Research Article   **Open Access**

Statistical Considerations of Length Bias for Evaluating Diagnostic Tests in Screening Studies

Kyunghee K Song*

Division of Biostatistics, Center for Devices and Radiological Health/Food and Drug, Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993

Abstract

A diagnostic test in a screening study detects a clinical condition of interest in its asymptomatic stage. Evaluating the diagnostic test in a screening study is a challenging task since a diagnostic test is evaluated based on its analytical and clinical performance. In order to evaluate clinical performance of a diagnostic test in a screening study, it is crucial to investigate clinical outcome studies such as survival studies. Furthermore, there are important biases to be considered in a screening study. In this paper, statistical issues associated with screening studies are discussed, and statistical adjustment for screening-related bias, which is called length bias, is presented. Both Vardi's bias-adjusted nonparametric maximum likelihood estimator and linear combination estimators have shown to adjust length bias successfully, and generate bias-adjusted survival curve close to the true survival curve. Finally, some practical issues associated with early detection are also presented.

Introduction

In general, a diagnostic test is evaluated based on its analytical and clinical performance. Clinical performance examines how diagnostic tests would perform in a clinical practice, which leads to how they would be used in a clinical practice. Diagnostic tests in these days have broad applications in terms of analytes as well as the utility of the test results. Diagnostic tests are used to determine the status of the condition of interest such as disease, infection, or any other specified condition of clinical interest.

The benefit of early detection of the disease has been recognized and addressed in many different clinical conditions, especially when treatment procedure following early detection leads to successful outcomes, whereas late detection of the disease could lead to much worse outcomes. Due to this reason, early detection of the disease has gained a lot of interest in clinical practice. Screening is the process by which asymptomatic people are tested to determine whether they are likely to have a particular disease [1]. The goal of screening is to detect and treat the disease early to benefit the patients. Therefore, successful screening is to detect the condition early and also to decrease cause-specific morbidity and mortality rates.

The phenomenon, called length biased sampling, is well-known in screening studies for chronic diseases. Length biased sampling introduces bias, called "length bias", in the study. In a screening study, diagnostic tests are applied to asymptomatic individuals to detect the condition of interest before symptoms appear. The longer the preclinical stage an individual has, the more likely the individual is to be detected in a screening study, leading to a length biased sample. Survival estimates based on length biased samples are overestimated and the success of screening would be overstated. Therefore, the bias should be adjusted to estimate the true survival curve.

In this paper, statistical issues associated with length biased sampling and bias-adjustment procedures are presented. Different bias-adjusted survival estimators are compared in the presence of

length biasing. Additional issues associated with the use of diagnostic test in a screening study are also discussed.

Length Biased Sampling

Length biased sampling occurs when the probability of selecting a sample is proportional to its lifelength. Suppose the original observation x has f(x) as the probability density function, then the length biased distribution of f(x), denoted as g(x), is written as

$$g(x) = \frac{xf(x)}{E(x)}$$

where

$$E(x) = \int xf(x)dx$$

Therefore, samples from g(x) are called "length biased samples."

Statistical Consideration for Length Bias and Bias Adjustment

A nonparametric maximum likelihood estimator (NPMLE) of survival in a length biased sample is given by [2]. This bias adjusted NPMLE is based on the mixture of the two independent samples: one from the original distribution and the other from the length biased distribution.

