A SCORE-TYPE TEST FOR HETEROGENEITY IN CURE RATE MODELS IN A STRATIFIED POPULATION

By

Shitong Gu

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biostatistics—Master of Science

2017

ABSTRACT

A SCORE-TYPE TEST FOR HETEROGENEITY IN CURE RATE MODELS IN A STRATIFIED POPULATION

By

Shitong Gu

In oncology research, with the advent of new therapeutic agents, many long-term survivors are deemed permanently cured. The cure rate survival model is a popular model that has been used to represent data generated from cancer studies where a substantial fraction of the population does not experience the event of interest even after a sufficient long observation time. This model is primarily used to evaluate the cure fraction in view of observed data. Existing testing methods have relied on the so-called constancy assumption that neglects potential covariate effect on the cure fraction under the alternatives to no cure. In this thesis, I extend the literature by proposing a score-type test that incorporates information from categorical covariates in detecting cure. I show that the limiting null distribution of the proposed test statistic is a mixture chi-squared distribution which, in practice, can be well approximated by a resampling method that uses normal variates to perturb the associated influence function. The numerical simulations show that the proposed test can greatly improve efficiency over the classical score test that neglects covariate information under heterogeneity, in settings where the true cure fraction depends on covariates. Ovarian cancer data collected in Los Angeles from the SEER registry is used to illustrate the real life application of the methodology.

TABLE OF CONTENTS

LIST OF TABLES iv

LIST OF FIGURES v

KEY TO SYMBOLS vi

Chapter 1 1
Introduction 1
1.1 Ovarian Cancer research 1
1.2 Mixture Cure Rate Models and Evaluation of Cure 1
1.3 Method of Silvapulle and Silvapulle 4
1.4 Organization of This Thesis 5

Chapter 2 6
Testing Procedure Method 6
2.1 Cure Rate Model 6
2.2 Likelihood of Cure Rate Model 7
2.3 Score Function and its Variance 8
2.4 Asymptotic Property 10
2.5 A Categorical Cure Rate Score Test to Evaluate Heterogeneity 11

Chapter 3 13
Numerical Results 13
3.1 Simulation Studies 13
3.2 Real Data: Ovarian cancer data from SEER registry database 16

Chapter 4 19
Discussion 19

APPENDICES 21
Appendix A: Details in Observed Covariance Matrix 22
Appendix B: Proof of Influence Function Theorems 23

REFERENCES 24

LIST OF TABLES

Table 1: Empirical size of the score test statistics at 5% significance level. 15

Table 2: Empirical power of the score test statistics at 5% significance level under constant alternative. 16

Table 3: Empirical power of the score test statistics at 5% significance level under categorical alternatives. 16

Table 4: The score-type test statistics(p-values) for ovarian cancer data 17

LIST OF FIGURES

Figure 1: Overall Kaplan-Meijer estimates for ovarian cancer data from SEER registry of Los Angeles with 95% confidence bands (blue areas). 3

Figure 2: Stratified Kaplan-Meijer estimates for ovarian cancer data from SEER registry of Los Angeles with 95% confidence bands. Four age strata contain 172, 780, 885 and 625 patients respectively in age-groups 1 to 4 going from younger to older groups. 4

KEY TO SYMBOLS

All English and Greek letter-based symbols used for mathematical expressions in this thesis are listed here in alphabetical order.

C	Potential Censoring Variable;
h_0	Baseline Hazard Function;
$h()$	Hazard Function;
$H()$	Influence function;
I	Fisher Information matrix;
K	Dimension of Categorical Variable z;
n	Sample Size of Data;
S_0	Baseline Survival Function;
$S()$	Survival Function;
t	Follow-up time of the Study;
T	Survival Time;
$u()$	Average Score Function;
$U()$	Score Function;
w, μ, α and β	Unknown Parameters;
X and Z	Covariates or Known Data;
$\pi()$	Uncured fraction Function;
ℓ	Log-likelihood function;
$\mathcal{L}$	Likelihood function;
$\mathcal{T}$	Limiting distribution.

Chapter 1

Introduction

1.1 Ovarian Cancer research

Ovarian cancer is one of the leading causes of death among women due to its high potential for metastasis and high relapse rate (Engel et al., 2002). The associated mortality rate varies greatly by the stage of cancer and age, with younger women exhibiting higher survival rates (Ries, 1993; Yancik, 1993). Albeit being largely fatal, recent studies have reported a rapid increase in five year survival for ovarian cancer, due to the development of modern therapeutic agents (Jemal et al., 2005). As a result, a non-negligible proportion of patients, especially those at the early stage of the disease, do not experience relapse or related death even after a sufficient follow-up time and may be regarded as "cured". Empirically, data generated from studies on ovarian cancer will typically exhibit heavy censoring and survival functions that do not degenerate to zero after a reasonably long period of time.

1.2 Mixture Cure Rate Models and Evaluation of Cure

A popular statistical approach to represent time to event data with a heavy censoring rate and a non-decaying survival function is the cure rate model which views censoring as mixture of a true censoring and cure, even though this dichotomy is unobservable. This class of models was originally developed by Boag (Boag, 1949), Berkson and Gage (Berkson and Gage, 1952). Earlier formulations of these models relied on parametric assumptions (Cantor and Shuster, 1992; Denham et al., 1996; Ghitany et al., 1994). Semiparametric models (Kuk and Chen, 1992; Wang

et al., 2012), and nonparametric mixture models (Peng and Dear, 2000) have also been discussed. An earlier application of these models in cancer research is due to Farewell (1986).

In real applications of cure rate models, one is often interested in evaluating the cure fraction. The existing methodology, for the most part, has relied on the restrictive assumption that neglects potential covariate effect on the cure fraction under the alternatives to homogeneity (Klebanov and Yakovlev, 2007; Li et al., 2007; Zhao et al., 2009). This approach is important but has limitations. A general limitation is that it may fail to detect cure in the population when the cure rate depends on covariates.

