Model free audit methodology for bias evaluation of tumour progression in oncology

Andrew Stone[a*], Euan Macpherson[a], Ann Smith[a], Christopher Jennison[b]

[a] AstraZeneca, Alderley Park, Macclesfield UK

[b] University of Bath, UK

[*] Corresponding author: andrew.stone@astrazeneca.com

Abstract

Many oncology studies incorporate a blinded, independent central radiological review (BICR) to make an assessment of the integrity of the primary endpoint, progression free survival (PFS). Recently it has been suggested that, in order to assess the potential for bias amongst investigators, a BICR amongst only a sample of patients could be performed; if evidence of bias is detected, according to a pre-defined threshold, the BICR is then assessed in all patients, otherwise it is concluded the sample was sufficient to rule-out meaningful levels of bias. In this paper, we present an approach that adapts a method originally created for defining futility bounds in group sequential designs. The hazard ratio ratio (HRR), the ratio of the hazard ratio (HR) for the treatment effect estimated from the BICR to the corresponding HR for the investigator assessments, is used as the metric to define bias. The approach is simple to implement, and ensures a high probability that a substantial true bias will be detected. In the absence of bias, there is a high probability of accepting the accuracy of local evaluations based on the sample, in which case an expensive BICR of all patients is avoided. The properties of the approach are demonstrated by retrospective application to a completed PIII trial in colorectal cancer. The same approach could easily be adapted for other disease settings, and for test statistics other than the hazard ratio.

Keywords: progression, sample, independent review, oncology

Introduction

Progression Free Survival (PFS) is often accepted as a valid endpoint in oncology both for assessing activity and for registration of drugs. PFS, defined as the earliest of disease progression or death, is a time-to-event endpoint which assesses the relative rate with which the disease worsens. Standard criteria, such as RECIST 1.1 [1] are applied to calculate the PFS time for each individual. The longest

diameters of a set of target lesions are measured repeatedly over time, together with an overall assessment of other non-target lesions and whether any new lesions appear. Disease progression occurs if either the sum of target lesions has increased by 20% from the nadir or there is, in the investigator's opinion, clear progression of non-target lesions or a new lesion detected.

Whilst the criteria appear largely objective there remains a degree of judgement and measurement error [2]. Furthermore, a high rate of disagreement, 50% in some cases [3], has been observed between readers in the timing of progression; much of this is attributed to different readers selecting different target lesions. This level of discordance has led to the widespread use of a blinded, independent, central review (BICR) to confirm and even replace the investigators' assessment of progression when this is the primary endpoint. Not only is a BICR expensive, up to $4-5M for a Phase III trial, it may also introduce new problems and can by itself introduce bias: if the investigator decides there is progression earlier than the BICR then no more tumour assessments will be available to the BICR and the only option for the BICR analysis is to censor patients at the time the investigator defines progression. This censoring is likely to be informative and thus, if the rate of such censoring differs between arms, then, whilst the BICR assessments remain informative, bias will be introduced in the estimation of the BICR hazard ratio (HR) [4].

We are most interested in whether the disagreement between readers in the time of progression for individual patients results in a biased estimate of a treatment effect. A number of reviews [4-6], have shown a high concordance between the local evaluation (LE) HR estimated by the investigator and HR estimated by the BICR, particularly in blinded trials, although there is some overlap in the trials considered in these reviews. Given the cost and complexity of a BICR, the idea of performing the independent review amongst a sample of patients has emerged: if the sample satisfactorily rules out the presence of bias then no more scans are re-read, otherwise the BICR is performed in all patients. An Oncology Drugs Advisory Committee (ODAC) meeting was convened in July 2012 [7] to discuss this concept and all the committee members supported the notion of a sample review.

There are currently two main methods for conducting a sample review, in this paper we present a third. In [8], the authors define θ_C to be the log hazard ratio when progression is evaluated by BICR and they test the null hypothesis H_0: $\theta_C \geq \gamma$, where the threshold γ is termed the "clinical irrelevance factor". The testing procedure uses estimates of θ_C based on (i) LE of the full set of patients plus BICR of a sample of patients or, if it is deemed appropriate, (ii) BICR of the full sample. The estimate of θ_C in (i) is a combination of HR estimates from LE and BICR data chosen to have minimum variance, given the correlation between LE and BICR estimates of HR (which can be estimated by bootstrapping the audited cases). Since H_0 can be tested twice, a multiple testing procedure is used to protect the overall type I error rate: it is a non-significant result in the first of these tests (when the upper limit of a $1-\alpha/2$ confidence interval is greater than γ) that leads to a BICR of the full data set.

