w_i from design weights d_i under some distance metric Φ_s. A common distance metric is the chi-squared distance measure:

$$\Phi_s = \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i}$$

for some set of known q_i's which are independent of d_i. This distance measure results in $\hat{\bar{Y}}_{Cal}$ being the GREG:

$$\hat{\bar{Y}}_{GREG} = \hat{\bar{Y}}_{\pi} + N^{-1} \left(\mathbf{X} - \hat{\mathbf{X}}_{\pi} \right)^T \hat{\mathbf{B}}, \tag{4}$$

where $\hat{\mathbf{X}}_{\pi} = \sum_{i \in s} d_i \mathbf{x}_i$ and $\hat{\mathbf{B}} = \left[\sum_{i \in s} d_i q_i \mathbf{x}_i \mathbf{x}_i^T \right]^{-1} \sum_{i \in s} d_i q_i \mathbf{x}_i y_i$.

The MCE is derived by finding the set of weights $\{w_i\}_{i \in s}$ that satisfy the constraints:

$$\sum_{i \in s} w_i = N, \tag{5}$$

$$\sum_{i \in s} w_i \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}). \tag{6}$$

The MCE estimator of the mean is

$$\hat{\bar{Y}}_{MC} = \hat{\bar{Y}}_{\pi} + N^{-1} \left(\sum_{i=1}^{N} \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \sum_{i \in s} d_i \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right) \hat{B}_N, \tag{7}$$

with

$$\hat{B}_N = \frac{\sum_{i \in s} d_i q_i \left(\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \bar{\mu} \right) \left(y_i - \bar{y} \right)}{\sum_{i \in s} d_i q_i \left(\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \bar{\mu} \right)^2},$$

where $\hat{\bar{Y}}_{\pi} = N^{-1} \sum_{i \in s} d_i y_i$, $\bar{y} = \sum_{i \in s} d_i q_i y_i / \sum_{i \in s} d_i q_i$ and $\bar{\mu} = \sum_{i \in s} d_i q_i \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) / \sum_{i \in s} d_i q_i$. Wu and Sitter also consider a MCE without the constraint (5). This new estimator $\hat{\bar{Y}}_{MC}^*$ replaces $\hat{B}$ with

$$\hat{B}_N^* = \frac{\sum_{i \in s} d_i q_i \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) y_i}{\sum_{i \in s} d_i q_i \left(\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^2}.$$

The GDE is a special case of MCE where $\hat{B}_N = 1$.

In simple random sampling without replacement (*srswor*), both $\hat{\bar{Y}}_{GD}$ and $\hat{\bar{Y}}_{MC}$ are design consistent in the sense that, as the sample size increases, the difference between the estimator of the mean and the population mean converges in probability to zero. This design-based property holds regardless of whether the model for y is specified correctly. Kennel[13] extended Wu and Sitter's *srswor* theory to multistage complex samples. Both estimators are also approximately model-unbiased in the sense that $E_M \left(\hat{\bar{Y}} - \bar{Y}_U \right) \approx 0$ where $\bar{Y}_U$ is the population mean and E_M denotes expectation with respect to model (2) if the working model is correctly specified. Thus, the GDE and MCE are doubly robust as in Robins, et. al[14] and Kang and Schafer[15].

1.2 | Time-to-Event Models

There are a number of time-to-event models, including proportional hazard models (PHM)[16], accelerated failure time models (AFTM)[17,18,19], and threshold regression models (TRM)[20,21]. We will cover only PHMs, although the theoretical results in Theorems 1 and 2 below do apply when a TRM is used to estimate $p(t \mid \mathbf{x}, \boldsymbol{\theta})$ (see Reist[22]).

1.2.1 | Estimating $p_N(t \mid \mathbf{x})$

A standard use of time-to-event models is to predict the failure probability for an individual at some time t given a vector of covariates $\mathbf{x}$. PHMs model time-to-event data through the hazard function $\lambda(t \mid \mathbf{x}, \boldsymbol{\theta})$. The failure probability is $p(t \mid \mathbf{x}, \boldsymbol{\theta}) = 1 - S(t \mid \mathbf{x}, \boldsymbol{\theta})$ where $S(t \mid \mathbf{x}, \boldsymbol{\theta}) = \exp(-\Lambda(t \mid \mathbf{x}, \boldsymbol{\theta}))$ is the survival function with $\Lambda(t \mid \mathbf{x}, \boldsymbol{\theta}) = \int_0^t \lambda(t \mid \mathbf{x}, \boldsymbol{\theta})$. The quantity $p(t \mid \mathbf{x}, \boldsymbol{\theta})$ can be estimated as follows:

$$p(t \mid \hat{\boldsymbol{\theta}}, \mathbf{x}) = 1 - \exp \left(- \int_0^t \lambda(t \mid \hat{\boldsymbol{\theta}}, \mathbf{x}) dt \right).$$

with $\hat{\boldsymbol{\theta}}$ being an estimator of $\boldsymbol{\theta}$. Both parametric and semiparametric PHMs are considered below.

The hazard function in a proportional hazards model is defined as

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\theta}) = \lambda_0(t) g \left(\boldsymbol{\theta}^T \mathbf{x} \right),$$

where $\lambda_0(t)$ is the baseline hazard when $\mathbf{x} = 0$ and $g(\boldsymbol{\theta}^T \mathbf{x})$ is a parametric function where $g(0) = 1$. When $\lambda_0(t)$ also has a parametric specification, the PHM is considered parametric. If $\lambda_0(t)$ is left unspecified, then the PHM is considered semiparametric. PHMs can be fit using a traditional maximum likelihood estimation (MLE) framework if both $\lambda_0(t)$ and $g\left(\boldsymbol{\theta}^T \mathbf{x}\right)$ are

parametric[23]. An example of a parametric hazard function for the PHM is the Weibull distribution specification:

$$\lambda(t, \mathbf{x} \mid \boldsymbol{\theta}) = \frac{\delta}{g(\boldsymbol{\theta}^T \mathbf{x})} \left[\frac{t}{g(\boldsymbol{\theta}^T \mathbf{x})} \right]^{\delta-1} = \left(\delta t^{\delta-1} \right) g\left(\boldsymbol{\theta}^T \mathbf{x} \right)^{-\delta},$$

where δ is known as the *shape parameter*[17]. One common way of specifying $g\left(\boldsymbol{\theta}^T \mathbf{x} \right)$ in a proportional hazard context is to let $g\left(\boldsymbol{\theta}^T \mathbf{x} \right) = \exp(\boldsymbol{\theta}^T \mathbf{x})$[23].

The semiparametric Cox version of a PHM[16] is one of the most widely used time-to-event models because of the flexibility gained by not needing to specify the distribution of the baseline hazard. These models are fit using partial likelihood[16,24].

1.2.2 | Accelerated Failure Time Models (AFTMs)

One straightforward way to consider modeling the time-to-event T is to consider a log-linear formulation

$$\ln(T) = \boldsymbol{\theta}^T \mathbf{x} + \sigma \epsilon. \tag{8}$$

A model that can be expressed in this form is called an AFTM because the effect of covariates is to accelerate or decelerate the time-to-event[17]. Wei[19] argues that AFTMs are easily interpreted since covariates have a direct effect on failure times.

Most of the commonly used parametric time-to-event models are AFTMs. The exponential, Weibull, log-normal, log-logistic, gamma, inverse gaussian, and generalized gamma models are all AFTMs. For example, if ϵ follows a logistic distribution, then the model in (8) becomes a log-logistic model[17]. Parametric AFTMs can be fit, under right censoring, using the same formulation of the likelihood used for the parametric PHM.

Louis[25] first developed a semiparametric formulation of the AFTM for a single treatment variable. Later Tsiatis[26] and Wei[19] generalized this to multiple treatments. Semiparametric AFTMs were put into a rank base inference framework by Jin[27]. In this paper we cover only parametric AFTMs and exclude semiparametric models from the theory and simulations.

2 | MODELS FOR SURVEY DATA

The standard MLE methods for estimating $\boldsymbol{\theta}$ will not be design consistent as defined in Fuller (Definition 1.3.1)[28] because they do not account for the way the sample was selected. Rather than maximizing a regular full or partial likelihood, pseudo maximum likelihood estimation (PMLE) methods are used which consist of maximizing a survey-weighted likelihood (e.g., see Binder[29],[2]). In the context of survival data, the PMLE method has been used for the Weibull AFTM model[30] and for TRMs[31]. This section will review the adjustments that need to be made to produce design consistent estimates for the time-to-event models presented in Section 1.

2.1 | Time-to-Event GDEs and MCEs

The combination of survival modeling and model-assisted estimation for finite populations can be used to construct GDE and MCE estimators to estimate $p_N(t)$ for a given $t \leq t_o$. The work of Wu and Sitter[10] and Kennel[13] can be used to construct GDE and MCE estimators of $p_N(t)$ when general linear models (GLM) are used to model $I_{\{T_i \leq t\}}$ or T. A GLM cannot be used to predict T and estimate $p_N(t)$ with the empirical distribution function of the T_i's if T is censored. Standard time-to-event models such as a PHM can be used to develop GDEs and MCEs for predicting $p_N(t)$ for $t \leq t_o$.

General difference estimators and model calibrated estimators can be constructed using the estimates of $p_N(t \mid \mathbf{X})$ as follows:

$$\hat{p}_{GD}(t) = N^{-1} \left(\sum_{i=1}^{N} p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) + \sum_{i \in s} d_i \left[I_{\{T_i \leq t\}} - p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right] \right) \tag{9}$$

and

$$\hat{p}_{MC}(t) = \hat{p}_\pi(t) + N^{-1} \left(\sum_{i=1}^{N} p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \sum_{i \in s} d_i p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right) \hat{B}, \tag{10}$$

where options for $\hat{B}$ are defined below. In (9) and (10), $p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})$ plays the role of $\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ in (3) and (7). Although we consider only the case of N known, in some special cases results can be extended to unknown N by using $\hat{N} = \sum_{i \in s} d_i$. The first term in each of the GDE and MCE is a sum over all N units in the population and requires the covariate values for each individual

unit. Suppose that strata can be formed based on combinations of the covariates so that every unit in a stratum has the same values of the covariates and that the strata exhaust the population. If the population count of units in each stratum is either known or estimated, then $\sum_{i=1}^{N} p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})$ can be evaluated without having a full population frame. Using estimates of N and the stratum counts does complicate variance estimation, although replication is one approach for reflecting uncertainty due to using estimated population counts.

Two alternatives for $\hat{B}$ will be considered which are adapted from Wu and Sitter. The first is derived subject to the following constraints:

$$\sum_{i \in s} w_i = N, \text{ and} \tag{11}$$

$$\sum_{i \in s} w_i p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}). \tag{12}$$

and is equal to

$$\hat{B} = \frac{\sum_{i \in s} d_i \left(p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \bar{p} \right) \left(I_{\{T_i \leq t\}} - \bar{I} \right)}{\sum_{i \in s} d_i \left(p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \bar{p} \right)^2}, \tag{13}$$

where $\bar{I} = \sum_{i \in s} d_i I_{\{T_i \leq t\}} \big/ \sum_{i \in s} d_i$, and $\bar{p} = \sum_{i \in s} d_i p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) \big/ \sum_{i \in s} d_i$. The second adjustment, $\hat{B}^*$, which can also be used in (10), is derived subject to only constraint (12) and can be calculated as

$$\hat{B}^* = \frac{\sum_{i \in s} d_i p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) I_{\{T_i \leq t\}}}{\sum_{i \in s} d_i \left(p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right)^2}. \tag{14}$$

The asymptotic variances of $\hat{p}_{GD}(t)$ and $\hat{p}_{MC}(t)$ and their estimators of variance are presented in Theorem 2 below.

2.2 | Theoretical Results

This section provides asymptotic results for both the $\hat{p}_{GD}(t)$ and $\hat{p}_{MC}(t)$ estimators and their respective variance estimators, where the underlying model is a time-to-event model. Specifically, results are shown for parametric and semiparametric PHMs. For the semiparametric PHMs, the case is addressed where the baseline hazard is estimated using a Breslow type estimator[32,33,34]. Both $\hat{p}_{GD}(t)$ and $\hat{p}_{MC}(t)$ are design consistent as shown below, and estimators of asymptotic variances are presented that are design consistent.

2.2.1 | Design Consistency of $\hat{p}_{GD}(t)$ and $\hat{p}_{MC}(t)$

To prove design consistency of $\hat{p}_{GD}(t)$ for a fixed t, the same asymptotic formulation is used as in Fuller[28] in which both the population and sample size become large. Consider a sequence of populations indexed by j in which both the sample size n_j and the population size N_j approach infinity as $j \to \infty$. To simplify the notation, we omit the subscript j below and in the Appendix. The following conditions are assumed to hold:

(i) $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_N + O_p\left(n^{-1/2}\right)$ and $\boldsymbol{\theta}_N \to \boldsymbol{\theta}$, where $\hat{\boldsymbol{\theta}}$ is the PMLE of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_N$ is the finite population value of the parameter, and $\boldsymbol{\theta}$ is its underlying constant value;

(ii) $\partial p(t \mid \mathbf{x}_i, \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}$, is continuous in $\boldsymbol{\gamma}$ where $\boldsymbol{\gamma}$ is in a neighborhood of $\boldsymbol{\theta}$. $|\partial p(t \mid \mathbf{x}_i, \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}| \leq c_1\left(\mathbf{x}_i, \boldsymbol{\theta}\right)$, a constant, for all values $\boldsymbol{\gamma}$ in a neighborhood of $\boldsymbol{\theta}$; and $N^{-1} \sum_{i=1}^{N} c_1\left(\mathbf{x}_i, \boldsymbol{\theta}\right) = O(1)$.

