Meta-analysis of prognostic studies for a biomarker with a study-specific cut-off value

Eiji Sadashima

Kurume University Graduate School of Medicine

67 Asahi-Machi, Kurume City, Fukuoka 830-0011, Japan

and

Shin-Koga Hospital, Medical Corporation Tenjinkai

120 Tenjin-Chyou, Kurume City, Fukuoka 830-8577, Japan

E-mail:a211gm012s@std.kurume-u.ac.jp

Satoshi Hattori

Biostatistics Center, Kurume University

67 Asahi-Machi, Kurume City, Fukuoka 830-0011, Japan

E-mail:hattori_satoshi@med.kurume-u.ac.jp

Kunihiko Takahashi

Department of Biostatistics, Nagoya University Graduate School of Medicine,

65, Tsurumai-cho Showa-ku, Nagoya, Aichi, 466-8550, Japan

E-mail:kunihiko@med.nagoya-u.ac.jp

Running title: Meta-analysis of prognostic studies.

Abstract

In prognostic studies, a summary statistic such as a hazard ratio is often reported between low-expression and high-expression groups of a biomarker with a study-specific cut-off value. Recently, several meta-analyses of prognostic studies have been reported, but these studies simply combined hazard ratios provided by the individual studies, overlooking the fact that the cut-off values are study-specific. We propose a method to summarize hazard ratios with study-specific cut-off values by estimating the hazard ratio for a 1-unit change of the biomarker in the underlying individual-level model. To this end, we introduce a model for a relationship between a reported log- hazard ratio for a 1-unit expected difference in the mean biomarker value between the low- and high- expression groups, which approximates the individual-level model, and propose to make an inference of the model by using the method for trend estimation based on grouped exposure data. Our combined estimator provides a valid interpretation if the biomarker distribution is correctly specified. We applied our proposed method to a dataset that examined the association between the biomarker Ki-67 and disease-free survival in breast cancer patients. We conducted simulation studies to examine the performance of our method.

Key words: Biomarker; Cut-off value; Finite-mixture model; Meta-analysis; Prognostic study

1 Introduction

Prognostic studies have been conducted to determine whether specific biomarkers are associated with the prognosis of various diseases. Such studies have contributed to the understanding of disease progression and to the identification of subgroups of patients with poor/good prognoses and are expected to play important roles in clinical decision making, healthcare policy and patient management (Hemingway et al., 2013). However, several authors have also raised important issues regarding the conducting and reporting of prognostic studies (Altman, 2001; Hemingway et al., 2010; McShane et al., 2005; Riley et al., 2003, 2013). For example, a fundamental problem with prognostic studies is that they are often conducted with only a small sample size in a single or a few facilities. Therefore, even if a single study identifies (potential) prognostic factors, it is unclear whether the findings hold in general and thus findings in prognostic studies should be further assessed. Meta-analysis is useful for this purpose (Riley et al., 2013). Meta-analysis is a powerful tool for identifying sound evidence by examining multiple independent studies and has been widely applied for evaluations of treatment effects in clinical trials, and the findings of well-conducted meta-analyses are regarded as highly reliable evidence (The American Society of Clinical Oncology, 1997). However, applications of meta-analyses have been very limited for prognostic studies. Recently, several meta-analyses of prognostic studies have been conducted including the meta-analysis by de Azambuja et al. (2007) for the antigen Ki-67 in early-stage breast cancer, Callagy et al. (2007) for the protein BCL-2 in breast cancer, and Pak et al. (2014) and Na et al. (2014) for FDG-PET in head and neck cancer and lung cancer, respectively.

In prognostic studies, an outcome measure such as a hazard ratio is often used to distinguish patients who show high-expression of a biomarker and those who show low-expression. The definitions of high- and low-expressions depend on a cut-off value for the biomarker. Although it has been pointed out that such a categorization may be suboptimal from a statistical point of view (Altman, 2001), this method is often used in the analyses of prognostic studies since it is easy for non-statisticians to understand. However, the use of study-specific cut-off values makes it difficult to conduct a meta-analysis of prognostic studies. The above-mentioned meta-analyses of prognostic studies by de Azambuja et al. (2007), Callagy et al. (2008), Pak et al (2014) and Na et al. (2014) applied standard meta-analysis techniques but did not take into account for the problem that the hazard ratio obtained in the prognostic study was calculated based on a study-specific cut-off value, making it difficult or impossible to accurately interpret the combined hazard ratio. This has been a pressing issue in meta-analyses of prognostic studies of biomarkers (Hemingway et al.,2010; Riley et al.,2003) and statistical methods have been less developed for the meta-analysis of prognostic studies (Sutton and Higgins, 2008).

Recently, Riley et al. (2015) attempted to address this issue. They proposed a multivariate meta-regression model, in which the study-specific cut-off value was incorporated as study-level covariates. Their focus was on examining the association between an outcome such as a hazard ratios and a cut-off value. Their method can then be used to obtain confidence intervals and prediction intervals for any given cut-off value. This is very useful for determining an appropriate cut-off value for dividing subjects into two prognostic groups.

In this paper, we develop a method to obtain a summary statistic for multiple hazard ratios in a literature-based meta-analysis, where literature-based means a meta-analysis using the summary statistics reported in published studies. Our focus is different from that of Riley et al. (2015); we summarize the published hazard ratios independently of their cut-off values by estimating the underlying relationship of an individual subject between a biomarker and the outcome. A fundamental difficulty in making this inference is that biomarkers for individual subjects are not observable in literature-based meta-analysis. To overcome this difficulty, we introduce an idea from methods for a meta-analysis evaluating the exposure-response relationship based on the outcomes of grouped exposure reported by Shi and Copas (2004) and Takahashi and Tango (2010). That is, we assume that unobservable measurements of the biomarker follow a continuous distribution. Its unknown parameters are estimated by the maximum likelihood method based on the number of patients in the low-expression group, that in the high-expression group and the cut-off value reported in each paper. With the estimated distributions in hand, we propose a method for combining the results of prognostic studies based on a fixed-effect model or a random-effect model.

In Section 2, we summarize the data reported by de Azambuja et al. (2007) and their results of a meta-analysis of prognostic studies of Ki-67, and we prepare the notations. In Section 3, we propose an inference procedure. For simplicity of presentation, we begin by explaining our proposal under the assumption that the distributions of the biomarker are common across studies (Subsection 3.1). As suggested by the data of de Azambuja et al. (2007), however, distributions may not be common across studies. We next apply a finite-mixture model for modeling distributions of Ki-67 (Subsection 3.2). In Section 3.3, some extensions are addressed. The data by de Azambuja et al. (2007) are re-analyzed using our proposed method in Section 5. In Section 6, we present the results of simulation studies to examine whether our proposed method works well in practice. Finally, several remaining issues are discussed in Section 7.

2 Motivating data and notations

The antigen Ki-67 is of great interest due to its role as a biomarker of proliferation (Brown et al., 1996). In particular, Cheang et al. (2009) showed that Ki-67 was useful to discriminate Luminal A and Luminal B, which were identified as breast cancer subtypes by gene-expression profiling, and then Ki-67 has been employed in the classification rule for the prognosis of breast cancer patients in the 2011 St. Gallen International Expert Consensus Report (Goldhirsch et al. 2011). Ki-67 is measured as a proportion of positive cells in a tumor specimen and thus it ranges from 0 to 1, or 0% to 100%. While de Azambuja et al. (2007) displayed percentage, we present it as a proportion from 0 to 1 in this paper. De Azambuja et al. (2007) conducted a meta-analysis of 45 prognostic studies of the significance of Ki-67 in early-stage breast cancer. See their Table 2 for a list of the studies included. They examined associations between the expression of Ki-67 and both disease-free survival and overall survival. In the present paper, we will deal only with disease-free survival. The data analyzed by de Azambuja et al. (2007) consisted of the hazard ratios of the 45 studies with their confidence intervals, along with the number of patients in the high-expression group, the number of patients in the low-expression group and the cut-off value of each study. De Azambuja et al. (2007) applied the standard meta-analysis techniques, including a fixed-effect and a random-effect model. They reported combined hazard ratios of 1.88 (95% confidence interval: 1.75, 2.02) by the fixed-effect model and 1.93 (1.74, 2.14) by the random-effect model, and concluded that Ki-67 positivity appears to be associated with a higher risk of relapse in patients with early-stage breast cancer. The cut-off values ranged from 0.035 to 0.32. However, their combined hazard ratios were obtained by simply applying the standard meta-analysis techniques without taking into account that the cut-off values were study-specific, and therefore were difficult to interpret.

Let us consider the cases of a literature-based meta-analysis of S prognostic studies for a continuous biomarker. Let $X_i^{(s)}$ be a measurement of a biomarker of the i-th patient of the s-th study. We assume that the range of the biomarker is [0,κ], where κ is a fixed constant. Here, we allow κ to be infinity and then [0,κ] is regarded as [0, ∞). As explained, Ki-67 has the range of [0,1], and there are many other biomarkers with the range [0,1], including BCL2 protein and p53 (Callagy et al., 2007). Biomarkers of an other range also have been examined in prognostic studies: Pak et al. (2014) reported a meta-analysis of Metabolic Tumor Volume (MTV) and Total Lesion Glycolysis (TLG) in head and neck cancer. MTV and TLG can be regarded as ranging from 0 to infinity. We suppose the following data available. For $s = 1, \ldots, S$, let $n_1^{(s)}$ and $n_0^{(s)}$ be the number of patients in the high-expression group and the number of patients in the low-expression group, respectively. The number of total patients in the s-th study is $N^{(s)} = n_1^{(s)} + n_0^{(s)}$. The high-expression and the low-expression groups are defined according to whether or not the measurement of the biomarker is equal to or more than a study-specific cut-off value $c^{(s)}$. We will use the logarithm of the hazard ratio of the high-expression group relative to the low-expression group of the s-th study represented by $y^{(s)}$ and its standard error denoted by $s^{(s)}$. These variables, including a cut-off value, are usually available in prognostic studies. $N^{(s)}$ and $c^{(s)}$ are regarded as fixed. Although $s^{(s)}$ is estimated from data, it is also regarded as fixed as is often done in meta-analysis studies (Normand, 1999).