Suppose that the original observation x has f(x) as the probability density function, and the length biased observation y has g(y) as the probability density function. Let η_{1i} be the multiplicity of the x's and η_{2i} be the multiplicity of the y's at t_i. After pooling samples, {$x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n$} and ordering the samples, denote the ordered observations

*Corresponding author: Kyunghee K. Song, Division of Biostatistics, Center for Devices and Radiological Health/Food and Drug, Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, E-mail: Kyunghee.Song@fda.hhs.gov

as $\{t_1, t_2,...., t_h\}$. Note that $h \leq m + n$ for tied observations. Then the probability of the data is written as

$$P\{t_i, \eta_{1i}, \eta_{2i}; i = 1,...,h\} = \prod_{i=1}^{h} \{dF(t_i)\}^{\eta_{1i}} \{\frac{t_i dF(t_i)}{\int_0^\infty u dF(u)}\}^{\eta_{2i}}$$

The likelihood function becomes,

$$L(p_1,...,p_h) = \prod_{i=1}^{h} p_i^{\eta_{1i}} (\frac{t_i p_i}{\sum_{j=1}^{h} t_j p_j})^{\eta_{2i}} \qquad (1)$$

with denoting $p_i = dF(t_i)$.

Using Lagrangean multiplier and a routine maximization procedure [2], the unique solution of (1) is

$$\widehat{p_i} = \frac{(\eta_{1i} + \eta_{2i})\widehat{\mu}}{nt_i + m\widehat{\mu}} \qquad (2)$$

where $\widehat{\mu}$ is the solution of

$$\sum_{i=1}^{h} \frac{(\eta_{1i} + \eta_{2i})t_i}{nt_i + m\widehat{\mu}} = 1 \qquad (3)$$

The unique Vardi's estimate is obtained by solving equation (3) for $\widehat{\mu}$ first and then plugging the value of $\widehat{\mu}$ into equation (2) to obtain $\widehat{p_i}$.

We also consider the linear combination of two estimators for mixture samples. Thus, compute the EDF (empirical distribution function) estimator based only on samples from F(x) and compute Cox estimator based only on samples from G(y). If we only obtain m samples from F, then the resulting survival function is the standard EDF estimator, $\widehat{F}_{EDF}$.

$$\widehat{F}_{EDF} = \frac{\sum_{i=1}^{m} I(x_i \leq x)}{m}$$

On the other hand, if we only obtain n length biased samples, then the resulting survival function turns out to be the Cox estimator, $\widehat{F}_{COX}$.

$$\widehat{F}_{COX}(y) = \sum_{i=1}^{n} \frac{\frac{1}{y_i} I(y_i \leq y)}{(\sum_{i=1}^{n} \frac{1}{y_i})}$$

These two estimators are combined linearly to obtain the linear combination estimator (LCE), $\widehat{F}_{LCE}$. The explicit form for LCE from the mixture distribution is written as

$$\widehat{F}_{LCE} = k_1 \widehat{F}_{EDF} + k_2 \widehat{F}_{COX}$$

where k_1 and k_2 indicate the proportions of samples from different distributions.

As a naïve estimator of F, ignoring length bias, we consider empirical distribution function estimate (EDF) from the observed mixture samples.

$$\widehat{F}(x) = \frac{number\ of\ x_i s \leq x}{h}$$

where $h \leq m + n$ (for tied observations).

Simulation Study

Suppose the observations are from gamma distribution with parameters α and β.

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Then its length biased distribution is also gamma distribution with parameters $\alpha+1$ and β.

$$g(x) = \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\beta x}$$

In this simulation study, the parameters, α and β, are set at α = 1.5 and β =1.0. Therefore, for a simulation, a gamma distribution with shape parameter 1.5 and scale parameter 1.0 is considered as f(x). Then, its length biased distribution, g(y), becomes a gamma distribution with shape parameter 2.5 and scale parameter 1.0.

Three estimators, $\widehat{F}_V$, $\widehat{F}_{EDF}$, and $\widehat{F}_{LCE}$, based on the mixture samples are obtained and compared through a series of simulations with sample size n=100. The EDF estimator, $\widehat{F}_{EDF}$, is considered as the worse case since length bias existing in the observations is completely ignored. The linear combination estimator, $\widehat{F}_{LCE}$, which is easier to compute compared to Vardi's NPMLE, is considered as a competitor to Vardi's NPMLE, $\widehat{F}_V$.

