THE MARKETING JOURNAL OF THE INSTITUTE FOR OPERATIONS RESEARCH AND THE MANAGEMENT SCIENCES

informs

MARKETING SCIENCE

Self-Signaling and Prosocial Behavior: A Cause Marketing Experiment
Jean-Pierre Dubé, Xueming Luo, Zheng Fang
161

Signaling Virtue: Charitable Behavior Under Consumer Elective Pricing
Minah H. Jung, Leif D. Nelson, Uri Gneezy, Ayelet Gneezy
187

A Cross-Cohort Changepoint Model for Customer-Base Analysis
Arun Gopalakrishnan, Eric T. Bradlow, Peter S. Fader
195

Is Advance Selling Desirable with Competition?
Gérard P. Cachon, Pnina Feldman
214

Can Retail Sales Volatility be Curbed Through Marketing Actions?
Mercedes Esteban-Bravo, Jose M. Vidal-Sanz, Gökhan Yildirim
232

The Effects of Product Line Breadth: Evidence from the Automotive Industry
Antonio Moreno, Christian Terwiesch
254

Entry of Copycats of Luxury Brands
Sarah Yini Gao, Wei Shi Lim, Christopher S. Tang
272

"Ten Million Readers Can't be Wrong!," or Can They?
On the Role of Information About Adoption Stock in New Product Trial
Coby Morvinski, On Amir, Eitan Muller
290

Could Good Intentions Backfire? An Empirical Analysis of the Bank Deposit Insurance
Weining Bao, Jian Ni
301

Focus on Authors
320

2016 Guest Editors, Guest Associate Editors, and Ad Hoc Reviewers
324

Volume 36 • Number 2 • March–April 2017

Marketing Science

Tricked by Truncation: Spurious Duration Dependence and Social Contagion in Hazard Models

Christophe Van den Bulte, Raghuram Iyengar,

To cite this article:
Christophe Van den Bulte, Raghuram Iyengar, (2011) Tricked by Truncation: Spurious Duration Dependence and Social Contagion in Hazard Models. Marketing Science 30(2):233-248. https://doi.org/10.1287/mksc.1100.0615

MARKETING SCIENCE
CELEBRATING 30 YEARS
Vol. 30, No. 2, March–April 2011, pp. 233–248
ISSN 0732-2399 | EISSN 1526-548X | 11 | 3002 | 0233
DOI 10.1287/mksc.1100.0615
© 2011 INFORMS

Tricked by Truncation: Spurious Duration Dependence and Social Contagion in Hazard Models

Christophe Van den Bulte, Raghuram Iyengar
The Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania 19104
{vdbulte@wharton.upenn.edu, riyengar@wharton.upenn.edu}

We show both analytically and through Monte Carlo simulations that applying standard hazard models to right-truncated data, i.e., data from which all right-censored observations are omitted, induces spurious positive duration dependence and hence can trick researchers into believing to have found evidence of social contagion when there is none. Truncation also tends to deflate the effect of time-invariant covariates. These results imply that not accounting for right truncation can lead managers to rely too much on word of mouth in generating new product adoption and to poorly identify the customers most likely to adopt early. Not accounting for right truncation can also lead to suboptimal pricing decisions and to erroneous assessments of variations in customer lifetime value. We assess the effectiveness of four possible solutions to the problem and find that only using an analytically corrected likelihood function protects one against truncation artifacts inflating coefficients of contagion and attenuating coefficients of time-invariant covariates.

Key words: hazard models; duration dependence; new product diffusion; social contagion
History: Received: May 4, 2009; accepted: September 19, 2010; Michel Wedel served as the guest editor-in-chief and Wayne DeSarbo served as associate editor for this article. Published online in *Articles in Advance* December 30, 2010.

1. Introduction

Marketers are often interested in understanding how long it takes for events to happen and what drives these durations to be short or long. What explains how long it takes for potential customers to adopt a new product? For firms to enter a new market? For current customers to churn? For salespeople to leave? Typically, these questions are investigated using hazard models that take into account that some cases are "right-censored," i.e., that they will ultimately make the transition but have not done so yet by the end of the observation window.

Hazard models are frequently used in studies of new product adoption and firm entry investigating some type of contagion. Convincingly documenting contagion effects in nonexperimental studies is difficult and subject to several threats to validity, stemming mostly from omitted variables and reverse causality. These challenges are by now well recognized in sociology, economics, and marketing (e.g., Aral et al. 2009, Erbring and Young 1979, Jackson 2008, Van den Bulte and Lilien 2001, Van den Bulte and Stremersch 2004). What has not been appreciated so far in any of those fields, however, is that right truncation in hazard modeling—i.e., not observing or ignoring right-censored cases, an issue faced by marketing scientists in several recent studies—can also trick one into seeing evidence of contagion when there is none.[1] Using both analytics and Monte Carlo simulations, we first show that omitting right-censored observations from one's data indeed generates spurious positive duration dependence and other artifacts, and we then assess the effectiveness of four possible corrections to protect one from such truncation artifacts.[2]

Standard hazard rate modeling properly accounts for the information about the right-censored cases: even though one does not know these cases' exact

[1] Truncation and censoring in duration data are related but distinct concepts. Censoring occurs whenever there are cases whose event times fall outside the observation window, whereas truncation occurs whenever such censored cases are excluded from the analysis. Right censoring occurs when one or more cases at risk of transitioning have not done so yet by the end of the observation window. Right truncation occurs when these right-censored cases are excluded from the analysis. Left censoring refers to situations where the unobserved transitions take place before the start of the observation window rather than after its end, and left truncation occurs when the left censored cases are excluded from the analysis. Unlike right truncation, left truncation does not create spurious duration dependence or spurious contagion (see §1 of the electronic companion to this paper, available as part of the online version that can be found at http://mktsci.pubs.informs.org/).

[2] We focus our analysis on right-truncation bias in hazard modeling of duration data, as opposed to the issues of truncation, response-based sampling, and selective data pruning in cross-sectional and panel data structures, which are already well documented (e.g., Andrews and Currim 2005, Zanutto and Bradlow 2006).

event time, one knows it falls after the end of the observation window. Still, there are several occasions in which researchers do not observe the censored cases at all and have genuinely truncated data (e.g., Moe and Fader 2002, Nam et al. 2010) or where they choose to delete such cases under the assumption that they were never at risk of adopting anyway, either in the main analysis (e.g., Manchanda et al. 2008) or as a robustness check (e.g., Bell and Song 2007). Researchers may also choose to delete right-censored cases from their data because their main interest lies in postadoption behavior, which is observed only for adopters (e.g., Prins et al. 2009). Excluding right-censored cases does not only lead to a loss of information and efficiency, but—unless appropriate steps are taken—also generates several biases that can lead to erroneous research conclusions.

It is rather intuitive that omitting right-censored cases without taking precautions in the estimation procedure generates a downward bias in the expected duration and hence an upward bias in the mean hazard rate. This has long been known in statistics (e.g., Deemer and Votaw 1955, Den Broeder 1955) and has not escaped the notice of some social scientists either (e.g., Helsen and Schmittlein 1993, Tuma and Hannan 1979). What has not been recognized, however, is that such practice also generates spurious positive duration dependence, i.e., an upward bias in how the hazard increases over time. Because positive duration dependence is often interpreted as evidence of contagion, applying standard hazard modeling techniques to truncated data can trick one into believing that there is evidence of contagion when there is none. This problem is similar to how underestimating the ceiling parameter is associated with overestimating the contagion parameter in macrolevel diffusion models (Van den Bulte and Lilien 1997).

Truncation artifacts have implications for other substantive issues besides contagion. For instance, it may inflate the estimated price sensitivity for a new product and so lead firms to charge prices that are too low. It can also deflate the importance of time-invariant drivers of customer churn and so lead to erroneous customer lifetime value assessments. Our focus, however, will mainly be on contagion.

We first present some new analytical results on spurious duration dependence induced by right truncation. We then present the results from simulation studies in which we generate spurious positive duration dependence and contagion when the true data-generating process has neither, but the model ignores right truncation. We also find that right truncation attenuates the coefficients of time-invariant covariates. Next, we illustrate that these biases can negatively affect marketing policies in new product advertising, pricing, and customer management. Finally, we assess the effectiveness of four simple approaches that researchers dealing with right-truncated data can use to protect themselves from being tricked by truncation: (i) using a likelihood function that correctly accounts for right truncation, (ii) using a nonparametric baseline hazard that absorbs the artificial positive duration dependence induced by truncation, (iii) using random effects, and (iv) including both a nonparametric baseline and random effects. We find that only the first approach protects against truncation artifacts—inflated coefficients of contagion and attenuated coefficients of time-invariant covariates. Finally, we show that uncertainty about who should be included in the risk set does not justify self-inflicted truncation because mover–stayer or split-population hazard models effectively handle that problem. For simplicity of exposition and relevance to marketers, all analyses pertain to situations where censoring occurs at some exogenous time that is constant across cases (Type I censoring).