Using time to death data from ovarian cancer in the SEER registry in the Los Angeles area, the overall Kaplan-Meier estimate (Figure 1) gives an indication of cure with an obvious plateau occurring at around 180 months, consistent with the 10-year mark that has been reported in the literature Tai et al. (2005). Our analysis based on the classical approach that neglects covariate information under alternative failed to detect cure for these data. However, an additional analysis that stratifies the Kaplan-Meier curve by age-group shows non-overlapping confidence bands for the survival curves, leading to the conjecture that the cure rate may depend on age group. More importantly, the estimated survival rate in the younger-age group is much higher and levels off more quickly than that of other age groups. These preliminary works imply that cure, if it exists, is mainly in the younger age group. Thus, if there is a substantial portion of older post-menopausal women as is the case in the source sample, cure may be hard to detect. This then necessitates the need for a statistical methodology that incorporates covariate information to detect cure.

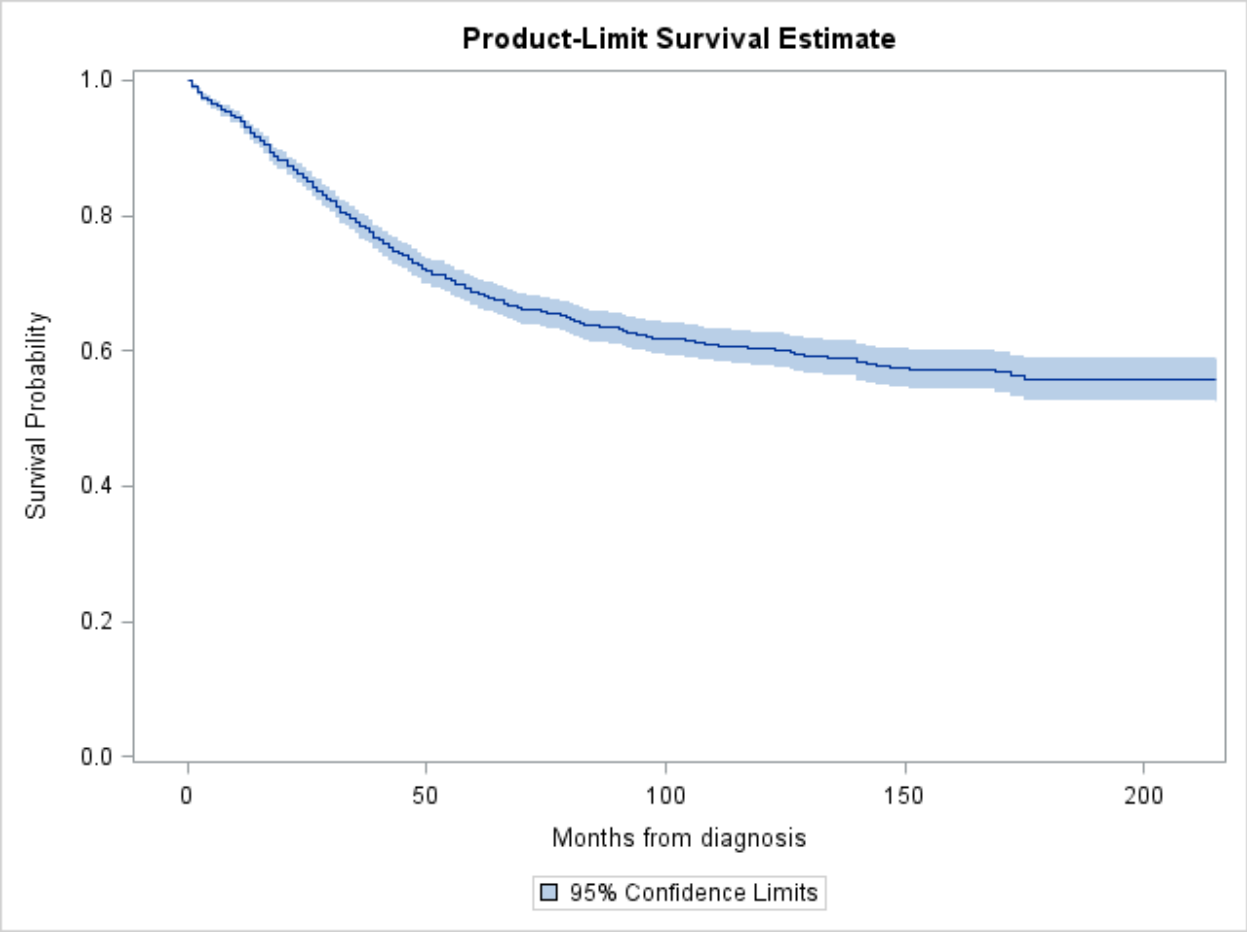

Figure 1: Overall Kaplan-Meijer estimates for ovarian cancer data from SEER registry of Los Angeles with 95% confidence bands (blue areas).