3 Inference procedure

3.1 Under a homogeneous distribution of a biomarker

We propose to summarize published hazard ratios by estimating the underlying individual-level association between a biomarker and an outcome based on the published hazard ratios. That is, we consider the following individual-level model for the association between a biomarker and a time-to-event:

$$\lambda(t|X_i^{(s)}) = \lambda_0(t)\exp{(\tilde{\beta}X_i^{(s)})}, \tag{1}$$

where $\lambda(t|X_i^{(s)})$ is a conditional hazard function given $X_i^{(s)}$, $\tilde{\beta}$ is a regression coefficient, and $\lambda_0(t)$ is a baseline hazard function. We propose to make a summary as $\tilde{\beta}$ in the individual-level model (1), which is a 1-unit change of the biomarker value X_i. However, since the biomarker $X_i^{(s)}$ is not observed in a literature-based meta-analysis, it is very hard to estimate $\tilde{\beta}$. To make an inference on the individual-level model (1), we assume that $X_i^{(s)}$ follows a common continuous distribution on $[0,\kappa]$ across S studies. We consider a parametric class of probability density functions for $X_i^{(s)}$. It is denoted by $\{f(x;\theta);\theta \in \Theta\}$, where θ is an unknown parameter and Θ is a parameter space. The cumulative distribution function for $f(x;\theta)$ is denoted by $F(x;\theta)$. Given that $N^{(s)}$, $n_0^{(s)}$ is regarded as the realization of a binomial random variable with the probability of "success", $P(0 \leq X_i^{(s)} \leq c^{(s)}) = F(c^{(s)};\theta)$. Then, the log-likelihood function for $\{n_0^{(s)}; s = 1,\ldots,S\}$ is given by

$$\log L \;=\; \sum_{s=1}^{S}\Big[n_0^{(s)}\log\{F(c^{(s)};\theta)\} + n_1^{(s)}\log\{1 - F(c^{(s)};\theta)\}\Big].$$

The parameter θ can be estimated by the maximum likelihood method (Shi and Copas, 2004; Takahashi and Tango, 2010) and the maximum likelihood estimator is denoted by $\hat{\theta}$. Although $\hat{\theta}$ is calculated from data, we regard it as fixed in order to construct our method. Define

$$d_0^{(s)} \;=\; E(X_i^{(s)} \mid 0 \leq X_i^{(s)} < c^{(s)}) = \frac{\int_0^{c^{(s)}} xf(x;\hat{\theta})dx}{\int_0^{c^{(s)}} f(x;\hat{\theta})dx},$$

and

$$d_1^{(s)} \;=\; E(X_i^{(s)} \mid c^{(s)} \leq X_i^{(s)} \leq \kappa) = \frac{\int_{c^{(s)}}^{\kappa} f(x;\hat{\theta})dx}{\int_{c^{(s)}}^{\kappa} f(x;\hat{\theta})dx} \qquad (2)$$

which are the expectations of the low-expression group and the high-expression group, respectively. By regarding the hazard ratio of the s-th study as that for the unit of $d^{(s)} = d_1^{(s)} - d_0^{(s)}$, as a working model for Model (1), we introduce a study-level model for the association between the log-hazard ratio and $d^{(s)}$. That is,

$$y^{(s)} = \beta d^{(s)} + s^{(s)}\epsilon^{(s)} \qquad (3)$$

is assumed, where β is an unknown parameter, and $\epsilon^{(s)}$ is a zero-mean normal random error with the variance $Var(\epsilon^{(s)}) = 1$. This is a fixed-effect model. The parameter β is interpreted as a log-hazard ratio for the unit change of $d^{(s)}$. Dividing both sides of model (3) by $d^{(s)}$, one obtains

$$\frac{y^{(s)}}{d^{(s)}} = \beta + \frac{s^{(s)}}{d^{(s)}}\epsilon^{(s)}. \qquad (4)$$

This is the standard fixed-effect model of a one-way layout, which frequently arises in meta-analyses of treatment effects in clinical trials. Thus one can estimate β by using the weighted least squared method. That is,

$$\hat{\beta} \;=\; \frac{\sum_{s=1}^{S} d^{(s)}y^{(s)}/\{s^{(s)}\}^2}{\sum_{s=1}^{S}\{d^{(s)}/s^{(s)}\}^2},$$

with $Var(\hat{\beta}) = [\sum_{s=1}^{S}\{d^{(s)}/s^{(s)}\}^2]^{-1}$.

Note that $y^{(s)} = \log\lambda(t|c^{(s)} \leq X_i^{(s)} \leq \kappa) - \log\lambda(t|0 \leq X_i^{(s)} < c^{(s)})$. Then, Model (3) does not agree with the individual-level model (1), which is a model for $\log\lambda(t|X_i^{(s)})$. Model (1) leads to

$$E[\log\{\lambda(t|X_i^{(s)})\}|c^{(s)} \leq X_i^{(s)} \leq \kappa] - E[\log\{\lambda(t|X_i^{(s)})\}|0 \leq X_i^{(s)} < c^{(s)}] = \tilde{\beta}d^{(s)}. \; (5)$$

Intuitively, $y^{(s)}$, an estimator of $\log \lambda(t|c^{(s)} \leq X_i^{(s)} \leq \kappa) - \log \lambda(t|0 \leq X_i^{(s)} < c^{(s)})$, approximates the left-hand side of (5) well, and thus we anticipate that $\hat{\beta}$ can be interpreted as an estimate of regression coefficient $\tilde{\beta}$ of the individual-level model (1), although it does not hold strictly in general. We will examine in the simulation study section whether $\hat{\beta}$ in Model (3) works well as an estimator of $\tilde{\beta}$.

One can also consider a random-effect model,

$$y^{(s)} = \beta^{(s)} d^{(s)} + s^{(s)} \epsilon^{(s)}, \tag{6}$$

where $\beta^{(s)}$ is a random-effect following $N(\beta, \tau^2)$ independently of $\epsilon^{(s)}$, and τ^2 is a between-study variance. Model (6) is equivalent to

$$\frac{y^{(s)}}{d^{(s)}} = \beta^{(s)} + \frac{s^{(s)}}{d^{(s)}} \epsilon^{(s)}. \tag{7}$$

Assuming that $s^{(s)}$ and $d^{(s)}$ are fixed, this is the standard random-effect model with a random-intercept, that is frequently seen in meta-analyses of treatment effects in clinical trials. Based on expression (7), one can estimate unknown parameters β and τ by the moment method (DerSimonian and Laird, 1986) or the restricted maximum likelihood (REML) method (Normand, 1999 among others). For the REML method, in addition to software packages specialized for meta-analyses, one can use standard softwares for the linear mixed-effect models such as the MIXED Procedure in SAS (Normand, 1999). Instead of relying on expression (7), one can create a SAS code that handle (6) directly with a slight modification of the SAS codes given by van Houwelingen et al. (2002).

3.2 Under heterogeneous distributions of a biomarker among studies

In the previous subsection, we introduced our proposed method under the assumption that the distributions of the biomarker were common across studies with a probability density function $f(x; \theta)$. However, this may be unrealistic in practice. In general, the reports of prognostic studies tend to provide the cut-off values used along with their rationales. Among the 45 studies in the meta-analysis by de Azambuja et al. (2007), 14 studies used the median value as the cut-off value. The medians of these 14 studies ranged from 0.035 to 0.286, indicating that the distributions of Ki-67 are unlikely to be common across studies. We propose the use of a finite mixture model (McLachlan and Peel, 2000) to model heterogeneity in the distribution of the biomarker among studies. Consider J subpopulations of studies and suppose that each study belongs to one of the subpopulations. Within each subpopulation, the distribution of the biomarker is assumed to be common. The probability density function of the j-th subpopulation is denoted by $f_j(x;\theta_j), j = 1, 2, \ldots, J$, where $f_j(x;\theta_j)$ is defined on $[0, \kappa]$ and θ_j is its unknown parameter. The cumulative distribution function corresponding to $f_j(x;\theta_j)$ is denoted by $F_j(x;\theta_j)$.

Let $Z_j^{(s)}$ be an indicator function for membership in the j-th subpopulation. That is, $Z_j^{(s)} = 1$ if the s-th study belongs to the j-th subpopulation and $Z_j^{(s)} = 0$ otherwise. If $Z_j^{(s)}$ were observable, one could specify the distribution of the biomarker. However, it is unknown. Regarding $\{Z_j^{(s)}\}$ as missing values, one can employ the expectation-maximization (EM) algorithm for parameter estimation. Let $\xi_{sj} = P(Z_j^{(s)} = 1)$ and $\phi = (\theta_1, \ldots, \theta_J, \xi_1, \ldots, \xi_J)$. The complete-data log-likelihood function is then

$$\begin{aligned}\log L^c &= \sum_{s=1}^{S}\sum_{j=1}^{J} Z_j^{(s)}\Big[n_0^{(s)} \log\{F_j^{(s)}(c^{(s)};\theta_j)\} + n_1^{(s)} \log\{1 - F_j^{(s)}(c^{(s)};\theta_j)\}\Big] \\ &+ \sum_{s=1}^{S}\sum_{j=1}^{J} Z_j^{(s)} \log \xi_j.\end{aligned}$$

McLachlan and Peer (2000) dealt with the EM algorithm for a mixture of generalized linear models. Although our model is not a generalized linear model, one can estimate unknown parameters by using the EM algorithm. That is, we iterate the following E-step and M-step until a conversion criterion is satisfied from an initial value $\phi^{[0]}$. Let the parameter at the k-th step be denoted by $\phi^{[k]}$.

Expectation step (E-step): Calculate the expected value of the log-likelihood function, with respect to the conditional distribution given $n_0^{(s)}$ and $n_1^{(s)}$ under the current estimate of the parameter $\phi^{[k]}$.

$$E_{\phi^{[k]}}(\log L^c \mid n_0^{(s)}, n_1^{(s)}) = \sum_{s=1}^{S}\sum_{j=1}^{J} \bar{Z}_j^{(s),[k]} \Big[n_0^{(s)} \log \{F_j^{(s)}(c^{(s)};\theta_j)\}$$

$$+ n_1^{(s)} \log \{1 - F_j^{(s)}(c^{(s)};\theta_j)\}\Big] + \sum_{s=1}^{S}\sum_{j=1}^{J} \bar{Z}_j^{(s),[k]} \log \xi_j,$$

where

$$\bar{Z}_j^{(s),[k]} = E_{\phi^{[k]}}(Z_j^{(s)} \mid n_0^{(s)}, n_1^{(s)}) = P_{\phi^{[k]}}(Z_j^{(s)} = 1 \mid n_0^{(s)}, n_1^{(s)}),$$

and $E_{\phi^{[k]}}$ implies the (conditional) expectation with respect to a distribution of the parameter $\phi^{[k]}$.

Maximization step (M-step): Find the parameter that maximizes $E_{\phi^{[k]}}(\log L^c \mid n_0^{(s)}, n_1^{(s)})$.

We define $\hat{\phi}$ by the convergence point of $\{\phi^{[k]}\}$ in the EM algorithm. Then, we assign each study to the population with the maximum posterior probability $P(Z_j^{(s)} \mid n_0^{(s)}, n_1^{(s)})$. Then $d^{(s)} = d_1^{(s)} - d_0^{(s)}$ is calculated according to the distribution of the assigned subpopulation. The fixed-effect model (3) and the random-effect model (6) can be applied using an approach similar to that in Subsection 3.1.

3.3 Extensions

Some extensions are addressed in this subsection. In practice, there are heterogeneity among studies. Here, we mention only extensions for the fixed-effect model (3). Extensions for the random-effect model are straightforward. For example, among 45 studies included in meta-analysis for Ki-67 data by de Azambuja et al. (2007), 22 and 15 studies employed anti-Ki-67 and anti-MIB-1 antibodies, respectively. The other 8 studies employed other antibodies, or used both the anti-Ki-67 and anti-MIB1 antibodies. For such a case, one may wish to evaluate whether prognostic capacity of Ki-67 is dependent on antibodies. The following meta-regression could be considered for this purpose:

$$y^{(s)} = \beta d^{(s)} + \gamma^T w^{(s)} + s^{(s)} \epsilon^{(s)},$$

where $w^{(s)}$ and γ^T are vectors of study-level covariates and unknown regression coefficients. For example, by incorporating dummy variables for the use of antibodies, it would be possible to evaluate and adjust for effects of antibodies in the meta-analysis for Ki-67. Unknown parameters can be estimated by the ML or REML methods by following the standard linear mixed effect model theory.