The second method [5] concludes that bias is absent if appropriately defined measures of discordance in progression times are similar between treatment arms. The philosophy of this second approach is to regard the sample review as an audit to assess whether there is evidence of bias in the local evaluation for that particular trial, rather than to re-test statistical significance. The discordance measures, late

and early discrepancy rate (LDR and EDR), are compared between treatment arms and were chosen as they were found to be sensitive to bias [9]. The LDR quantifies the frequency that the LE declares progression later than the BICR as a proportion of the total number of discrepancies in the timing of progression. The EDR quantifies the frequency with which the LE declares progression early relative to BICR as a proportion of the total number of investigator assessed progressions. Initially, the authors proposed accepting the sample if the observed values of LDR and EDR were less than a fixed acceptance threshold but later proposed modifying the approach [10] to allow the acceptance thresholds to vary by design in order to guarantee the same high probability that bias would be detected if the LE and BICR HRs differed by a fixed proportion. As a result, in order to utilise the Amit method an error model must first be set-up [9] by the user to define the appropriate sample acceptance thresholds, and this can make transferring the method between different researchers a challenge. The performance of these two existing methods has been compared [11].

The model free audit approach presented in this paper is based on an approach to futility analyses developed to be used in group sequential designs. The approach has features in common with both the Dodd and Amit methods; it is simple to implement and reliably identifies bias. In common with the Dodd method it utilises the HRs directly and in common with the Amit method it aims to detect bias in terms of differences in treatment effect estimates rather than to re-test statistical significance. A key advantage of the approach lies in its simplicity and hence the ease with which it can be applied by different researchers.

The paper is structured as follows: firstly the methodology is outlined, followed by a results section identifying the likely sample sizes required to have appropriate sensitivity and specificity. The approach is then retrospectively applied to data from a trial in metastatic colorectal cancer, where the BICR was performed in all patients in order to confirm the analytical findings. Finally the paper discusses potential applications and practical considerations.

Methods

The primary inference of the model-free audit procedure concerns the point estimate for the hazard ratio ratio (HRR) in the full data set, which is equal to the point estimate of BICR HR divided by the point estimate of the LE HR. In the model-free approach, absence of bias, or more precisely lack of evidence of meaningful bias, is concluded if there is a low conditional probability that the HRR seen in the random sample of patients would have been observed if, in fact, the point estimate of the HRR in the full trial were unacceptably high, 1.25 for example. In the discussion section we explore this choice in more detail.

If no bias is found the sample is accepted and no further scans are assessed by the BICR, otherwise the BICR is performed in all patients. We propose that the estimate of treatment effect should be based on the local evaluation if either the sample is accepted or if the BICR is performed in all patients and there is insufficient evidence of bias, but if the BICR in the full trial indicates the presence of bias then inference about the treatment effect should be based on the BICR. In practical terms, the sample for

BICR assessment is drawn at completion of the trial. The patients that form this sample are randomly selected within each treatment arm, with separate sampling within patients with progression events and with censored times to event according to the LE. All scans from sampled patients are then assessed by the BICR.

The proposed process for generating the BICR sample and its evaluation is set out in Figure 1.