(iii) The π-estimators (divided by N) for certain population means are asymptotically normally distributed.

Theorem 1. If $\hat{p}_{GD}(t)$ is constructed using a time-to-event model and (i)-(iii) hold, then for a fixed time t

$$\hat{p}_{GD}(t) = \hat{p}_{\pi}(t) + O_p\left(n^{-1/2}\right),$$

where $\hat{p}_{\pi}(t)$ is the π-estimator in (1) of the finite population proportion $p_N(t)$. Thus, $\hat{p}_{GD}(t)$ is design consistent since $\hat{p}_{\pi}(t)$ is.

The proof is in the Appendix. Note that $\hat{p}_{GD}(t)$ is a special case of $\hat{p}_{MC}(t)$, where $\hat{B}_N = 1$. Because of this, Theorem 1 can be generalized to show that $\hat{p}_{MC}(t)$, which uses $\hat{B}$, and $\hat{p}_{MC}^*(t)$, which uses $\hat{B}^*$, are design consistent by noting that $\hat{B}$ and $\hat{B}^*$ are both $O_p(1)$ and that $\hat{p}_{\pi}(t)$ is design consistent.

2.2.2 | **Design Consistency of $\hat{V}[\hat{p}_{GD}(t)]$ and $\hat{V}[\hat{p}_{MC}(t)]$**

To show design consistency of the variance estimators below an additional condition is necessary :

(iv) $\partial^2 p(t \mid \mathbf{x}_i, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'$ is continuous in $\boldsymbol{\gamma}$ for each $\mathbf{x}_i$, $|\partial^2 p(t \mid \mathbf{x}_i, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'| \leq c_2\left(\mathbf{x}_i, \boldsymbol{\theta}\right)$, a constant, and $N^{-1}\sum_{i=1}^{N} c_2\left(x_i, \boldsymbol{\theta}\right) = O(1)$.

Theorem 2. If $\hat{p}_{GD}(t)$ is constructed using a time-to-event model where (i)-(iv) hold, then for a fixed time t, the approximate design variance estimator of $\hat{p}_{GD}(t)$ is

$$V\left(\hat{p}_{GD}(t)\right) \doteq N^{-2} \sum_{j}^{N} \sum_{i<j}^{N} \left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2, \qquad (15)$$

where π_{ij} is the joint probability of selecting the ith and j^{th} units and $e_i = I_{\{T_i \leq t\}} - p(t \mid x_i, \boldsymbol{\theta}_N)$. This can be estimated by

$$\hat{V}\left(\hat{p}_{GD}(t)\right) = N^{-2} \sum_{j \in s} \sum_{i<j \in s} \left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{\hat{e}_i}{\pi_i} - \frac{\hat{e}_j}{\pi_j}\right)^2, \qquad (16)$$

where $\hat{e}_i = I_{\{T_i \leq t\}} - p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})$.

The proof is in the Appendix. Further, Theorem 2 can be generalized to $\hat{V}\left(\hat{p}_{MC}(t)\right)$ by noting that $\hat{B} = B_N + o_p(1)$ and $\hat{B}^* = B_N + o_p(1)$ and substituting $I_{\{T_i \leq t\}} - p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right)B_N$ for e_i in the variance formula and $I_{\{T_i \leq t\}} - p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)\hat{B}$ or $I_{\{T_i \leq t\}} - p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)\hat{B}^*$ into the variance estimator for $\hat{e}_i$.

In the special case of stratified simple random sampling without replacement used in the simulation study in section 3 the design variance and its estimator can be simplified. Let $h = 1, \ldots, H$ index the strata, N_h be the population count of units in stratum h, and $W_h = N_h/N$. The variance of $\hat{p}_{GD}(t)$ and its estimator are then

$$V\left(\hat{p}_{GD}(t)\right) = \sum_{h=1}^{H} W_h^2\left(1 - f_h\right)S_{he}^2/n_h$$

$$\hat{V}\left(\hat{p}_{GD}(t)\right) = \sum_{h=1}^{H} W_h^2\left(1 - f_h\right)\hat{S}_{he}^2/n_h \qquad (17)$$

where n_h is the sample size in stratum h, $f_h = n_n/N_h$ is the proportion of the units in stratum h that are in the sample, $S_{he}^2 = (N_h - 1)^{-1}\sum_{i=1}^{N_h}(e_{hi} - \bar{e}_h)^2$, $e_{hi} = I_{\{T_{hi} \leq t\}} - p\left(t \mid \mathbf{x}_{hi}, \boldsymbol{\theta}_N\right)$, T_{hi} is the event time for individual hi, and $\mathbf{x}_{hi}$ is the vector of covariates for individual hi. The other terms are $\bar{e}_h = N_h^{-1}\sum_{i=1}^{N_h} e_{hi}$, $\hat{S}_{he}^2 = (n_h - 1)^{-1}\sum_{i=1}^{n_h}(\hat{e}_{hi} - \hat{\bar{e}}_{hs})^2$ with $\hat{e}_{hi} = I_{\{T_{hi} \leq t\}} - p\left(t \mid \mathbf{x}_{hi}, \hat{\boldsymbol{\theta}}\right)$, and $\hat{\bar{e}}_{hs} = n_h^{-1}\sum_{i=1}^{n_h}\hat{e}_{hi}$ for the GDE. Similar forms apply for the variance and variance estimator for the MCE with the estimated residual defined as $\hat{e}_{hi} = I_{\{T_{hi} \leq t\}} - p\left(t \mid \mathbf{x}_{hi}, \hat{\boldsymbol{\theta}}\right)\hat{B}$ if constraints (11) and (12) are used. If only (12) is used, the residual is defined with $\hat{B}^*$.

Although it is difficult to say theoretically whether the GDE or MCE will have a smaller variance, an intuitive comparison is possible. The variances of both depend on residuals of the form, $I_{\{T_i \leq t\}} - \hat{I}_{\{T_i \leq t\}}$ where $\hat{I}_{\{T_i \leq t\}} = p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)$ for the GDE and $\hat{I}_{\{T_i \leq t\}} = p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)\tilde{B}$ where $\tilde{B}$ is one of the slope estimators for the MCE. The estimator with smaller squared residuals will be more precise. In cases where the hazard function is modeled accurately, $\hat{B}$ or $\hat{B}^*$ will only add extra noise without reducing bias. On the other hand, if $\lambda(t \mid \mathbf{x}, \theta)$ is poorly modeled, injecting a slope estimator into the residuals for an MCE can help reduce its variance.

3 | **SIMULATION STUDY USING GENERATED DATA**

To evaluate the performance of the estimators, a simulation study was conducted in which the following were manipulated:

- Correlation between $\ln(T)$ and predictor $\mathbf{Z}$,
- Distribution of T, the time to an event,

- Amount of censoring, $\%C$,

- Sample size, n,

- Prevalence of the event at time t in the finite population, $p_N(t)$.

The simulations are limited to GDEs and MCEs constructed using lognormal, Weibull, semiparametric PH, and logistic models. These estimators are compared to more traditional estimators: the π-estimator and the GREG with a linear model.

3.1 | Populations

Three types populations were generated based on the log of event time: lognormal (LN), Weibull with a common baseline hazard (WCB), and Weibull with a mixture of two baseline hazards (WMB). Finite populations with $N = 100,000$ were generated as independent and identically distributed samples from:

$$\ln(T) = 1 + \theta_1 X + Z + W, \tag{18}$$

where Z was generated from a gamma distribution with shape and scale parameters equal to 1. For the LN populations, W was drawn from a normal distribution with mean zero, standard deviation σ (discussed below), and $X = 0$. For the WCB and the WMB populations, W was drawn from a generalized extreme value (GEV) distribution with the location parameter and shape parameters set to zero and shape parameter σ. For the WCB populations that have a common baseline hazard, $X = 0$. For populations with a mixture of two baseline hazards, X was drawn from a Bernoulli distribution with $p = 0.4$. Parameter values used in generating the $\ln(T)$'s are summarized in Table 1 . For each population the proportion of units that have experienced the event at or before any time t is computed as $p_N(t) = N^{-1} \sum_{i=1}^{N} I_{T_i \leq t}$ where each T_i was generated from (18).

TABLE 1 Parameters used to generate $\ln(T)$ in (18)

Population	θ_1	X	Z	W
Lognormal (LN)	0	0	$\Gamma(1, 1)$	$N(0, \sigma^2)$
Weibull, common baseline (WCB)	0	0	$\Gamma(1, 1)$	$GEV(0, 0, \sigma)$
Weibull, mixture baseline (WMB)	1	$B(0.4)$	$\Gamma(1, 1)$	$GEV(0, 0, \sigma)$

In all cases, σ was set to generate finite populations in which the correlation between $\ln(T)$ and Z was a given ρ. Nine populations were generated by crossing the LN, WCB, and WMB distributions with the correlations $\rho = 0.8, 0.6$, and 0.4. For each population, three sets of censored values of T and censor indicators were derived as follows:

$$\tilde{T}_i^{(j)} = min(T_i, Q_j), \tag{19}$$
$$c_i^{(j)} = I_{\{T_i \leq Q_j\}} \tag{20}$$

for $j = 1, 2, 3$, where Q_j is the jth finite population quartile of T. This generated censored values of T such that 75%, 50%, or 25% of the cases in the population were censored in the sense that there is no observation after time $t_o = Q_j$. These censoring times can be thought of as the times at which survey data collection ends. Although such high levels of censoring would be unusual in clinical studies, they would be more common in longitudinal, sample surveys when analyses are done periodically throughout the life of the study. For example, the HRS recruits cohorts of persons when they are in the age range 50-56. Assuming that death is the event, in the early years in which a cohort is in the sample there can be very high levels of censoring since most people are still alive.

3.2 | Sample Design

For this simulation a stratified simple random sample design was used with strata based on the values of Z. The units were sorted in ascending order based on Z, then the first 10,000 were assigned to stratum 1, the next 20,000 were assigned to stratum 2, the next 30,000 were assigned to stratum 3, and the last 40,000 were assigned to stratum 4. Two sample sizes were used: $n = 200$ and 1000. The sample was allocated equally to each stratum, i.e., $n_h = n/4$ for all h. For each population-sample size combination, $L = 10,000$ samples were drawn.

3.3 | **Estimators**

For each sample, four time-to-event models for estimating $p_N(t)$ were paired with the general difference and model-calibrated estimators. All of the time-to-event models were fit with an intercept and one predictor, Z. The models were lognormal, Weibull, semiparametric PH, and logistic. Each of these was tested with the censoring conditions 75%, 50%, and 25%. The $\hat{\theta}$ parameter estimates were PMLEs. For the semiparametric PH estimator the baseline hazard was estimated using a Breslow estimator[33]. For each model, three estimates of $p_N(t)$ were calculated, a GDE in (9) and two versions of the MCE in (10). These are denoted as GD, MC1, and MC2 in the subsequent tables. MC1 is the MCE with one constraint defined by (12) and uses $\hat{B}^*$ as the slope estimate, and MC2 is the MCE estimator with two constraints defined by (11) and (12) and uses $\hat{B}$. To distinguish in the discussion which time-to-event model was used for estimating $p_N(t)$, LN (lognormal) and LG (logistic) are paired with the estimator labels. For example, LN-GD denotes the general difference estimator using the lognormal time-to-event model; LG-GD is the GDE paired with logistic.

Since there are three types of populations (LN, WCB, and WMB) and four models used for estimating each $p(t \mid \mathbf{x}_i, \boldsymbol{\theta})$, we summarize which combinations of population and model give correctly specified time-to-event models:

- LN population with lognormal model used to estimate the probability, $p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})$, that an event occurs at or before time t

- WCB with either Weibull or semiparametric PH time-to-event model

- WMB with semiparametric PH time-to-event

All other combinations are ones where the time-to-event model is misspecified compared to how the population was generated. Note that using a logistic time-to-event model is always a misspecification in our simulation.

Estimates of $p_N(t)$ were then generated for three values of t. The three values were selected so that the finite population value of $p_N(t)$ was 0.75, 0.50, or 0.25. Note that for 75% censoring only $p_N(t) = 0.25$ could be estimated. Likewise for 50% censoring, only $p_N(t) = 0.50$ or $p_N(t) = 0.25$ could be estimated. Additionally, to compare these alternatives with existing methods, we computed the π-estimator and a GREG based on a linear model with an intercept and one predictor, Z.

3.4 | **Evaluation Criteria**

A number of criteria were used to evaluate the performance of the time-to-event based GDE and MCE related to efficiency, bias, and performance of variance estimators. These criteria are also used in evaluating the simulation results in Section 4. The simulated RMSE at a fixed time t was estimated as follows:

$$RMSE(t) = \left(L^{-1} \sum_{k=1}^{L} \left[\hat{p}_k(t) - p_N(t) \right]^2 \right)^{1/2},$$

where L represents the 10,000 simulations, and $\hat{p}_k(t)$ is an estimate of $p_N(t)$ for the kth simulation. To compare the simulated RMSE of an estimator A with the RMSE of the π-estimator, the percent reduction in RMSE (Δ_{RMSE}) was calculated as follows:

$$\Delta_{RMSE}(t) = 100 \left[1 - \left(\frac{RMSE_A(t)}{RMSE_\pi(t)} \right) \right],$$

Two measures were calculated to evaluate the bias of GDE and MCE that were derived from time-to-event models. The first measure is the simulated Relative Bias (RB) calculated as:

$$RB(t) = \frac{1}{L} \sum_{k=1}^{L} \left(\frac{\hat{p}_k(t) - p_N(t)}{p_N(t)} \right).$$

The second measure is the Bias Ratio (BR). The BR compares the magnitude of the simulated bias of an estimator to the magnitude of the simulated standard error of the same estimator. The BR is calculated as follows:

$$BR(t) = \frac{L^{-1} \sum_{k=1}^{L} \left[\hat{p}_k(t) - p_N(t) \right]}{\left(L^{-1} \sum_{k=1}^{L} \left[\hat{p}_k(t) - \bar{p}(t) \right]^2 \right)^{\frac{1}{2}}},$$

where $\bar{p}(t) = L^{-1} \sum_{k=1}^{L} \hat{p}_k(t)$. For confidence intervals to cover at the desired rate, BR must converge to 0 with increasing sample size in addition to $\hat{p}_k(t) - p_N(t)$ converging to 0.