2. Mathematical Analysis

2.1. Homogeneous Case in Continuous Time

Assume that adoption is a continuous-time process, and let $h(t)$, $f(t)$, and $F(t)$ denote the hazard function, probability density function (pdf), and cumulative density function (cdf) of the true adoption times, with $F(0) = 0$. Let $T > 0$ denote the censoring time, such that by the end of the observation window $F(T)$ adoptions have taken place (the distinction between the actual and expected level of censoring at T is not essential for our argument). Let $h_T(t)$, $f_T(t)$, and $F_T(t)$ denote the hazard function, pdf, and cdf of the truncated adoption times given truncation at T ($t \leq T$; $F_T(T) = 1$). The standard relations apply:

$$h(t) = f(t)/[1 - F(t)], \quad 0 \leq t < \infty, \tag{1a}$$

$$h_T(t) = f_T(t)/[1 - F_T(t)], \quad 0 \leq t \leq T. \tag{1b}$$

In addition, the following relation between the two pdfs holds as well (e.g., Klein and Moeschberger 2003):

$$f_T(t) = f(t)/F(T), \quad 0 \leq t \leq T. \tag{2}$$

Substituting (2) into (1b), and taking into account that (2) implies $F(T)F_T(t) = F(t)$, we can express $h_T(t)$ as

$$\begin{aligned} h_T(t) &= \frac{f(t)}{F(T) - F(t)} \\ &= \frac{f(t)}{1 - F(t)} \times \frac{1 - F(t)}{F(T) - F(t)} \\ &= h(t)D_T(t), \quad 0 \leq t \leq T. \end{aligned} \tag{3}$$

Hence, the factor $D_T(t) = (1 - F(t))/(F(T) - F(t))$ is a multiplier indicating to what extent the hazard function of the right-truncated process is different from that of the true process. It simply is the ratio of the number of survivors in the true population and that in the truncated population, and it can also be expressed in terms of the survival function, $S(t) = 1 - F(t)$:

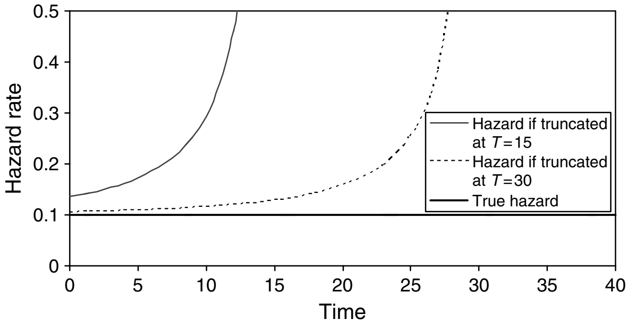
$$D_T(t) = \frac{1 - F(t)}{F(T) - F(t)} = \frac{S(t)}{S(t) - S(T)} = \frac{1}{1 - S(T)/S(t)}, \quad 0 \le t \le T. \tag{4}$$

This divergence has the following notable characteristics.

- The divergence is positive ($D_T(t) > 1$) as long as $F(T) < 1$. This is consistent with prior results that truncation induces an upward bias in the hazard. Also, the divergence becomes larger the more severe the truncation ($dD_T(t)/dF(T) < 0$).
- Assuming $f(t) > 0$ for all t, the divergence becomes larger over time ($dD_T(t)/dt > 0$). Hence, *right truncation induces spurious duration dependence.*
- The amount of spurious duration dependence goes to infinity as time reaches the truncation point T ($D_T(t) \to \infty$ as $F(t) \to F(T) < 1$).
- The divergence is convex with respect to time whenever $df(t)/dt \ge -2[f(t)]^2/[F(T) - F(t)]$. This condition is met at every point in time in a process with hazard $h(t) = p + qF(t)$, regardless of the rate parameters ($p > 0$, $q \ge 0$) and the truncation level, although it need not be for every conceivable process.

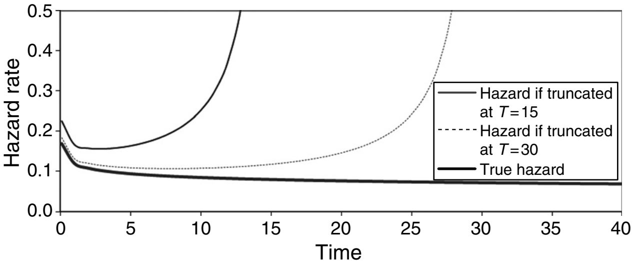
Figure 1 plots the hazard function for an exponential process with a constant hazard rate of 0.10 as well as hazard functions for the process truncated at $T = 15$ and $T = 30$, corresponding to truncation at penetration levels of 79% and 96%, respectively. Because the true hazard is constant, the shape of the hazards for the truncated process is identical to that of the divergence multiplier. Figure 2 plots the hazard function for a Weibull process with the scale parameter equal to 0.10 and a shape parameter of 0.85 as well as hazard functions for the process truncated at $T = 15$ and $T = 30$. As shown in Figure 2, truncation can make a process with negative duration dependence appear as if it has U-shaped duration dependence.

Figure 1 **Truncation Induces Positive Duration Dependence Even When the True Process Has a Constant Hazard**



Note. Exponential process with rate = 0.10.

Figure 2 **Truncation Can Induce U-Shaped Duration Dependence When the True Process Has Negative Duration Dependence**



2.2. Homogeneous Case in Discrete Time

When adoption is a discrete-time process evolving over periods $1, 2, 3, \ldots$, the hazard rate is a probability bounded between 0 and 1 and is defined as

$$h(t) = \frac{F(t) - F(t-1)}{1 - F(t-1)}, \quad 1 \le t < \infty. \tag{5}$$

The hazard of the truncated process now becomes

$$h_T(t) = \frac{F(t) - F(t-1)}{1 - F(t-1)} \times \frac{1 - F(t-1)}{F(T) - F(t-1)}, \quad 1 \le t \le T,$$
$$= h(t)D_T(t), \quad 1 \le t \le T. \tag{6}$$

The multiplicative divergence behaves in the same fashion is in the continuous case, except that in the last period T, it reaches a maximum of $[h(T)]^{-1}$, so the hazard of the truncated process is forced to equal 1.

2.3. Heterogeneity in Hazard Rates

The derivations above did not make any assumption on the shape of the hazard function but did assume that it was homogeneous. When the hazard varies across members i of the population ($i = 1, \ldots, N$), so does the divergence factor. It equals $D_{iT}(t) = (1 - F_i(t))/(F_i(T) - F_i(t))$ for continuous-time processes and $D_{iT}(t) = (1 - F_i(t-1))/(F_i(T) - F_i(t-1))$ for discrete-time processes.

Cases with a consistently higher hazard will have a lower divergence at first, since $D_{iT}(0)$ in continuous time and $D_{iT}(1)$ in discrete time both equal $1/(F_i(T))$, given that $F(0) = 0$. It is difficult to say anything definite about differences at later time points without imposing some parametric structure on the process. Hence, we focus on two instances of special interest: constant additive heterogeneity and constant multiplicative heterogeneity. For ease of exposition, we contrast the divergence of two cases, where one has a higher hazard than the other, $h_2(t) > h_1(t)$. In both instances, we rely on the general property that any continuous-time survival function can be written in terms of the integrated or cumulative hazard function: $S(t) = \exp(-\int_0^t h(u)\,du)$.

Constant additive heterogeneity, where $h_2(t) = \theta + h_1(t)$ and $\theta > 0$, corresponds to adoption processes where some adopters have a higher time-invariant tendency to adopt early or, in Bass model terminology, have a higher coefficient of innovation. In this situation, we can write the survival function of the second customer as

$$S_2(t) = \exp\left(-\int_0^t h_2(u)\,du\right) = \exp\left(-\int_0^t (\theta + h_1(u))\,du\right)$$
$$= \exp\left(-\theta t - \int_0^t h_1(u)\,du\right) = \exp(-\theta t) S_1(t). \quad (7)$$

Substituting this expression in Equation (4) gives

$$D_{2T}(t) = \frac{1}{1 - S_2(T)/S_2(t)} = \frac{1}{1 - e^{-\theta(T-t)} S_1(T)/S_1(t)}$$
$$< \frac{1}{1 - S_1(T)/S_1(t)} = D_{1T}(t), \quad 0 \le t < T. \quad (8)$$

Hence, when the hazards of adoption differ between cases by an additive constant, then the amount of divergence between the true and truncated hazard differs between cases too as long as $t < T$ (as $t \to T$, all divergences tend to infinity in continuous-time processes). Specifically, cases with a higher hazard exhibit a lower divergence from truncation than cases with a lower hazard. Also, the difference in divergence from truncation is not constant over time.

We now turn to the instance of constant multiplicative heterogeneity, where $h_2(t) = \eta h_1(t)$ and $\eta > 1$. This corresponds to the class of proportional hazards models, where time-invariant variables affect the hazard by a constant ratio regardless of time. This class includes popular models such as the Cox, Weibull, and exponential models. In this situation, we can write the survival function of the second customer as

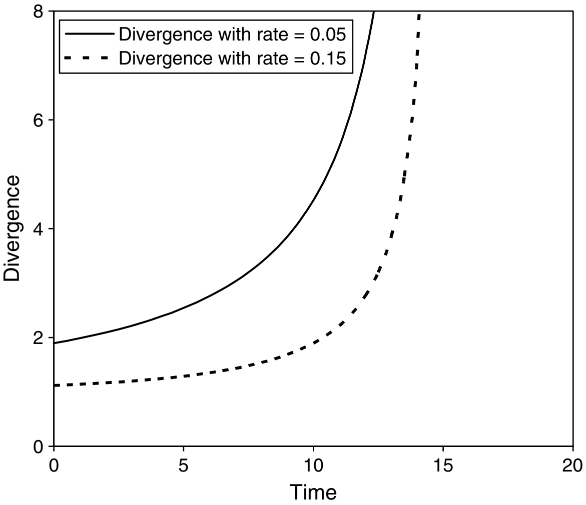
$$S_2(t) = \exp\left(-\int_0^t h_2(u)\,du\right) = \exp\left(-\int_0^t \eta h_1(u)\,du\right)$$
$$= \left(\exp\left(-\int_0^t h_1(u)\,du\right)\right)^{\eta} = (S_1(t))^{\eta}. \quad (9)$$

Substituting this expression in Equation (4) gives

$$D_{2T}(t) = \frac{1}{1 - S_2(T)/S_2(t)} = \frac{1}{1 - (S_1(T)/S_1(t))^{\eta}}$$
$$< \frac{1}{1 - S_1(T)/S_1(t)} = D_{1T}(t), \quad 0 \le t \le T. \quad (10)$$

Equations (7)–(10) establish two additional characteristics of the truncation-induced divergence when the hazards of adoption differ among cases by additive or multiplicativeconstants:

Figure 3 **Divergence for Two Exponential Processes with Rates of 0.05 and 0.15, Both Truncated at $T = 15$**



- Cases with a higher hazard exhibit a lower divergence induced by truncation than cases with a lower hazard.
- The difference in the amount of divergence from truncation is not constant over time.