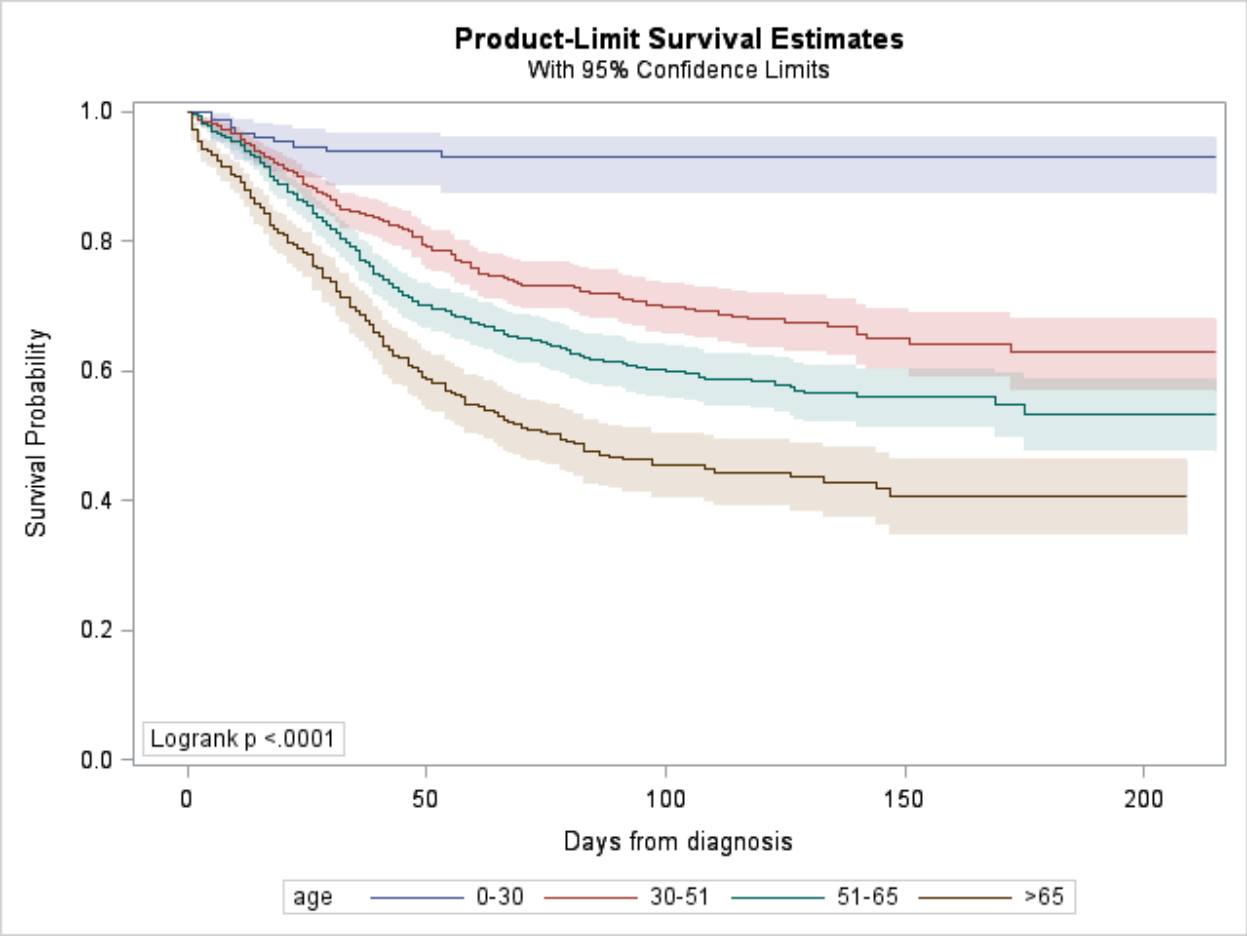

Figure 2: Stratified Kaplan-Meijer estimates for ovarian cancer data from SEER registry of Los Angeles with 95% confidence bands. Four age strata contain 172, 780, 885 and 625 patients respectively in age-groups 1 to 4 going from younger to older groups.

1.3 Method of Silvapulle and Silvapulle

In this thesis, I consider the evaluation of cure when the cure fraction depends on a categorical covariate under the alternative hypothesis. Incorporating covariate information in the evaluation of cure is nontrivial, owing to the multiplicity of constrained (non-negative) parameters resting on the boundary of the support space under the null hypothesis. Silvapulle and Silvapulle (1995), using the asymptotic normality of the partial score function coupled with the likelihood ratio test approach, have proposed a generalized score test for this testing problem. I

apply this method to detect a cure fraction for ovarian cancer using surveillance data from the SEER registry.

1.4 Organization of This Thesis

The remainder of this thesis is organized as follows. Chapter 2 provides a brief introduction to the class of cure rate models and specifies the associated likelihood function. I also show that the evaluation of the cure fraction when the latter depends on categorical covariate can be conducted using the generalized one-sided score-type test proposed by Silvapulle and Silvapulle (1995). In Chapter 3, I conduct numerical studies to evaluate the finite sample performance of the proposed testing procedure. I used the ovarian cancer data from the SEER registry to illustrate the real life applications of this methodology. In Chapter 4, I summarize and discuss further extensions of the methodology.

Chapter 2

Testing Procedure Method

2.1 Cure Rate Model

Suppose T is the survival time, e.g. the time since diagnosis of ovarian cancer, and C the potential censoring variable, e.g. the duration of study or death results from other causes. Empirically only the follow-up time $t = \min(T, C)$ and the censoring indicator $\delta = \mathbf{1}(T = t)$ are observed. The mixture cure rate model (Berkson and Gage, 1952) postulates an heterogeneous population, among which proportion $1 - \pi$ are deemed cured and proportion π are not cured. Specifically, the mixture cure rate model assumed that the overall survival is $S_{\text{pop}}(t|x) = 1 - \pi + \pi S(t|x)$ with $0 < \pi \leq 1$ where x is a potential covariate, $S_{\text{pop}}(t|x)$ denotes the population (overall) survival function and $S(t|x)$ denotes the survival function for the uncured population. The survival function $S(t|x)$ for the uncured population degenerates to zero when t goes to infinity that is $\lim_{t\to\infty} S(t|x) = 0$, unlike the overall survival function $S_{\text{pop}}(t|x)$ which converge to the fraction $1 - \pi$ at infinity. In the initial formulation of this class of models, the cure fraction $1 - \pi$ was assumed constant and the baseline intensity assumed to be fully parametric. Farewell (1982) extended this class of models by allowing the cure fraction to depend on covariates through a logistic transformation. Based on these works, Kuk and Chen (1992) proposed a flexible class of semiparametric cure models that only imposes the proportionality assumption of the hazard function for the failure time distribution of uncured patients coupled with a nonparametric estimation of baseline hazard.

In this thesis, I entertain the class of cure survival models in the finite dimensional setting. I consider the Cox-type model $h(t|x) = h_0(t)\exp(x^T\beta)$ for the uncured population with h_0 denoting the baseline hazard function of the uncured and β represents the vector regression coefficients for covariate vector x. I assume that the baseline intensity h_0 is the two-parameter Weibull model and further assume that the cure fraction $1-\pi$ depends on a categorical variable say z, taking K plausible values. I denote $\pi(z)$ to highlight the dependence on z.