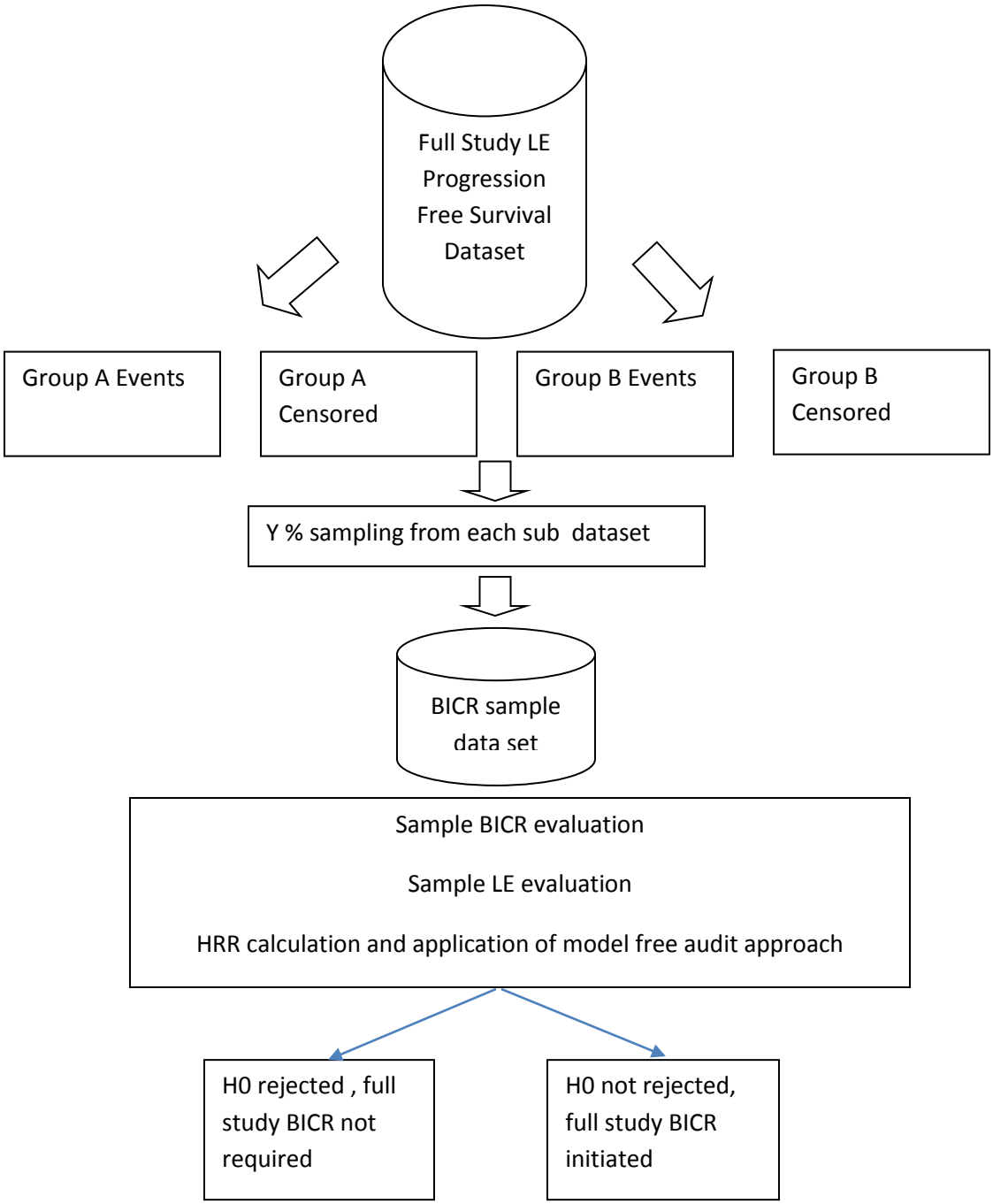


Figure 1 Process for selecting the sample for BICR and evaluating the sample results

Statistical model and assumptions:

Under the assumption of proportional hazards, denote the hazard ratio between the control and experimental treatment by

$$HR = \frac{Hazard\ rate\ of\ experimental\ treatment}{Hazard\ rate\ of\ control\ arm},$$

so a value of HR below 1 indicates the new treatment is superior to control.

Denote the estimate of HR based on the full data set and BICR evaluations of progression by

$$\widehat{HR}_{BICR,F}$$

and the estimate of HR based on the full data set and local evaluations of progression by

$$\widehat{HR}_{LE,F}.$$

We suppose that, as the gold standard, the BICR evaluations provide an unbiased estimate of the true HR, while the local evaluations may be biased. The estimated Hazard Ratio Ratio based on the full data set is

$$\widehat{HRR}_F = \frac{\widehat{HR}_{BICR,F}}{\widehat{HR}_{LE,F}}$$

and we write its large sample distribution, expressed on the log scale, as

$$\ln(\widehat{HRR}_F) \sim N(\ln(HRR),\ I_F^{-1}).$$

Here the true Hazard Ratio Ratio, HRR, is defined through the equation $\ln(HRR) = E(\ln(\widehat{HR}_{BICR,F}/\widehat{HR}_{LE,F}))$ and I_F denotes the Fisher information for $\ln(HRR)$ in the full data set. After assessment of the sample of the data, we have the estimate of HR based on BICR evaluations, $\widehat{HR}_{BICR,S}$, and the estimate of HR based on local evaluations of sampled subjects, $\widehat{HR}_{LE,S}$. The estimate of the Hazard Rate Ratio based on the sample is