The observed mixture samples are from two different sets of samples – one from f(x) and the other from g(y) which is the length biased distribution of f(x). The proportion of the two samples is indicated as k_1 and k_2 (k_1 = 0.85, 0.70, 0.40 and k_2 = 0.15, 0.30, 0.60, respectively).

One thousand samples are generated for each configuration. For each simulation, the following percentiles are computed: 95, 90, 80, 70, 60, 50, 40, 30, 20, 10 and 5.

Plots are drawn based on the percentiles obtained from the simulation. In (Figure 1) through 3, three sets of k_1 and k_2 values are considered for plots (Figure 1: k_1 = 0.85, k_2 = 0.15; Figure 2: k_1 = 0.70, k_2 = 0.30; Figure 3: k_1 = 0.40, k_2 = 0.60). Note that by increasing k_2, the proportion of length biased observations is increased in the observed samples. These figures show that both $\widehat{F}_V$ and $\widehat{F}_{LCE}$ approximate the true values quite well for various proportions of length biasing. It is also clear $\widehat{F}_{EDF}$ overestimates the true survival curve. The difference between $\widehat{F}_{EDF}$ and true value becomes larger as length biasing is increased. With 60% length biasing, there are noticeable differences between $\widehat{F}_{EDF}$ and other estimates (Figure 3).

In summary, the simulation study shows that length bias in observed samples are successfully adjusted using both Vardi's estimator and linear combination estimator which includes Cox estimator. Both estimators successfully adjust bias and generate survival curve close to the true survival under different proportions of length biasing.

Discussion

Screening tests have a tendency to detect more slowly growing (less aggressive) cancers because they are in the asymptomatic population longer than the more rapidly growing ones, which quickly become symptomatic and no longer need screening to be detected. Cases detected in a screening study are likely to have a better prognosis, resulting in overestimated survival. These cases are called length biased samples.

Understanding the natural history of the disease is crucial for evaluating the performance. Biological heterogeneity of the disease is found to be associated with length bias [3] and it is discussed in the study
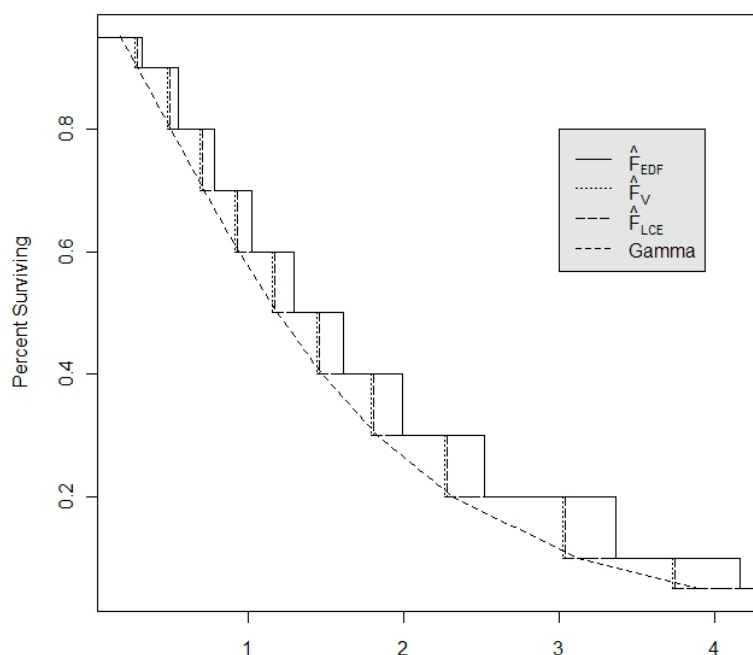
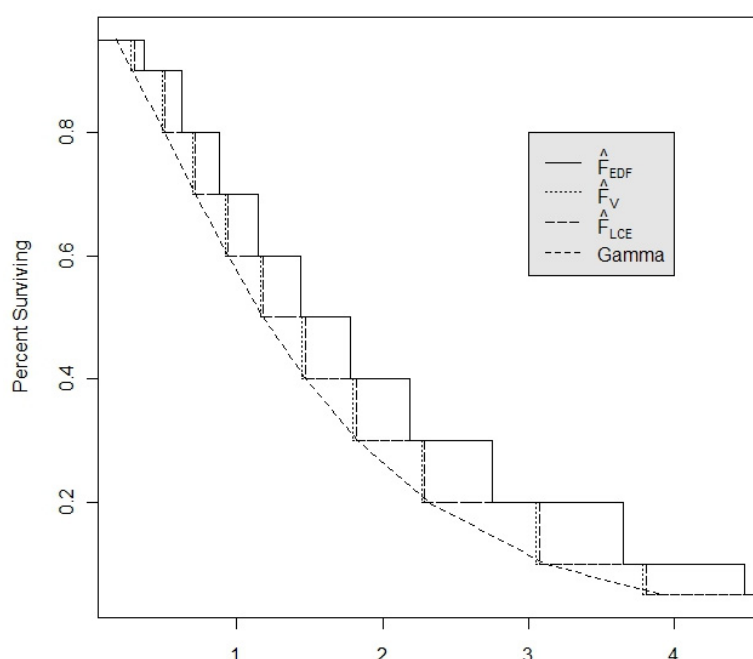**Figure 1:** Gamma (1.5), 15% length biasing, N=100.