2.2 Likelihood of Cure Rate Model

For the ith individual, $i = 1, 2, \ldots, \mathrm{n}$, the follow-up time is specified as t_i, the covariates (e.g. number of primary tumors) related to the uncured survival function are specified as x_i, the categorical covariates (e.g. age-groups) associated with cure fraction are specified as z_i, and the censoring indicator is denoted by δ_i. In general, we will assume that x contains z. Suppose that each observation is random independent copy from the quantity $\{t, x, z, \delta\}$. The conditional likelihood (given covariates x, z) for two component survival data is

$$\mathcal{L} = \prod_{i=1}^{n}\left\{(1-\pi(z_i)+\pi(z_i)S(t_i|x_i))^{1-\delta_i} \times (\pi(z_i)h(t_i|x_i)S(t_i|x_i))^{\delta_i}\right\}$$

where $S(t_i|x_i)$ and $h(t_i|x_i)$ represent the survival and hazard function for subject i respectively. If the two-parameter Weibull model is assumed as the baseline intensity, the associated conditional log likelihood of the data is then

$$\ell(\theta) = \sum_{i=1}^{n} \Big\{ (1-\delta_i)\log\Big(1-\pi(z_i)+\pi(z_i)\{S_0(t_i)\}^{\exp\{\alpha(-x_i^T\beta)\}}\Big) + \delta_i \log\Big(\pi(z_i)\exp\{\alpha(-x_i^T\beta)\}\{h_0(t_i)\}\{S_0(t_i)\}^{\exp\{\alpha(-x_i^T\beta)\}}\Big)\Big\}$$

where $S_0(t_i) = \exp(-({t_i}^{\alpha}))^{\exp\{\alpha(-\mu)\}}$ and $h_0(t_i) = \alpha {t_i}^{\alpha-1}(e^{-\mu})^{\alpha}$, $\alpha > 0$, denote the baseline survival function and the baseline hazard function respectively for subject i. Here α and μ are parameters for the two-parameter Weibull distribution. Assume that the population under the study consists of K distinct strata and let $w = [w_1,\ w_2, \dots, w_K]^T$ be a K dimensional vector such that the uncured fraction is parameterized as $\pi(z_i) = \frac{1}{1+\sum_1^K w_k \mathbf{1}(z_i=k)}$ where $1 \le k \le K$, $w_k \ge 0$ and $\mathbf{1}(z_i = k)$ is the indicator function for the event that subject i belongs to stratum k with $\sum_1^K \mathbf{1}(z_i = k) = 1$. Let $\theta = \{w, \mu, \alpha, \beta\}$ be the collection of all model parameters.

2.3 Score Function and its Variance

Suppose w is the main parameter vector of scientific interest and let $\gamma = \{\mu, \alpha, \beta\}$ denote the collection of nuisance parameters, yielding $\theta = \{w, \gamma\}$. The score function $U_n(\theta)$ of the sample with n observations can be partitioned into two partial score functions $\Big(U_{nw}(\theta), U_{n\gamma}(\theta)\Big)$ that in turn are consistent with the partition $\{w, \gamma\}$ of θ. Let $\theta^* = \{w^*, \gamma^*\}$ be the true value of θ with $E\big(U_n(\theta^*)\big) = 0$. The observed Fisher information $I(\theta)$ can be partitioned as follows $\Big(I_{ww}(\theta), I_{w\gamma}(\theta), I_{\gamma w}(\theta), I_{\gamma\gamma}(\theta)\Big)$. $I(\theta^*)$ is the true information matrix, which cannot be obtained in closed form in practice. Some conditions should be assumed in advance.

Condition A:

A1: $I(\theta^*)$ is a nonsingular Fisher information matrix.

A2: As n $\to \infty, n^{-\frac{1}{2}}U_n(\theta^*) \xrightarrow{d} N(0, I(\theta^*))$.

A3: For any $\alpha > 0$, $\sup_{\|\rho\| \le \alpha} \left\{ n^{-\frac{1}{2}} \left[U_n\left(\theta^* + n^{-\frac{1}{2}}\rho\right) - U_n(\theta^*) \right] + I(\theta^*)\rho \right\} = o_p(1)$, where $o_p(1)$ represents the convergence to 0 in probability as n$\to\infty$.

The conditions above ensure that the asymptotic normality of the 'single' observation $u_w(\theta^*) = n^{-\frac{1}{2}}U_{nw}(\theta^*)$ has the variance-covariance matrix

$$G^{ww}(\theta^*)^{-1} = I_{ww}(\theta^*) - I_{w\gamma}(\theta^*)I_{\gamma\gamma}^{-1}(\theta^*)I_{\gamma w}(\theta^*).$$

Let $\widehat{\theta_0}$ denote the estimator of θ under the null hypothesis and $\widehat{\theta_0}$ converges to θ_0^*, which is not necessarily θ^*. Instead of $G^{ww}(\theta^*)^{-1}$, the observed variance under the null is $G^{ww}(\theta_0^*)^{-1} = I_{ww}(\theta_0^*) - I_{w\gamma}(\theta_0^*)I_{\gamma\gamma}^{-1}(\theta_0^*)I_{\gamma w}(\theta_0^*)$. However, the estimator $G^{ww}\left(\widehat{\theta_0}\right)^{-1}$ sometimes may be non-positive definite if the alternative is true and θ_0^* is far away from global maximum θ^* (Freedman, 2012).

Instead of using the observed variance of the score function, another approach is to use the influence function to obtain a more accurate estimation of the variance for score function, $\Lambda(\theta_0^*)$, where

$$\Lambda(\theta_0^*) = Var\left(u_w(\widehat{\theta_0})\right) = E(H_1(\theta_0^*)H_1(\theta_0^*)^T).$$

$H_i(\theta_0^*) = U_{iw}(\theta_0^*) - I_{w\gamma}(\theta_0^*)I_{\gamma\gamma}^{-1}(\theta_0^*)U_{i\gamma}(\theta_0^*)$ is defined as the influence function of $U_{iw}(\widehat{\theta_0})$ (proof see Appendix), where $U_{iw}(\theta_0^*)$ and $U_{i\gamma}(\theta_0^*)$ are the contributions of subject i to the partial score function $U_{nw}(\theta)$ and $U_{n\gamma}(\theta)$ evaluated at θ_0^*, respectively. A basic Taylor expansion can

be applied to infer that the average partial score function $u_w(\theta)$ evaluated at $\widehat{\theta_0}$ converges to the influence function $H_1(\theta)$ evaluated at θ_0^*. That is $u_w(\widehat{\theta_0}) = H_1(\theta_0^*) + \rho_n(0)$, where $H_1(\theta_0^*) = n^{-\frac{1}{2}}\sum_{i=1}^{n} H_i(\theta_0^*)$ and $\rho_n(0) \xrightarrow{p} 0$ as $n \to \infty$. It implies that $u_w(\widehat{\theta_0})$ converges to a centered normal distribution with asymptotic covariance matrix $E(H_1(\theta_0^*)H_1(\theta_0^*)^T)$.