Figure 3 illustrates these properties for two exponential processes with constant hazards $\lambda_1 = 0.15$ and $\lambda_2 = 0.05$, and divergences $D_{iT}(t) = 1/(1 - e^{-\lambda_i(T-t)})$.

These two characteristics hold for discrete-time processes as well, where the probability of survival up to period t equals $1 - F_i(t-1) = \prod_{j=1}^{t-1}(1 - P_{ij})$, with P_{ij} being the hazard in period j. With constant additive heterogeneity, where $P_{2j} = \theta + P_{1j}$ and $\theta > 0$, we have

$$D_{2T}(t) = \frac{1}{1 - (1 - F_2(T))/(1 - F_2(t-1))}$$
$$= \frac{1}{1 - (\prod_{j=1}^{T}(1 - P_{2j}))/(\prod_{j=1}^{t-1}(1 - P_{2j}))}$$
$$= \frac{1}{1 - \prod_{j=t}^{T}(1 - P_{2j})}$$
$$= \frac{1}{1 - \prod_{j=t}^{T}(1 - \theta - P_{1j})}$$
$$< \frac{1}{1 - \prod_{j=t}^{T}(1 - P_{1j})}$$
$$= D_{1T}(t), \quad 0 \le t \le T. \quad (11)$$

Similarly, with constant multiplicative heterogeneity, $P_{2j} = \eta P_{1j}$ and $\eta > 1$, we have

$$D_{2T}(t) = \frac{1}{1 - \prod_{j=t}^{T}(1 - \eta P_{1j})}$$
$$< \frac{1}{1 - \prod_{j=t}^{T}(1 - P_{1j})}$$
$$= D_{1T}(t), \quad 0 \le t \le T. \quad (12)$$

As with continuous time, a higher hazard is associated with a lower divergence, and the difference between the divergences varies over time.

The time-varying heterogeneity in divergence factors suggests that trying to control for right-truncation bias by controlling for unobserved time-invariant heterogeneity (e.g., through random effects) will be only partially successful at best. Although it may capture the average difference in divergence over time, it will not capture how this difference evolves over time.[3]

3. Monte Carlo Simulation Analysis

We have shown that deleting right-censored cases from the data creates spurious positive duration dependence in the hazard of the truncated process. Because positive duration dependence, and, more specifically, a hazard that increases with the cumulative number of prior adoptions, is often interpreted as evidence of contagion, our results show that truncation can lead one to believe contagion is at work when, in fact, there is none.

From a research practice point of view, our analytical results raise two new questions. First, is the spurious effect large enough to reject unwarrantedly the null hypothesis of no contagion, or create other problems of substantive interpretation? Second, what can one do to protect oneself against such Type I errors? In this section, we focus on the first question. The answer is not obvious a priori. For instance, Figure 1 shows that, when truncation occurs at $T = 30$ and only 5% of the cases are truncated, the upward divergence becomes meaningfully large only around period 20, at which time only 14% of the original population and approximately 9% of the estimation sample has not adopted yet.[4] Is the divergence in those last 9% of observed adoptions large enough to lead to biased estimates and erroneous rejection of the (true) null hypothesis?

We use Monte Carlo simulation to assess whether truncation effects can be large enough to create spurious evidence of contagion effects and create other problems of substantive interpretation. Because spurious duration dependence is especially problematic for research on social contagion, and discrete-time rather than continuous-time models are typically being used in such investigations (e.g., Bell and Song 2007, Iyengar et al. 2011, Manchanda et al. 2008, Nam et al. 2010, Van den Bulte and Lilien 2001), we also use discrete-time hazard rate models in our simulations.

For each data structure we create, we estimate the hazard model not only on the truncated sample but also on the nontruncated sample to check that we indeed recover the true parameter values when the data and model are correctly specified. We use standard maximum likelihood estimation. Let t_i be the duration over which case i is observed, and let δ_i be the censoring indicator, such that $\delta_i = 1$ if t_i is not censored and adoption is observed and $\delta_i = 0$ if t_i is censored such that adoption is not observed. The log-likelihood function for the sample of $i = 1, \ldots, N$ is then given by

$$\mathrm{LL} = \sum_{i=1}^{N} \left[\delta_i \ln f_i(t_i) + (1-\delta_i) \ln S_i(t_i)\right]. \qquad (13)$$

In discrete time, this log-likelihood can be expressed in terms of a binary dependent variable model, where the adoption indicator variable y_{ij} is set to 0 if i has not adopted by period j and is set to 1 if it has. The discrete-time hazard of adoption equals the probability $P(y_{ij} = 1 \mid y_{ij-1} = 0)$, or P_{ij} for short. The log-likelihood in Equation (13) can then be expressed as

$$\mathrm{LL} = \sum_{i=1}^{N} \left[\delta_i \ln\left(P_{it_i} \prod_{j=1}^{t_i-1} (1-P_{ij})\right) + (1-\delta_i) \ln\left(\prod_{j=1}^{t_i} (1-P_{ij})\right)\right], \qquad (14)$$

which, in turn can be expressed as the log-likelihood for a panel data set with y_{ij} as the dependent variable:

$$\mathrm{LL} = \sum_{i=1}^{N} \sum_{j=1}^{t_i} \left[y_{ij} \ln P_{ij} + (1-y_{ij}) \ln(1-P_{ij})\right]. \qquad (15)$$

Estimating a standard hazard rate model to truncated data amounts to omitting the terms multiplied by $(1-\delta_i)$ in Equations (13) and (14) without recognizing that the original population pdf $f_i(t_i)$ should be replaced by the truncated pdf $f_{T_i}(t_i) = f_i(t_i)/F_i(T_i)$. It also amounts to estimating Equation (15) after excluding all cases for which the dependent variable y_{ij} remains 0 from 1 to t_i.

3.1. Homogeneous Hazard

We start with the simplest situation where the true data-generating process has a hazard rate that is constant across cases as well as over time. We generate the data, estimate standard hazard models on both truncated and nontruncated data, and assess to what extent truncation—i.e., omitting censored cases from the estimation sample—generates spurious evidence of duration dependence and contagion.

[3] There is another more obvious reason to suspect that controlling for time-invariant unobserved heterogeneity in the hazards will not be an effective way to handle right-truncation biases: whereas we have just shown that right truncation creates spurious *positive* duration dependence, it is well known that unobserved heterogeneity creates spurious *negative* duration dependence in hazard rates.

[4] In a discrete-time process with a constant hazard rate of 0.10, the probability of adoption by period 20 is approximately 86%, and hence the probability of not having adopted is approximately 14%. In addition, the probability of adoption by the end of the 30 periods is approximately 95%. Thus, between periods 20 and 30, there are approximately 9% more adoptions.

We generate 1,000 data sets for each of six conditions in a 2×3 design: the sample size (N) is either 500 or 1,000, and censoring occurs either at $T=15$, $T=20$, or $T=30$. In each condition, we specify a discrete-time process with hazard $h(t)=\Lambda(-2.25)$, where Λ is the logistic cdf. This corresponds to $h(t)=0.09535$, and leads to, on average, 78% of all people having adopted $T=15$, 87% at $T=20$, and 95% at $T=30$.[5]

In every cell of the design, each person is assumed to be part of a neighborhood (physical neighborhood or egocentric network) with $N-1$ other members, each of which adopts stochastically at the same constant hazard rate. After generating all the data, we estimate two models: (1) model 1, with contagion (the number of adopters in one's neighborhood, divided by N) to check for spurious contagion, and (2) model 2, with time and time squared to check for spurious duration dependence in general. In each of the six cells, we estimate each of the two models both with and without truncation.

When the data are not truncated and people who have not adopted by the end of the observation window are kept in the data set, the null hypotheses of no contagion, no time trend, and no quadratic trend are each rejected with 95% confidence in only about 5% of the cases. This is so, regardless of sample size and observation window. In short, censoring in itself is not a problem, as expected of a proper hazard model specification.

Truncation, however, leads to entirely different results. Figure 4 reports the histograms for the coefficients of contagion in model 1 across all 1,000 replications for each of the six cells. Three conclusions can be drawn. First, there is an upward bias in the contagion effect in each condition, as all distributions of estimated coefficients are shifted to the right of the true effect, 0. In actuality, each and every one of the 6,000 parameter estimates is larger than 0. Second, as expected, the bias decreases with the time of truncation (going from left to right in Figure 4) but is not affected by the sample size. For $N=500$, the distribution mean decreases from 1.77 when $T=15$, to 1.19 when $T=20$, and to 0.58 when $T=30$. For $N=1{,}000$, the distribution means are 1.78, 1.19, and 0.58, respectively. Finally, again as one would expect, the variance is larger in smaller samples (top row versus bottom row) and in more highly truncated samples (left versus right columns). In short, there is very strong evidence of bias, and it is a function of the truncation level. Even 5% truncation ($T=30$) generates a clear upward bias in contagion. Repeating this analysis for hazard values of approximately 5% and 15% produced the same patterns and qualitative insights.

Additional insights can be gained from performing significance tests on the coefficients of the included covariates (contagion in model 1, and time and time squared in model 2). Because their true effect is zero, in a correctly specified statistical model, one should reject the null hypothesis of no effect with 95% confidence in only 5% of the cases; in 2.5% of the cases, the Wald test statistic should indicate that the estimated coefficient is significantly smaller than zero; and in 2.5% of the cases, it should indicate that it is significantly larger than zero. As the first two double columns in Table 1 show, this is indeed what happens when censored observations are not truncated. Truncation, however, leads to rejecting the hypothesis of no contagion at a much higher frequency than the nominal level. As the third double column in Table 1 shows, truncation would lead one to believe to have found significant evidence of contagion in 92%–100% of all cases, even though the true effect is nil. The fourth column indicates that this indeed because of spurious positive (and convex) duration dependence induced by truncation, consistent with our analytical results and Figure 1.

Our analytical results of truncation artifacts do not rely on restrictions on how the true hazard function evolves over time, and hence they also apply to processes with genuine positive or negative duration dependence. To illustrate that the resulting biases can be quite significant in those cases as well, we performed a few additional small-scale simulations where the true data-generating process has genuine positive duration dependence, genuine negative duration dependence, or genuine contagion (see §2 of the electronic companion).