Another important extension is to relax the log-linear association in Model (1). That is, when the linear association (1) between a biomarker and a log-hazard ratio may not hold, one may consider a non-linear relationship. Suppose we are interested in making an inference on the model,

$$\lambda(t|X_i^{(s)}) = \lambda_0(t) \exp{(\tilde{\eta}^T h(X_i^{(s)}))},$$

where $h(x) = (h_1(x), h_2(x), ..., h_q(x))^T$ is a q-dimensional vector-valued known function, $h_j(x), j = 1, 2, .., q$ is a scaler-valued known function and $\tilde{\eta}$ is a q-dimensional vector of unknown regression coefficients. For example, one may consider a model with $q = 1$ and $h_1(x) = \log{(1 + x)}$. Or, one may consider more complicated non-linear models using spline functions such as B-spline functions. We then define

$$\begin{aligned} g_{j0}^{(s)} &= E(h_j(X_i^{(s)}|0 \leq X_i^{(s)} < c^{(c)}), \\ g_{j1}^{(s)} &= E(h_j(X_i^{(s)}|c^{(c)} \leq X_i^{(s)} \leq \kappa), \end{aligned}$$

and $g_j^{(s)} = g_{j1}^{(s)} - g_{j0}^{(s)}$. Having determined the distribution of the biomarker identified, one can calculate $g_j^{(s)}$ using an approach similar that used for $d^{(s)}$. To make an inference for $\tilde{\eta}$, we fit the model,

$$y^{(s)} = \eta^T g^{(s)} + s^{(s)}\epsilon^{(s)},$$

where $g^{(s)} = (g_1^{(s)}, g_2^{(s)}, ..., g_q^{(s)})^T$ and η is a vector of regression coefficients. The unknown parameter η can be estimated by the ML or REML methods. Or, by divided by $d^{(s)}$, similarly to the models (3) and (6), one can utilize softwares allowing the standard meta-regression.

4 Application

We re-analyze this dataset to illustrate our proposed method. See Section 2 for more details on the data by de Azambuja et al. (2007). As presented in Figure 1 of Billgren et al. (2002), in early-stage breast cancer patients, Ki-67 is likely to distribute around 0 and is unlikely to distribute close to 1. Let $f^*(x; a, b) = b^a x^{a-1} \exp(-bx) I\{x > 0\}/\Gamma(a)$, which is the probability density function of the gamma distribution, where $I(.)$ is the indicator function and $\Gamma(a)$ is the gamma function. Denote the corresponding cumulative distribution function by $F^*(x; a, b)$. We introduce a distribution on [0,1] from the gamma distribution by truncating at $x = 1$. That is, we define a probability density function on [0,1] by $f(x; a, b) = f^*(x; a, b)I(0 \leq x \leq 1)/F^*(1; a, b)$. This is called the truncated gamma distribution (Johnson et al., 1994). We use the truncated gamma distribution as a model of the distribution of Ki-67.

We consider a single truncated gamma distribution and a mixture of truncated gamma distributions of two, three or four components. The EM-algorithm did not converge in a mixture of four truncated gamma distributions. To select the number of components, one may try to use the Akaike information criterion

(AIC) (Akaike, 1973) based on the binomial likelihood. However, this is not a good idea since our objective is to construct a good model to estimate the conditional expectations $d_0^{(s)}$ and $d_1^{(s)}$, whereas the AIC based on the binomial likelihood measures the fitting of a model to cell frequencies of $\{(n_0^{(s)}, n_1^{(s)})\}$. Indeed, as will be mentioned in the last paragraph of this section, the AIC did not work well in this example. Instead, we utilized a sample mean of the biomarker. Among the 45 studies, 7 studies reported a sample mean $m^{(s)}$ of Ki-67 across the entire study population (rather than separate means for the high- and the low-expression). A model-based estimate of the mean of Ki-67 is given by $\hat{m}^{(s)} = (d_0^{(s)} \times n_0^{(s)} + d_1^{(s)} \times n_1^{(s)})/(n_0^{(s)} + n_1^{(s)})$. In Figure 1, scatter plots of $m^{(s)}$ and $\hat{m}^{(s)}$ based on a mixture of truncated gamma distributions are presented. The beta distribution is representative as a distribution on [0,1]. It is very flexible and is widely used in practice. For reference, we also applied the mixture of beta distributions as the distribution of Ki-67. A scatter plot for the mixture of three beta distributions is also presented in Figure 1. As shown in the figure, the mixture of three truncated gamma distributions seems to fit best. Indeed, this mixture had the minimum mean squared discrepancies between $\hat{m}^{(s)}$ and $m^{(s)}$ (MD), which is defined as $MD = \sum_{s=1}^{S} \Delta_m^{(s)} (\hat{m}^{(s)} - m^{(s)})^2 / \sum_{s=1}^{S} \Delta_m^{(s)}$, where $\Delta_m^{(s)}$ is 1 if the sth study reports $m^{(s)}$ and is 0 otherwise. We therefore employed a mixture of three truncated gamma distributions. Three, 17 and 25 studies were classified into components 1, 2 and 3, respectively. Figure 2 presents three estimated distributions in the mixture. With the mixture of three truncated gamma distributions, we applied the fixed-effect model (3) and the random-effect model (6). The regression coefficient β was estimated as 1.38 (standard error: 0.08) and 1.48 (1.23) for the fixed- and the random-effect models, respectively. The hazard ratio for 0.2 units of $d^{(s)}$ is given by $\exp(\beta/5)$. Then the hazard ratios for 0.2 units of $X_i^{(s)}$ was estimated as 1.32 (95% confidence interval: 1.28, 1.36) and 1.34 (1.28, 1.41) with the fixed- and the random-effect models, respectively, as given in Table 1, in which for reference, estimates with other distributions of Ki-67 are also presented.

It is very important in a meta-analysis to determine whether between-study heterogeneity exists. From expression (7) above, one can assess between-study heterogeneity by using a forest plot, Cochran's Q-test or the I^2-index (Viechtbauer, 2010) based on $y^{(s)}/d^{(s)}$. The forest plot given in Figure 3, which was created by the *metafor* package in R (Viechtbauer, 2010), indicates heterogeneity among the studies. The p-value of the likelihood ratio test was 0.002 and the I^2-index was 43.02%, indicating that between-study heterogeneity cannot be ignored. The estimate by the random-effect model is thus more appealing.

In Figure 4, we present hazard ratios relative to baseline over Ki-67 according to Model (1). To assess the appropriateness of the linear relationship between the log-hazard and $d^{(s)}$, we conducted residual analysis for Model (1). By using the best linear unbiased predictor for the random-effect (Laird and Ware 1982), we predict the residual ϵ_i. The predicted residual is denoted by $\hat{\epsilon}_i$. In Figure 5A, we give a plot of $\hat{\epsilon}_i$s over $d^{(s)}$ and an estimated mean profile over $d^{(s)}$ by Gaussian kernel smoothing. In Figure 5B, plots of the sample quantile and the quantile based on the standard normal distribution for $\hat{\epsilon}_i$ are presented. These figures indicate that the estimated mean profile has a tendency to decrease slightly as $d^{(s)}$ increases, but there seems no substantial systematic departure from the linearity and thus that Model (1) fit well.

[Insert Figures 1,2,3,4 and 5 around here.]

[Insert Table 1 around here.]

It is also important to evaluate whether or not estimates may suffer from small study effects including publication bias (Sterne et al. 2011 among others).

One can assess and adjust for small study effects by applying techniques such as Egger's regression test (Egger et al., 1997), the funnel plot (Light and Pillemer, 1984) and the trim-and-fill method (Duval and Tweedie, 2000) to $y^{(s)}/d^{(s)}$. Here, Egger's regression test provided a p-value of 0.002. Although Egger's regression may not keep its nominal level in some situations (Jin et al. 2015 and references therein), this small p-value may suggest concerns to publication bias. In Figure 6, we present a funnel plot of the 45 studies, which was created by the *metafor* package in R (Viechtbauer, 2010). The funnel plot seems to be highly asymmetric, and thus some studies may not have been reported. To determine whether publication bias strongly influenced our estimate of the hazard ratio, we applied the trim-and-fill method. As shown in Figure 6, eight studies (represented by open circles) were suggested not to be reported. The trim-and-fill estimate of the hazard ratio for a 0.2 unit was 1.30 (1.24,1.38) for the random-effect model. Recall that the unadjusted hazard ratio with the random-effect model was 1.32 (1.28, 1.41) and then influence of the publication bias was very small.

[Insert Figure 6 around here.]

We also fit an alternative non-linear model;

$$\lambda(t|X_i^{(s)}) = \lambda_0(t) \exp{(\tilde{\eta}^{(s)}(\log X_i^{(s)} + 1))}, \tag{8}$$

where $\tilde{\eta}^{(s)}$ is a random-effect. To make an inference on Model (8), we followed the method given in Section 3.3. In Figure 4, we present hazard ratios relative to baseline over Ki-67 based on this model, and in Figures 5C and 5D, we show results of residual analysis. They indicate that Model (8) seems to have slightly better fit than Model (1). Figure 4 indicates that the profile of the hazrd ratios relative to baseline was similar between Models (1) and (8). Except for the linear model (3) and (6), estimated results can not be summarized using only a single quantity as shown in Table 1. Accordingly, graphical displays like Figure 4 are very important. In agreement with our results, de Azambuja et al. (2007) observed a statistical significance in association between Ki-67 and the progression-free-survival. However, their hazard ratios are hard to interpret since they depend on the cut-off values of the studies included. The hazard-biomarker relationship presented in Figure 4 can be interpreted more easily as the underlying individual-level association between Ki-67 and the hazard, which is free from the cut-off values.

We also tried to apply distributions other than a mixture of truncated gamma distributions. We could not obtain convergence in the estimation of unknown parameters when we applied a truncated log-normal distribution or mixtures of two, three or four truncated log-normal distributions. We will close this section with an interesting point concerning performance of the AIC, which we observed while applying a mixture of beta-distributions. By the scatter plots given in Figure 1, we concluded that a mixture of three truncated gamma distributions are more suitable than that of three beta distribution. Figure 7 presents the estimated distributions of the mixture of three beta distributions. They have a peak at $x = 1$ and are far from the histogram given in Figure 1 of Billgren et al. (2002). In our experience, the histogram given by Billgren et al. (2002) seems a typical distribution of Ki-67. Indeed, measurements of Ki-67 obtained by the first author of the present paper from 228 patients in Shin-Koga Hospital in Japan had a distribution similar to that reported by Billgren et al. (2002) (data not shown). The mixture of three beta distributions thus seems not to be relevant to the estimation of $d^{(s)}$ since a peak around $x = 1$ causes an over-estimation of $d_1^{(s)}$. Indeed, the assigned $d^{(s)}$ values were very close to 1 in almost all of the studies and are substantially different from those obtained with the mixture of three truncated gamma distributions. As a result, the estimate of β by the mixture of three beta distributions is much smaller than that by the mixture of three truncated gamma distributions as shown in Table 1. On the other hand, we observed that the AIC for the mixture of three beta distributions was 7248, which is smaller than the AIC of 7272 for the mixture of three truncated gamma distributions. It suggests that the AIC may select a distribution inappropriate to our end. Note that the AIC is defined based on the log-likelihood of the binomial distribution instead of likelihood of (unobservable) $\{X_i^{(s)}\}$. This mode of defining AIC may be responsible for the observed poor performance of the AIC. The largest cut-off value among the 45 studies was 0.32 and it is thus difficult to determine the shape of the density function in a range from 0.32 to 1 based on data grouped by a cut-off value.

[Insert Figure 7 around here.]

5 Simulation study

5.1 Setting

We conducted a simulation study to determine whether $\hat{\beta}$ in model (3) works well as an estimator of $\tilde{\beta}$, which is the parameter in the individual-level model (1). We also examined whether our proposed estimator is sensitive to or robust against specification of the distribution of the biomarker, and evaluated performance of some criterion to identify the distribution of the biomarker.