2.4 Asymptotic Property

The information matrix can be defined as $I(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{U_i(\theta)^T U_i(\theta) - g_i(\theta)\}$ where $g(\theta) = \frac{1}{\mathcal{L}}\frac{\partial^2}{\partial\theta\partial\theta^T}\mathcal{L}(\theta)$. It can be known that

$$I(\widehat{\theta_0}) \to \frac{1}{n}E_{\theta^*}\{U_n(\theta_0^*)U_n(\theta_0^*)^T - g(\theta_0^*)\} = \psi(\theta^*, \theta_0^*)$$

$$I(\hat{\theta}) \to \frac{1}{n}E_{\theta^*}\{U_n(\theta^*)U_n(\theta^*)^T - g(\theta^*)\} = I(\theta^*).$$

Only when the null is true, $\widehat{\theta_0}$ converges to θ^*, $E_{\theta^*}\{g(\theta_0^*)\}$ is equal to zero for any θ, and therefore, the above two information matrices are equivalent, i.e., $\psi(\theta^*, \theta_0^*) = I(\theta^*)$. The estimate of the Fisher information matrix under the null will be accurate and positive definite. However, when the alternative holds, although $E_{\theta^*}\{U_n(\theta_0^*)U_n(\theta_0^*)^T\}$ is positive definite, $E_{\theta^*}\{g(\theta_0^*)\}$ is no longer equal to zero and $\psi(\theta^*, \theta_0^*)$ may be non-positive definite with some negative eigenvalues. Therefore, $I(\widehat{\theta_0})$ may not be consistent with $I(\hat{\theta})$ and will be less robust. When estimating the observed variance $G^{ww}(\widehat{\theta_0})^{-1} = I_{ww}(\widehat{\theta_0}) - I_{w\gamma}(\widehat{\theta_0})I_{\gamma\gamma}^{-1}(\widehat{\theta_0})I_{\gamma w}(\widehat{\theta_0})$, $I_{ww}(\widehat{\theta_0})$, a block matrix of $I(\widehat{\theta_0})$, may also have negative eigenvalues which will make the observed variance non-positive definite and inconsistent with the true covariance matrix (Freedman, 2012). Nevertheless, $\Lambda(\theta_0^*)$ is always positive definite because of its quadratic form, and thus retains consistency.

2.5 A Categorical Cure Rate Score Test to Evaluate Heterogeneity

To derive a score-type test in evaluating the homogeneity under each stratum, I consider the following hypotheses:

$$H_0: \pi(z) = 1 \text{ for all } z \quad vs. \quad H_1: \pi(z) < 1 \text{ for some } z$$

According to the model of $\pi(z_i) = \frac{1}{1+\sum_1^K w_k I(z_i=k)}$, the hypothesis above can be transformed as follows:

$$H_0: w_k = 0 \text{ for all } k \quad vs. \quad H_1: w_k > 0 \text{ for some } k$$

Silvapulle and Silvapulle used the likelihood ratio approach to derive the general one-sided score statistic on the basis of the asymptotic normality of the single realization of $u(\theta) = n^{-\frac{1}{2}} U_n(\theta)$, $T_n = u(\theta)^T I^{-1}(\theta) u(\theta) - inf\{(u(\theta) - b)^T I^{-1}(\theta)(u(\theta) - b): b \geq 0\}$ (Silvapulle and Silvapulle, 1995). Due to the elements of w being non-negative (on the boundary of the parameter space), the Silvapulle test statistic can be applied to evaluate these hypotheses. It takes the form

$$T_n = u_w(\widehat{\theta_0})^T \widehat{\Lambda(\theta_0^*)}^{-1} u_w(\widehat{\theta_0}) - \inf_{\widetilde{w} \geq 0} \left\{ \left(u_w(\widehat{\theta_0}) - \widetilde{w}\right)^T \widehat{\Lambda(\theta_0^*)}^{-1} \left(u_w(\widehat{\theta_0}) - \widetilde{w}\right) \right\}$$

with $\widetilde{w} = [\widetilde{w}_1, \widetilde{w}_2, \dots, \widetilde{w}_K]^T$ and $\widetilde{w} \geq 0$ interpreted elementwise, that is $\widetilde{w_k} \geq 0, k = 1, 2, \dots, K$. Here $\widehat{\Lambda(\theta_0^*)}$ represents the estimate of variance for partial score function $u_w(\widehat{\theta_0})$, $\widehat{\Lambda(\theta_0^*)} =$

$\frac{1}{n}\sum_i^n \left(H_i(\widehat{\theta_0})H_i(\widehat{\theta_0})^T\right)$. When the population is not stratified, $K = 1$, the test statistics has a simple form $T_n = \left[u_w(\widehat{\theta_0})^T \widehat{\Lambda(\theta_0^*)}^{-1} u_w(\widehat{\theta_0})\right] \mathbf{1}(u_w(\widehat{\theta_0}) > 0)$ where $\mathbf{1}(u_w(\widehat{\theta_0}) > 0)$ is an indicator function. The statistic T_n has a limiting distribution $\mathcal{T} = Z^T \Lambda(\theta_0^*)^{-1} Z - \inf_{\widetilde{w} \geq 0}\{(Z - \widetilde{w})^T \Lambda(\theta_0^*)^{-1}(Z - \widetilde{w})\}$ with $Z \sim N(0, \Lambda(\theta_0^*))$ for large n and is asymptotically equivalent to a simple mixture chi-squared distribution $\frac{1}{2}(\chi_0^2 + \chi_1^2)$ when the dimension of parameter w is 1. If w is not a scalar, the limiting distribution would be a mixture of chi-squared distributions from 0 to K degree of freedom. The associated mixing weights cannot be derived analytically but can be approximated using numerical methods (Shapiro, 1988).