Two measures were calculated to evaluate the performance of the variance estimators. The first is the Variance Ratio (VR), which is the ratio of the simulation mean of the estimated sampling variance to the empirical variance of the estimator. This was calculated as:

$$VR(t) = \frac{L^{-1} \sum_{k=1}^{L} \hat{V}(\hat{p}_k(t))}{L^{-1} \sum_{k=1}^{L} \left[\hat{p}_k(t) - \bar{p}(t)\right]^2}$$

where for the GDE $\hat{V}(\hat{p}_k(t))$ is defined by (17) for the kth sample. The variance estimator for MC1 has the same form as (17) with the residual modified by $\hat{B}$ as described below that equation; for MC2 the residual is defined using $\hat{B}^*$. The *BR* and *VR* measures will be used in sec. 4. The second measure is confidence interval coverage. For the kth sample, the normal approximation, 95% confidence interval was calculated as

$$CI_i = \left(\hat{p}_k(t) - 1.96\sqrt{\hat{V}(\hat{p}_k(t))}, \quad \hat{p}_k(t) + 1.96\sqrt{\hat{V}(\hat{p}_k(t))} \right).$$

The proportion of times that the confidence intervals included the population value was then tabulated across the simulations and is labeled "Coverage" when it is used in tables.

3.5 | Results

Biases and relative biases for all estimators in all scenarios in the generated populations were minimal and are not presented here. Differences in RMSEs were more substantial. An estimator's reduction in RMSE when compared to the π-estimator was affected by three conditions: the correlation between $\ln(T)$ and predictor Z, sample size, and the proportion of the individuals in the finite population that have experienced the event by time t.

Tables 2 and 3 provide the simulated reductions in RMSE when compared to the π-estimator for the LN lognormal population with $n = 200$ and $n = 1000$, respectively. Comparisons for other populations (Weibull with a common baseline hazard, WCB, and Weibull for a mixture of baseline hazards, WMB) were similar. The rows in these tables are sorted by ρ, then $p_N(t)$, and finally % censored. These parameters were described in section 3.1. The time-to-event estimators, i.e., ones in which $p_N(t)$ is estimated based on Weibull, lognormal, or proportional hazards models, are shown in columns 5-13 (Weibull/GD through Proportional hazard/MC2); for each of these models the final three columns labeled "Logistic" are cases where $p_N(t)$ is estimated via a logistic model. Some of the conclusions about RMSEs that can be drawn from these tables are:

1. The time-to-event model-based estimators never underperformed the π-estimator.

2. The GREG is less precise than the π-estimator for a number of combinations—$\rho = 0.8$ plus $p_N(t) = 0.25$ and 0.50 and all levels of censoring; $\rho = 0.6$ plus $p_N(t) = 0.25$ and all levels of censoring. For other combinations the GREG achieves some small reductions in RMSE.

3. The time-to-event model-based estimators never underperformed, and in many cases outperformed, the GREG (col. 4) and logistic-based GD, MC1, and MC2 estimators (LG-GD, LG-MC1, LG-MC2 in cols. 14-16).

4. The reductions in RMSE for the nine estimators based on time-to-event models (Weibull, lognormal, and proportional hazard) were similar.

5. The reduction in RMSE for the nine estimators based on time-to-event models and the GREG generally increased as the failure rate $p_N(t)$ increased. For example, when $p_N(t) = 0.75$ with $\rho = 0.8$ and 25% censoring, reductions in RMSE are nearly 15% for both $n = 200$ and 1000.

6. RMSE reductions for the nine estimators based on time-to-event models, the GREG, and LG-MC1 increased with increasing ρ.

7. Reductions in RMSE for the nine estimators based on time-to-event models and the GREG were similar at both sample sizes.

8. Reduction in RMSE for GD, MC1, and MC2 were substantially decreased or eliminated when prevalences were estimated using the logistic model when $n = 200$. MC1 and MC2 generally reduce RMSEs compared to the π-estimator and are more efficient than the GD, which often has a larger RMSE than the π-estimator. This is consistent with the observation in sec. 2.2.2 that inclusion of a slope estimator in the MCEs can reduce variances when the hazard function used in the estimators is misspecified.

9. The amount of censoring had a limited effect on all of the estimators. However, we assume a uniform censoring time for all units due to stoppage of data collection. If censoring could occur at random times, results could be affected.

TABLE 2 Simulated Percent Reduction of RMSE Relative to the π-estimator: Lognormal Population, $N=100{,}000$; $n=200$; $L=10{,}000$

ρ	$p_N(t)$	% censored	GREG	Weibull			Lognormal			Proportional Hazard			Logistic		
				GD	MC1	MC2	GD	MC1	MC2	GD	MC1	MC2	GD	MC1	MC2
0.8	0.25	0.25	-3.70	0.96	0.86	0.96	1.06	1.03	1.07	0.95	0.81	0.93	-1.15	0.44	0.30
		0.50	-3.70	1.00	0.95	1.00	1.05	1.02	1.05	1.00	0.94	0.99	-1.15	0.44	0.30
		0.75	-3.70	1.01	1.01	1.01	1.03	1.03	1.04	1.01	1.01	1.01	-1.15	0.44	0.30
	0.50	0.25	-0.96	4.42	4.42	4.48	4.68	4.61	4.68	4.51	4.50	4.50	-2.85	1.77	0.35
		0.50	-0.96	4.43	4.38	4.41	4.60	4.57	4.59	4.41	4.37	4.40	-2.85	1.77	0.35
	0.75	0.25	9.78	14.24	14.15	14.22	14.42	14.30	14.42	14.12	14.04	14.13	1.80	9.56	5.57
0.6	0.25	0.25	-0.71	0.87	0.80	0.89	0.95	0.92	0.97	0.88	0.79	0.88	-4.85	0.23	-0.22
		0.50	-0.71	0.89	0.84	0.91	0.94	0.90	0.95	0.90	0.84	0.90	-4.85	0.23	-0.22
		0.75	-0.71	0.88	0.87	0.88	0.91	0.89	0.91	0.88	0.87	0.88	-4.34	0.89	-0.45
	0.50	0.25	1.93	3.32	3.36	3.39	3.56	3.49	3.56	3.37	3.41	3.38	-4.34	0.89	-0.45
		0.50	1.93	3.40	3.34	3.38	3.49	3.45	3.49	3.38	3.34	3.38	-4.34	0.89	-0.45
	0.75	0.25	6.42	7.70	7.54	7.66	7.75	7.60	7.75	7.64	7.48	7.63	-5.74	4.59	-0.52
0.4	0.25	0.25	0.17	0.71	0.68	0.74	0.79	0.75	0.81	0.72	0.67	0.72	-1.89	0.31	0.05
		0.50	0.17	0.66	0.65	0.66	0.66	0.66	0.66	0.66	0.65	0.66	-1.89	0.31	0.05
		0.75	0.17	0.75	0.74	0.75	0.76	0.75	0.76	0.75	0.74	0.75	-1.89	0.31	0.05
	0.50	0.25	1.61	2.05	2.03	2.10	2.20	2.07	2.19	2.07	2.04	2.08	0.05	1.90	1.60
		0.50	1.61	2.13	2.07	2.11	2.16	2.09	2.16	2.11	2.07	2.11	0.05	1.90	1.60
	0.75	0.25	2.93	3.38	3.28	3.37	3.34	3.23	3.33	3.38	3.27	3.38	-8.95	2.08	-3.14

TABLE 3 Simulated Percent Reduction of RMSE Relative to the π-estimator: Lognormal Population, $N=100,000$; $n=1000$; $L=10,000$

ρ	$p_N(t)$	% censored	GREG	Weibull			Lognormal			Proportional Hazard			Logistic		
				GD	MC1	MC2	GD	MC1	MC2	GD	MC1	MC2	GD	MC1	MC2
0.8	0.25	0.25	-3.22	0.97	0.89	0.95	1.02	1.01	1.02	0.93	0.84	0.92	1.04	1.04	1.04
		0.50	-3.22	1.01	0.98	1.00	1.02	1.01	1.02	0.99	0.96	0.99	1.04	1.04	1.04
		0.75	-3.22	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.04	1.04	1.04
	0.50	0.25	-0.87	4.21	4.25	4.26	4.35	4.34	4.36	4.28	4.29	4.28	4.33	4.33	4.33
		0.50	-0.87	4.27	4.24	4.25	4.35	4.35	4.35	4.26	4.23	4.25	4.33	4.33	4.33
	0.75	0.25	11.00	14.80	14.76	14.76	15.02	14.99	15.02	14.66	14.63	14.66	15.00	15.00	15.00
0.6	0.25	0.25	-0.57	0.86	0.80	0.86	0.88	0.85	0.87	0.85	0.78	0.85	-0.97	0.40	0.31
		0.50	-0.57	0.87	0.84	0.87	0.87	0.85	0.87	0.87	0.84	0.87	-0.97	0.40	0.31
		0.75	-0.57	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	-0.97	0.40	0.31
	0.50	0.25	2.08	3.10	3.17	3.16	3.23	3.23	3.24	3.16	3.20	3.16	3.23	3.22	3.23
		0.50	2.08	3.19	3.16	3.18	3.23	3.23	3.23	3.18	3.16	3.17	3.23	3.22	3.23
	0.75	0.25	6.73	7.55	7.49	7.53	7.63	7.60	7.63	7.51	7.44	7.50	-5.03	5.31	0.05
0.4	0.25	0.25	0.21	0.64	0.63	0.65	0.67	0.65	0.67	0.64	0.62	0.65	-1.75	-0.07	-0.15
		0.50	0.21	0.66	0.65	0.66	0.66	0.66	0.66	0.66	0.65	0.66	-1.75	-0.07	-0.15
		0.75	0.21	0.67	0.67	0.67	0.66	0.67	0.66	0.67	0.67	0.67	-1.75	-0.07	-0.15
	0.50	0.25	1.61	1.80	1.83	1.84	1.86	1.84	1.86	1.82	1.84	1.83	1.85	1.85	1.85
		0.50	1.61	1.85	1.83	1.85	1.85	1.84	1.85	1.85	1.83	1.84	1.85	1.85	1.85
	0.75	0.25	3.41	3.56	3.51	3.55	3.58	3.56	3.58	3.54	3.50	3.54	-10.31	2.18	-3.95

(a) $n = 200$ **(b)** $n = 1000$
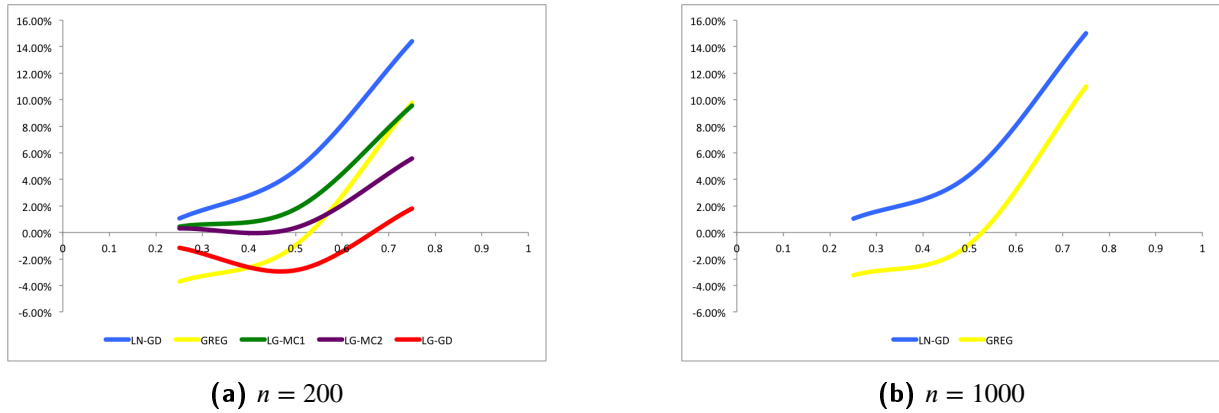
FIGURE 1 Percent reduction in RMSE as a function of $p_N(t)$: Lognormal population, $\rho = 0.8$, 25% censoring. (In 1(b) only LN-GD and GREG are presented, because the LN-GD, LG-MC1, LG-MC2, and LG-GD curves were indistinguishable.)