Much as in the analysis by de Azambuja et al. (2007), we considered a meta-analysis of 45 studies. Let $T_i^{(s)}$, $C_i^{(s)}$ and $X_i^{(s)}$ be a failure time, a censoring time and an observation of a biomarker of interest of the i-th patient of the s-th study, respectively. The number of patients of the s-th study $n^{(s)}$ was set as the Ki-67 dataset reported by de Azambuja et al. (2007) for $s = 1, \ldots, S$. We assume that each study has a distribution of the biomarker, which is one of three truncated gamma distributions. That is, we generated $\{X_i^{(s)}; i = 1, \ldots, n^{(s)}\}$ from one of three gamma distributions, $GMM(0.083, 4615)$, $GMM(0.205, 394611)$ and $GMM(0.383, 592020)$, where $GMM(a, b)$ denotes the gamma distribution with the parameters of (a, b), truncated at 1. This dataset is denoted by $GMM3$. These truncated gamma distributions are the same as those estimated in the Application section. In addition, we generated $\{X_i^{(s)}; i = 1, \ldots, n^{(s)}\}$ from a mixture of three log-normal distributions $LN(-0.96, 1.28)$, $LN(-0.16, 1.77)$ and $LN(-2.50, 2.58)$ truncated at 1, where $LN(u, v)$ is the log-normal distribution with the mean u and the variance v of the log-transformed variable, and from a mixture of three normal distributions $N(-0.1, 0.2)$, $N(0.15, 0.2)$ and $N(0.3, 0.2)$ truncated at 0 and 1. These datasets are denoted as $LN3$ and $NRM3$, respectively. The failure time $T_i^{(s)}$ was generated from a Cox proportional hazards model,

$$\lambda(t|X_i^{(s)}, b^{(s)}) = 1 \times \exp\{(b^{(s)} + \log\rho) \times X_i^{(s)}/0.2\}, \tag{9}$$

where $b^{(s)}$ is a random-effect following a normal distribution of mean zero and the variance 0.01, and the parameter ρ is the hazard ratio for a change of 0.2 of the biomarker. We set ρ as 1.35, 1.2 or 1. The censoring time $C_i^{(s)}$ was generated from an exponential distribution with a mean of 2. In each study, $\tilde{T}_i^{(s)} = min(T_i^{(s)}, C_i^{(s)})$ and $\Delta_i^{(s)} = I(T_i^{(s)} \leq C_i^{(s)})$ are available, and based on these data, a log-hazard ratio $y^{(s)}$ of the high-expression group relative to the low-expression group is estimated by a Cox regression. In generating the datasets $GMM3$, $LN3$ and $NRM3$, 27.0%, 24.7% and 28.0% of observations were censored for $\rho = 1.35$, 27.2%, 27.7% and 30.0% for $\rho = 1.2$ and 33.3%, 33.3% and 33.3% for $\rho = 1$, respectively. The high-expression and low-expression groups were defined according to whether or not Ki-67 is less than a cut-off value, where the cut-off value was set to be identical to that from the Ki-67 data reported by de

Azambuja et al. (2007). We assume that $(\tilde{T}_i^{(s)}, \Delta_i^{(s)})$ is not available, but $y^{(s)}$ with its standard error and the cut-off value are available.

5.2 Sensitivity of the proposed method to specification of the number of components in the mixture model

In this subsection, we examine whether our method is sensitive to or robust against the specification of the number of components in the mixture model. We applied our proposed method to the dataset $GMM3$. For the assignment of $d^{(s)}$, we applied a mixture of the gamma distributions truncated at 1 with the number of components $1, 2, \ldots, 5$. For the estimation of a combined hazard ratio, we employed the random-effect model (6). We generated 1,000 meta-analyses and empirically evaluated averages and mean-squared-errors (MSEs) for the hazard ratio $\exp(\tilde{\beta})$. Table 2 summarizes the results of the simulation study. With five components, we did not obtain successful convergence of the EM-algorithm in almost all the realizations. The proposed method with the three correctly specified components had about 2% bias for $\rho = 1.35$. When $\rho = 1.2$, bias was very small. We observed that the biases and MSEs of the truncated gamma distributions of incorrectly specified numbers of components were very similar to those of the correctly specified mixture of the three truncated gamma distributions in this simulation. We also evaluated empirical coverage probabilities for the two-sided 95% confidence intervals and power/size for the one-sided 2.5%-level test for the null hypothesis $\rho = 0$ (HR=1) of no association between the biomarker and the hazard ratio. Table 2 indicates that the empirical coverage probabilities could be far from the nominal level of 95%. This may have been due to the discrepancy between Models (1) and (3), which may introduces biases. Although in practice the amount of bias would not be very serious, it may nonetheless cause poor coverage probabilities with large sample size in particular when the hazard ratio is relatively large, in which case the discrepancy between the two model is likely to be large. We observed that empirical sizes were close to the one-sided nominal level of 2.5%. Thus, even though the coverage probabilities may be far from the nominal level, the proposed method provides a valid test of the association between the biomarker and the hazard ratio. We also applied our method to $LN3$ with the mixture of the log-normal distribution for the biomarker; the results are also shown in Table 2. We did not obtain successful convergence with the mixture of the five components in almost all the realizations. With three or four correctly-specified components, only about 3% bias exists for $\rho = 1.35$ and only negligible bias exists for $\rho = 1.2$. With two components, a larger bias exists. These results indicate that when the true hazard ratio is rather large, the proposed estimator has a small bias as an estimator of the parameter $\tilde{\beta}$ for the individual-level model (1), and when the true hazard ratio is rather small, it has only a negligible bias. Furthermore, specifying a smaller number of components may lead to biased estimates. We also conducted a simulation study for a fixed-effect model; we generated failure times from model (9) without $b^{(s)}$ and applied a fixed-effect model (3). The results were very similar to those of the random-effect model and are not shown here.

[Insert Table 2 around here.]

5.3 Sensitivity of the proposed method to the specification of the biomarker distributions in the mixture model

We next examined whether our proposed method is sensitive to the specification of distributions in the mixture model. To the datasets described above, we applied mixture models of the same number of components as the true distribution, but with possibly misspecified distributions: a mixture of three gamma distributions, that of three log-normal distributions and that of three beta-distributions. The results are presented in Table 3. When the distribution is correctly specified, our method has only negligible biases, and when misspecified, it may have considerable biases and larger MSEs. Table 3 also provides the results when the true distribution of the biomarker is a mixture of three normal distributions. In this case, both of the fitted models are misspecified, and lead to a biased estimation.

[Insert Table 3 around here.]

5.4 Performance of criteria for selecting the biomarker distribution

Finally, we evaluated the performance of two criteria to select the distribution of the biomarker. One was AIC based on the binomial distribution and the other is the mean squared discrepancies (MD) criterion, which was used in the Application section. As shown in the Application section, $m^{(s)}$ may only be observed in part of studies. We considered three situations, in which 30, 15, and 7 studies provide $m^{(s)}$, respectively, and these cases are denoted by $MD30$, $MD15$ and $MD7$, respectively. For each criterion, we counted the number of realizations for which each model had the minimum value of the criterion and show them in Tables 2 and 3. Table 2 indicates that for the mixture of the truncated gamma distributions, both AIC and MD protected to select a model of less components than the true one, which may cause biased estimation. We observed that AIC was likely to select right number of components, and MD was likely to overestimate the n umber of components. This property of MD, however, seems not to be so problematic in practice since, as seen in Table 2, the models of four components (overestimated number of components) were not likely to lead biased estimation. Table 2 also indicate that the performance of MD is better with more studies reporting $m^{(s)}$. For a mixture of the log-normal distributions, similar tendency was observed. In Table 3, comparisons of performance of AIC and MD to select a distributional form in the finite mixture model were presented. When the true biomarker distribution is a mixture of three truncated Gamma distributions, AIC often selected a mixture of beta distributions incorrectly, which was observed in the Application section. On the other hand, MD can select a mixture of the truncated Gamma distributions more frequently and can protect to select a mixture of beta distributions. When the true biomarker distribution is a mixture of the truncated log-normal distributions, both AIC and MD select the right distributions frequently and MD outperformed AIC.

6 Discussion

We proposed a method for a meta-analysis of prognostic studies of a biomarker. The guide in Figure 8 summarizes the important steps of our proposed method. Our estimator can be interpreted as a hazard ratio per the difference of expectation of the biomarker between the high- and the low- expression groups. Our simulation study found that the estimator $\exp(\hat{\beta})$ does not have considerable biases as an estimator of the hazard ratio for the individual-level model (1), provided that the distribution of the biomarker is correctly specified, although these two quantities do not agree mathematically. Thus, our method is useful to estimate the individual-level parameters from a literature-based meta-analysis of prognostic studies, and is more appealing than simply combining hazard ratios with a study-specific cut-off value as done by de Azambuja et al. (2007). Recently, Hamingway et al. (2010) reported a meta-analysis of prognostic studies for C-reactive protein in stable coronary artery disease. They adjusted observed log-hazard ratios with study-specific cut-off values by using a scaling factor given based on log-normal distributions. Their scaling factor was calculated based on the sample mean and standard deviation or the sample median and upper and lower percentiles of the biomarker for each study. However, in general such information is not reported in prognostic studies. We observed that among 45 studies enrolled in the meta-analysis by de Azambuja et al. (2007), only 6 studies reported this information, and thus the method by Hemingway et al. (2010) could not be applied. Similar to Hemingway et al. (2010), we assume a parametric model for the biomarker. To account for heterogeneity of distributions among studies, we introduced a finite-mixture model of the truncated gamma distributions (or the truncated log-normal distributions). However, a more flexible models may be needed in some cases. If so, a mixture of distributions of different forms could be considered such as a mixture of the truncated gamma and log-normal distributions, or some random-effect models, where heterogeneity is modeled with random-effects. One limitation in our method is that we assume that $d^{(s)}$ is fixed (without error), although it is estimated with data. It is valuable to develop methods accounting for uncertainty in estimation of $d^{(s)}$.

Through our real data analysis and simulation study, we observed that the EM-algorithm for inference of the biomarker distribution may fail to converge. This is one limitation of our method. We applied an algorithm similar to that for the finite-mixture model of the generalized linear model discussed in McLachlan and Peer (2000). Our model has a much complicated link-function than those handled by McLachlan and Peer(2000) and this may cause non-convergence of the EM-algorithm. More research are warranted to improve numerical stability in making an inference on the biomarker distribution.

Prognostic studies are often analyzed based on two group comparison. Some prognostic studies may report comparison of more than 2 groups. For example, Pinder et al. (1995) classified subjects into three groups with two cut-off values of 0.17 and 0.34, and showed the Kaplan-Meier curves of overall survival. Although Pinder et al. (1995) did not provide hazard ratios of the middle or the high-expression groups relative to the low-expression group, these values can be extracted from Kaplan-Meier plots by means of Parmar et al. (1998). The method to calculate the group-specific mean of the biomarker $d_0^{(s)}$ and $d_1^{(s)}$ can be extended to the case of more than 2 groups by the method of Shi and Copas (2004). One may consider a model $y_k^{(s)} = \beta(d_k^{(s)} - d_0^{(s)}) + s^{(s)}\epsilon^{(s)}$, which is a natural extension of Model (3), where $y_k^{(s)}$ is the log-hazard ratio of the kth group relative to the lowest expression group and $d_k^{(s)}$ is an estimated mean of the biomarker for the kth group. Since the data of the lowest expression group are shared by $y_k^{(s)}$s, within-study correlations among $y_k^{(s)}$s are required to make an inference on this model, and it would be hard to estimate the correlation among log-hazard ratios within a study. Furthermore, in practice, some prognostic studies may not report cut-off values. Our method can not incorporate such studies, and further researches are warranted to address these important issues.

For Models (3) and (6), we applied the funnel plot and the trim-and-fill method to $y^{(s)}/d^{(s)}$ to detect and adjust for the influence of publication bias. It remains unclear how such intuitive applications of the funnel plot and the trim-and-fill method work in meta-analysis of prognostic studies. When the underlying relationship between the log-hazard ratio and the biomarker is far from linear, some non-linear modeling should be considered. As given in Subsection 3.3, our inference procedure can be easily extended to the non-linear functional form of $d^{(s)}$. It is unclear how to apply the funnel plot and the trim-and-fill method if the non-linear model has two or more regressors. It is important to develop methods for handling publication bias appropriately for the meta-analysis of prognostic studies, since publication bias could greatly influence the estimation in a meta-analysis of prognostic studies as it does in meta-analyses for treatment effects in clinical trials or indeed could have an even more pronounced impact on estimation.