Chapter 3

Numerical Results

3.1 Simulation Studies

A numerical study is conducted to evaluate the empirical size and power for the proposed score test. This performance is further compared with that of the classical score test that assumes constancy of the cure rate. Throughout the study, a Weibull model $h_0(t) = \alpha t^{\alpha-1}(e^{-\mu})^{\alpha} = 0.8t^{-0.2}e^{-0.8\times3}$ with scale parameter $\alpha = 0.8$ and intercept $\mu = 3$ is imposed on the baseline hazards function. Consistent with the real data setting, four strata roughly of equal size are generated. More specifically, for each subject one uniform random variate on interval (0, 1) is generated and the subject is assigned to stratum, say k taking values in the set {1, 2, 3, 4}, if the generated random value falls in the interval $[(k-1)\ /\ 4,\ k\ /\ 4]$. The non-cure fraction is set at $(1+w_k)^{-1}$ for each stratum k under the group dependent model and $(1+w)^{-1}$ for the constant non-cure fraction model. The regression component of the hazards function for the uncured population $h(t) = h_0(t)\exp\{\alpha(-x^T\beta)\}$, where $x = (x_1, x_2, x_3, x_4)$, in which (x_1, x_2, x_3) are indicator variables for the first three stratum memberships (treating the last stratum as a reference) and x_4 is an independently generated variate from a truncated standard normal distribution on the interval [-1, 1]. Coefficient vector $(\beta_1, \beta_2, \beta_3, \beta_4)$ takes the value $(0.05, 0.1, 0.15, 0.5)$. An exponential survival model with hazards rate λ_c, taking values in the set {0.0075, 0.015, 0.031} representing a light, moderate, and heavy censoring scheme, is used to generate the censoring variable C_i.

These censoring mechanisms empirically yield censoring rates of about 15%, 25% and 40%. Varying the censoring rate aims at evaluating the impact of censoring on the proposed test. One thousand Monte Carlo (MC) replicates of sizes, n=400, 600, and 800 were generated.

Under the working covariate dependent cure fraction model, the observed test statistic for the a^{th} $(a = 1, 2, \ldots, 1000)$ MC sample is $T_n{}^{(a)} = u_w\left(\widehat{\theta_0}^{(a)}\right)^T \Lambda\left(\widehat{\theta_0^{*(a)}}\right)^{-1} u_w\left(\widehat{\theta_0}^{(a)}\right) - \inf_{\widetilde{w}\geq 0}\left\{\left(u_w\left(\widehat{\theta_0}^{(a)}\right) - \widetilde{w}\right)^T \Lambda\left(\widehat{\theta_0^{*(a)}}\right)^{-1}\left(u_w\left(\widehat{\theta_0}^{(a)}\right) - \widetilde{w}\right)\right\}$. Under the working constant cure fraction model, this test statistic simply reduces to $T_n{}^{(a)} = u_w\left(\widehat{\theta_0}^{(a)}\right)^T \Lambda\left(\widehat{\theta_0^{*(a)}}\right)^{-1} u_w\left(\widehat{\theta_0}^{(a)}\right)\mathbf{1}\left(u_w\left(\widehat{\theta_0}^{(a)}\right) > 0\right)$. The limiting null distribution of both test statistics is approximated by the following resampling method.

For $b = 1, \cdots, B$ replicates;

Step 1: Generate $\{\varepsilon_1^{(b)}, \ldots, \varepsilon_n^{(b)}\}$ from standard normal $\mathrm{N}(0,1)$;

Step 2: Perturb the influence function $\widetilde{u_w}^{(b)}\left(\widehat{\theta_0}\right) = n^{-\frac{1}{2}}\sum_{i=1}^n H_i\left(\widehat{\theta_0}\right)\varepsilon_i{}^{(b)}$;

Step 3: Calculate the resample test statistic $\widetilde{T_n}^{(b)} = \widetilde{u_w}\left(\widehat{\theta_0}^{(b)}\right)^T \widehat{\Lambda(\theta_0^*)}^{-1}\widetilde{u_w}\left(\widehat{\theta_0}^{(b)}\right) - \inf_{\widetilde{w}\geq 0}\left\{\left(\widetilde{u_w}\left(\widehat{\theta_0}^{(b)}\right) - \widetilde{w}\right)^T \widehat{\Lambda(\theta_0^*)}^{-1}\left(\widetilde{u_w}\left(\widehat{\theta_0}^{(b)}\right) - \widetilde{w}\right)\right\}$.

According to the Glivenko–Cantelli theorem, if B is large enough, the p-value of the test could be approximated by $\frac{1}{B}\sum_{b=1}^B \mathbf{1}\left(\widetilde{T_n}^{(b)} > T_n\right)$ where T_n is the observed value of the test statistic. Here I set B to 1,000. The empirical power of the test statistic at 5% significance level is the proportion of rejections of the null when the alternative is true. Likewise, the empirical size of the test statistic at 5% significance level is the proportion of rejections of the null when the null is true.

To investigate if the classical test and the stratum dependent cure rate test maintain their size, the cure rate was set to zero. Results from these analyses (see Table 1) show that the two tests have a very conservative behavior in that they tend to reject the null less often than anticipated. And this behavior did not recede with increasing sample sizes. I observe however that the tests can be forced to converge to the 5% nominal level by drastically increasing the censoring rate. Indeed, when there is no cure and a substantial censoring rate, some of the censored observations may be misclassified as cured. This is unrealistic in real many applications where the censoring rate is typical kept at the minimum level.

Table 1: Empirical size of the score test statistics at 5% significance level.

$\boldsymbol{w^* = 0}$									
censoring	**0.0075 (light)**			**0.015 (moderate)**			**0.031 (heavy)**		
Sample size	400	600	800	400	600	800	400	600	800
Constant Cure Rate Test	0.0007	0.0005	0.0008	0.0037	0.0032	0.0038	0.0090	0.0095	0.0088
Categorical Cure Rate Test	0.0003	0.0006	0.0005	0.0024	0.0028	0.0033	0.0102	0.0097	0.0093

I further investigate the empirical power of both the classical test and the stratum dependent cure rate test, when the true cure rate does and does not depend on the covariate. Table 2 shows the results of this analysis when the true cure rate depends on the covariate. As expected, when the true cure fraction is a constant, the classical approach exhibits more power than its stratum dependent counterpart. In contrast, when the true cure rate depends on stratification variable, the proposed score test clearly outperforms the classical test in detecting cure (Table 3).

Table 2: Empirical power of the score test statistics at 5% significance level under constant alternative.