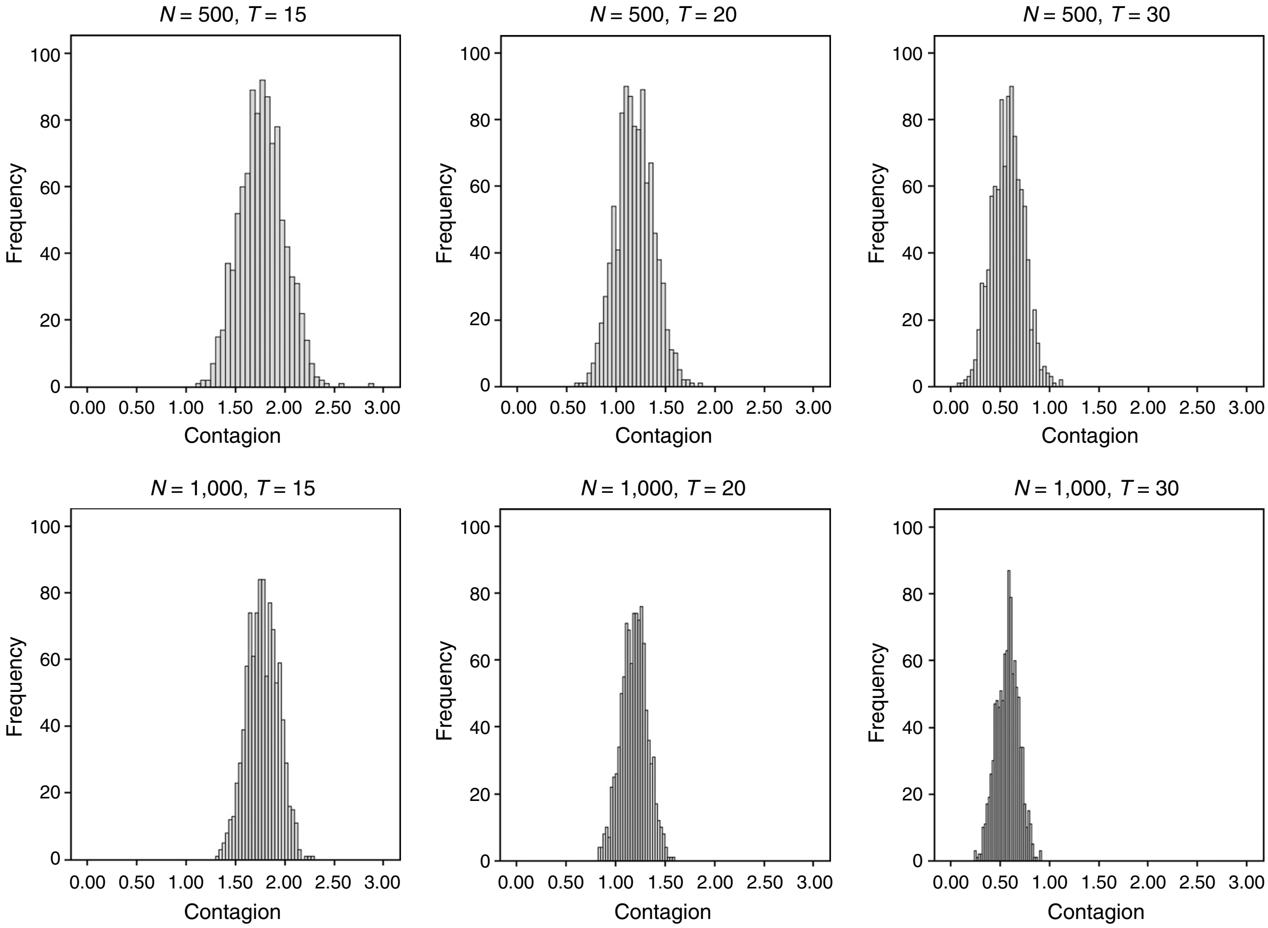
3.2. Attenuation of the Effect of Time-Invariant Covariates

As the literature on unobserved heterogeneity in hazard models shows, spurious negative duration dependence can create biases in the coefficients of time-invariant covariates (Lancaster 1990, Vaupel and Yashin 1985). We now illustrate how right truncation and the resulting spurious positive duration dependence can also lead to erroneous research conclusions on such covariates.

First, we generated 1,000 instances of a discrete-time process with the case-specific hazard specified as $h_i(t)=\Lambda(-2.25+x_i)$, where Λ is the logistic cdf, x_i is an observed time-invariant covariate distributed

[5] Put differently, the average level of censoring and truncation is 22% when $T=15$, 13% when $T=20$, and 5% when $T=30$. For representativeness, the samples sizes, the adoption hazard, and observation windows are chosen such that they cover the sample size ($N\approx 500$), number of monthly observations ($T=34$), and censoring levels (at least 5%) reported in a recent study by Manchanda et al. (2008). The same holds for choosing a logit discrete-time hazard model rather than another equally appropriate stochastic model of adoption.

Figure 4 **Histograms of Estimated Contagion Coefficients in Models Without Duration Dependence**



Note. One thousand estimates in each graph.

$N(0,1)$, and t goes from 0 to $T = 20$. Estimating the model on the nontruncated data led to successful recovery of the parameters, whereas estimating the model on the truncated sample led to a pronounced downward bias in the coefficient of x_i from its true value of 1 to an average estimate of 0.53, with all 1,000 estimates being significantly smaller than the true value ($p < 0.05$). A variant of the analysis set x_i to be a binary variable taking values $(-1,1)$ equally distributed across cases. The average of the 1,000 estimates was 0.54, with all estimates significantly smaller than the true value ($p < 0.05$). Thus, again, right truncation generated a downward bias in the effect of the time-invariant covariate. This attenuation is similar to that in truncated linear regression models estimated with standard ordinary least squares ignoring the truncation.

3.3. Joint Effects of Truncation and Unobserved Heterogeneity

As we have documented, right truncation leads to spurious *positive* duration dependence. It is well known that unobserved heterogeneity creates spurious *negative* duration dependence in hazard rates. Here, we explore the interaction between these two effects by considering the effect of truncation on duration dependence when the true model has unobserved heterogeneity. As the negative effect of unobserved heterogeneity is the largest early in the process when the heterogeneity is the greatest, but the positive effect of truncation is the largest late in the process when $F(t)$ approaches $F(T) < 1$, one would expect the joint effect of truncation and unobserved heterogeneity to be a *U-shaped distortion* of the true duration dependence.

To document this phenomenon, we generated 1,000 instances of a discrete-time process with the time-invariant, case-specific hazard specified as $h_i(t) = \Lambda(-2.25 + x_i)$, where x_i is a time-invariant covariate distributed $N(0, 1/2)$ or $N(0,1)$, and t goes from 0 to $T = 20$. As expected, estimating a model excluding the effect of x_i with a flexible baseline hazard (time dummies) on the nontruncated data led to the well-known spurious negative duration, with

Table 1 Frequency in Percent of Erroneously Rejecting the Null Hypothesis ($\alpha = 0.05$) with Constant Homogeneous Hazards

	No truncation				Truncation			
	Model 1		Model 2		Model 1		Model 2	
	Contagion		Duration dependence		Contagion		Duration dependence	
Parameter	<0	>0	<0	>0	<0	>0	<0	>0
$N = 500, T = 15$								
Contagion	2.5	1.9	—	—	0.0	100.0	—	—
Time	—	—	1.7	1.9	—	—	53.7	0.0
Time squared	—	—	2.2	2.0	—	—	0.0	99.3
$N = 500, T = 20$								
Contagion	2.4	2.5	—	—	0.0	100.0	—	—
Time	—	—	2.0	2.2	—	—	47.1	0.0
Time squared	—	—	2.2	2.8	—	—	0.0	98.8
$N = 500, T = 30$								
Contagion	2.6	2.8	—	—	0.0	92.3	—	—
Time	—	—	2.3	2.3	—	—	33.8	0.0
Time squared	—	—	2.4	2.3	—	—	0.0	90.3
$N = 1{,}000, T = 15$								
Contagion	2.7	2.3	—	—	0.0	100.0	—	—
Time	—	—	2.3	2.3	—	—	82.2	0.0
Time squared	—	—	2.6	2.4	—	—	0.0	100.0
$N = 1{,}000, T = 20$								
Contagion	2.1	1.8	—	—	0.0	100.0	—	—
Time	—	—	2.5	2.7	—	—	76.7	0.0
Time squared	—	—	2.5	2.6	—	—	0.0	100.0
$N = 1{,}000, T = 30$								
Contagion	1.3	3.2	—	—	0.0	99.9	—	—
Time	—	—	2.3	1.8	—	—	58.7	0.0
Time squared	—	—	1.6	2.3	—	—	0.0	99.8

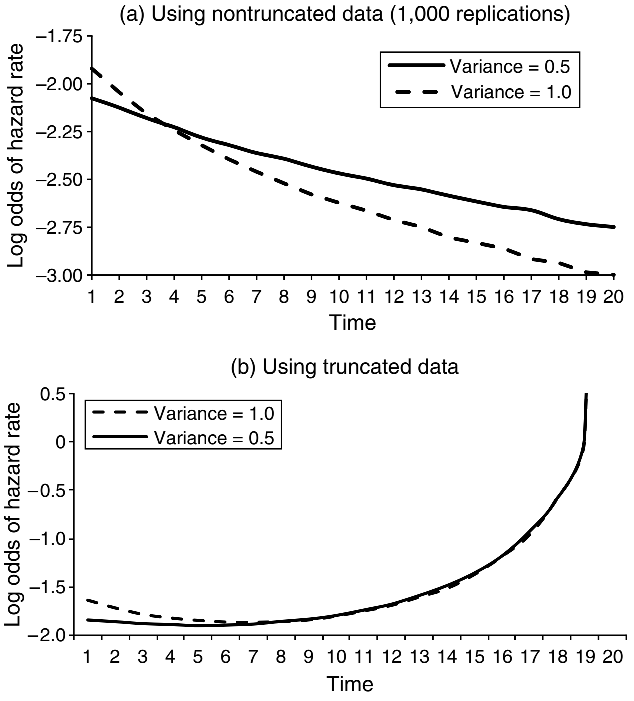
Note. All percentages are based on 1,000 simulated data sets.

the problem being more pronounced in the data with the greater variance in the unobserved variable (see Figure 5(a)). Truncation dramatically changes the situation. As expected, estimating a model with unobserved heterogeneity and truncation produces spurious U-shaped duration dependence (see Figure 5(b)). Note, the effect of truncation is larger than that of unobserved heterogeneity, and the U shape is visibly pronounced only in the case with the higher level of heterogeneity. A variant of the analysis with binary unobserved heterogeneity and x_i taking values $(-1/2, 1/2)$ or $(-1, 1)$ produces similar conclusions.