We applied the maximum likelihood method for the grouped data proposed by Shi and Copas (2004) and Takahashi and Tango (2010). This method assumes a parametric family of unobservable measurements of a biomarker. As shown by the simulation studies, our estimator may be sensitive to the specification of the number of components and that of the distributions in the mixture model of the biomarker. However, as observed in the Application and Simulation sections, selecting the distributions of a biomarker by simply relying on the AIC may not be relevant. As demonstrated in the Application section, a sample mean of the biomarker over an entire study population may be reported in some studies. As indicated in our simulation study, even if the number of studies that provide a sample mean is limited, contrasting the sample mean with model-based counterpart may be very useful for identifying the distribution of the biomarker. As observed in our simulation study, the MD is improved with more studies reporting the overall mean $m^{(s)}$. On the other hand, one concern is that if reporting process of $m^{(s)}$ is subject to some selection bias, the MD criteria may not work well. Therefore, it is encouraged for authors of a prognostic study to report the overall mean of the biomarker. External information or indevidual pateint data (IPD) at hand may be useful to identify the distribution of the biomarker even if IPD are available only for one study or is out of enrolled studies. Furthermore, if the means of the low- and the high-expression groups are reported, one can use them for $d_0^{(0)}$ and $d_1^{(s)}$ and then can avoid uncertainty in modeling the biomarker distributions. Thus, authors of a prognostic studies are encouraged to report group-specific means of the low- and the high-expression groups.

Conducting an individual patient data (IPD) meta-analysis is a promising solution to the cut-off value issue (Altman, 2001; Riley et al., 2003; Riley et al., 2013; Sutton and Higgins, 2008). However, it is difficult to obtain all IPD for prognostic studies, since no study registration system such as the Cochran Library for clinical trials is available for prognostic studies. Literature-based meta-analyses are thus also important for prognostic studies. In addition, even if the IPD can be collected, doing so can be overly time-consuming and very costly (Abo-Zaid et al., 2012; Altman et al., 2006). It is thus important to use great care when designing an IPD meta-analysis. The results of a literature-based meta-analysis may provide helpful information when planning an IPD meta-analysis of prognostic studies. Statistical methods for mixed IPD and aggregated data meta-analysis have been developed for treatment effects in clinical trials (Riley et al., 2008a) and for diagnostic studies (Riley et al., 2008b). Developing methods for a mixed IPD and aggregated data meta-analysis is a particularly attractive approach for prognostic studies since IPD are useful in the identification of the distribution of a biomarker. Some papers on prognostic studies may report a hazard ratio for a 1-unit change of the biomaker. Na et al. (2008) reported the hazard ratio for a 1-unit change of SUV for non-small cell lung cancer. However, it is excluded in the meta-analysis by Paesmans et al. (2010) since almost all the papers relied on the hazard ratio for the high-/low-expression groups. One potential approach to this issue is to combine the estimated hazard ratio obtained by our method for studies relying on cut-off values and hazard ratios for a 1-unit change reported in literatures or those derived from IPD by means of the standard meta-analysis techniques. It is valuable to examine whether this or other related methods work well in practice.

Acknowledgment

The second author's research was partly supported by grant from the Kurume University Millennium Box Foundation for the Promotion of Science and by a Grant-in-Aid (C) from the Ministry of Education, Science, Sports and Technology of Japan (#21500286). All the authors thank the Associate Editor and two reviewers for their insightful comments. They also thank Professor Richard Riley for providing us a copy of their paper even while it was in press.

References

Abo-Zaid, G., Sauerbrei, W., and Riley, R. D. (2012). Individual participant data meta-analysis of prognostic factor studies: state of the art? *BMC Medical Research Methodology* **12**, 56.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood princile. *2nd Inter. Symp. on Information Theory* (Petrov, B. N. and Csaki, F., eds.), Budapest, Akademiai Kiado, 267–281. (Reproduced in Breakthroughs in Statistics, **1** (Kotz, S. and Johnson, N. L. eds.), Springer-Verlag, New York (1992) 610–624.

Altman, D. G. (2001). Systematic reviews of evaluations of prognostic variables. *British Medical Journal* **323**, 224–228.

Altman, D. G., Trivella, M., Pezzella, F., Harris, A., and Pastorino, U. (2006). Systematic review of multiple studies of prognosis: the feasitility of obtaining individual patient data. *Advances in Statistical Methods for the Health Sciences*, Boston, Birkhäuser, 3–18.

Billgren, A-M., Tani, E., Liedberg, A., Skoog, L. and Rutqvist, L. E. (2002). Prognostic significance of tumor cell proliferation analyzed in fine needle aspirates from primary breast cancer. *Breast Cancer Research and Treatment* **71**, 161–170.

Brown, R.W., Allred, C.D., Clark, G.M., Osborne CK, Hilsenbeck SG. (1996) Prognostic Value of Ki-67 Compared to S-Phase Fraction in Axillary Node-negative Breast Cancer. *Clin Cancer Res* **2**, 585–592.

Callagy, G. M., Webber, M. J., Pharoa, P. D. P. and Carldas, C. (2008). Meta-analysis confirms BCL2 is an independent prognostic marker in breast cancer. *BMC Cancer* **8**, 153–162.

Cheang, M. C. U., Chia, S. K. C., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P. S., Parker, J. S., Perou, C. M., Ellis, M. J., Nielsen, T. O. (2009). Ki67 index, HER2 status, and prognosis of patients with Luminal B breast cancer. *Journal of National Cancer Institute* **101**, 736–750.

de Azambuja, E., Cardoso, F., de Castro Jr, G., Colozza, M., Mano, M. S., Durbecq, V., Sotiriou, C., Larsimont, D., Piccart-Gebhart, M.J., and Paesmans, M. (2007). Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12155 patients. *British Journal of Cancer* **96**, 1504–1513.

DerSimonian, R. and Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled Clinical Trials* **7**, 177–188.

Duval, S. and Tweedie, R. (2000). A nonparametric ”Trim and Fill” method of accounting for publication bias in meta-analysis. *Journal of American Statistical Association* **95**, 89–98.

Egger, M., Smith, D. G., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by simple, graphical test. *British Medical Journal* **315**, 629–634.

Goldhirsch, A., Wood, W. C., Coates, A. S., Gelber, R. D., Thurlimann, B., Senn, H. J., and Panel members. (2011). Strategies for subtypes-dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer. *Annals of Oncology* **22**, 1736–1747.

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., Lindsay, A. B., Udumyan, R., Moons, K. G. M., Steyerberg, E. W., Robert, I., Schroter, S., Altman, D. G., Riley, R. D., for the PROGRESS Group (2013). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes.. *British Medical Journal* **346**, e5595.

Hemingway, H., Riley, R. D. and Altman, D. G. (2010). Ten steps towards improving prognosis research. *British Medical Journal* **340**, 410–414.

Hemingway, H., Philipson, P., Chen, R., (2010). Evaluating the quality of research into a single prognostic biomarker: a systematic review and meta-analysis of 83 studies of C-Reactive protein in stable coronary artery disease. *PLoS Med* **7**, e1000286.

Jin, Z. C., Zhou, X. H., and He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine* **34**, 343–360.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963–974.

Light, R. and Pillemer, D. (1984). Summing up: The Science of Reviewing Research.. *Cambridge.*

McLachlan, G, J. and Krishnan, T. (2008). The EM Algorithm and Extensions. Second Edition. John Wiley and Sons, Inc, Hoboken, New Jersey.

McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models. John Wiley and Sons, Inc, New York.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M. and Clark, G.M. (2005). Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of National Cancer Institute* **97**, 1180–1184.

Na, F., Wang, J., Li, C., Deng,L., Xue, J., and Lu, Y. (2014). Primary tumor standardized uptake value measured on F18-Fluorodeoxyplucose positron emission tomography is of prediction value for survival and local control in non-small-cell lung cancer receiving radiotherapy: meta-analysis. *Journal of Thoracic Oncology* **9**, 834–842.

Na, I.I., Cheon, G. J., Choe, D. H., Byun, B. H., Kang, H. J., Koh, J. S., Park, J. H., Baek, H. J., Ryoo, B. Y., Lee, J. C., and Yang, S. H. (2008). Clinical significance of ^{18}F-FDG uptake by N2 lymph nodes in patients with resected stage IIIA N2 non-small-cell lung cancer: A retrospective study. *Lung Cancer* **60**, 69–74.

Normand, T. S-L. (1999). Formulating, Evaluating, Combining, and Reporting. *Statistics in Medicine* **18**, 321–359.

Paesmas M., Berghmans T., Dusart M., Garcia, C., Hossein-Foucher, C., Lafitte, J. J., Mascaux, C., Meert, A. P., Roelandts, M., Scherpereel, A., Munoz, A.

P., and Sculier, J. P. for the European Lung Cancer Working Party, and on behalf of the IASLC Lung Cancer Staging Project (2010). Primary tumor standardizing uptake measured on fluorodeoxyglucose positron emission tomograph is of prognostic value for survival in non-small cell lung cancer. *Journal of Thoracic Oncology* **5**, 612–619.

Pak, K., Cheon, G. J., Nam, H. Y., Kim, S. J., Kang, K. W., Chung, J. K., Kim, E. E., and Lee, D. S. (2014). Prognostic value of metabolic tumor volume and total lesion glycolysis in head and neck cancer: a systematic review and meta-analysis. *The Journal of Nuclear Medicine* **55**, 884–890.

Parmar, M., Torri, V., and Stewart, L. (1998). Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine*, **17**, 2815–2834.

Pinder, S. E., Wencyk, P., Sibbering, D. M., Bell, J. A., Elston, C. W., Nicholson, R., Robertson, J. F. R., Blamey, R. W. and Ellis, I. O. (1995). Assessment of the new proliferation marker MIB1 in breast carcinoma using image analysis: associations with other prognostic factors and survival. *British Journal of Cancer* **71**, 146–149.

Riley, R. D., Elia, E. G., Malin, G., Hemming, K., Price, M. P., (2015). Multivariate meta-analysis of prognostic factor studies with multiple cut-points and/or methods of measurement. *Statistics in Medicine* **34**, 2481–2496.

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., Jones, D. R. (2003). Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *British Journal of Cancer* **88**, 1191–1198.

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G. M., Abrams, K. R., Kyzas, P.A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., Hemingway, H. for the PROGRESS Group (2013). Prognosis research strategy (PROGRESS) 2: Prognostic factor research. *PLOS Medicine*, **10**, issue 2, e1001380.

Riley, R. D., Lambert, P. C., Staessen, J. A., Wang, J., Gueyffier, F., Thijs, L. and Boutitie, F. (2008a). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* **27**, 1870–1893.

Riley, R. D., Dodd, S. R., Craig, J. V., Thompson, J. R. and Williamson, P. R. (2008b). Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine* **27**, 6111–6136.

Shi, Q. J. and Copas, J.B. (2004). Meta-analysis for trend estimation. *Statistics in Medicine* **23**, 3–19.

Sterne, J. A. C, Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rcker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, Moher, D. G. D. and Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* **342**, d4002.

Sutton, A. J. and Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine* **27**, 625–650.

Takahashi, K. and Tango, T. (2010). Assignment of grouped exposure levels for trend estimation in a regression analysis of summarized data. *Statistics in Medicine* **29**, 2605–2616.