$w^* = 0.1$									
censoring	**0.0075 (light)**			**0.015 (moderate)**			**0.031 (heavy)**		
Sample size	400	600	800	400	600	800	400	600	800
Constant Test	0.614	0.771	0.824	0.564	0.707	0.779	0.251	0.458	0.602
Categorical Test	0.341	0.738	0.802	0.214	0.555	0.730	0.050	0.168	0.305

Table 3: Empirical power of the score test statistics at 5% significance level under categorical alternatives.

$w^* = (0, 0.05, 0.3, 0.6)$									
censoring	**0.0075 (light)**			**0.015 (moderate)**			**0.031 (heavy)**		
Sample size	400	600	800	400	600	800	400	600	800
Constant Test	0.237	0.451	0.626	0.151	0.277	0.383	0.066	0.104	0.148
Categorical Test	0.582	0.905	0.974	0.491	0.869	0.960	0.147	0.415	0.658

Overall, if the sample size gets large, the power of the test with the same settings will increase. However, increasing censoring rates negatively impact the power of the tests. This last result implies that if the censoring rate is minimal after a long enough follow-up, there is a greater power in detecting cure.

3.2 Real Data: Ovarian cancer data from SEER registry database

The proposed score-type test is performed to detect cure for ovarian cancer in the population with the first malignant ovarian tumor diagnosed between 1992 and 2009 in Los Angeles. The population consists of 2462 patients with survival status and diagnosis information collected by the SEER registry database. To evaluate whether the follow-up time (time since diagnosis) for ovarian cancer is long enough, a nonparametric test proposed by Maller and Zhou

(1994) is performed. With sample size n, let Q_n^* denote the largest uncensored survival time, Q_n the largest survival time and N_n the number of uncensored patients between the time $2Q_n^* - Q_n$ to time Q_n^*. According to Maller and Zhou (1994), if $\widehat{\beta_n} = (1 - N_n/n)^n$ is below 0.05, the follow-up time would be considered sufficient. For the Los Angeles ovarian cancer data, the estimate of β_n is much lower than 0.05, which gives some evidence that the follow-up time in these data satisfies this condition.

To perform the proposed test for the cure fraction, a Cox-type model of the form $h(t) = h_0(t)\exp\{\alpha(-x^T\beta)\}$, was entertained for the hazards function of the uncured population. Here $x = (x_1, x_2, x_3, x_4, x_5)$ represents the vector of covariates with x_1, x_2, x_3 denoting the indicator variables for the age groups 0-30, 30-51, 51-65 (treating the age group 65+ as a reference), x_4 the number of primary tumors detected at diagnosis, and x_5 the number of lymph nodes detected. We assume that the baseline intensity $h_0(t)$ for the uncured population is a two-parameter Weibull model. The limiting distribution of the proposed test was approximated using 10,000 resamples.

Table 4: The score-type test statistics(p-values) for ovarian cancer data

	Constant Cure Rate Test	Categorical Cure Rate Test
Test Statistics	1.885293	11.01644
P value	0.0832	0.0062

Using the classical test that assumes a constant w across all the age groups, the null hypothesis failed to be rejected at 5% significance level (Table 4). In contrast, the proposed test that incorporates the age-group information into the testing procedure, gives a strong evidence of cure for at least one of the age groups. Further analyses (results not shown) show that the evidence of cure is primarily due to the younger population, as anticipated form the preliminary

analysis in Figure 2. The failure to detect cure in the classical testing procedure is consistent with the preliminary analysis in that the cure depends on age. This analysis is an excellent example where covariates information can improve power of detecting heterogeneity in the population.

Chapter 4

Discussion

The evaluation of the cure fraction in the cure survival models has attracted considerable attention in the statistical literature, but the existing testing procedures have relied on a restrictive assumption. Typically the working model entertained in these methods assumes that the cure fraction being evaluated does not depend on covariates. In this thesis, I considered the realistic situation where the cure fraction depends on a categorical covariate. Under this assumption, the null hypothesis results in multiple parameters resting on the boundary of the parameter space. To accommodate this complication, the Silvapulle and Silvapulle generalized score-test was performed. The limiting null distribution of this test was rigorously approximated by a resampling procedure which does not require the null model to be fitted repeatedly. The result from the simulation studies showed that, if the true cure fractions are stratum-specific, ignoring this information may have a negative impact on the power of the test. However, the constant cure fraction based test may outperform the proposed test when the true cure fraction does not depend on covariates. The key contribution of this thesis is that in real life applications, a conservative approach that assumes heterogeneity of the cure fractions should be entertained given that the true model is unknown to the analyst.

If heterogeneity of the cure fraction is suspected, one trivial approach would be to conduct a constant cure fraction based test within each stratum. This approach is important but has limitations. While it is theoretically feasible to conduct a constant cure rate test within each specified stratum, it is not practical to do so. In the proposed test, all the parameters but w

throughout strata are constant and are estimated jointly under the null hypothesis. If patients in a particular stratum represent a small component of the overall sample as is the case of younger pre-menopausal patients for ovarian cancer data, the parameter estimated under the null in this particular stratum would be less reliable which may result in loss of efficiency.

The main limitation of the proposed test is that I assume that the cure fraction depends on a categorical covariate with a finite number of levels. However, in practice the cure fraction may also depend on a categorical covariate with an infinite number of levels consistent with a continuous covariate. This then begs for a testing procedure that can accommodate both categorical and continuous covariates. Since the true model for the cure fraction is usually unknown to the analyst in real applications, a conservative approach would be to entertain a more general specification of the working model. This recommendation has been proposed in earlier work by (Cao et al., 2014).

APPENDICES

Appendix A: Details in Observed Covariance Matrix

The observed Fisher information $I(\theta)$ can be partitioned as $\left(I_{ww}(\theta), I_{w\gamma}(\theta), I_{\gamma w}(\theta), I_{\gamma\gamma}(\theta)\right)$.