$$\widehat{HRR}_S = \frac{\widehat{HR}_{BICR,S}}{\widehat{HR}_{LE,S}}$$

and the large sample distribution of this estimate is given by

$$\ln(\widehat{HRR}_S) \sim N(\ln(HRR),\ I_S^{-1}),$$

where I_S denotes the information for $\ln(HRR)$ in the sample data. We proceed on the assumption that the estimates $\ln(\widehat{HRR}_S)$ and $\ln(\widehat{HRR}_F)$ have the canonical form of joint distribution described in Jennison & Turnbull, Ch. 11 [12]. Specifically, the two estimates are bivariate normal with means and variances as stated above and their covariance is I_F^{-1}. It follows that the conditional distribution of $\ln(\widehat{HRR}_S)$ given $\widehat{HRR}_F = \widehat{HRR}_F$ is

$$\ln(\widehat{HRR}_S) \mid \widehat{HRR}_F = \widehat{HRR}_F \sim N\left(\ln\left(\widehat{HRR}_F\right), I_S^{-1} - I_F^{-1}\right).$$

Standardised test statistics are defined as

$$Z_F = \ln(\widehat{HRR}_F)\sqrt{I_F} \text{ and } Z_S = \ln(\widehat{HRR}_S)\sqrt{I_S}$$

and the conditional distribution of Z_S given $Z_F = \tilde{Z}_F$ (so $\ln(\widehat{HRR}_F) = \tilde{Z}_F/\sqrt{I_F}$) is

$$Z_S \mid Z_F = \tilde{Z}_F \sim N\left(\tilde{Z}_F \frac{\sqrt{I_S}}{\sqrt{I_F}}, \frac{I_F - I_S}{I_F}\right) \qquad (1).$$

When analysing the sample data, we specify a maximum acceptable value HRR_U (for example, as suggested **Error! Reference source not found.**) for $\widehat{HRR}_F$ and test the null hypothesis H_0: $\widehat{HRR}_F \geq HRR_U$ against the alternative H_1: $\widehat{HRR}_F < HRR_U$. Note that these hypotheses concern $\widehat{HRR}_F$, the final *estimate* of HRR. The distribution of Z_S given $\widehat{HRR}_F = HRR_U$, the case at the boundary of the null hypothesis, is given by (1) with $\tilde{Z}_F = \ln(HRR_U)\sqrt{I_F}$ and Z_S will tend to take lower values under H_1. So, for a level α test, we stop and reject H_0 based on the sample of data if

$$Z_S < \ln(HRR_U)\sqrt{I_S} - \Phi^{-1}(1-\alpha)\sqrt{\frac{I_F - I_S}{I_F}}, \quad (2)$$

where Φ is the standard normal cumulative distribution function. This criterion can be expressed as a bound on the estimated $\widehat{HRR}_S$ from the data sample:

$$\ln(\widehat{HRR}_S) < \ln(HRR_U) - \Phi^{-1}(1-\alpha)\sqrt{\frac{I_F - I_S}{I_S I_F}}$$

or, equivalently,

$$\widehat{HRR}_S < exp\left[\ln(HRR_U) - \Phi^{-1}(1-\alpha)\sqrt{\frac{(I_F - I_S)}{I_S I_F}}\right] = AT, \text{ say,} \quad (3)$$

where AT indicates the "acceptance threshold" for the Hazard Ratio Ratio observed in the sample data.

If the above test does not reject H_0, BICR is conducted for the full set of data so $\widehat{HRR}_F$ is known exactly and there is then no error in determining whether or not H_0 is true. Thus, the type I error probability α assigned to the analysis of the sample data is the total type I error probability for testing H_0: $\widehat{HRR}_F \geq HRR_U$.

Suppose now that the full data estimate of HRR takes the value $\widehat{HRR}_F = 1$. We refer to the probability of stopping to reject H_0 after analysing the sample data in this case as the "specificity" of the method. Conditionally, given $\widehat{HRR}_F = 1$, $Z_S \sim N\left(0, \frac{I_F - I_S}{I_F}\right)$ and the probability of satisfying (2), the specificity, is

$$\Phi\left[\ln(HRR_U)\sqrt{\frac{I_S I_F}{(I_F - I_S)}} - \Phi^{-1}(1-\alpha)\right] = \Phi\left[\ln(AT)\sqrt{\frac{I_S I_F}{(I_F - I_S)}}\right]. \quad (4)$$