Figure 2: Gamma (1.5), 30% length biasing, N=100.

of neuroblastoma [4]. Length bias results from biological heterogeneity of disease. For example, some patients have rapidly growing aggressive cancers and the others have slowly growing and less aggressive cancers. In order to account for different subtypes (heterogeneity) of disease in asymptomatic population, the multiple-type heterogeneous model was presented and discussed in [3].

Extensive statistical work has been done in the area of length biased samples and length biased distributions [5,6,7,8,9] which includes adjusting for bias adjustment [2,10,11] However, to date, these statistical approaches have not been widely implemented in study design and in evaluating screening studies in practice.

Early detection capability is the important feature of screening. For example, [12], discussed a potential utility of CA-125 for early detection of ovarian cancer. For colorectal cancers, there were fecal
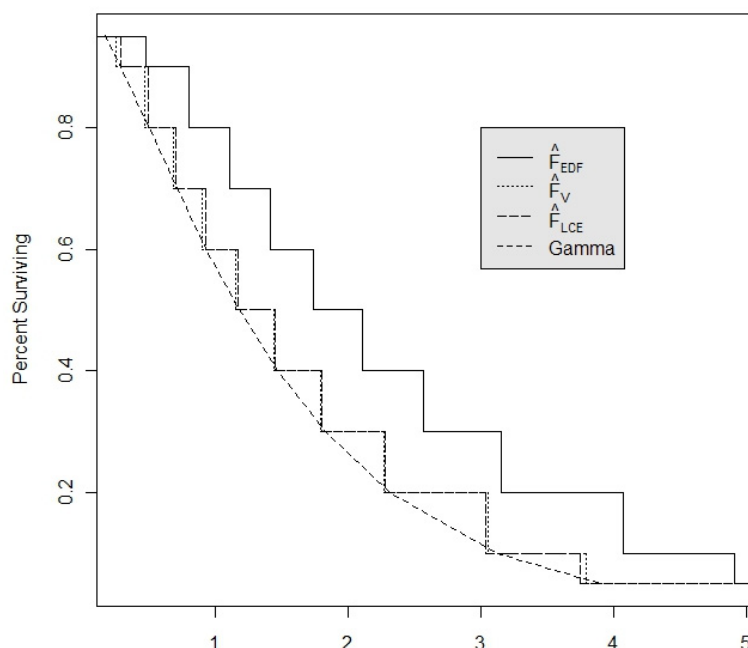
Figure 3: Gamma (1.5), 60% length biasing, N=100.