$I(\theta) = I(w,\gamma) = \begin{bmatrix} I_{ww}(\theta) & I_{w\gamma}(\theta) \\ I_{\gamma w}(\theta) & I_{\gamma\gamma}(\theta) \end{bmatrix}$ where the elements $I_{ww}, I_{w\gamma} = I_{\gamma w}^{T}$ and $I_{\gamma\gamma}$ are, respectively,

$$-\frac{\partial^2}{\partial w \partial w^T} l(\theta), \qquad -\frac{\partial^2}{\partial w \partial \gamma} l(\theta), \qquad -\frac{\partial^2}{\partial \gamma \partial \gamma^T} l(\theta).$$

$I^{-1}(\theta) = \begin{bmatrix} G^{ww}(\theta) & G^{w\gamma}(\theta) \\ G^{\gamma w}(\theta) & G^{\gamma\gamma}(\theta) \end{bmatrix}$ where $G^{ww} = \left(I_{ww} - I_{w\gamma} I_{\gamma\gamma}^{-1} I_{\gamma w}\right)^{-1}$. Condition A ensure that the asymptotic normality of the 'single' realization U_{1w} has the variance-covariance matrix $\{G^{ww}\}^{-1} = I_{ww} - I_{w\gamma} I_{\gamma\gamma}^{-1} I_{\gamma w}$.

Appendix B: Proof of Influence Function Theorems

A basic Taylor expansion is applied to $U_{nw}(\widehat{\theta_0})$ around the true value θ_0^* to get $U_{nw}(\widehat{\theta_0}) \approx U_{nw}(\theta_0^*) + \frac{\partial}{\partial\gamma}U_{nw}(\theta_0^*)\{\hat{\gamma} - \gamma^*\}$. Similarly, we can apply Taylor expansion to $U_{n\gamma}(\widehat{\theta_0})$ around the true value θ_0^* to have $U_{n\gamma}(\widehat{\theta_0}) = 0 \approx U_{n\gamma}(\theta_0^*) + \frac{\partial}{\partial\gamma}U_{n\gamma}(\theta_0^*)\{\hat{\gamma} - \gamma^*\}$. These two Taylor expansion could lead to $U_{nw}(\widehat{\theta_0}) \approx U_{nw}(\theta_0^*) - \frac{\partial}{\partial\gamma}U_{nw}(\theta_0^*)\left\{\frac{\partial}{\partial\gamma}U_{n\gamma}(\theta_0^*)\right\}^{-1}U_{n\gamma}(\theta_0^*)$ which can be rewritten as $U_{nw}(\widehat{\theta_0}) \approx U_{nw}(\theta_0^*) - I_{w\gamma}(\theta_0^*)I_{\gamma\gamma}^{-1}(\theta_0^*)U_{n\gamma}(\theta_0^*)$. Define function $H(.)$ evaluated at θ_0^* is $H_n(\theta_0^*) = U_{nw}(\theta_0^*) - I_{w\gamma}(\theta_0^*)I_{\gamma\gamma}^{-1}(\theta_0^*)U_{n\gamma}(\theta_0^*)$ and the contribution of subject i to this function is the influence function of average partial score function $U_{1w}(\widehat{\theta_0})$ which is $H_i(\theta_0^*) = U_{iw}(\theta_0^*) - I_{w\gamma}(\theta_0^*)I_{\gamma\gamma}^{-1}(\theta_0^*)U_{i\gamma}(\theta_0^*)$. Based on these, we then can have $U_{1w}(\widehat{\theta_0}) \approx \sum_{i=1}^{n} H_i(\theta_0^*)$. Notice that, unlike $U_{iw}(\widehat{\theta_0})$, influence function $H_i(\theta_0^*)$ is independent with covariance $\Lambda(\theta_0^*) = \frac{1}{n}\sum_i^n E(H_i(\theta_0^*)H_i(\theta_0^*)^T)$.

REFERENCES

REFERENCES

Berkson, J., and Gage, R.P. (1952). Survival curve for cancer patients following treatment. J. Am. Stat. Assoc. *47*, 501–515.

Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J. R. Stat. Soc. Ser. B Methodol. 11, 15–53.

Cantor, A.B., and Shuster, J.J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. Stat. Med. 11, 931–937.

Cao, G., Hsu, W.-W., and Todem, D. (2014). A score-type test for heterogeneity in zero-inflated models in a stratified population. Stat. Med. 33, 2103–2114.

Denham, J., Denham, E., Dear, K., and Hudson, G. (1996). The follicular non-Hodgkin's Lymphomas—I. The possibility of cure. Eur. J. Cancer 32, 470–479.

Engel, J., Eckel, R., Schubert-Fritschle, G., Kerr, J., Kuhn, W., Diebold, J., Kimmig, R., Rehbock, J., and Hölzel, D. (2002). Moderate progress for ovarian cancer in the last 20 years: prolongation of survival, but no improvement in the cure rate. Eur. J. Cancer 38, 2435–2445.

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

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

Freedman, D.A. (2012). How can the score test be inconsistent? Am. Stat.

Ghitany, M., Maller, R.A., and Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. J. Multivar. Anal. 49, 218–241.

Hsu, W.-W., Todem, D., and Kim, K. (2016). A sup-score test for the cure fraction in mixture models for long-term survivors. Biometrics.

Klebanov, L.B., and Yakovlev, A.Y. (2007). A new approach to testing for sufficient follow-up in cure-rate analysis. J. Stat. Plan. Inference 137, 3557–3569.

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

Maller, R.A., and Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. J. Am. Stat. Assoc. 89, 1499–1506.

Peng, Y., and Dear, K.B. (2000). A nonparametric mixture model for cure rate estimation. Biometrics 56, 237–243.

Ries, L.A.G. (1993). Ovarian cancer: survival and treatment differences by age. Cancer 71, 524–529.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. Int. Stat. Rev. Int. Stat. 49–62.

Silvapulle, M.J., and Silvapulle, P. (1995). A score test against one-sided alternatives. J. Am. Stat. Assoc. 90, 342–349.

Tai, P., Yu, E., Cserni, G., Vlastos, G., Royce, M., Kunkler, I., and Vinh-Hung, V. (2005). Minimum follow-up time required for the estimation of statistical cure of cancer patients: verification using data from 42 cancer sites in the SEER database. BMC Cancer 5, 1.

Wang, L., Du, P., and Liang, H. (2012). Two-Component Mixture Cure Rate Model with Spline Estimated Nonparametric Components. Biometrics 68, 726–735.

Yancik, R. (1993). Ovarian cancer: age contrasts in incidence, histology, disease stage at diagnosis, and mortality. Cancer 71, 517–523.

Zhao, Y., Lee, A.H., Yau, K.K., Burke, V., and McLachlan, G.J. (2009). A score test for assessing the cured proportion in the long-term survivor mixture model. Stat. Med. 28, 3454.