Values of I_F and I_S

For a two-treatment comparison with randomisation ratio $k:1$ between treatment arms, we use the result

$$\widehat{Var}\left(\ln(\widehat{HR})\right) \cong \frac{(k+1)^2}{k\,n}$$

from [12]. In the full data with $n_{L,F}$ LE events and $n_{B,F}$ BICR events, we have, approximately,

$$Var\left(\ln(\widehat{HR}_{BICR,F})\right) = \frac{(k+1)^2}{k n_{B,F}} \quad \text{and} \quad Var\left(\ln(\widehat{HR}_{LE,F})\right) = \frac{(k+1)^2}{k n_{L,F}},$$

so if $Corr\left(\ln(\widehat{HR}_{BICR,F}),\ \ln(\widehat{HR}_{LE,F})\right) = \rho$, we obtain

$$Var(\ln(\widehat{HRR}_F)) = Var\left(\ln(\widehat{HR}_{BICR,F}) - \ln(\widehat{HR}_{LE,F})\right) = \frac{(k+1)^2}{k n_{B,F}} + \frac{(k+1)^2}{k n_{L,F}} - 2\,\rho\,\frac{(k+1)^2}{k}\sqrt{\frac{1}{n_{B,F} n_{L,F}}}.$$

Defining $r = n_{B,F}/n_{L,F}$, we have

$$I_F = [Var(\ln(\widehat{HRR}_F))]^{-1} = \frac{k\,n_{L,F}}{(k+1)^2}\frac{r}{(1+r-2\rho\sqrt{r})}. \quad (5)$$

In the data sample, let $n_{L,S}$ denote the number of LE events and $n_{B,S}$ the number of BICR events. By a scaling argument, we expect the ratio $n_{B,S}/n_{L,S}$ to be close to r and, approximately,

$$I_S = [Var(\ln(\widehat{HRR}_S))]^{-1} = \frac{k\,n_{L,S}}{(k+1)^2}\frac{r}{(1+r-2\rho\sqrt{r})}. \quad (6)$$

One practical consideration is how to estimate the correlation ρ between $\ln(\widehat{HR}_{BICR})$ and $\ln(\widehat{HR}_{LE})$. We have followed [8], and estimated the correlation using a bootstrap approach. In the sample there are n_{sample} patients of whom $n_{L,S}$ have events according to the LE. In the bootstrap calculations, the n_{sample} subjects are sampled with replacement, stratified by treatment arm and whether the patients had an event, to create a sample of size n_{sample}. Using both the LE and BICR determined PFS times, $\ln(\widehat{HR}_{BICR})$ and $\ln(\widehat{HR}_{LE})$ are computed in the bootstrap sample. This is repeated *b* times and the sample correlation coefficient of $\ln(\widehat{HR}_{BICR})$ and $\ln(\widehat{HR}_{LE})$ provides the estimate of ρ. Results presented in the supplementary appendix support the assumption that this correlation is independent of the size of the sample and, in particular, that $\mathrm{Corr}(\ln(\widehat{\mathrm{HR}}_{\mathrm{BICR,S}}), \ln(\widehat{\mathrm{HR}}_{\mathrm{LE,S}})) = \mathrm{Corr}(\ln(\widehat{\mathrm{HR}}_{\mathrm{BICR,F}}), \ln(\widehat{\mathrm{HR}}_{\mathrm{LE,F}}))$.

Results

We have investigated the methods described above in an example with a total sample size of N_{study}*= 500* patients and a selection of values for the audit sample size n_{sample}. We have assumed 60% of patients

have an event according to local evaluation and 55% according to BICR. The lower event rate for BICR reflects the fact that any BICR progressions occurring after local evaluation progression are unlikely to be captured. The acceptance threshold, *AT*, is calculated from (3) using $\alpha = 0.1$ and values of I_F from (5) and I_S from (6) with *k=1* and *r=0.55/0.6=0.92*. The specificity, the probability of accepting local evaluations based on the sample if $\widehat{HRR}_F = 1$, is found from (4).

Results for different scenarios are shown in Figures 2 to 5. By construction, with $\alpha = 0.1$ the sensitivity, defined as the probability of accepting H_0 when $\widehat{HRR}_F = HRR_U$, is 90% in all cases. For a given total sample size N_{study}, the acceptance threshold and the specificity change with the correlation ρ between local evaluation and BICR (Figure 2), the size n_{sample} of the audit sample (Figure 3), and the value HRR_U of $\widehat{HRR}_F$ used to define H_0 (Figure 4). We see that the acceptance threshold and specificity increase with each of ρ, n_{sample} and HRR_U. Figure 5 demonstrates how the acceptance threshold and specificity vary with total sample size N_{study} when n_{sample} is fixed at a value of 200.