3.4. Conclusions from Simulations

Our simulations show that the analytical results presented earlier are consequential. The right-truncation bias is often large enough to induce (i) spurious positive duration dependence, (ii) spurious evidence of social contagion, (iii) overestimation of genuine contagion (see §2 of the electronic companion), and (iv) deflated effects of time-invariant covariates.

Figure 5 Spurious Duration Dependence with Unobserved Heterogeneity


(a) Using nontruncated data (1,000 replications)

(b) Using truncated data

Note. In panel a, the lines are not perfectly smooth because they show the average across 1,000 replications.

4. Managerial Implications of Right-Truncation Artifacts

We now document some managerial implications of these biases. We consider three substantive issues: new product pricing without contagion, new product advertising with contagion, and customer lifetime value.

4.1. New Product Pricing

Prices of new technologies and products often decrease over time. Optimal dynamic pricing policies require managers to forecast the price sensitivity of adoption for such innovations. Often, this can be done by analyzing data of previously introduced analogous products. Our results imply that if the analysis of the analog uses truncated data, the price coefficient will be inflated. This, in turn, will lead the firm to set prices that are too low.

Consider a new product with a market potential of 50,000 customers. Management wants to set the prices for the first 20 periods that will maximize the net present value of the profits over this time span, given a marginal cost of zero, a fixed cost per period of $250, and a 10% discount rate. The company has data on an analogous product, covering 50,000 potential adopters adopting with a true discrete-time hazard $h_i(t) = \Lambda(-1.75 - 0.05p_t)$, where Λ is the logistic cdf, i goes from 1 to 50,000, t goes from 1 to $T = 20$, and p_t is the price variable. The price of the analog product decreased linearly from a price of \$20 to \$1. The censoring level is 13%.

We simulated such data and found that estimating the model on the censored but nontruncated data leads to successful recovery of the parameters. In contrast, estimating the model on the truncated data resulted in a pronounced overestimation of the price sensitivity from its true value of -0.05 to an estimate of -0.11 ($\text{SE} = 0.001$).

Using the estimates from the censored but nontruncated data, the optimal policy for the new product is to decrease the price over time from \$28.41 at launch to \$21.20 in period 20.[6] This generates a true expected net present value (NPV) of \$380,254. Using the estimates from the truncated data, the optimal price declines from \$17.34 to \$10.72. Inserting this pricing policy in the true model with a price sensitivity of -0.05 generates a true expected NPV of \$331,413, which is 13% less than what could be gained from the policy based on the correct analysis. Note that the optimization and profit calculations in both cases assume 50,000 potential adopters for the new product and differ only in the hazard equation used as input for the optimization. These results illustrate how truncation bias can lead managers to leave substantial profits on the table.

4.2. New Product Advertising

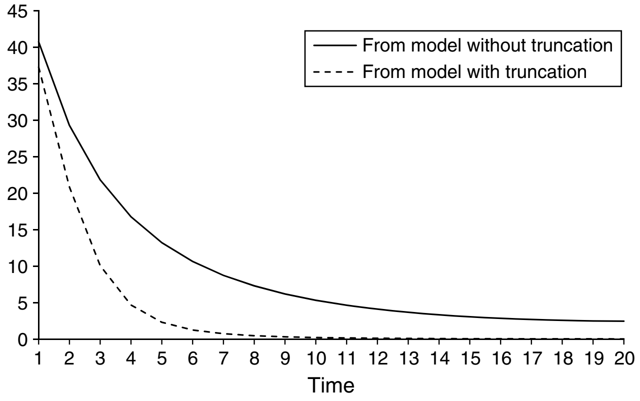
New product managers are often keenly interested in identifying effective combinations of paid-for marketing communications and free word of mouth. Both intuition and formal analysis suggest that the optimum is to start with high advertising (or other marketing effort) to trigger initial adoptions and then reduce spending over time as the "snowball" of free word of mouth takes over (e.g., Horsky and Simon 1983). Our results imply that managers basing their advertising decisions on truncated data are likely to overestimate the importance of word of mouth and hence to cut their advertising and other marketing efforts too quickly.

Consider a new product with a market potential of 50,000 customers. Management wants to determine advertising levels for the first 20 periods that will maximize the net present value of the profits over this time span, given a unit price with zero marginal cost, an advertising cost per period that is 50 times the level of advertising (e.g., gross rating points), and a 10% discount rate. The company has historical adoption data on an analogous product. The data cover 50,000 potential adopters with a true discrete-time adoption hazard $h_i(t) = \Lambda(-3.00 + 0.50 \ln(a_t))$, where Λ is the logistic cdf, i goes from 1 to 50,000, t goes from 1 to $T = 20$, and a_t is the level of advertising. For the analog product, the company used advertising pulsing, alternating between an advertising level of 1 unit or 1.5 units every period. The censoring level is 34%.

The company is not aware of the true hazard and erroneously believes that adoption might be subject to word of mouth. The analysts therefore estimate a hazard model that includes not only an intercept and the log of advertising but also a contagion variable (the fraction of all potential adopters that has already adopted in the last period), consistent with Horsky and Simon (1983). Estimating this model on the nontruncated data recovers the true parameters and provides no evidence of a significant contagion effect. In contrast, estimating the model on the truncated sample leads to a sizable spurious contagion effect of 2.64 ($p < 0.05$), whereas the effect of advertising is slightly inflated at 0.61—still close to its true value of 0.50.

Figure 6 shows the optimal advertising policies identified from both models. Using the proper estimates from the nontruncated data, the optimal policy is to decrease advertising gradually from 40.73 at launch to 2.78 in period 20. Inserting this policy in the true model generates a true expected NPV of \$23,326. Using the estimates from the truncated data leads to an advertising path that declines steeply from 37.29 at launch to 1.27 in period 6, and then it further declines to 0.05 in period 20. Inserting this second policy in the true model generates a true expected NPV of only \$19,810. This is 15% less than from the policy based on the correct analysis. As with the pricing example, the optimization and profit calculations in both cases assume 50,000 potential adopters for the new product. The differences in advertising policies and NPVs illustrate how truncation biases in contagion estimates can lead to suboptimal advertising decisions.

Figure 6 Optimal Advertising from Models With and Without Truncation



[6] The decrease in optimal price over time stems from the link between current price level, current adoptions, and the remaining market potential in the future. Myopic sequential period-by-period optimization or using a constant market potential both lead to a constant optimal price.

4.3. Customer Lifetime Value

Hazard models are also used for predicting churn, a critical input in calculating customer lifetime value (CLV) that puts a ceiling on the cost a firm would be willing to incur to acquire a customer. A recent study by Schmitt et al. (2011), for instance, found that customers acquired through a referral program were approximately 20% less likely to churn than customers acquired through other means, and calculated the corresponding maximum referral fee that was economically justified. Truncation would induce spurious duration dependence and so lead managers to erroneously believe that loyalty is deteriorating over time. It would also result in underestimating the effect of time-invariant covariates and so depress differences in churn between referred and nonreferred customers, leading the firm to spend too little on the referral fee.

To illustrate these problems, we generate data on 50,000 customers with a discrete-time hazard of defection $h_i(t) = \Lambda(-2.00 - 0.20x_i)$, where Λ is the logistic cdf, i goes from 1 to 50,000, x_i is a dummy that indicates whether customer i is a referred customer, and t goes from 1 to $T = 20$. The data consist of 25,000 referred and nonreferred customers each. Given the hazard model, the per-period defection rate for the nonreferred customers is approximately 12%, whereas that for referred customers is approximately 10%. The censoring level is 34%. Estimating the model on the nontruncated data leads to successful recovery of the parameters. Adding a flexible baseline hazard produced does not produce any significant effect ($p > 0.05$) among the time dummies. In contrast, estimating the model with flexible baseline hazard on the truncated sample leads to a pronounced bias in the coefficient of x_i from its true value of -0.20 to an average estimate of -0.12. In addition, we find evidence of an increasing baseline hazard of defection. Thus, there are two erroneous conclusions from the analysis on the truncated sample—decreasing customer loyalty and too small a difference in churn rates between referred and nonreferred customers.

We calculate the economic relevance of the latter using the NPV of the referred and nonreferred customer bases over 20 periods. We assume a unit margin and a discount rate of 10%. Based on the estimates from the nontruncated data, the predicted NPV of referred and nonrefereed customers is $110,749 and $98,949, respectively. The difference is $11,800. As there are 25,000 customers of each type, a company should be willing to pay a premium of $0.47 to acquire a referral customer. Based on the estimates from the truncated data, the predicted NPV of referred and nonrefereed customers is $96,067 and $89,422, respectively. The difference of $6,645 implies that the company should be willing to pay only an additional $0.27 for a referred customer. This is approximately 43% lower than the true difference of $0.47 and illustrates how truncation leads to underappreciating differences in customer lifetime value.