(a) $n = 200$ **(b)** $n = 1000$
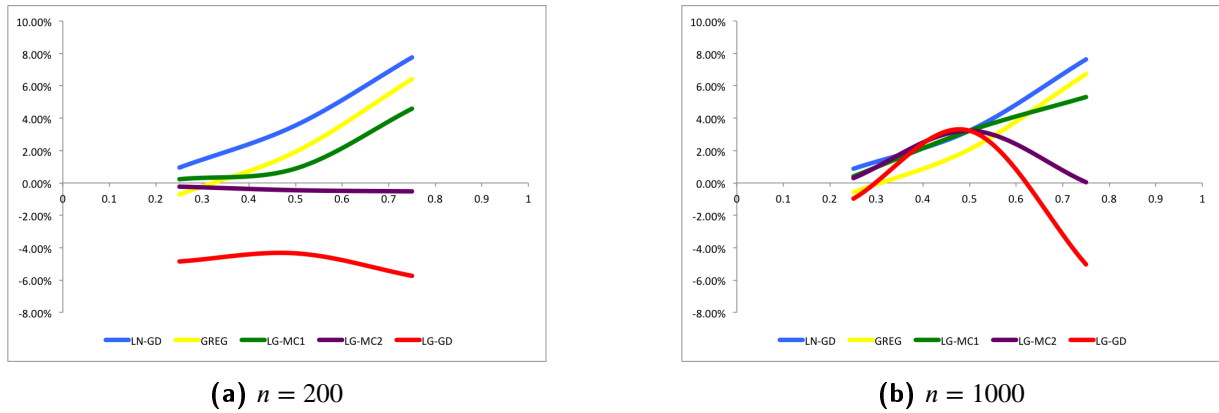
FIGURE 2 Percent reduction in RMSE as a function of $p_N(t)$: Lognormal population, $\rho = 0.6$, 25% censoring

(a) $n = 200$ **(b)** $n = 1000$
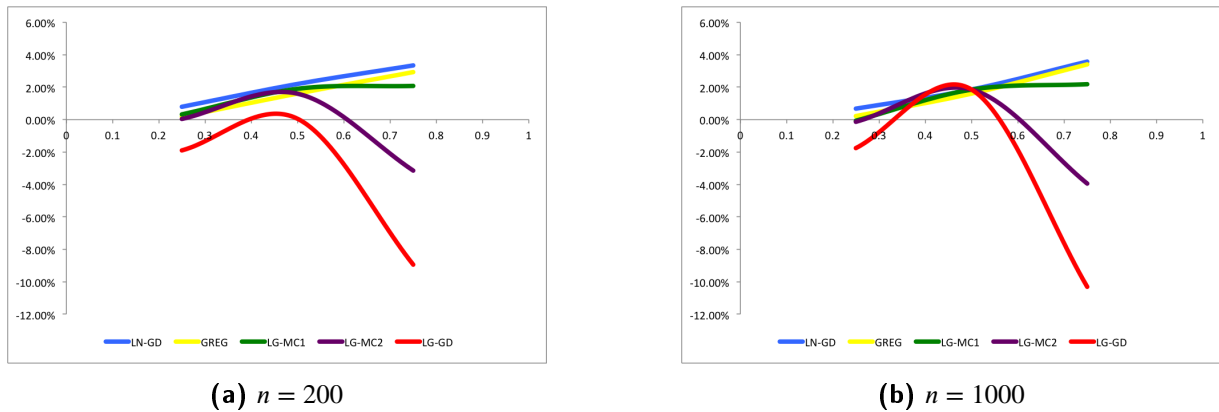
FIGURE 3 Percent reduction in RMSE as a function of $p_N(t)$: Lognormal population, $\rho = 0.4$, 25% censoring

Figures 1 -3 plot the percent reduction in RMSE compared to that of the π-estimator versus values of $p_N(t)$ for the lognormal GD (LN-GD), logistic GD (LG-GD), LG-MC1, LG-MC2, and GREG estimators. Nonparametric smoothers are plotted to make the patterns more apparent. All three figures are for 25% censoring only. Comparisons for other levels of censoring were similar. We include only LN-GD in the figures since all of the estimators based on time-to-event models (LN, WB, and PH paired with GD, MC1, and MC2) perform about the same. Subplots (a) and (b) show results for $n = 200$ and $n = 1000$. In Figure1 (b) only the LN-GD and GREG are presented, because the LN-GD, LG-MC1, LG-MC2, and LG-GD curves were indistinguishable.

In the three figures, the logistic general difference estimator (LG-GD) is generally the poorest performer relative to the π-estimator. Regardless of the size of the correlation between the log failure time, $\ln(T)$, and the covariate, Z, LN-GD is the best or nearly best performer. The RMSE reduction for LN-GD increases as $p_N(t)$ and ρ increase. Although there are cases where the logistic models reduce the RMSE slightly, estimators based on the underlying lognormal model are generally more efficient. In particular, the LN-GD estimator performs best because it correctly models $p_N(t)$. (This is also true for the model-calibrated estimators, LN-MC1 and LN-MC2, not shown in the figures.) The LG estimators are inferior because they use the wrong model for $p_N(t)$—a problem that is especially clear for the smaller values of ρ. Although the GREG is reasonably efficient compared to LN-GD, it is limited by requiring that the model for $p_N(t)$ must be linear.

As noted at the beginning of this section, the ratios of the variance estimators to the empirical variance and the coverage of the 95% normal approximation confidence intervals were also evaluated. In all cases the variance estimators were approximately unbiased and the confidence intervals covered at the desired rate. Thus, we do not report the details here.

4 | NURSES' HEALTH STUDY APPLICATION

To test the estimators on a real population we used data from the Nurses' Heath Study (NHS)[35]. We estimate the proportion of a population who have experienced death using only a sample of the population. A subset of the nurses' data serves as a simulation population that uses the same estimators and evaluation criteria as in section 3. We do not attempt to make estimates for the full population of nurses.

4.1 | About the Nurses' Health Study

The NHS is based on a panel of over 120,000 female nurses that has been followed since the mid-1970s. Originally, the NHS focused on the long-term effects of oral contraceptives. Although this is still a main focus of the NHS, the NHS now also focuses on smoking, cancer, and heart disease. It asks about lifestyle factors, such as nutrition and quality of life and also collects information on more than 30 diseases.

The target population for the NHS is female registered nurses in the 11 most populated states who were married and ages 30-55 in 1976. The frame was constructed using membership roles from nursing boards who agreed to participate in the NHS. In 1976, the 238,026 nurses on the frame were mailed an initial questionnaire. Of these, 121,700 nurses returned a completed questionnaire and were enrolled in the study. Every other year since 1976, study participants have received a follow-up questionnaire to collect information about disease and health-related topics. In addition, biological samples have been collected from subsamples of the panel. More information about the NHS can be found at http://www.nurseshealthstudy.org.

4.2 | Finite Population Creation

The finite population used in this application is a subset of the NHS population. The population is similar to other studies that used time-to-event models to study the incidence of lung disease (Bain et al.[36], Lee et al.[37]). This extract contained information from 1986 through 2012. To be eligible for the population, a panel participant had to meet the following criteria:

- Alive in 1986,

- Not diagnosed with cancer prior to 1986 (with the exception of non-melanoma skin cancer)

- Known smoking status in 1986,

- Known pack years in 1986,

- Known body mass index (BMI) for at least one year during 1986 to 2012

Pack years is calculated by multiplying the packs of cigarettes smoked per day for a year's time by the number of years that a person smoked. One pack year is equal to smoking 20 cigarettes per day for one year. BMI is equal to a person's weight in kilograms divided by the square of the person's height in meters. These restrictions resulted in a finite population of 103,878 nurses. The following variables were retained on the file:

- Death indicator (died between 1986 and 2012)

- Age at death (in years, to the tenth of a year)

- Age in 1986 (in years, to the tenth of a year)

- BMI for every observation between 1986 and 2012 (based on height reported in 1976)

- Smoking status in 1986 (Current Smoker, Past Smoker, Never Smoked)

- Pack years smoked as of 1986

The following variables were derived from these variables:

- BMI in 1986, where missing values of BMI were imputed using the BMI closest to 1986 that was observed

- A six level classification of BMI (Underweight, Normal, Overweight, Class 1 Obesity, Class 2 Obesity, Class 3 Obesity)

- A four level classification of BMI, which groups all three levels of obesity into one category (Underweight, Normal, Overweight, Obese)

- A three level classification of age in 1986 (<50, 50 to 60, >60)

- Years to death after 1986 calculated to the tenth of a year (with a value of 26 if alive in 2012)

4.3 | Sample Design

Two stratified simple random sample designs were used in this simulation study. The first had three strata based on the three levels of smoking status. The second had 36 strata formed by crossing smoking status, 3-level age group, and 4-level BMI. Both of these designs used strata that are related to death, with the 36 strata design expected to be more effective in reducing variance for estimates of the proportion of persons experiencing the event. Tables 4 -6 show the counts and row percentages of smoking status, age group, and six level BMI—all in 1986—crossed with the death-by-2012 indicator in the finite population.

TABLE 4 Smoking Status by Death Indicator: Counts and Row Percentages (as of 2012)

Status in 1986	Alive Count	%	Deceased Count	%	Total Count	%
Never Smoked	37,789	80.0	9,445	20.0	47,234	100
Current Smoker	13,698	61.8	8,463	38.2	22,161	100
Past Smoker	26,277	76.2	8,206	23.8	34,483	100
All nurses	77,764	74.9	26,114	25.1	103,878	100

TABLE 5 Age Group by Death Indicator: Counts and Row Percentages (as of 2012)

	Alive		Deceased	
Age in 1986	Count	%	Count	%
<50	36,077	91.1	3,531	8.9
50-60	31,286	61.8	11,243	26.4
>60	10,401	47.8	11,349	52.2

TABLE 6 Six-level BMI by Death Indicator: Counts and Row Percentages (as of 2012)

|		Alive		Deceased	
BMI in 1986	Classification	Count	%	Count	%
<18.5	Underweight	816	56.8	621	43.2
18.5-24.9	Normal Weight	42,302	77.8	12,079	22.2
25.0-29.5	Overweight	22,991	74.1	8,043	25.9
30.0-34.9	Class 1 Obesity	8,133	70.6	3,392	29.4
35.0-39.9	Class 2 Obesity	2,532	66.2	1,292	33.8
≥ 40.0	Class 3 Obesity	990	59.0	687	41.0

For all three tables, the chi-squared test of independence rejected the null hypothesis of independence for $\alpha = 0.01$. The finite population sample size was large, meaning that very small differences could be detected. However, there is variation in the percentage of nurses who have died across subgroups, which suggests that these variables do have some value in predicting death by 2012 and, thus, also time to death.

Two sample sizes were used to mimic the simulation study in section 3. For each sample design, samples of 216 and 1,008 were selected. These total samples were allocated equally to each of the strata. For example, for the case of 36 strata and the total sample size of 216, simple random samples of 6 persons were selected without replacement from each stratum. For the total sample size of 1008, 28 persons were selected from each of the 36 strata. This design creates sampling weights, d_i, that vary among strata.

4.4 | Model Development

As with the simulation study in section 3, five different models were fit to estimate the proportion of the population who had died at or before time t, which in this study was the year 2012 or 26 years after the recruitment of the nurses population. The five models were a linear model, logistic model, Weibull model, lognormal model, and semiparametric proportional hazards model. All five models were fit using the same set of predictor variables: smoking status, continuous BMI, BMI squared, continuous age, pack years, and pack years squared. The squared term for BMI was used to account for the fact that both small and large values of BMI result in higher risk of death. In an attempt to reduce collinearity between BMI and BMI squared, mean BMI was subtracted from BMI before it was squared.
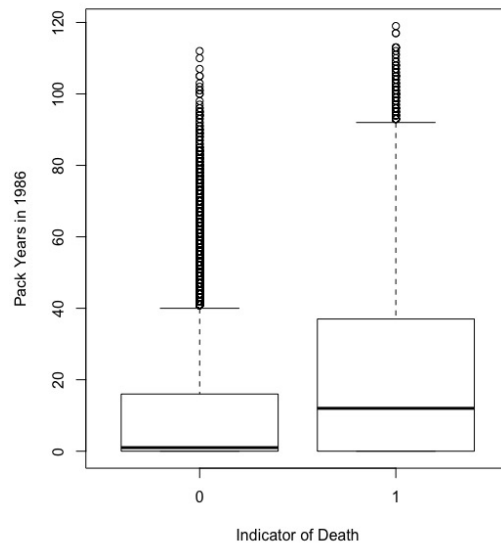
FIGURE 4 Pack Years by Death Indicator

The box plot of pack years is displayed in Figure 4 . This box plot shows that death generally seems more likely among nurses with more pack years by 1986. A squared term was introduced, because in similar studies it was thought that an increase in smoking has a negative effect on time to death, but this effect moderates for higher levels of pack years[37]. As with BMI squared, mean pack years was subtracted from pack years before it was squared to reduce collinearity between pack years and pack years squared.

4.5 | Results

A total of 10,000 samples were drawn for each of the four sample design-sample size combinations. The same estimators as in section 3 were used here to estimate the percentage of the population that had died by the end of 2012, i.e., $p_N(26) \approx 0.25$. Table 7 shows the results using the same five metrics as in section 3.4 for each estimator and sample design–sample size combination.

All of the estimators were approximately unbiased. (See the rows in Table 7 for %RB.) (Note that the RBs in this application were much smaller that those in Wu and Sitter[10], who reported RBs as high as 5.71% in a different population). Because all estimators were essentially unbiased, the RMSEs and standard errors are nearly equal. Thus, selection of an estimator can be based on RMSE and confidence interval coverage, at least in this application.
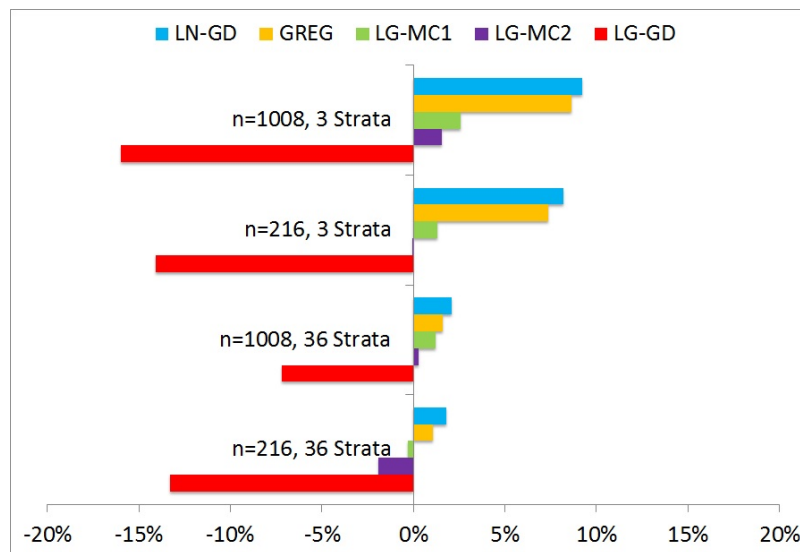
FIGURE 5 Nurses population: Simulated percent reduction of RMSE relative to the π-estimator by sample size and number of strata

The RMSE performance of the nine time-to-event model-based estimators was similar. (See the rows in Table 7 for Δ_{RMSE}) Therefore, for simplicity, only the LN-GD is compared in this discussion to the other methods when examining efficiency. Figure 5 shows the percent reduction in RMSE of each of the estimators compared to the π-estimator. Negative values mean that an estimator had a larger RMSE than the π-estimator. The LN-GD and GREG outperformed the estimators based on logistic models for every condition. The LG-GD estimator had significantly larger RMSEs than the π-estimator. Similar to the section 3 simulation study, the LG-MC2 slightly under-performed the π-estimator for two combinations (3 or 36 strata, $n = 216$) and had little if any gains for the other combinations. This finding is also in contrast to those of Wu and Sitter[10].