occult blood tests [13] and fecal DNA tests [14]. For breast cancer, mammography has been widely used. However, early detection could lead to overdiagnosis of disease. Overdiagnosis detects 'cases' that do not cause symptoms or increase morbidity/mortality. Jorgensen and Gotzsche [15] said that there was a 30% overdiagnosis for lung cancer after long term follow-up of patients screened by radiography. Witte [16] said that in prostate cancer testing using PSA (prostate specific antigen), between 20-60% of early stage prostate cancers detected using PSA might be considered "overdiagnosis."

Another important type of screen-related bias is lead-time bias. With lead-time bias, the early detection seems to prolong survival even when there is no actual improvement in survival. This is due to the fact that the cases are found at an earlier point in the natural history.

In this paper, a length bias, which is one of major statistical challenges in a screening study, has been discussed, and bias adjustment procedures using statistical methods are presented. The goal of screening is to improve patient outcomes by detecting conditions of interest early and treating/managing them appropriately. In a screening study, the basic requirement of the screening test (diagnostic test) should be accuracy, that it should detect cases without error. However, for the screening procedure to be successful, the test outcomes should lead to the successful patient outcomes. It is crucial to identify and adjust length bias in a screening study in order to successfully evaluate the clinical performance of the diagnostic tests.

References

1. Goldstein MM, Messing EM (1998) Prostate and bladder cancer screening. J Am Coll Surg 186: 63-74.

2. Vardi Y (1982) Nonparametric estimation in the presence of length bias. Ann Statist 10: 616-620.

3. Yoshimoto Y, Tanaka Y (2008) Biological heterogeneity and length-biased sampling in asymptomatic neurological patients. British Journal of Neurosurgery 22: 368-372.

4. Woods WG, Gao RN, Shuster JJ, Robison LL, Bernstein M, et al. (2002) Screening of infancts and mortality due to neuroblastoma. N Eng J Med 34: 1041-1046.

5. Blumenthal S (1967) Proportional sampling in life length studies. Technometrics 9: 205-218.

6. Cox DR (1969) Some sampling problems in technology. In: New Developments in Survey Sampling, Johnson NL, Smith H, Jr eds, Wiley, New York, 506-527.

7. Patil GP, Rao CR (1978) Weighted distribution and size-biased sampling with applications to wildlife populations and human families. Biometrics 34: 179-189.

8. Lui KJ, Peterman TA, Lawrence DN, Allen JR (1988) A model based approach to characterize the incubation period of pediatric transfusion associated acquired immunodeficiency syndrome. Statistics in medicine 7: 395-401.

9. Wang MC (1996) Hazards regression analysis for length biased data. Biometrika 83: 343-354.

10. Song KK, Weissfeld LA (1999) Performance of Nonparametric Maximum Likelihood Estimator in a Length Biased Sample. Communications in Statistics -SIMULA 28: 637-655.

11. Song KK, Weissfeld LA (2010) Comparison of two distributions in the presence of length bias. Model Assisted Statistics and Applications 5: 179-186.

12. Karlan BY, McIntosh M (2007) The quest for ovarian cancer's holy grail: can CA-125 still be chalice of early detection?. J Clin Oncol 25: 1303-1304.

13. Morikawa T, Kato J, Yamaji Y, Wada R, Mitsushima T, et al. (2005) A comparison of immunochemical fecal occult blood test and total colonoscopy in the asymptomatic population. Gastroenterology 129: 422-428.

14. Imperiale TF, Ransohoff DF, Itzkowitz SH, Turnbull BA, Ross ME, et al. (2004) Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. N Eng J Med 351: 2704-2714.

15. Jorgensen KJ, Gotzche PC (2009) Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. BMJ 339: b2587.

16. Witte JS (2010) Personalized prostate cancer sreening: Improving PSA tests with genomic information. Sci Transl Med 2: 62ps55.