5. How to Handle Right Truncation

Having documented the problem, we now turn to what one can do to protect oneself from being tricked by truncation. One path of action is obviously to avoid the problem by carefully defining the population at risk, drawing a proper sample, keeping all censored cases in the sample, and handling censoring through standard hazard modeling. For instance, when studying the adoption of the third prescription drug launched in a specific therapeutic category, Iyengar et al. (2011) defined the population of physicians at risk as those who had prescribed one or both of the two incumbent drugs at least once within the last two years prior to launch of the focal drug. However, occasions arise where researchers do not observe the censored cases and have genuinely truncated data (e.g., Nam et al. 2010), or in which they are not certain as to whether censored cases are indeed at risk of adoption or not (e.g., Bell and Song 2007, Manchanda et al. 2008). To assist researchers facing the first problem, we assess the effectiveness of four procedures to protect oneself against right-truncation artifacts: (i) using an analytical correction to the discrete-time hazard likelihood function, (ii) including a nonparametric baseline hazard, (iii) including random effects, and (iv) including both a nonparametric baseline hazard and random effects. To assist researchers facing the second problem, we suggest applying mover–stayer hazard or split-population hazard models to use the information in the data to determine how large the risk set is (e.g., Dekimpe et al. 1998, Sinha and Chandrashekaran 1992), and we briefly assess their effectiveness in correctly recovering the true data-generating process. In this section, we focus on the first problem: how to accommodate one's model for right truncation.

5.1. Properly Correcting the Likelihood Function

Researchers working with genuinely truncated data can modify their likelihood functions to properly take into account truncation (e.g., Moe and Fader 2002). The standard likelihood expression for a discrete-time hazard model set up as a panel model estimated on truncated data is Equation (15), limited to cases with $\delta_i = 1$:

$$\mathrm{LL} = \sum_{i=1}^{N} \delta_i \sum_{j=1}^{t_i} \big[y_{ij} \ln P_{ij} + (1 - y_{ij}) \ln(1 - P_{ij}) \big],$$

where P_{ij} is i's hazard of adoption in period j. The analytical expression for the divergence factors can be used to correct this likelihood for the divergence induced by right truncation:

$$\mathrm{LL} = \sum_{i=1}^{N} \delta_i \sum_{j=1}^{t_i} \big[y_{ij} \ln P_{ij} D_{ijT} + (1 - y_{ij}) \ln(1 - P_{ij} D_{ijT}) \big] \tag{16a}$$

$$= \sum_{i=1}^{N} \delta_i \sum_{j=1}^{t_i} \left[y_{ij} \ln\left(\frac{P_{ij}}{1 - \prod_{k=j}^{T}(1 - P_{ik})} \right) + (1 - y_{ij}) \ln\left(1 - \frac{P_{ij}}{1 - \prod_{k=j}^{T}(1 - P_{ik})} \right) \right], \tag{16b}$$

where (16b) uses the fact that the discrete-time survival function up to period j of case i equals $\prod_{k=1}^{j-1}(1 - P_{ik})$ and so $D_{ijT} = 1/(1 - \prod_{k=j}^{T}(1 - P_{ik}))$, as developed earlier (Equation (11)). An alternative derivation of (16b) is to start with the proper log-likelihood function for a truncated process:

$$\mathrm{LL} = \sum_{i=1}^{N} \delta_i \ln \frac{f_i(t_i)}{F_i(T)} \tag{17a}$$

$$= \sum_{i=1}^{N} \delta_i \ln \left(\frac{P_{it_i} \prod_{k=1}^{t_i - 1}(1 - P_{ik})}{1 - \prod_{k=1}^{T}(1 - P_{ik})} \right), \tag{17b}$$

and note that (17b) can be rewritten as (16b), just as (14) can be rewritten as (15).

There are three caveats. First, the analytical correction requires one to compute the hazards from periods 1 through T even for the cases who adopted before T. That means that one cannot use covariates in the model that are not observed after adoption. Time-invariant covariates are obviously not a problem, nor are variables that are determined by what goes on in i's environment or social network. Second, the correction need not be quite accurate when truncation can happen randomly before the end of the observation window T. Because the true random time of truncation is by definition not observed, however, using the latest possible time of random truncation, which is T, when computing the corrected pdf and likelihood is the best one can do (Kalbfleisch and Lawless 1992). Finally, note that the second term in Equations (16a) and (16b) involves ln(0) in the last period $j = T$, which is not defined, but this quantity is actually never computed since $(1 - y_{ij})$ always equals zero in the last period in a truncated data set. Similarly, the first term involves $\ln(1) = 0$ in the last period $j = T$, which is at it should be because the cases who have survived until the last period in a truncated sample must adopt by definition. Because their adoptions provide no information, they should not contribute to the likelihood.

5.2. Flexible Baseline Hazard

When the model has no observed or unobserved heterogeneity in the hazards, then all cases will have the same divergence factor. Simply adding time dummies for each and every period will then capture the divergence between the true and the truncated hazard.[7] This is quite easy to implement with standard software packages. There are three caveats, though. First, the procedure induces a slight loss of degrees of freedom compared with the analytical correction for which no additional parameters need to be estimated. Second, the flexible baseline will absorb all cross-time variation that is common across cases, so the effect of common time-varying covariates cannot be estimated anymore. Finally, the correction will not be perfect when there are covariates that vary across the cases. As shown earlier, heterogeneity in hazards implies heterogeneity in divergences, so a flexible but common baseline will control only for the average divergence in each time period but not the heterogeneity in divergences across cases within each time period. As a result of this unobserved heterogeneity in true divergences, the flexible baseline hazard will overcorrect the cases with high hazards and low divergence and undercorrect the cases with low hazards and high divergence. As a result, the coefficients of variables that vary across cases will be biased toward zero.

5.3. Random Effects

Our analytical results imply that simply adding random effects is not an effective way to protect oneself from truncation artifacts. When there is no observed or unobserved heterogeneity in the true process, then there will be no heterogeneity in the divergence either. When there is heterogeneity in the true hazards, in contrast, then there will also be heterogeneity in the divergences, but this heterogeneity will be time varying and will hence not be accommodated by controlling for time-invariant heterogeneity.

5.4. Flexible Baseline Hazard with Random Effects

Including both a flexible baseline hazard and random effects may perform better than using each separately, as the model now controls for the average convex time trend and allows for constant heterogeneity around that trend. Still, one would expect this procedure to be less effective than the analytical fix because it does not capture the time-varying nature of the heterogeneity in the divergences. One would also expect it to be less efficient because it requires the estimation of several additional baseline parameters and one variance parameter.

[7] In analyses not reported here, we found that simply controlling for a quadratic trend in the baseline hazard by adding time and its squared term as additional covariates did not effectively correct for truncation-induced spurious contagion, even in samples without any heterogeneity in the hazards or divergences.

5.5. Monte Carlo Simulation

We assess the performance of these four correction procedures in four different data structures. Results are reported in §3 of the electronic companion. The key result is that only the analytical correction provides effective protection against truncation artifacts, especially when truncation amounts to not more than 50%. A caveat, however, is that even though the procedure holds the bias in check, truncation still decreases the amount of information in the data and so increases the variance in the estimates. Thus, whereas the analytical correction holds the attenuation of covariate effects and overestimation of contagion in check quite effectively *on average*, it is difficult to indicate a priori how good the parameter estimates will be for a *particular* empirical application with an extreme level of truncation, such as 80%.

5.6. Empirical Applications

We now illustrate the effectiveness of the four correction methods on two data sets that have been used in past research. For each application, we first estimate a model using the nontruncated data (which we call "original" estimates). Next, we remove the right-censored observations, estimate the model on the truncated data set using each of the four correction methods, and check how close those estimated parameters are to the original values. Note that what matters for the present research purpose is not whether each model includes the "correct" set of covariates representing the true data generating process but whether truncation affects one's conclusions about the effect of those covariates.

5.6.1. *Medical Innovation.* The first data set is the famous *Medical Innovation* data (Burt 1986) from the original study by Coleman et al. (1966). The data set contains information on the adoption of tetracycline by 125 physicians in four small cities in Illinois from its launch in November 1953 to March 1955. As there are missing covariates for 5 physicians, the final data cover 120 physicians, with 104 of them having adopted the drug by the end of the data period (13% censoring rate). Thus, the nontruncated data covers 120 physicians, whereas the truncated data covers 104 physicians.

The covariates are defined as follows. *Contagion* is the fraction of one's direct network contacts who have adopted previously. We have information on this variable for all 120 physicians over the entire 17 months, as required to implement the analytical correction. *Log journals* is the logarithm of the number of journals a physician receives or subscribes to. *Science* is a binary attitudinal measure coded as 1 if the physician agreed with the statement that it is more important for a physician to "keep himself informed of new scientific developments [than to] devote more time to his patients," and as 0 otherwise. *Chief* is also a dummy variable, capturing whether a physician has a chief or honorary position in his hospital.

The results in Table 2 show that the model without correction attenuates the coefficients of time-invariant covariates and inflates that of contagion. The analytical fix recovers the parameters, especially that of contagion, rather well. Simply adding a flexible baseline attenuates the impact of contagion so much so that it is not significant anymore (see also Van den Bulte and Lilien 2001). Adding random effects does not give any improvement over the model without correction. Finally, adding both random effects and a flexible baseline is better than either of the two corrections separately, but it still does not recover the original parameter estimates. In short, not accounting for truncation leads one to overestimate the amount of contagion and underestimate the importance of independent drivers of adoption, and it would lead managers to rely too much on word of mouth and to poorly identify the best seeding points to start a viral campaign.

Table 2 ***Medical Innovation***

Parameter	Original estimates	No correction	Analytical correction	Flexible baseline	Random effects[a]	Random effects and flexible baseline
Intercept	−4.04	−3.37	−4.19	−4.16	−3.37	−4.42
	(0.48)	(0.51)	(0.73)	(0.65)	(0.51)	(0.99)
Contagion	1.04	1.38	1.09	0.47*	1.38	0.52*
	(0.32)	(0.32)	(0.41)	(0.46)	(0.32)	(0.50)
Log journals	0.87	0.68	1.08	1.02	0.68	1.11
	(0.28)	(0.30)	(0.42)	(0.32)	(0.30)	(0.43)
Science	1.08	0.62	0.83	0.79	0.62	0.86
	(0.23)	(0.23)	(0.29)	(0.24)	(0.23)	(0.33)
Chief	−1.29	−0.83*	−1.19*	−0.84*	−0.83*	−0.95*
	(0.46)	(0.49)	(0.75)	(0.51)	(0.49)	(0.62)

[a] For the model with random effects only, the estimated variance of unobserved heterogeneity was the lower bound of 1E-8. Thus, there is no evidence of unobserved heterogeneity.