Figure 5 shows the importance of number of strata on the LN-GD and the GREG. For both estimators, the percent reduction in RMSE relative to the π-estimator for samples with three strata is four times larger than the RMSE for samples with 36 strata. Although this may seem counterintuitive, there are two explanations for it. First, an equal allocation is likely not optimal for predicting death. Second, the 3-strata design uses covariates in estimates that are some of the same as those used to form the 36 strata. Hence, for the GREG and LN-GD to see significant reductions in RMSE in the 36-strata design, either a more efficient allocation would be needed or covariates would need to be used that are not in the sample design. In the 36 strata design, BMI, age, and smoking status were used to define the strata. Besides the fact that continuous versions of BMI and age were used in the model, pack years was the only new information. In the 3 strata design, only smoking status was used to define the strata. This means that BMI, age, and pack years were all providing new information to the estimators that was not part of the sample design.

Table 7 shows that when $n = 1008$ the VRs were close to 1 for all of the estimators. Although not reported there, this was also the case for the simulations in section 3 and tells us that on average the asymptotic variance estimator was unbiased for the empirical variance of the estimator. Additionally, the simulated 95% confidence interval provided approximately nominal coverage, especially at the larger sample size.

TABLE 7 Nurses' Health Study Simulation Results for two sample designs: (3 strata, $n = 216$) and (36 strata, $n = 1,008$)

Metric	n	# Strata	π	GREG	Weibull GD	Weibull MC1	Weibull MC2	Lognormal GD	Lognormal MC1	Lognormal MC2	Proportional Hazard GD	Proportional Hazard MC1	Proportional Hazard MC2	Logistic GD	Logistic MC1	Logistic MC2
% Δ RMSE	216	3	0.00%	7.38%	8.24%	8.13%	8.17%	8.22%	8.04%	8.15%	8.20%	8.13%	8.16%	-14.07%	1.30%	-0.01%
		36	0.00%	1.07%	1.68%	1.62%	1.64%	1.79%	1.65%	1.71%	1.65%	1.60%	1.62%	-13.31%	-0.30%	-1.91%
	1008	3	0.00%	8.61%	9.58%	9.57%	9.58%	9.20%	9.38%	9.35%	9.59%	9.58%	9.58%	-15.97%	2.59%	1.57%
		36	0.00%	1.62%	2.22%	2.22%	2.22%	2.11%	2.12%	2.14%	2.23%	2.23%	2.23%	-7.21%	1.21%	0.27%
% RB	216	3	-0.02%	0.10%	0.00%	0.00%	0.00%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.90%	0.27%	0.41%
		36	-0.02%	0.14%	-0.07%	-0.07%	-0.07%	0.02%	0.03%	0.02%	-0.07%	-0.07%	-0.07%	0.53%	0.19%	0.31%
	1008	3	0.02%	0.02%	0.01%	0.01%	0.01%	0.02%	0.02%	0.02%	0.01%	0.01%	0.01%	0.20%	0.02%	0.06%
		36	0.05%	0.10%	0.07%	0.07%	0.07%	0.08%	0.08%	0.08%	0.07%	0.07%	0.07%	0.21%	0.08%	0.11%
BR	216	3	-0.001	0.009	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	0.035	0.023	0.035
		36	-0.001	0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.014	0.022
	1008	3	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.031	0.004	0.012
		36	0.008	0.016	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.032	0.013	0.018
VR	216	3	1.01	0.96	0.95	0.95	0.95	0.96	0.95	0.96	0.95	0.95	0.95	0.98	0.98	0.97
		36	1.00	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.96	0.97	0.98	0.97
	1008	3	1.01	1.01	1.01	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	0.99
		36	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99
Coverage	216	3	0.949	0.941	0.942	0.940	0.941	0.943	0.940	0.942	0.941	0.940	0.941	0.944	0.945	0.943
		36	0.940	0.937	0.935	0.935	0.935	0.937	0.937	0.937	0.935	0.935	0.935	0.942	0.938	0.939
	1008	3	0.952	0.951	0.952	0.952	0.952	0.951	0.953	0.952	0.952	0.952	0.952	0.950	0.951	0.949
		36	0.947	0.948	0.950	0.950	0.950	0.948	0.948	0.948	0.950	0.950	0.950	0.945	0.948	0.947

4.6 | Computational Problems with the Model Calibrated Logistic Estimators

Finally, it is worth noting that fitting logistic models to estimate $p_N(t)$ can lead to computational problems in small samples. This was due to the *separation* phenomenon, which is well-known[38]. In both the section 3 simulation and the nurses simulation, the LG-MC1 and LG-MC2 had some simulated samples that were excluded from analysis, because $\hat{p}(t)$ was less than 0 or greater than 1. This affected only a small proportion of the samples. This issue did not affect any of the time-to-event model-based MCEs. Table 8 shows the number of samples thrown out for each set of conditions in the NHS study. The problems with the logistic model-calibrated approach are caused by some combinations of covariates all having the event or not having the event. The fitting alogrithm sends one or more of the parameter estimates to $\pm\infty$. A potential fix is to combine levels of factors to create combinations where there is a mixture of events and non-events.

TABLE 8 Number of Samples out of 10,000 where the Model Calibrated Logistic Estimate was Greater than 1 or Less than 0 in the NHS simulation

n	Strata	LG-MC1	LG-MC2
216	3	146	146
216	36	81	81
1,008	3	12	12
1,008	36	6	6

The number of samples excluded was influenced by number of strata and sample size. A smaller sample size and fewer strata resulted in more excluded simulates, i.e., a less efficient design resulted in more samples being excluded. The most severe problem was with $n = 216$ and 3 strata, where 146 (or 1.46%) of the samples could not be included. Although this computational problem was rare, the fact that it happened at all is another reason not to use a logistic time-to-event model paired with the model calibrated estimators, LG-MC1 and LG-MC2, to estimate $p_N(t)$.

5 | CONCLUSION

This article introduced general difference estimators (GDE) and model-calibrated estimators (MCE) of failure probabilities using time-to-event models for the failure rates of individual cases. The new point estimators, which make use of covariates, and their variance estimators are design-consistent whether the time-to-event model is correctly specified or not. If the time-to-event model is correct, then the estimators are doubly robust. Two simulation studies showed that, for all of the conditions tested, the time-to-event based GDEs and MCEs performed as well, if not better, than the survey-weighted failure estimator that ignores covariates and a general regression estimator based on a linear model that incorporates the same covariates. However, in small samples the estimators are more sensitive to the choice of time-to-event model. A logistic model, in particular, can cause computational problems while lognormal, Weibull, and proportional hazards models did not. In the nurses simulation, the logistic-based estimators performed poorly under every condition with the smaller sample size, where the logistic general difference (LG-GD) estimator had RMSEs that were noticeably higher than those of the π-estimator. In the nurses population, a logistic model is a poorer approximation than the lognormal to $p_N(t)$, the proportion of the population that experiences an event at or before time t. Considering their statistical inefficiency and computational issues, it is clear that estimators based on a logistic time-to-event model will not a good choice for some datasets.

The time-to-event MCE did not perform better than the GDE, even when the relationship between the predictor Z and $\ln(T)$ was weak. This is contrary to the results in Wu and Sitter[10], who did not study time-dependent events. In their study MCE outperformed GDE for all values of the correlation between a covariate and an analysis variable. In our study, reductions in RMSE, compared to that of the π-estimator, were positively correlated with $p_N(t)$, which is consistent with the results in Wu and Sitter.

An important practical finding from the nurses' data simulation is that the time-to-event based GDEs and MCEs performed particularly well, compared to the basic survey-weighted π-estimator, when model information was not also used in the sample

design. Therefore, these estimators will perform best when covariates are available that are both predictive of the time-to-event and not used in the sample design. This might occur if good covariate information is not available at the time of data collection but is available afterwards, or if the sample is not specifically designed to estimate times to events. For example, when covariate information is obtained from administrative records, the lag time between the survey data collection and the acquisition, preparation, and linking of administrative data can be lengthy. Another example is a longitudinal survey where the sample is drawn at the beginning of a panel and covariates are collected sometime after the panel is originally fielded as in the Health and Retirement Study (HRS) or the Panel Study of Income Dynamics (PSID).

Additional work can be done in applying survival models to complex survey data. Although we covered only single-stage sampling, the theory can be extended to multistage sampling using standard methods in Fuller[28]. Multistage sampling is used in many household surveys like the HRS and PSID and will affect the form of variance estimators. The variance estimator presented here did perform well in simulations, but it does treat the estimated failure rate, $\hat{p}(t)$, at a particular time t as fixed when it, in fact, is estimated. Theoretical and empirical work is needed to determine whether replication estimators, like the bootstrap, can reflect this extra source of variation and would be preferable, especially in multistage samples. Adapting existing diagnostics or developing new ones for assessing model fit when using survey data is another important area for research. Finally, work is needed on additional time-to-event models. Threshold regression models, in particular, have been shown to have advantages when a proportional hazard assumption is incorrect.

DATA AVAILABILITY

The data that support the findings in Section 3, based on artificial data, are available from the corresponding author upon reasonable request. The rights to use the data in the simulations in Section 4 can be purchased from the Nurses Health Study, https://www.nurseshealthstudy.org/researchers. Restrictions apply to the availability of these data, which were used under license for this study.

ACKNOWLEDGMENTS

References

1. Irving SK, Loveless TA. *Dynamics of economic well-being: Participation in government programs, 2009-2012: Who gets assistance?*. Research Report: P70-141: U.S. Census Bureau; 2015.

2. Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika.* 1992;79(1):139–147.

3. Lin D. On fitting Cox's proportional hazards models to survey data. *Biometrika.* 2000;87(1):37–47.

4. Boudreau C, Lawless J. Survival analysis based on the proportional hazards model and survey data. *Can J Stat.* 2006;34(2):203–216.

5. Kong Lan, Cai Jianwen, Sen Pranab K. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika.* 2004;91(2):305–319.

6. Kong Lan, Cai Jianwen. Case–Cohort Analysis with Accelerated Failure Time Model. *Biometrics.* 2009;65(1):135–142.

7. Chiou Sy Han, Kang Sangwook, Yan Jun. Fitting accelerated failure time models in routine survival analysis with R package aftgee. *J Stat Softw.* 2014;61(11):1–23.

8. Chiou Sy Han, Kang Sangwook, Yan Jun. Semiparametric accelerated failure time modeling for clustered failure times from stratified sampling. *J Am Stat Assoc.* 2015;110(510):621–629.

9. Heeringa SG, West BT, Berglund PA. *Applied Survey Data Analysis*. CRC Press; 2010.

10. Wu C, Sitter RR. A model-calibration approach to using complete auxiliary information from survey data. *J Am Stat Assoc.* 2001;96(453):185–193.

11. Särndal CE, Swensson B, Wretman J. *Model Assisted Survey Sampling*. New York: Springer Science & Business Media; 1992.

12. Deville JC, Särndal CE. Calibration Estimators in Survey Sampling. *J Am Stat Assoc.* 1992;87(418):376-382.

13. Kennel T. Topics in model-assisted point and variance estimation in clustered samples. Phd thesis, University of Maryland, 2013. https://drum.lib.umd.edu/handle/1903/14064.

14. Robins JM, Hernan MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology.* 2000;11(5):550–560.

15. Kang JDY, Schafer JL. Demystifying double robustness: A comaprison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci.* 2007;22(4):523–539.

16. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Series B Methodol.* 1972;34(2):187–200.

17. Hosmer DW, May S, Lemeshow S. *Applied Survival Analysis*. Wiley-Interscience; 2008.

18. Lin D, Wei LJ, Ying Z. Accelerated failure time models for counting processes. *Biometrika.* 1998;85(3):605–618.

19. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med.* 1992;11(14-15):1871–1879.

20. Lee M-L T, Whitmore GA. Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Anal.* 2010;16(2):196–214.

21. Lee M-L T, Whitmore GA. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Stat Sci.* 2006;21:501–513.

22. Reist BM. Model-Assisted Estimators for Time-to-Event Data. Phd thesis, University of Maryland, 2017. https://drum.lib.umd.edu/handle/1903/20303.

23. Lawless JF. *Statistical Models and Methods for Lifetime Data (Wiley Series in Probability & Statistics)*. New York: John Wiley & Sons; 2nd ed.2003.

24. Cox DR. Partial likelihood. *Biometrika.* 1975;62(2):269–276.

25. Louis TA. Nonparametric analysis of an accelerated failure time model. *Biometrika.* 1981;68(2):381–390.

26. Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *Ann Stat.* 1990;18(1):354–372.

27. Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika.* 2003;90(2):341–353.

28. Fuller WA. *Sampling Statistics*. New York: John Wiley & Sons; 2011.

29. Binder DA. On the variances of asymptotically normal estimators from complex surveys. *Int Stat Rev.* 1983;51(3):279–292.

30. Lawless JF. Event History Analysis and Logitudinal Surveys. In: Chambers Ray L, Skinner Chris J, eds. *Analysis of Survey Data*, John Wiley & Sons 2003 (pp. 221–243).

31. Li Y, Liao D, Lee M-L T. Using threshold regression to analyze survival data from complex surveys: with application to mortality linked NHANES III Phase II Genetic Data. In: American Statistical Association; 2015.

32. Breslow N. Discussion on Regression models and life-tables (by D.R. Cox). *J. Roy. Statist. Soc. Ser. B.* 1972;34:216–217.

33. Breslow N. Covariance analysis of censored survival data. *Biometrics.* 1974;30:89–99.

34. Lin D. On the Breslow estimator. *Lifetime Data Anal.* 2007;13(4):471–480.

35. Nelson NJ. Nurses's Health Study: Nurses Helping Science and Themselves. *J Natl Cancer Inst.* 2000;92(8):597–599.

36. Bain C, Feskanich D, Speizer FE, et al. Lung cancer rates in men and women with comparable histories of smoking. *J Natl Cancer Inst.* 2004;96(11):826–834.

37. Lee M-L T, Whitmore GA, Rosner B. Benefits of threshold regression: A case-study comparison with Cox proportional hazards regression. In: Mathematical and Statistical Models and Methods in Reliability, Springer 2010 (pp. 359–370).

38. Albert A, Anderson JA. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika.* 1984;71:1–10.

39. Cochran WG. *Sampling Techniques*. New York: John Wiley & Sons, Inc.; 1977.

How to cite this article: Reist B., and Valliant, R., (2020), Model-Assisted Estimators for Time-to-Event Data, *Statistics in Medicine*, *2019;xx:xx–xx*.

APPENDIX

Proof of Theorem 1. Since (9) can be rewritten as

$$\hat{p}_{GD}(t) = \hat{p}_{\pi}(t) + N^{-1}\left(\sum_{i=1}^{N} p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) - \sum_{i\in s} d_i p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})\right)$$

and $\hat{p}_{\pi}(t)$ is design-consistent, it suffices to show that

$$\left(N^{-1}\sum_{i=1}^{N} p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) - N^{-1}\sum_{i\in s} d_i p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})\right) = O_p\left(n^{-1/2}\right).$$

Using assumptions (i) and (ii) and applying a Taylor series approximation to $p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)$ at $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_N$, we get

$$p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) = p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) + \left[\frac{\partial p\left(t \mid \mathbf{x}_i, \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\theta}^*}\right]^T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N\right), \tag{1}$$

where $\boldsymbol{\theta}^* \in (\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_N)$ or $(\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}})$. Now by (1) and assumptions (i) and (ii),

$$N^{-1}\sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) = N^{-1}\sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) + O_p\left(n^{-1/2}\right), \tag{2}$$

and

$$N^{-1}\sum_{i\in s} d_i p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) = N^{-1}\sum_{i\in s} d_i p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) + O_p\left(n^{-1/2}\right). \tag{3}$$

Note that because of condition (iii)

$$N^{-1}\sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) - N^{-1}\sum_{i\in s} d_i p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) = O_p\left(n^{-1/2}\right). \tag{4}$$

Now, by putting together (2), (3), and (4), we get

$$N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) - N^{-1} \sum_{i \in s} d_i p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) = O_p\left(n^{-1/2}\right),$$ (5)

as desired. $\square$

Proof of Theorem 2. Using assumptions (i), (ii), (iv) and applying a Taylor series second order approximation to $p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)$ at $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_N$, we get

$$
\begin{aligned}
p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) = p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) + \left[\frac{\partial p\left(t \mid \mathbf{x}_i, \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\theta}^*}\right]^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) \\
+ \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N)^T \left[\frac{\partial^2 p(t \mid \mathbf{x}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\Big|_{\boldsymbol{\theta}^*}\right](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N),
\end{aligned}
$$ (6)

where $\boldsymbol{\theta}^* \in (\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_N)$ or $(\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}})$ and $\left[\frac{\partial^2 p(t \mid \mathbf{x}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\Big|_{\boldsymbol{\theta}^*}\right]$ is the $p \times p$ matrix of second derivatives evaluated at $\boldsymbol{\theta}^*$. Now, by (6) and assumption (iv),

$$
\begin{aligned}
N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) =& N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) \\
& + \left\{N^{-1} \sum_{i=1}^{N} \frac{\partial p\left(t \mid \mathbf{x}_i, \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\theta}^*}\right\}^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) \\
& + O_p\left(n^{-1}\right)
\end{aligned}
$$ (7)

and

$$
\begin{aligned}
N^{-1} \sum_{i \in s} d_i p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) =& N^{-1} \sum_{i \in s} d_i p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) \\
& + \left\{N^{-1} \sum_{i \in s} d_i \frac{\partial p\left(t \mid \mathbf{x}_i, \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\theta}^*}\right\}^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) \\
& + O_p\left(n^{-1}\right).
\end{aligned}
$$ (8)

By assumptions (i) and (iii), we have

$$\left\{N^{-1} \sum_{i=1}^{N} \frac{\partial p\left(t \mid \mathbf{x}_i, \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\theta}^*}\right\} - \left\{N^{-1} \sum_{i \in s} d_i \frac{\partial p\left(t \mid \mathbf{x}_i, \boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\theta}^*}\right\} = O_p\left(n^{-1/2}\right).$$ (9)

Therefore, by subtracting (8) from (7), and using assumption (i) that $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) = O_p\left(n^{-1/2}\right)$, we get

$$
\begin{aligned}
& N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) - N^{-1} \sum_{i \in s} d_i p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) \\
& \qquad = N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) - N^{-1} \sum_{i \in s} d_i p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) + O_p\left(n^{-1}\right).
\end{aligned}
$$ (10)

Using Theorem 1 and (10) to replace $\hat{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}_N$ in $\hat{p}_{GD}(t)$ gives

$$
\begin{aligned}
\hat{p}_{GD}(t) =& \hat{p}_{\pi}(t) + \left(N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) - N^{-1} \sum_{i \in s} d_i p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right)\right) + O_p\left(n^{-1/2}\right) \\
=& N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right) + N^{-1} \sum_{i \in s} d_i \left[I_{\{T_i \leq t\}} - p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right)\right] + O_p\left(n^{-1/2}\right).
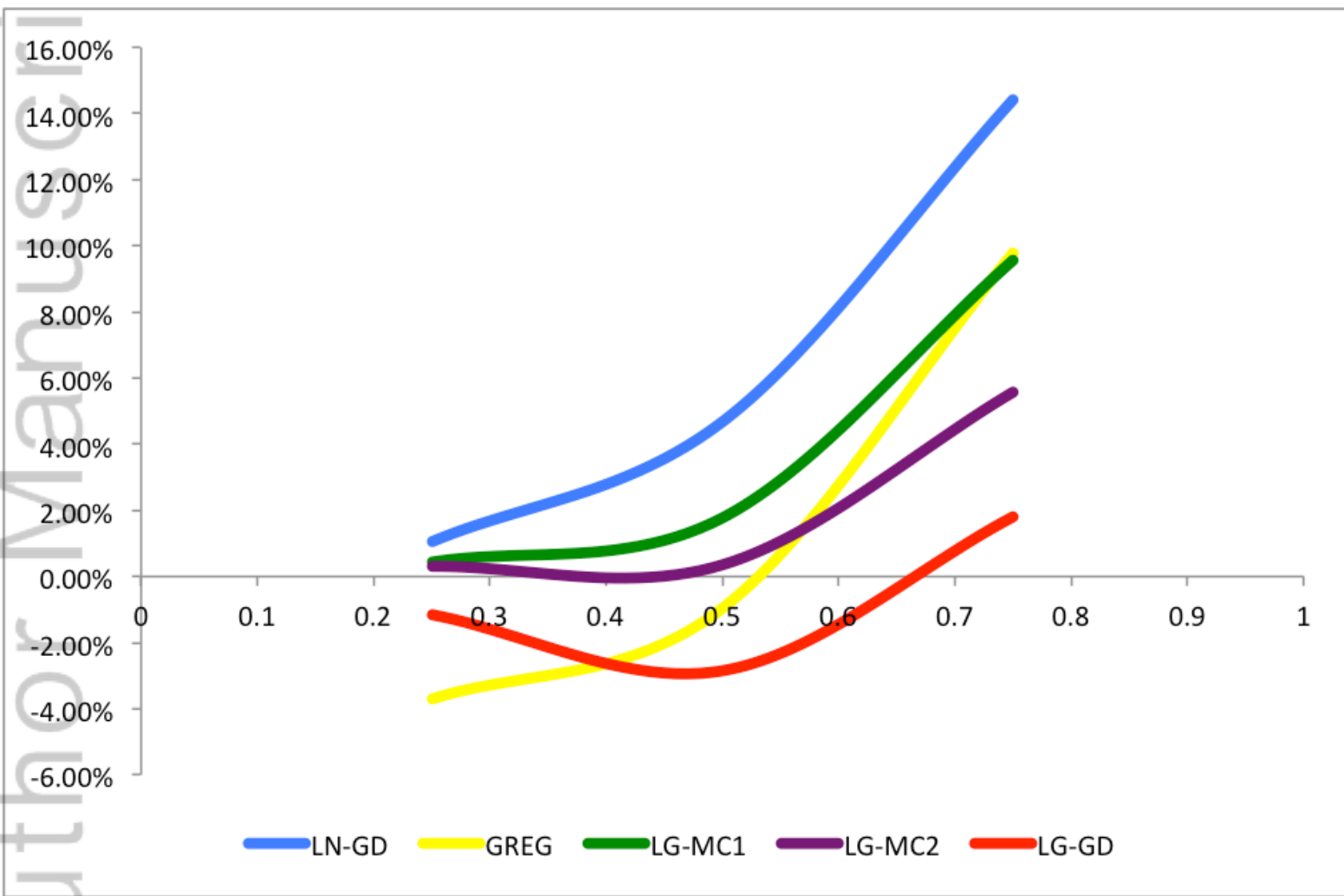\end{aligned}
$$ (11)

Finally, by noticing that $N^{-1} \sum_{i=1}^{N} p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right)$ is constant, the asymptotic variance of $\hat{p}_{GD}(t)$ is the asymptotic variance of the π-estimator of the population total of the $e_i = I_{\{T_i \leq t\}} - p\left(t \mid \mathbf{x}_i, \boldsymbol{\theta}_N\right)$. It now follows that the asymptotic variance estimator of $\hat{p}_{GD}(t)$ is the asymptotic variance estimator evaluated using the $\hat{e}_i = I_{\{T_i \leq t\}} - p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)$. The design-based formula for

whatever sample design was used then applies for estimating the design-variance of $N^{-1} \sum_{i \in s} d_i \left[I_{\{T_i \leq t\}} - p\left(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) \right]$, which is an estimated total (e.g., see Cochran, sec. 9.14[39]) for the formula used in Theorem 2. □

The case in which an estimator, $\hat{N} = \sum_s d_i$, is used in $\hat{p}_{GD}$ can be handled by using approximations similar to those above. We sketch the result here. Using a first order Taylor series approximation gives