The American Society of Clinical Oncology (1997). Clinical practice guidelines for the treatment of unresectable non-small-cell lung cancer. *Journal of Clinical Oncology* **15**, 2996-3018.

van Houwelingen, H. C., Arends L. R. and Stijnen T. (2002). Advanced methods in meta-analysis: multivarite approach and meta-regression.. *Statistics in Medicine* **21**, 589–624.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36**, 1–48.

Table 1: Summary of the meta-analysis of the dataset by de Azambuja et al. (2007): # implies the number of components of a mixture model.

Modeling for the distribution of Ki-67			Fixed-effect	Mixed-effect		
Distribution	# of components	AIC	HR(95%CI)	HR(95%CI)	tau2(95%CI)	I^2
Truncated-gamma	1	7547	1.32 (1.27, 1.36)	1.34 (1.28, 1.40)	1.01 (1.00, 1.04)	39.3%
Truncated-gamma mixture	2	7334	1.32 (1.28, 1.36)	1.35 (1.28, 1.42)	1.01 (1.00, 1.04)	40.2%
	3	7272	1.32 (1.28, 1.36)	1.34 (1.28, 1.41)	1.01 (1.00, 1.04)	43.0%
	4	EM failed	NA	NA	NA	NA
Beta mixture	3	7248	1.15 (1.13, 1.17)	1.16 (1.13, 1.19)	1.00 (1.00, 1.01)	37.50%

Table 2: Summary of the simulation study for evaluation of influence of the number of components in the mixture model: CP coverage probability, HR hazard ratio, MSE mean squared error

	Distribution		# of		Empirical HR			Selection of the biomarker distribution			
True HR	True dist.	Fitted dist.	converged	Average	MSE ($\times 10^2$)	CP	Power/size	AIC	MD30	MD15	MD7
1.35	GAMM3	GAMM1	999	1.32	0.12	66.9	100.0	0	0	0	1
		GAMM2	780	1.32	0.11	47.8	100.0	1	4	36	92
		GAMM3	781	1.32	0.11	48.8	100.0	992	743	726	679
		GAMM4	799	1.32	0.12	24.7	100.0	7	253	238	228
1.2	GAMM3	GAMM1	998	1.19	0.04	92.4	100.0	0	0	0	1
		GAMM2	772	1.19	0.02	91.8	100.0	0	6	27	86
		GAMM3	750	1.19	0.02	91.9	100.0	992	728	695	671
		GAMM4	755	1.19	0.02	91.0	100.0	8	266	278	242
1	GAMM3	GAMM1	997	1.00	0.02	95.2	2.4	0	0	0	0
		GAMM2	723	1.00	0.01	94.9	2.9	3	10	28	101
		GAMM3	721	1.00	0.01	94.9	3.5	992	708	701	661
		GAMM4	728	1.00	0.01	95.6	2.9	5	282	271	238
1.35	LN3	LN1	997	1.26	1.06	7.2	99.9	0	0	0	0
		LN2	682	1.39	0.31	67.4	100.0	0	11	18	13
		LN3	780	1.31	0.24	60.4	100.0	842	470	446	457
		LN4	631	1.32	0.14	72.4	100.0	158	519	536	530
1.2	LN3	LN1	997	1.16	0.31	50.0	99.3	0	0	0	0
		LN2	594	1.23	0.19	68.5	100.0	0	7	17	12
		LN3	718	1.19	0.07	79.7	100.0	832	462	431	436
		LN4	635	1.19	0.05	91.8	100.0	168	531	552	552
1	LN3	LN1	995	1.00	0.08	94.8	2.4	0	0	0	0
		LN2	491	1.00	0.04	95.9	0.8	0	11	10	12
		LN3	713	1.00	0.02	95.5	1.5	844	440	469	454
		LN4	606	1.00	0.03	95.2	1.5	156	549	521	534

Table 3: Summary of the simulation study for evaluation of influence of the distributional assumptions in the mixture model: CP coverage probability, HR hazard ratio, MSE mean squared error

	Distribution		# of		Empirical HR			Selection of the biomarker distribution			
True HR	True dist.	Fitted dist.	converged	Average	MSE ($\times 10^2$)	CP	Power/size	AIC	MD30	MD15	MD7
1.35	GMM3	GMM3	781	1.32	0.11	48.8	100.0	478	929	884	832
		LN3	638	1.70	12.53	0.0	100.0	158	41	37	54
		BETA3	946	1.25	1.21	15.5	100.0	364	30	79	114
1.2	GMM3	GMM3	750	1.19	0.02	91.9	100.0	485	934	894	846
		LN3	606	1.38	3.53	0.5	100.0	153	36	35	50
		BETA3	936	1.15	0.36	34.1	100.0	362	30	71	104
1	GMM3	GMM3	721	1.00	0.01	94.9	3.5	453	933	897	835
		LN3	442	1.00	0.13	90.3	6.3	189	34	38	52
		BETA3	674	1.00	0.01	95.7	3.3	358	33	65	113
1.35	LN3	GMM3	897	1.21	1.95	0.0	100.0	68	38	24	39
		LN3	780	1.31	0.25	60.4	100.0	897	920	934	920
		BETA3	959	1.21	2.12	3.5	100.0	35	42	42	41
1.2	LN3	GMM3	795	1.13	0.52	0.0	100.0	76	40	42	47
		LN3	718	1.19	0.08	79.7	100.0	893	910	905	892
		BETA3	815	1.10	0.97	0.0	100.0	31	50	53	61
1	LN3	GMM3	744	1.00	0.01	95.2	1.5	68	32	33	42
		LN3	713	1.00	0.02	95.5	1.5	897	929	922	908
		BETA3	682	1.00	0.01	94.7	1.6	35	39	45	50
1.35	NRM3	GMM3	803	1.09	6.70	0.0	100.0	713	0	0	0
		LN3	748	1.18	2.79	0.0	100.0	108	385	387	394
		BETA3	818	1.19	3.09	15.8	100.0	179	615	613	606
1.2	NRM3	GMM3	769	1.05	2.19	0.0	100.0	665	0	0	0
		LN3	752	1.11	0.89	5.1	99.9	110	335	341	367
		BETA3	738	1.12	0.90	36.6	99.6	225	665	659	633
1	NRM3	GMM3	756	1.00	0.01	95.4	0.0	680	0	0	0
		LN3	681	1.00	0.05	95.0	1.6	105	339	334	347
		BETA3	703	1.00	0.06	95.7	1.6	215	661	666	653

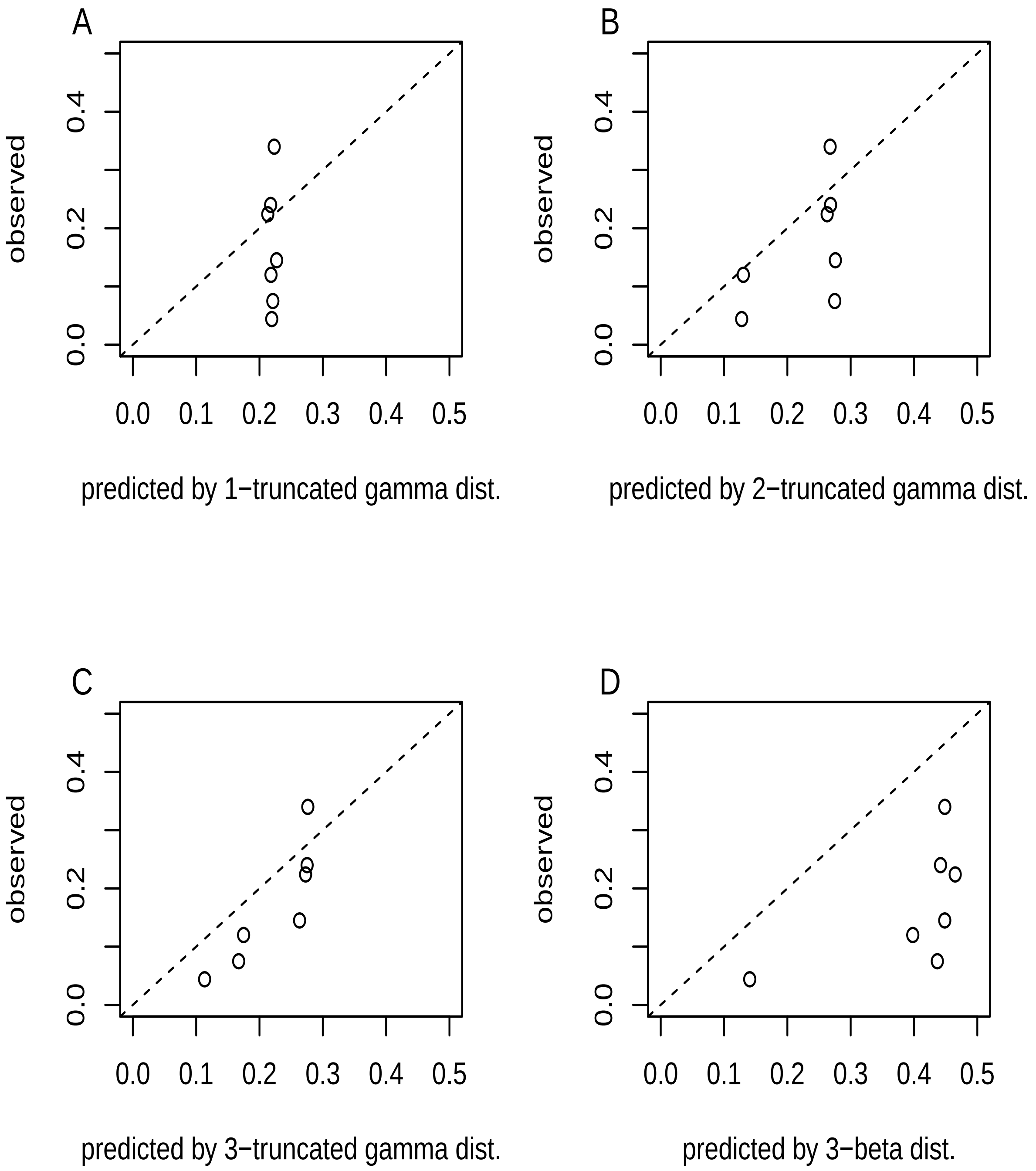


Figure 1: *Scatter plots of observed means of Ki-67 and empirical ones with (A) a truncated gamma distribution, a mixture of (B) two and (C) three truncated gamma distributions, and (D) that of three beta distributions.The broken lines indicate perfect fit (y=x).*