*$p > 0.05$. All the remaining parameters are significant at $p < 0.05$.

5.6.2. Netgrocer.com. The second data set is from a study by Bell and Song (2007). It contains information on zip code-level adoptions of the Internet retailer Netgrocer.com in 29,701 U.S. residential zip codes covering a period of 45 months from its launch in May 1997 to January 2001. As many as 11,791 zip codes did not see any adoption during the data window, amounting to 40% censoring. The main analysis by Bell and Song uses all the data from all 29,701 zip codes and does not suffer from right truncation. It is only in a robustness check that the authors use the truncated data with only 17,730 zip codes.

The zip code-specific contagion variable used by Bell and Song (2007) was available only until adoption but missing afterwards (this was confirmed in a personal communication with the authors), which precludes the use of the analytical correction. To circumvent this problem, we created a new contagion variable based on the cumulative fraction of all zip codes that had adopted previously. As this variable is common across zip codes, its effect is not identified in a model when a flexible baseline is included. Thus, we investigate only the effectiveness of the analytical correction and of adding random effects.

Table 3 shows the results for a very simple model controlling only for the size of the population in the zip code. Right truncation leads to overestimating the importance of contagion and underestimating that of the time-invariant covariate. The analytical correction does very well and produces estimates that are close to the original values, though the contagion parameter is somewhat inflated. Including a random effect, in contrast, is not effective at all. Extending the model with variables capturing differences in household demographics across zip codes (described in detail in the original study) does not change this conclusion (see Table 4). Here again, not accounting for truncation would make researchers overestimate the extent of contagion and underestimate the importance of demographic drivers of adoption, leading managers to rely too much on word-of-mouth dynamics and to poorly identify the best zip codes to target initially.

Table 3 Simple Model for Netgrocer.com

Parameter	Original estimates	Without correction	Analytical correction	Random effects
Intercept	−11.78	−8.17	−11.01	−9.41
	(0.07)	(0.07)	(0.11)	(0.16)
Contagion	3.13	4.89	4.04	6.26
	(0.05)	(0.05)	(0.06)	(0.18)
Log population	0.86	0.47	0.77	0.57
	(0.01)	(0.01)	(0.01)	(0.01)

Note. All parameters are significant at $p < 0.05$.

Table 4 Extended Model for Netgrocer.com

Parameter	Original estimates	Without correction	Analytical correction	Random effects
Intercept	−11.64	−8.11	−10.85	−9.27
	(0.08)	(0.08)	(0.14)	(0.06)
Contagion	3.59	5.26	4.53	6.75
	(0.05)	(0.05)	(0.06)	(0.17)
Log population	0.92	0.49	0.82	0.59
	(0.01)	(0.01)	(0.01)	(0.01)
Black	−0.69	−0.35	−0.62	−0.39
	(0.05)	(0.06)	(0.07)	(0.07)
Foreign	2.56	3.21	3.02	4.01
	(0.12)	(0.13)	(0.16)	(0.17)
Hispanic	−1.46	−1.66	−2.04	−1.97
	(0.11)	(0.11)	(0.16)	(0.14)
Large family	−5.09	−3.36	−4.65	−4.19
	(0.19)	(0.20)	(0.31)	(0.26)
Solo female	−7.14	−4.63	−6.35	−6.60
	(0.44)	(0.48)	(0.69)	(0.62)
Solo male	8.43	7.79	8.56	10.98
	(0.40)	(0.52)	(0.63)	(0.72)

Note. All parameters are significant at $p < 0.05$.

Note, the model that Bell and Song (2007) estimated on both original and truncated data included a flexible baseline, and the estimated coefficient of contagion in their robustness analysis on the truncated data (0.15) was actually slightly *smaller* than that from the main analysis (0.17) on the full data. This decrease in a model using a flexible baseline is consistent with our own simulation results (see Tables A5–A7 in §3 of the electronic companion) and with our reanalysis of the *Medical Innovation* data (see Table 2).

5.7. Managerial Applications

Finally, we revisit the three illustrations of managerial implications for pricing, advertising, and CLV calculations. Specifically, we now also estimate each model on the truncated data using the appropriately corrected likelihood function. As Table 5 shows, this allows one to recover the true parameter values. This in turn leads to pricing and advertising decisions that generate NPVs as high as the analysis using nontruncated data, and it also leads to a correct assessment of the CLV differential and hence justifiable acquisition cost differential.

6. Avoiding Self-Inflicted Truncation Using Mover–Stayer Modeling

As we have shown, analytically correcting the likelihood function for right truncation helps in protecting oneself from reaching erroneous conclusions when dealing with such data. There are some studies, however, where the original data were not truncated, yet the researchers chose to delete all nonadopters

Table 5 Parameter Estimates and Decision Outcomes for Three Managerial Applications

		Censored nontruncated		Without correction		Analytical correction	
Parameter	θ	Mean	SE	Mean	SE	Mean	SE
Pricing							
Intercept	−1.75	−1.75	0.01	−0.54	0.01	−1.73	0.05
Price	−0.05	−0.05	0.00	−0.11	0.00	−0.05	0.00
NPV ($)		380,254		331,413		380,249	
Advertising							
Intercept	−3.00	−3.03	0.01	−2.91	0.01	−3.00	0.03
Log ad	0.50	0.56	0.03	0.61	0.03	0.56	0.03
Contagion	0.00	0.02	0.03	2.67	0.03	0.12	0.07
NPV ($)		23,326		19,810		23,325	
Customer value							
Intercept	−2.00	−2.00	0.01	−1.92	0.01	−1.99	0.01
Referral	−0.20	−0.20	0.01	−0.12	0.01	−0.20	0.02
Acquisition cost differential ($)		0.47		0.27		0.47	

Note. The entries in the column θ denote the true value.

prior to estimation, either in the main analysis (e.g., Manchanda et al. 2008) or as a robustness check (e.g., Bell and Song 2007) because they were uncertain about whether censored cases were indeed at risk of adoption. Note that deleting censored cases and estimating a standard hazard model amounts to assuming that these cases were never at risk of adopting to begin with. This is quite a strong assumption.

Researchers who are uncertain about what cases to include in their analysis, however, need not rely purely on their own judgment. Instead, they can use mover–stayer or split-population hazard models and exploit the information in the data (e.g., Colombo and Morrison 1989, Dekimpe et al. 1998, Sinha and Chandrashekaran 1992, Srinivasan et al. 2006). In these models, the population is assumed to consist of two unobserved groups or latent classes: a mover group following a transition process specified by the hazard model, and a stayer group with a zero probability of change (in our case, adoption). One possible concern with this approach, however, is that the best-fitting fraction of "stayers" might be too close to the fraction of nonadopters observed by the end of the data window (Dekimpe et al. 1998). This would imply that, according to the model estimates, the diffusion process has come to closure by the end of the data window, which need not be true. To the extent that the fraction of stayers who will never adopt is overestimated, the fraction of "movers" who will ultimately adopt is underestimated, leading again to truncation bias. Investigating whether mover–stayer models suffer from this problem is important to researchers who are uncertain about the size of the population at risk and who might decide to right truncate their data instead (e.g., Manchanda et al. 2008) and to researchers who wonder whether using split-population hazard models with an unknown fraction of ultimate adopters is subject to the same problems as macrolevel diffusion models with an unknown ceiling (Bemmaor and Lee 2002, Van den Bulte and Lilien 1997).

A simulation study reported in §4 of the electronic companion indicates that, in large samples without extreme right censoring, well-specified mover–stayer models quite effectively recover the true data-generating process and hence avoid the need for one to truncate one's data. The parameter recovery is excellent even with an overparametrized class membership equation. Researchers who are uncertain about the risk set in their analysis can, and we believe should, avoid self-inflicted right truncation and its ensuing biases by using mover–stayer or split-population hazard models instead.

7. Conclusion

7.1. Truncation Artifacts

Marketing researchers are aware that omitted variables and reverse causality can generate spurious evidence of contagion. What has not been appreciated so far, however, is that right truncation in hazard modeling—an issue faced in several recent studies—can also trick one into seeing evidence of contagion, or more generally a positive trend in the hazard, when there is none. We have shown the following six consequences of right truncation analytically:

1. Right truncation inflates the hazard rate, and the divergence becomes larger the more severe the truncation.

2. Assuming $f(t) > 0$ for all t, the divergence becomes larger over time. Hence, right truncation induces spurious duration dependence.

3. As time reaches the truncation point T, the amount of spurious duration dependence goes to infinity in continuous-time analysis and goes to the reciprocal of the true hazard in discrete-time analysis.

4. The divergence is convex with respect to time whenever $df(t)/dt \geq -2[f(t)]^2/[F(T)-F(t)]$. This condition is met at every point in time in a process with hazard $h(t) = p + qF(t)$ regardless of the rate parameters ($p > 0$, $q \geq 0$) and the truncation level, although it need not be for every conceivable process.

5. When there is population heterogeneity, cases with a higher hazard exhibit a lower divergence induced by truncation than cases with a lower hazard.

6. When there is population heterogeneity, the difference in the amount of divergence from truncation is not constant over time.

These analytical results hold regardless of how the true hazard function evolves over time and hence apply to processes with genuine positive or negative duration dependence. To our knowledge, all these results apart from the first one are new.

Simulation analyses show that the biases induced by right truncation are large enough to result in seriously misleading inferences. First, unless the model includes a flexible baseline, deleting censored cases from the analysis will lead one to overestimate the amount of contagion.[8] More generally, it will inflate or deflate the effect of trending variables, depending on the sign of the coefficient and the trend of the variable. Second, deleting censored cases from the analysis will underestimate the importance of time-invariant characteristics in the transition process.