$$\hat{p}_{GD}(t) = \hat{N}^{-1} \left(\sum_{i=1}^N p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) + \sum_{i \in s} d_i \left[I_{\{T_i \leq t\}} - p(t \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}}) \right] \right)$$
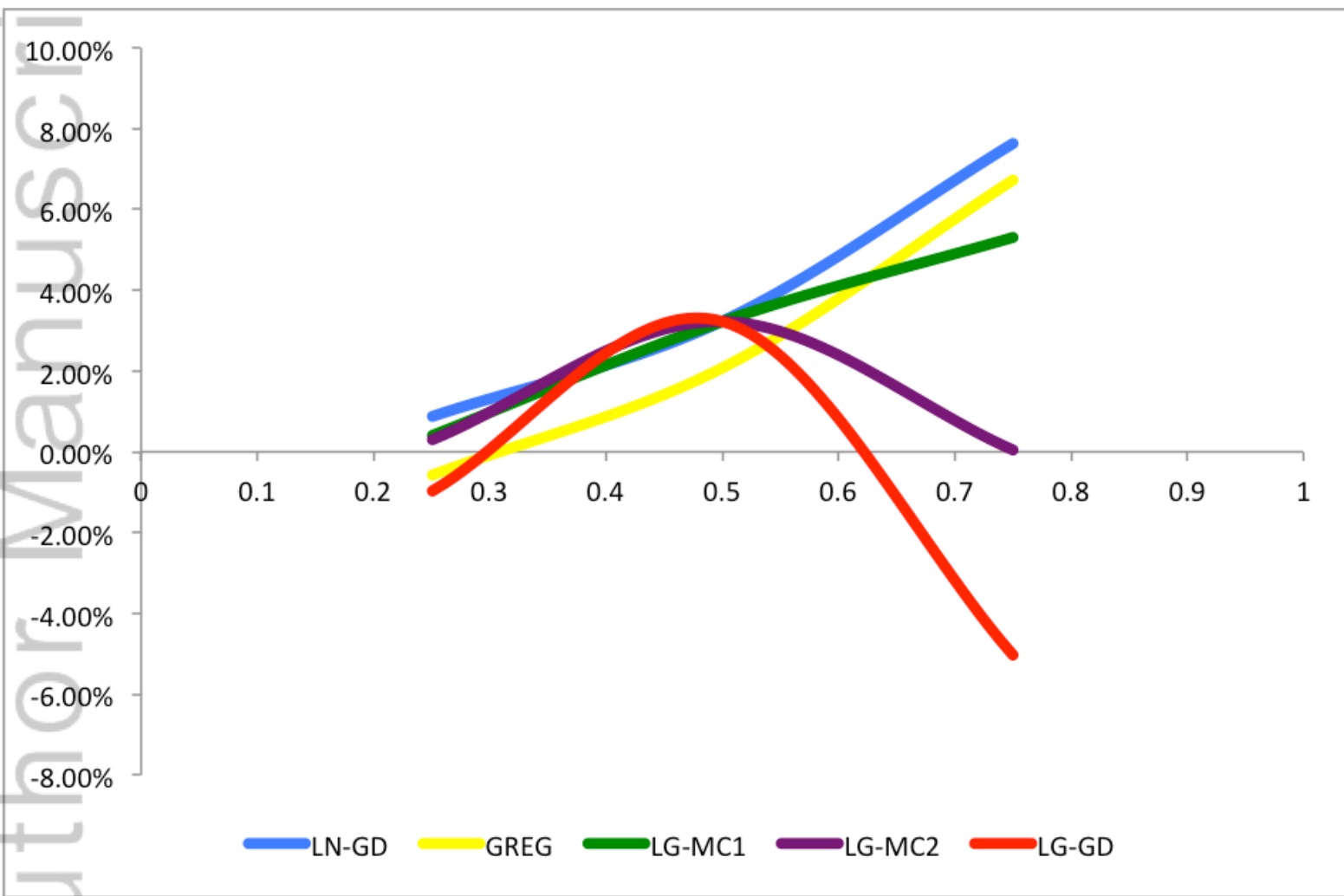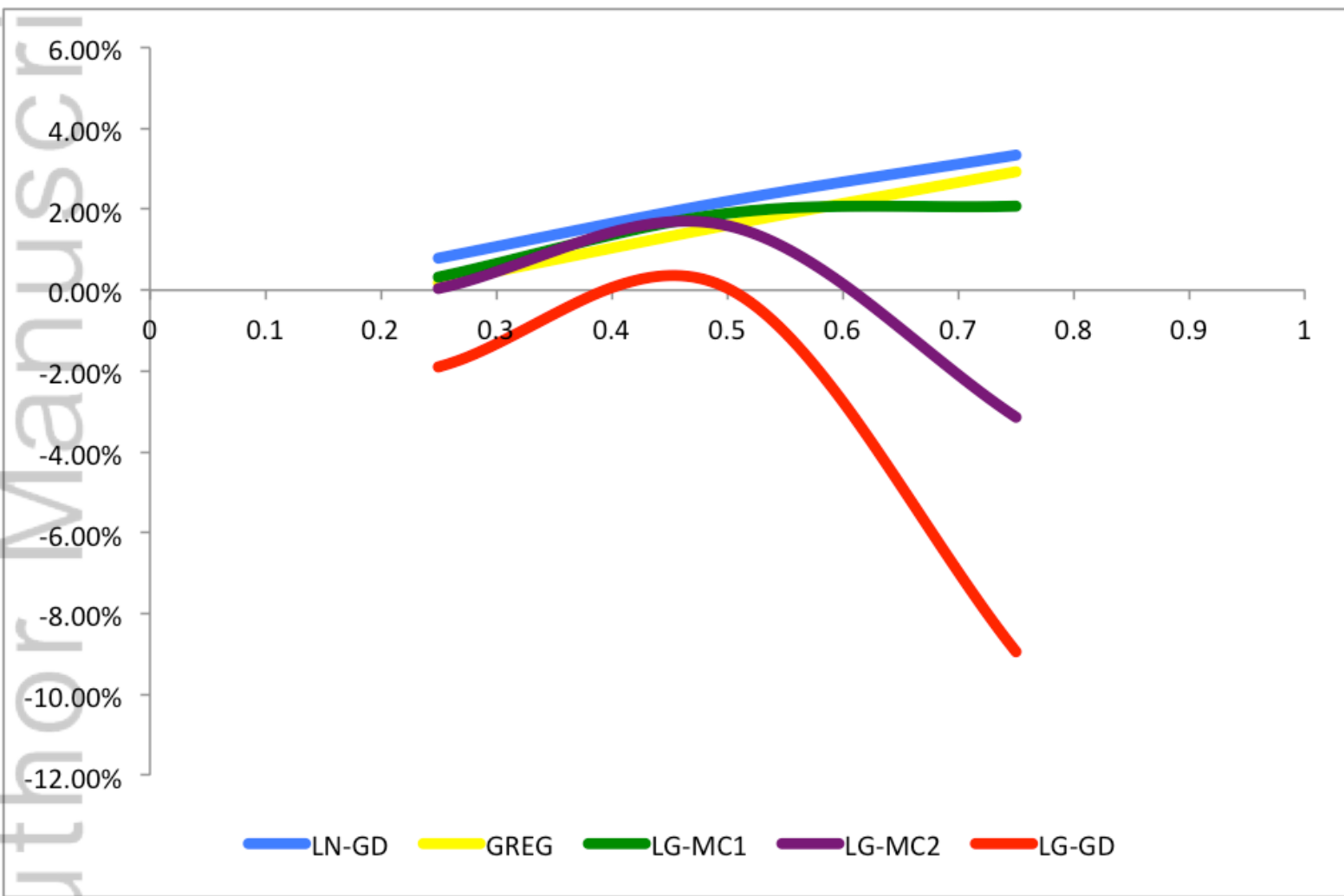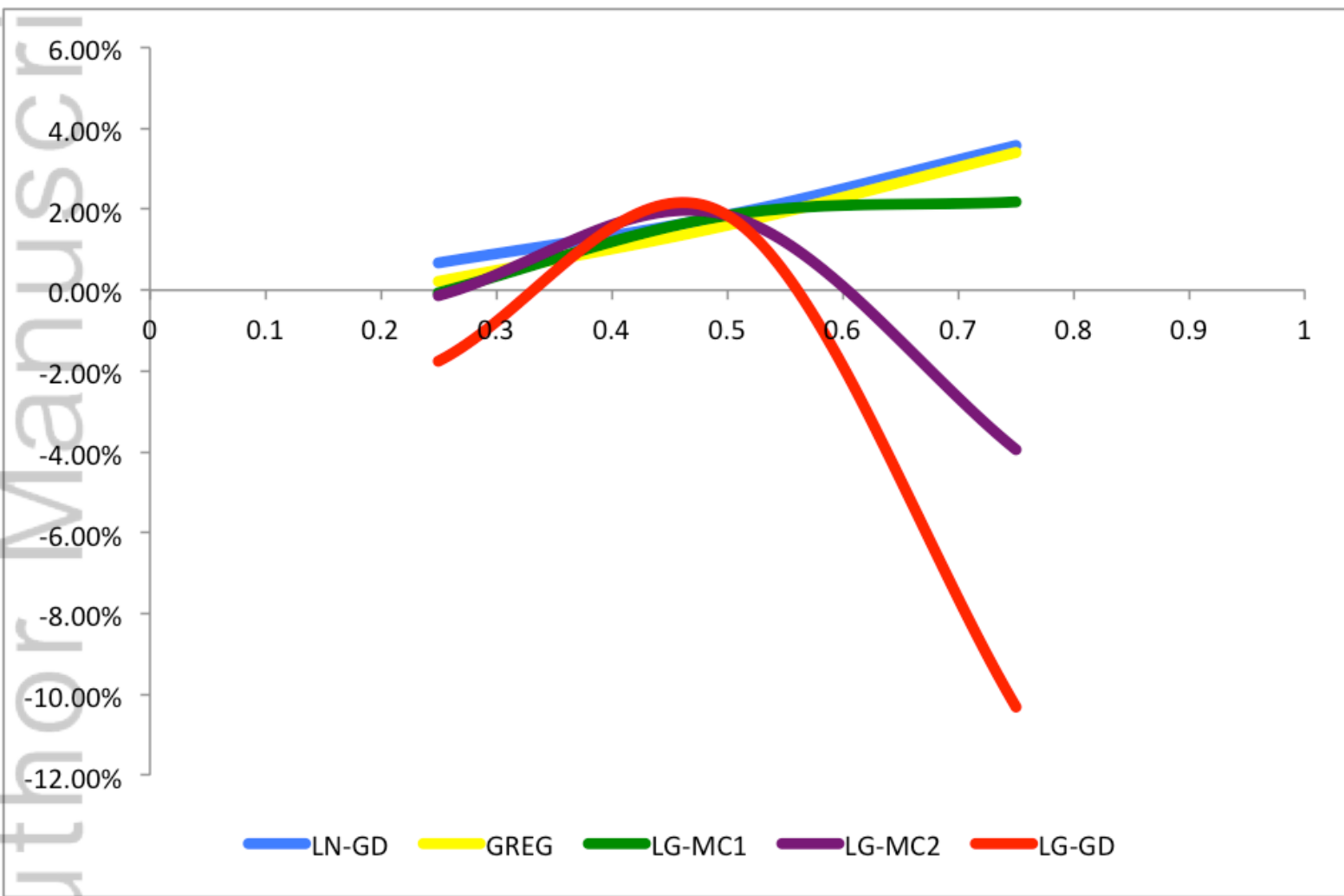
SIM_8728_Figure_1_n200.png
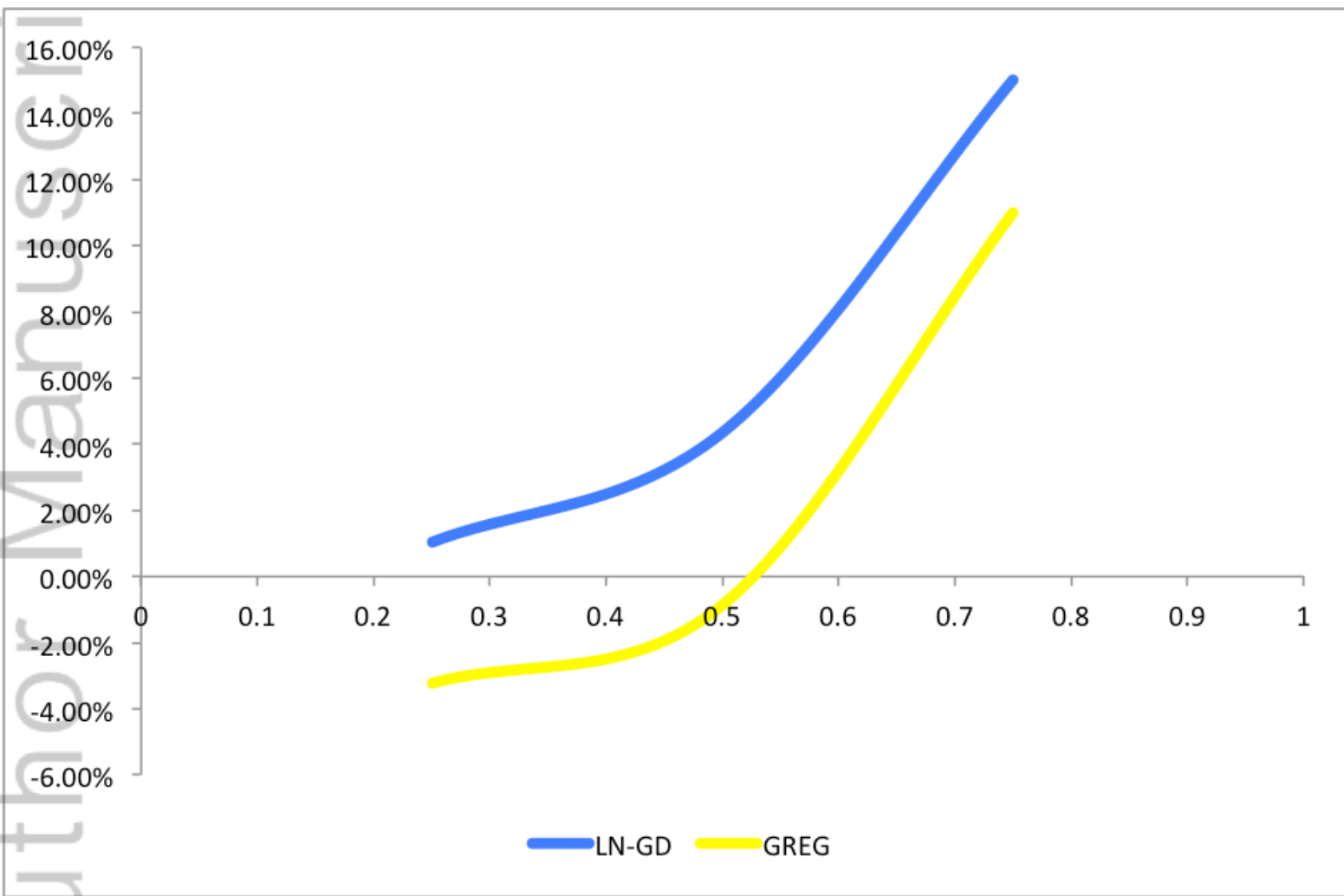
SIM_8728_Figure_1_n1000.png
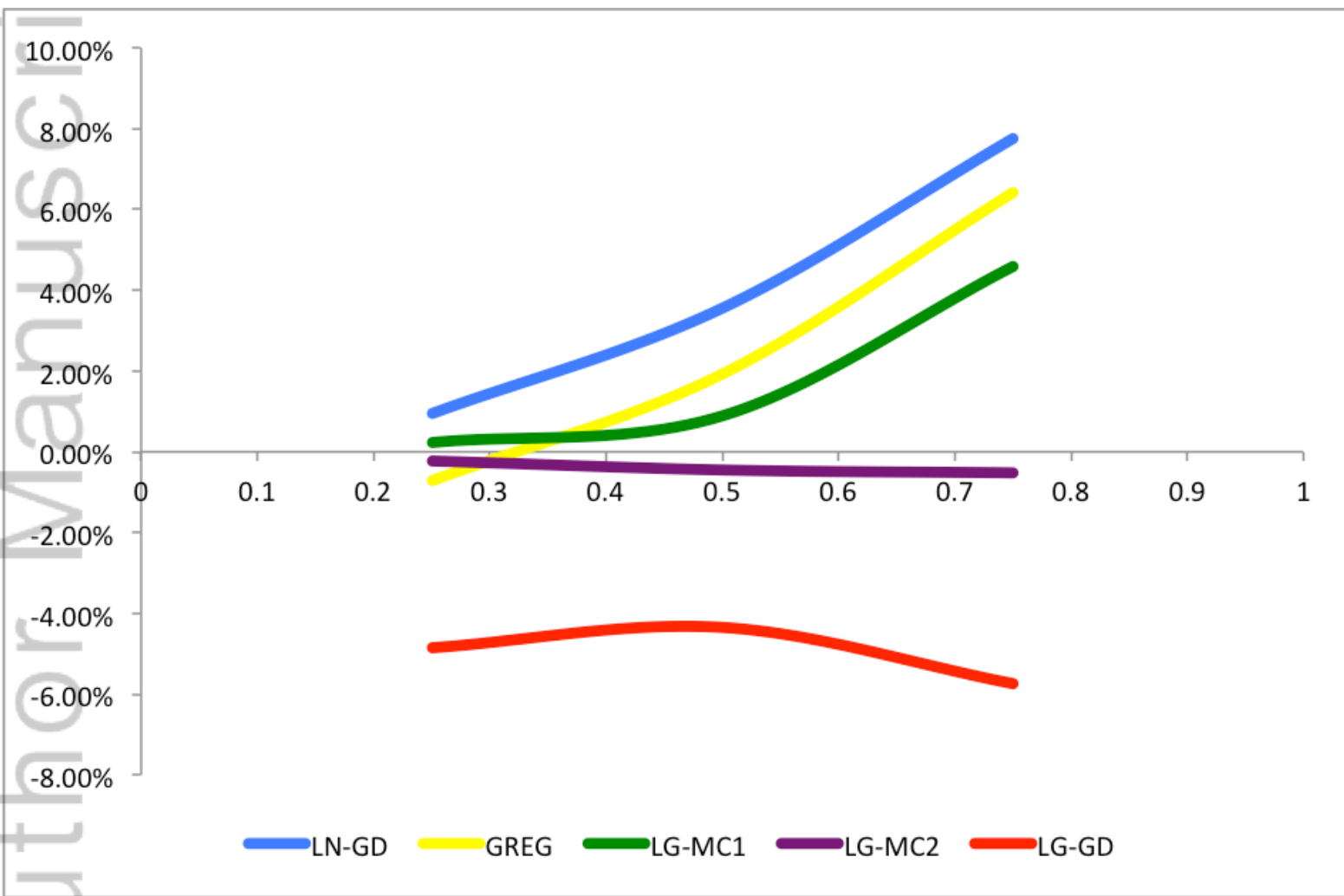
SIM_8728_Figure_2_n200.png
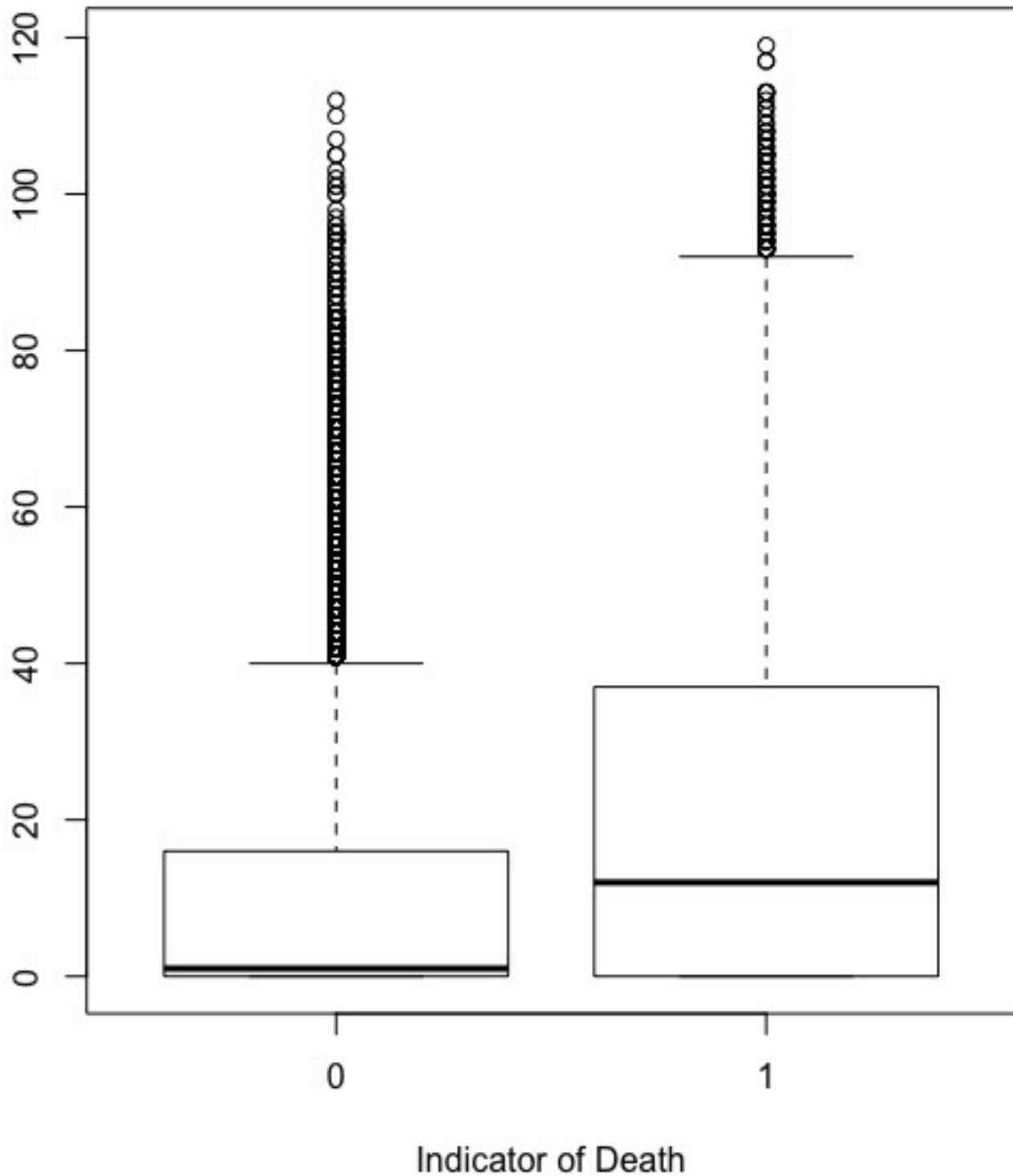
SIM_8728_Figure_2_n1000.png