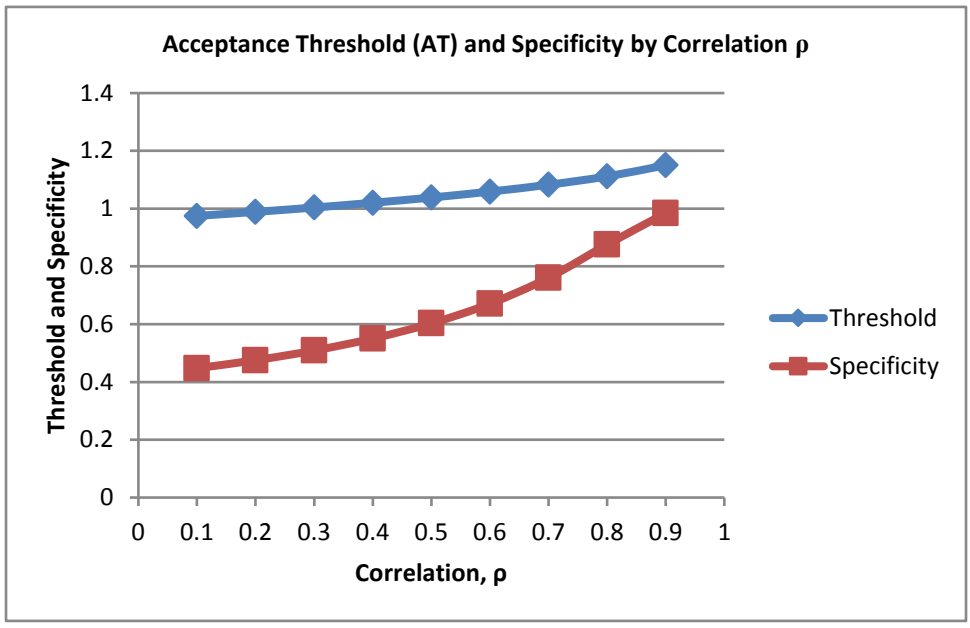
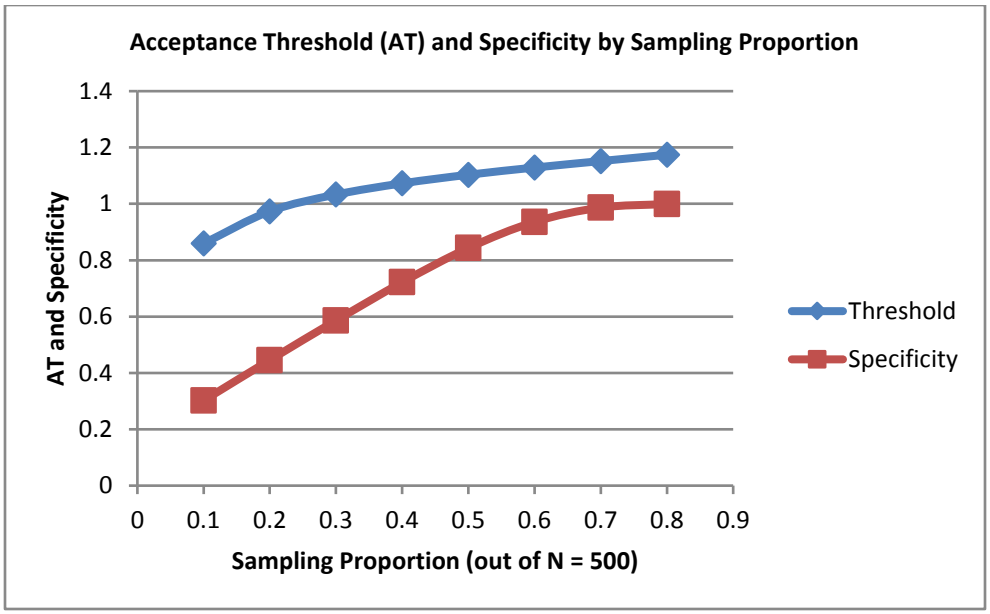
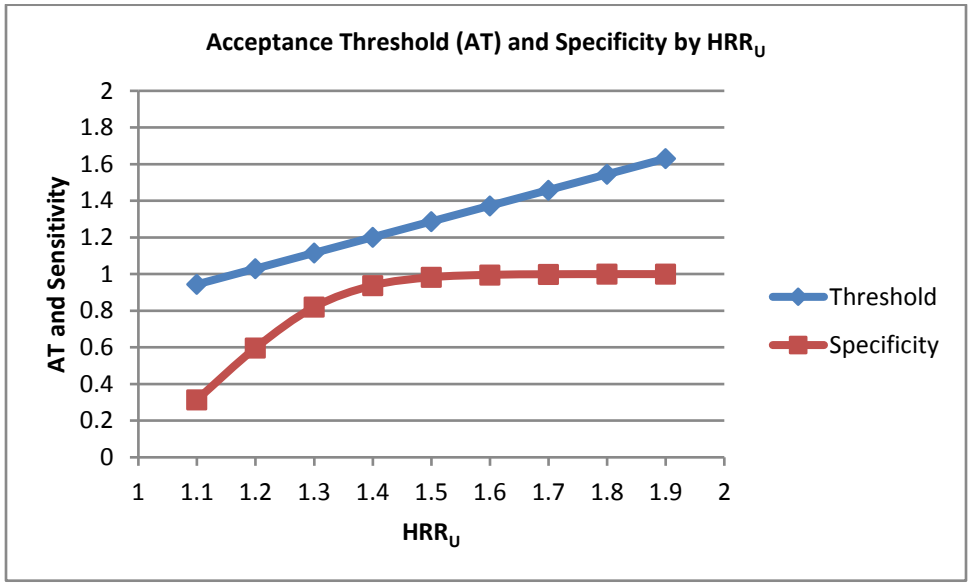