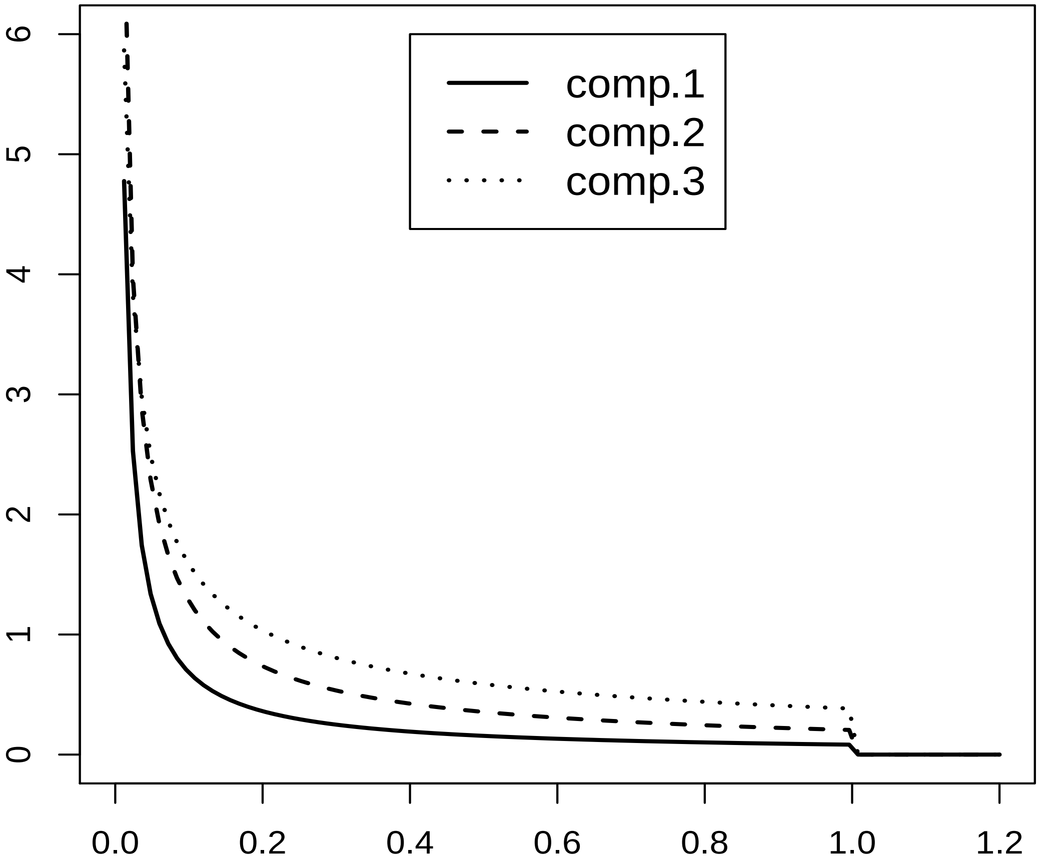


Figure 2: *Estimated probability density functions of three components in the mixture of truncated gamma distributions.*

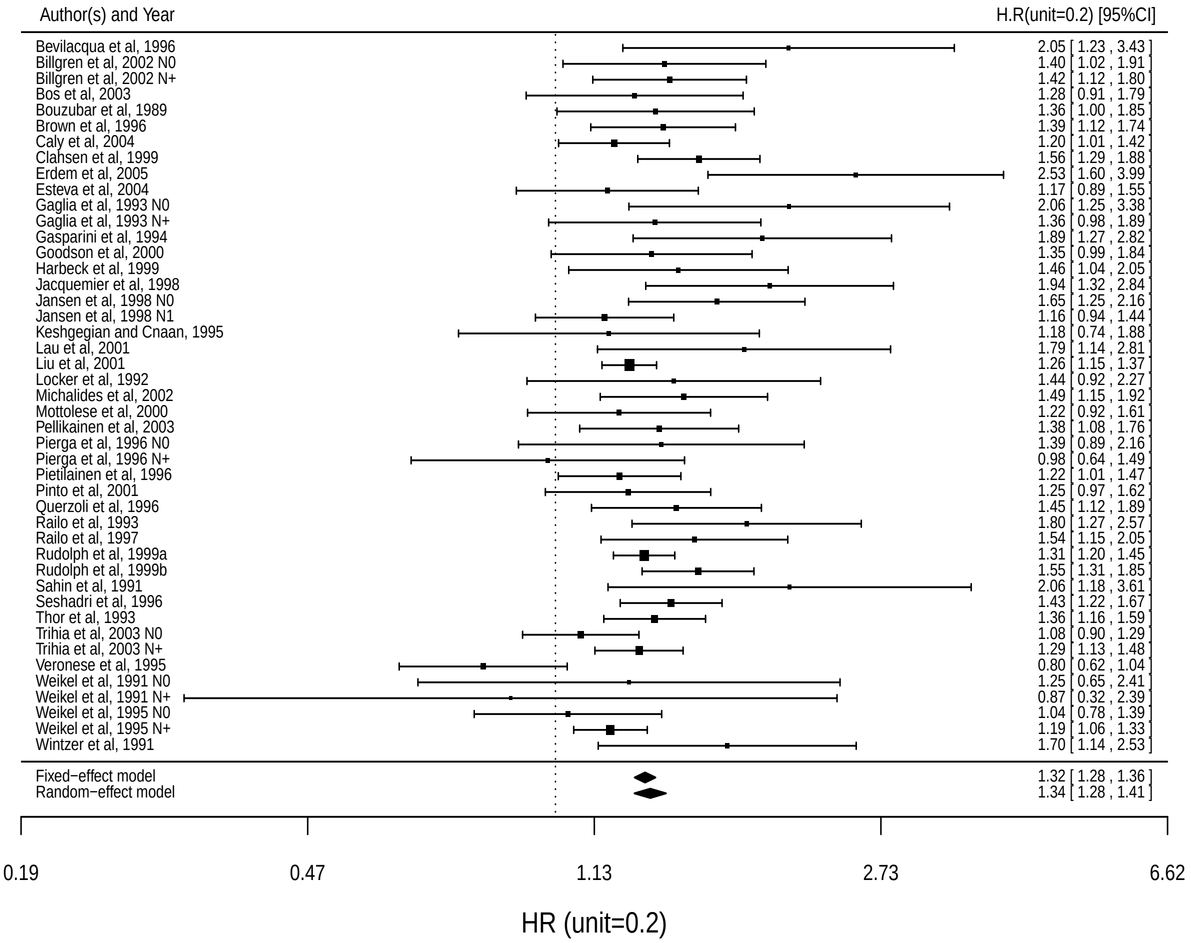


Figure 3: *A forest plot for assessing heterogeneity of* β *applied to* $\exp(0.2 \times y^{(s)}/d^{(s)})$*, which is a hazard ratio for 0.2 unit change based on Model (6)*

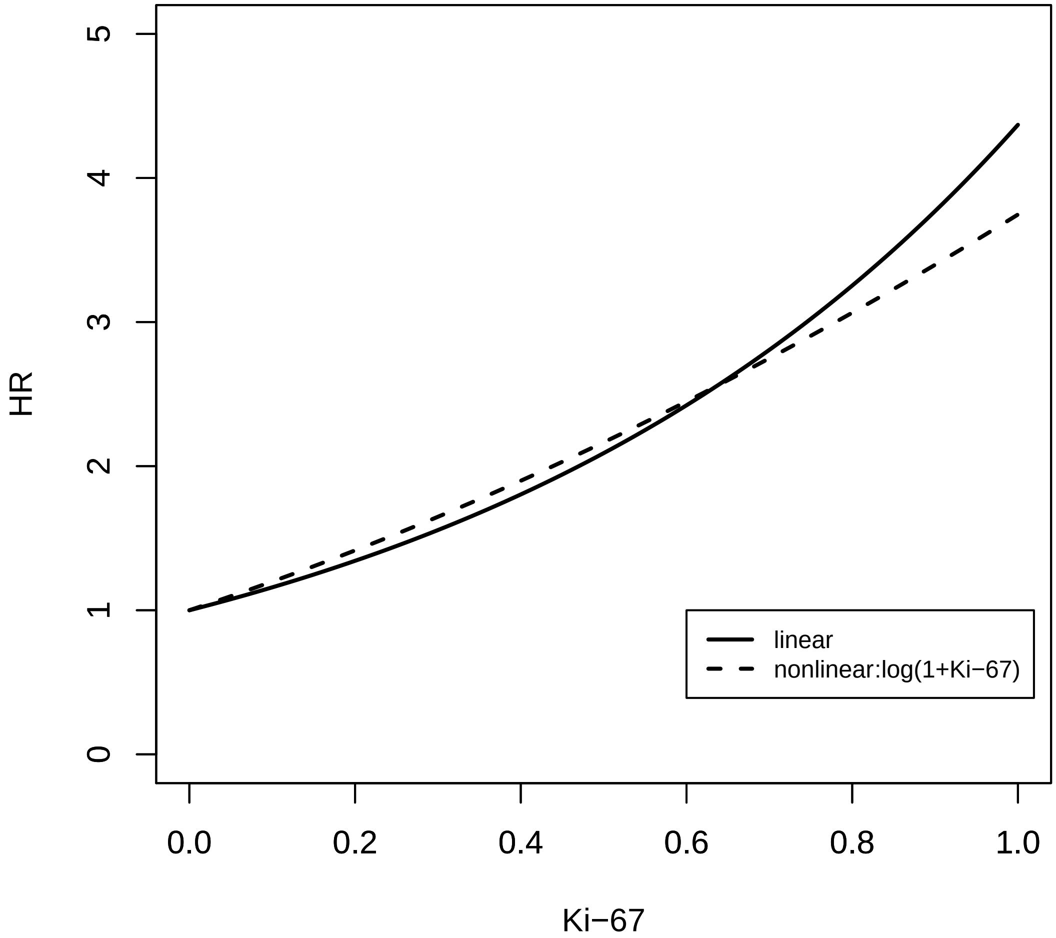


Figure 4: *Hazard ratios relative to baseline over Ki-67 estimated according to Model (1) (solid curve) and Model (8) (broken curve) with a mixture of three truncated gamma distributions.*

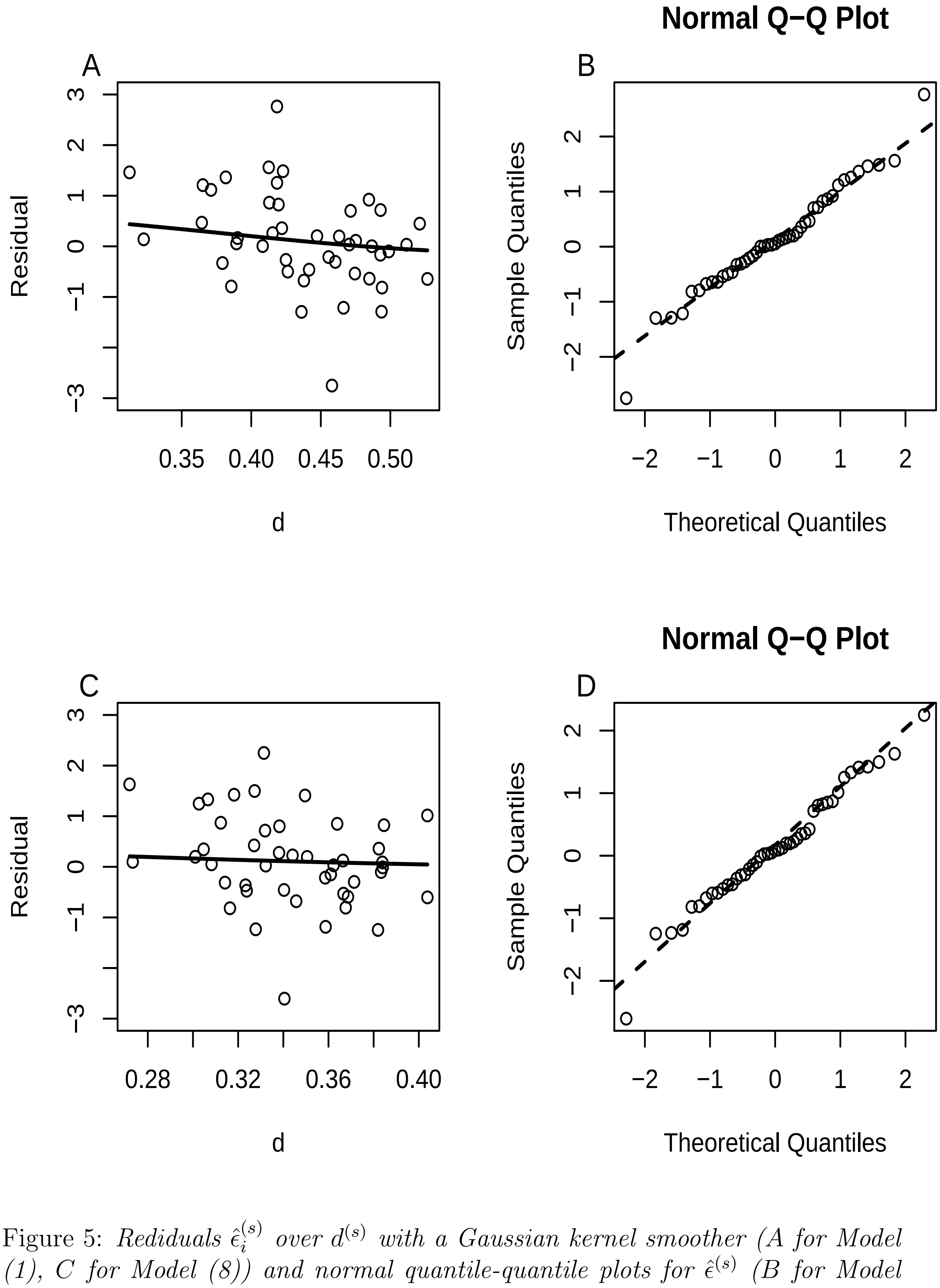


Figure 5: *Rediduals $\hat{\epsilon}_i^{(s)}$ over $d^{(s)}$ with a Gaussian kernel smoother (A for Model (1), C for Model (8)) and normal quantile-quantile plots for $\hat{\epsilon}^{(s)}$ (B for Model (1), D for Model (8)).*