SIM_8728_Figure_3_n200.png

SIM_8728_Figure_3_n1000.png

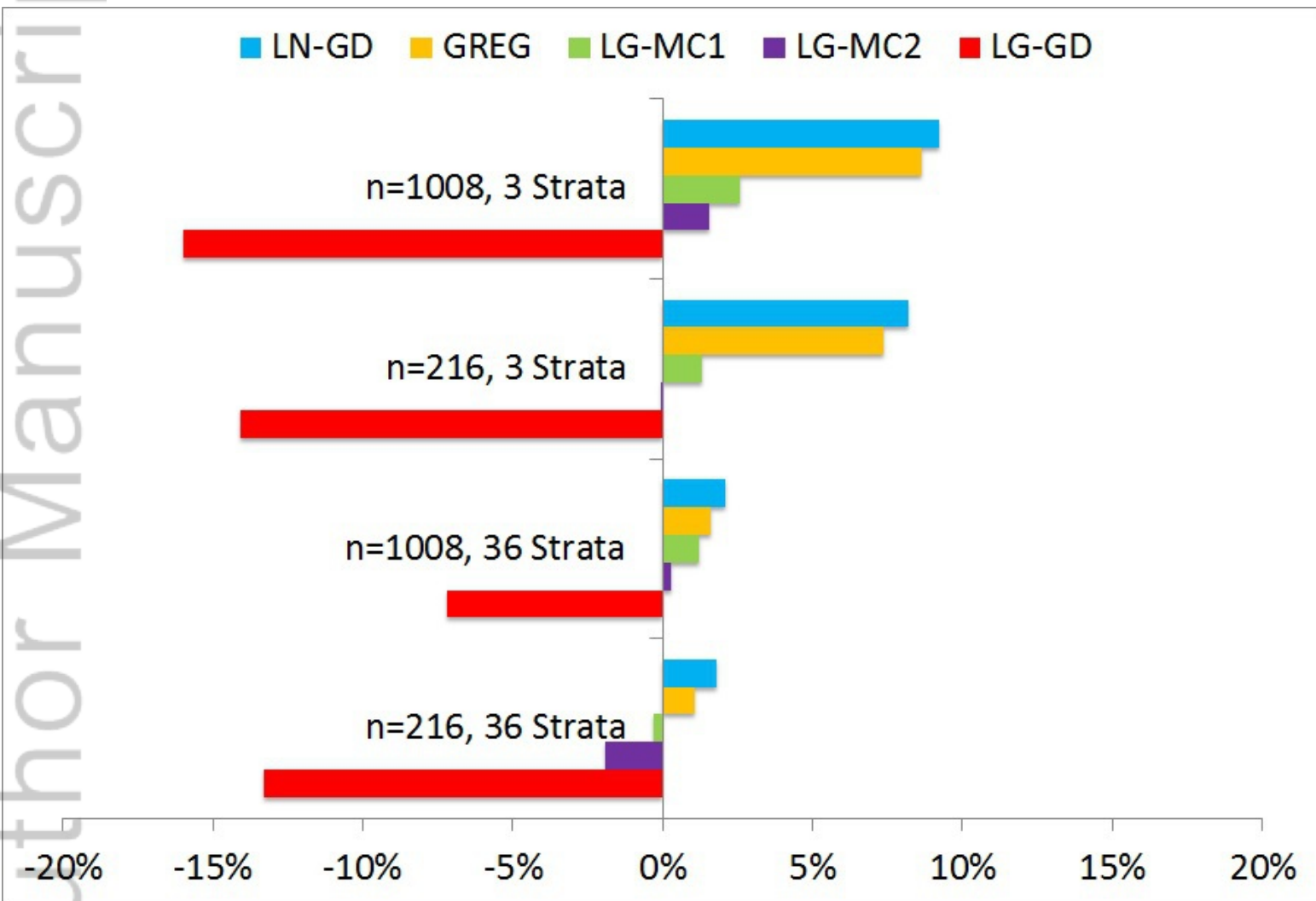SIM_8728_Pack_Years_Box_Plot.jpeg

SIM_8728_RMSE_NHS.jpg

Metric	n	# Strata	π	GREG	Weibull			Lognormal			Proportional Hazard			Logistic		
					GD	MC1	MC2	GD	MC1	MC2	GD	MC1	MC2	GD	MC1	MC2
% Δ RMSE	216	3	0.00%	7.38%	8.24%	8.13%	8.17%	8.22%	8.04%	8.15%	8.20%	8.13%	8.16%	-14.07%	1.30%	-0.01%
		36	0.00%	1.07%	1.68%	1.62%	1.64%	1.79%	1.65%	1.71%	1.65%	1.60%	1.62%	-13.31%	-0.30%	-1.91%
	1008	3	0.00%	8.61%	9.58%	9.57%	9.58%	9.20%	9.38%	9.35%	9.59%	9.58%	9.58%	-15.97%	2.59%	1.57%
		36	0.00%	1.62%	2.22%	2.22%	2.22%	2.11%	2.12%	2.14%	2.23%	2.23%	2.23%	-7.21%	1.21%	0.27%
% RB	216	3	-0.02%	0.10%	0.00%	0.00%	0.00%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.90%	0.27%	0.41%
		36	-0.02%	0.14%	-0.07%	-0.07%	-0.07%	0.02%	0.03%	0.02%	-0.07%	-0.07%	-0.07%	0.53%	0.19%	0.31%
	1008	3	0.02%	0.02%	0.01%	0.01%	0.01%	0.02%	0.02%	0.02%	0.01%	0.01%	0.01%	0.20%	0.02%	0.06%
		36	0.05%	0.10%	0.07%	0.07%	0.07%	0.08%	0.08%	0.08%	0.07%	0.07%	0.07%	0.21%	0.08%	0.11%
BR	216	3	-0.001	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.023	0.035
		36	-0.001	0.011	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	0.035	0.014	0.022
	1008	3	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.031	0.004	0.012
		36	0.008	0.016	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.032	0.013	0.018
VR	216	3	1.01	0.96	0.95	0.95	0.95	0.96	0.95	0.96	0.95	0.95	0.95	0.98	0.98	0.97
		36	1.00	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.96	0.97	0.98	0.97
	1008	3	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	0.99
		36	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99
Coverage	216	3	0.949	0.941	0.942	0.940	0.941	0.943	0.940	0.942	0.941	0.940	0.941	0.944	0.945	0.943
		36	0.940	0.937	0.935	0.935	0.935	0.937	0.937	0.937	0.935	0.935	0.935	0.942	0.938	0.939
	1008	3	0.952	0.951	0.952	0.952	0.952	0.951	0.953	0.952	0.952	0.952	0.952	0.950	0.951	0.949
		36	0.947	0.948	0.950	0.950	0.950	0.948	0.948	0.948	0.950	0.950	0.950	0.945	0.948	0.947