Figure 2 Acceptance threshold, *AT*, and specificity by correlation, ρ (N_{study}=500, n_{sample} = 200, proportion of patients with events = 0.6 for LE and 0.55 for BICR, testing H_0: $\widehat{HRR}_F \geq HRR_U = 1.25$)



237

238 **Figure 3 Acceptance threshold, *AT*, and specificity by sampling proportion, n_{sample} (ρ = 0.7, N_{study}=500, proportion of patients**
239 **with events = 0.6 for LE and 0.55 for BICR, testing H_0: $\widehat{HRR}_F \geq HRR_U = 1.25$)**



240

241 **Figure 4 Acceptance threshold, *AT*, and specificity by the value of HRR_U used to specify H_0 (ρ = 0.7, N_{study}=500, n_{sample}=200,**
242 **proportion of patients with events = 0.6 for LE and 0.55 for BICR)**

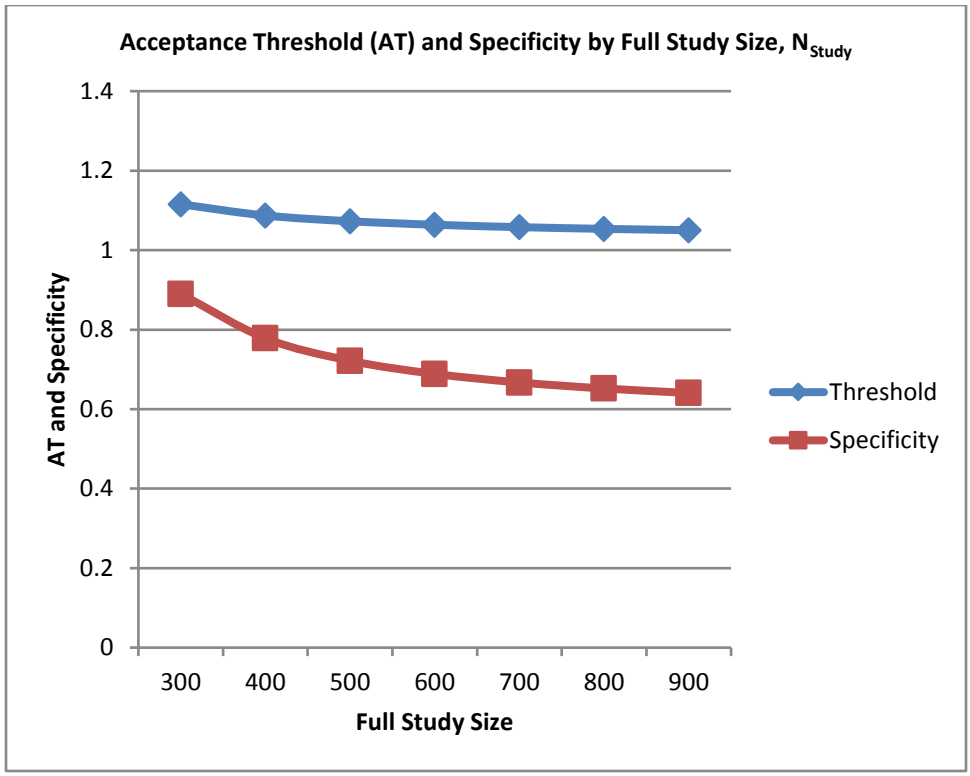


Figure 5 Acceptance threshold, *AT*, and specificity by full study population size N_{study} (ρ = 0.7, n_{sample}=200, proportion of patients with events = 0.6 for LE and 0.55 for BICR, testing H_0: $\widehat{HRR}_F \geq = HRR_U = 1.25$)