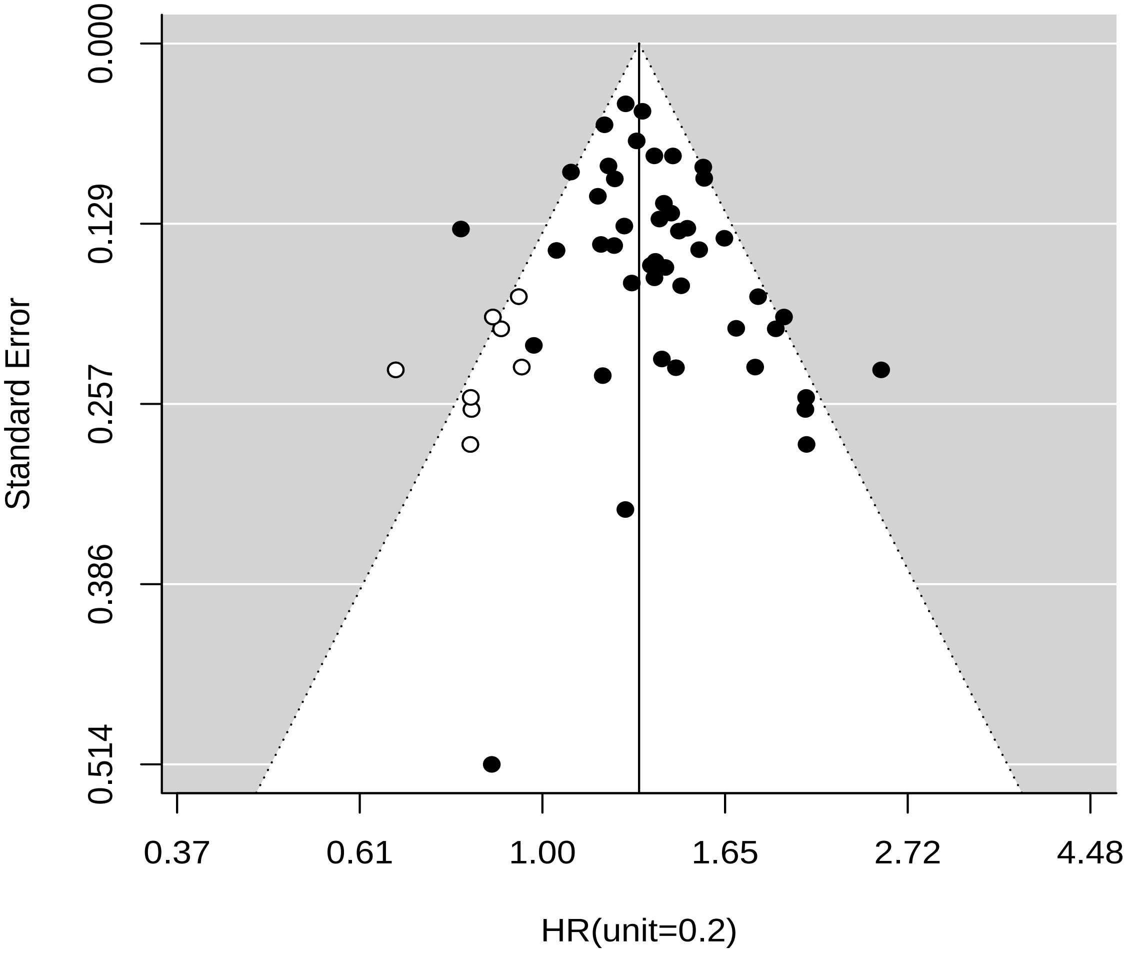


Figure 6: *A funnel plot for assessing of publication bias applied to* $\exp\left(0.2 \times y^{(s)}/d^{(s)}\right)$*, which is a hazard ratio for 0.2 unit change based on Model (6): open circles present imputed studies by the trim-and-fill method.*

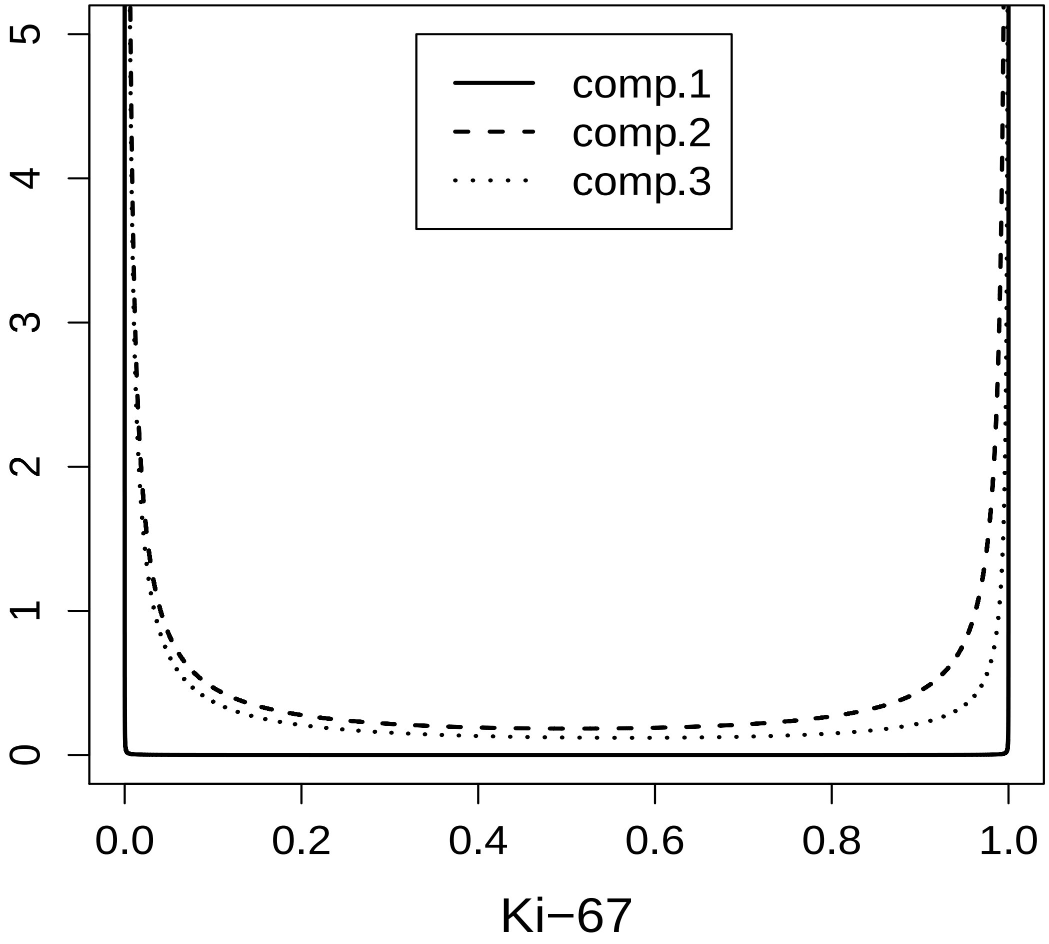
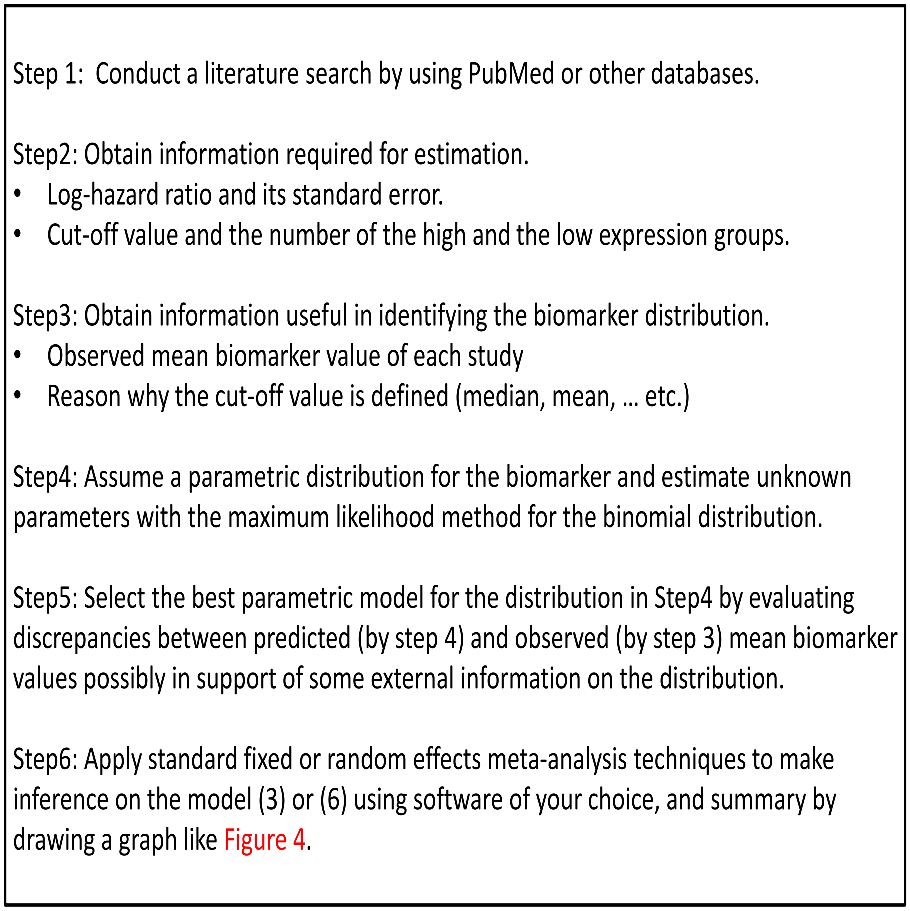


Figure 7: *Estimated probability density functions of three components in the mixture of beta distributions.*



Step 1: Conduct a literature search by using PubMed or other databases.

Step2: Obtain information required for estimation.

- Log-hazard ratio and its standard error.
- Cut-off value and the number of the high and the low expression groups.

Step3: Obtain information useful in identifying the biomarker distribution.

- Observed mean biomarker value of each study
- Reason why the cut-off value is defined (median, mean, … etc.)

Step4: Assume a parametric distribution for the biomarker and estimate unknown parameters with the maximum likelihood method for the binomial distribution.

Step5: Select the best parametric model for the distribution in Step4 by evaluating discrepancies between predicted (by step 4) and observed (by step 3) mean biomarker values possibly in support of some external information on the distribution.

Step6: Apply standard fixed or random effects meta-analysis techniques to make inference on the model (3) or (6) using software of your choice, and summary by drawing a graph like Figure 4.

Figure 8: *Step-by-step guide for our proposed method.*