Truncation can affect decisions pertaining to pricing, advertising, and customer acquisition. The profit impact, we have shown, can be significant. The specific consequences will vary across application areas. Researchers interested in managing contagion for new products, for instance, will note that the erroneous inferences will make managers rely too much on word of mouth in their market development efforts and poorly identify the customers most likely to adopt (and start generating buzz) early. We trust that thoughtful researchers of firm entry patterns, customer churn, sales force turnover, purchase acceleration, and other areas where hazard models are used will readily identify what our findings mean to their own area of application.

7.2. Recommendations for Research Practice

Researchers should avoid deleting censored cases. They can do so by carefully defining the population at risk, drawing a proper sample, keeping all censored cases in the sample, and handling censoring through standard hazard modeling (e.g., Iyengar et al. 2011). However, occasions arise where researchers do not observe the censored cases and have genuinely truncated data (e.g., Moe and Fader 2002, Nam et al. 2010). In such cases, using an analytical correction to the likelihood function of the hazard model is quite effective in recovering the true data-generating process, provided right truncation is not too high (in our simulation study, not more than 50%). Three alternative methods, (i) a nonparametric baseline hazard, (ii) random effects, and (iii) both a nonparametric baseline hazard and random effects, are all ineffective.

Some adoption researchers have chosen to right truncate their data because they were uncertain that censored cases were indeed at risk of adoption. Our analyses show that mover–stayer or split-population hazard models can recover the true data-generating structure very well, making self-inflicted right truncation and all the ensuing biases unnecessary and unjustifiable. An important caveat, however, is that we analyzed the performance of mover–stayer modeling under rather benign conditions. The statistical literature emphasizes that such models should not be used indiscriminately as the empirical estimates can be unstable in field applications (Farewell 1982, 1986; Meeker 1987; Miley 1978). Statisticians recommend that the approach be used with caution, especially when dealing with small to moderate samples, when the time spacing between late adoptions is not much larger than that between early adoptions and when there is no good theoretical or empirical support for the existence of a stayer segment (Farewell 1986, Kuk and Chen 1992, Li et al. 2001). The latter condition again points to the need for carefully delineating the population of interest when designing one's study.

In conclusion, we recommend researchers do the following:

1. Ideally, define the population at risk carefully, sample it appropriately, keep all censored cases in the sample, and handle censoring through standard hazard modeling.
2. When being justifiably concerned that not all cases included in the data are truly at risk, use a mover–stayer model rather than truncating their data voluntarily.
3. When facing genuinely right-truncated data, use the proper analytically corrected likelihood.

Our work also provides a compelling argument against the use of case-specific fixed effects to control for unobserved heterogeneity in discrete-time hazard models. The maximum likelihood estimate of the fixed effect for a right-censored case is minus infinity (e.g., Chamberlain 1980), meaning that such cases do not contribute to the model's likelihood value. Our work shows that such truncation leads to serious biases. Thus, rather than protecting against artifacts, using fixed effects generates spurious contagion in discrete-time hazard models.

7.3. Envoy

Estimating regular hazard models on right-truncated data, we have shown, can trick researchers into inflating duration dependence, inflating contagion, and deflating the importance of heterogeneity. Social contagion and heterogeneity are fundamental drivers of diffusion and adoption processes yet can easily be confounded. Recent research underscores the importance of properly quantifying each for both theoretical and practical purposes (e.g., Aral et al. 2009, Choi et al. 2010, Iyengar et al. 2011, Manchanda et al. 2008, Van den Bulte and Stremersch 2004). Our results on the deleterious effects of right truncation and on how to avoid them will help marketing scientists make greater progress in this endeavor.

[8] The bias may be negative in the presence of a flexible baseline hazard and a contagion variable that varies not only over time but also across cases, because the flexible baseline hazard overcorrects the cases with high hazards and undercorrects the cases with low hazards.

8. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mktsci.pubs.informs.org/.

Acknowledgments

The authors thank Yogesh Joshi for checking one of their results; David Bell and Sangyoung Song for making their data available for reanalysis; Puneet Manchanda for sharing some unpublished results; and Asim Ansari, David Bell, Gary Lilien, Renana Peres, Eric Schwartz, Catherine Tucker, the review team, and attendees at the 2010 INFORMS Marketing Science Conference for their comments.

References

Andrews, R. L., I. S. Currim. 2005. An experimental investigation of scanner data preparation strategies for consumer choice models. *Internat. J. Res. Marketing* **22**(3) 319–331.

Aral, S., L. Muchnik, A. Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA* **106**(51) 21544–21549.

Bell, D. R., S. Song. 2007. Neighborhood effects and trial on the Internet: Evidence from online grocery retailing. *Quant. Marketing Econom.* **5**(4) 361–400.

Bemmaor, A. C., Y. Lee. 2002. The impact of heterogeneity and ill-conditioning on diffusion model parameter estimates. *Marketing Sci.* **21**(2) 209–220.

Burt, R. S. 1986. The *Medical Innovation* network data. Technical Report 3, Center for the Social Sciences, Columbia University, New York.

Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Rev. Econom. Stud.* **47**(1) 225–238.

Choi, J., S. K. Hui, D. R. Bell. 2010. Spatio-temporal analysis of imitation behavior across new buyers at an online grocery retailer. *J. Marketing Res.* **47**(1) 75–89.

Coleman, J. S., E. Katz, H. Menzel. 1966. *Medical Innovation: A Diffusion Study*. Bobbs-Merrill Company, Indianapolis.

Colombo, R. A., D. G. Morrison. 1989. A brand switching model with implications for marketing strategies. *Marketing Sci.* **8**(1) 89–99.

Deemer, W. L., Jr., D. F., Votaw Jr. 1955. Estimation of parameters of truncated or censored exponential distributions. *Ann. Math. Stat.* **26**(3) 498–504.

Dekimpe, M. G., L. M. Van de Gucht, D. M. Hanssens, K. I. Powers. 1998. Long-run abstinence after narcotics abuse: What are the odds? *Management Sci.* **44**(11) 1478–1492.

Den Broeder, G. G., Jr. 1955. On parameter estimation for truncated Pearson Type III distributions. *Ann. Math. Stat.* **26**(4) 659–663.

Erbring, L., A. A. Young. 1979. Individuals and social structure: Contextual effects as endogenous feedback. *Sociol. Method. Res.* **7**(4) 396–430.

Farewell, V. T. 1982. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**(4) 1041–1046.

Farewell, V. T. 1986. Mixture models in survival analysis: Are they worth the risk? *Canadian J. Statist.* **14**(3) 257–262.

Helsen, K., D. C. Schmttlein. 1993. Analyzing duration times in marketing: Evidence for the effectiveness of hazard rate models. *Marketing Sci.* **12**(4) 395–414.

Horsky, D., L. S. Simon. 1983. Advertising and the diffusion of new products. *Marketing Sci.* **2**(1) 1–17.

Iyengar, R., C. Van den Bulte, T. W. Valente. 2011. Opinion leadership and social contagion in new product diffusion. *Marketing Sci.* **30**(2) 195–212.

Jackson, M. O. 2008. *Social and Economic Networks*. Princeton University Press, Princeton, NJ.

Kalbfleisch, J. D., J. F. Lawless. 1992. Some useful statistical methods for truncated data. *J. Quality Technol.* **24**(3) 145–152.

Klein, J. P., M. L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed. Springer, New York.

Kuk, A. Y. C., C.-H. Chen. 1992. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**(3) 531–541.

Lancaster, T. 1990. *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge, UK.

Li, C.-S., J. M. G. Taylor, J. P. Sy. 2001. Identifiability of cure models. *Stat. Probab. Lett.* **54**(4) 389–395.

Manchanda, P., Y. Xie, N. Youn. 2008. The role of targeted communication and contagion in product adoption. *Marketing Sci.* **27**(6) 961–976.

Meeker, W. Q. 1987. Limited failure population life test: Application to integrated circuit reliability. *Technometrics* **29**(1) 51–65.

Miley, A. D. 1978. Stability of parameter estimates in the split population exponential distribution. *Eval. Quart.* **2**(4) 646–649.

Moe, W. W., P. S. Fader. 2002. Using advance purchase orders to forecast new product sales. *Marketing Sci.* **21**(3) 347–364.

Nam, S., P. Manchanda, P. K. Chintagunta. 2010. The effect of signal quality and contiguous word of mouth on customer acquisition for a video-on-demand service. *Marketing Sci.* **29**(4) 690–700.

Prins, R., P. C. Verhoef, P. H. Franses. 2009. The impact of adoption timing on new service usage and early disadoption. *Internat. J. Res. Marketing* **26**(4) 304–313.

Schmitt, P., B. Skiera, C. Van den Bulte. 2011. Referral programs and customer value. *J. Marketing* **75**(1) 46–59.

Sinha, R. K., M. Chandrashekaran. 1992. A split hazard model for analyzing the diffusion of innovations. *J. Marketing Res.* **29**(1) 116–127.

Srinivasan, R., G. L. Lilien, A. Rangaswamy. 2006. The emergence of dominant designs. *J. Marketing* **70**(2) 1–17.

Tuma, N. B., M. T. Hannan. 1979. Approaches to the censoring problem in analysis of event histories. K. Schuessler, ed. *Sociological Methodology*, Vol. 10. Jossey-Bass, San Francisco, 209–240.

Van den Bulte, C., G. L. Lilien. 1997. Bias and systematic change in the parameter estimates of macro-level diffusion models. *Marketing Sci.* **16**(4) 338–353.

Van den Bulte, C., G. L. Lilien. 2001. *Medical Innovation* revisited: Social contagion versus marketing effort. *Amer. J. Sociol.* **106**(5) 1409–1435.

Van den Bulte, C., S. Stremersch. 2004. Social contagion and income heterogeneity in new product diffusion: A meta-analytic test. *Marketing Sci.* **23**(4) 530–544.

Vaupel, J. W., A. I. Yashin. 1985. Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *Amer. Statistician* **39**(3) 176–185.

Zanutto, E. L., E. T. Bradlow. 2006. Data pruning in consumer choice models. *Quant. Marketing Econom.* **4**(3) 267–287.