With a sample of 200 patients from a total of N_{study}=500, testing H_0: $\widehat{HRR}_F \geq HRR_U = 1.25$, Figure 2 shows that under an assumed correlation of ρ = 0.7, the acceptable sample threshold is *AT*=1.08 and the specificity is 0.76. As the correlation increases the specificity increases sharply, while the impact on the threshold is smaller, with *AT* rising from 0.97 for ρ = 0.1 to 1.15 for ρ = 0.9. Figure 3 shows that specificity increases with the size of the sample, n_{sample}, for example with n_{sample}=300, we have *AT*=1.14 and the specificity is 0.96. However, specificity decreases steeply as the sample size is reduced below 200, for example, specificity is only 0.47 for n_{sample}=100. Figure 4 shows that the acceptance threshold and specificity increase with HRR_U, with specificity close to 1 by the time HRR_U reaches 1.4. Analyses of previous trials at AstraZeneca have indicated a fairly stable estimate of correlation between local evaluation and BICR around 0.7 . In planning to apply the methodology described in this paper, the sample size can be calculated for an estimated value of the correlation ρ. While it is possible, in principle, to adjust the sample size in the light of observed data and an updated estimate of ρ, requesting additional central reviews could cause delays, making this approach impractical. A simpler option is to aim to err in the direction of under-estimating ρ then, as seen in Figure 2, if the true value of ρ is higher than this estimate, specificity will be higher than the design value.

Operating Characteristics by Simulated Retrospective Application to a Phase III Trial in First Line Metastatic Colorectal Cancer

In this section, we demonstrate that when the proposed method is applied in practice, the observed sensitivity and specificity align closely with the theory presented in the previous sections. This is achieved by repeated simulation of sample BICR results for a large clinical trial dataset.

Study Background

The proposed sample audit BICR approach was simulated by repeatedly applying it to data from a large randomised double blind study in first-line metastatic colorectal cancer (mCRC) with 1:1 randomisation in 1422 patients [14]. In this study, the duration of progression free survival for all patients was derived according to a local investigator evaluation (LE) and according to a supportive blinded independent central review (BICR). The primary results from analyses of the LE and BICR data are summarised in Table 1 and Table 2 below.

Table 1 Local Evaluation of PFS in a study of mCRC

Randomised Treatment Arm	Number of Patients (Number of Progression Events)	Hazard Ratio	95% Confidence Interval For Hazard Ratio
Active Treatment	709 (471)	1.103	(0.97,1.25)
Control Treatment	713 (453)		

Table 2 BICR Evaluation of PFS in a study of mCRC

Randomised Treatment Arm	Number of Patients (Number of Progression Events)	Hazard Ratio	95% Confidence Interval for Hazard Ratio
Active Treatment	709 (377)	1.041	(0.90, 1.20)
Control Treatment	713 (377)		

Demonstrating Sensitivity for a given Null Distribution

In the full study data reported above, the value of $\widehat{HRR}_F$ is $\widehat{HR}_{BICR,F}/\widehat{HR}_{L,F} = 0.944$. With both $\widehat{HR}_{BICR,F}$ and $\widehat{HR}_{L,F}$ above 1, there is no evidence of a beneficial treatment effect. Since $\widehat{HRR}_F < 1$, there is no indication of bias in the LE in favour of the active therapy. In order to use this example to demonstrate the theoretical properties introduced in the Methods section, we suppose that HRR_U is set to be 0.944 – even though a value greater than 1 would usually be specified. We, therefore, wish to test the null hypothesis

$$H_0: \widehat{HRR}_F \geq 0.944.$$

This H_0 is true for the given data set, so H_0 should be accepted and a full sample audit initiated based on the BICR sample with probability $1 - \alpha = 0.9$. Our objective is to demonstrate that the conclusion that a full sample audit should be conducted arises with this probability in simulations of the proposed method.

In each of 10000 simulations, we created a BICR sample dataset in the manner described in Figure 1. We first used a 30% sampling rate, so each sample contained 30% of patients with events and 30% of patients with censored events within each treatment arm.

For 1000 of the simulated datasets, we created 100 bootstrap samples and used these to estimate the correlation ρ between $\ln(\widehat{HR}_{LE})$ and $\ln(\widehat{HR}_{BICR})$. The median value obtained in the 1000 datasets was $\rho = 0.66$ and we have taken this as our overall estimate of ρ. Table 1 shows a total of $n_{L,F} = 924$ LE events and Table 2 a total of $n_{B,F} = 754$ BICR events, so

$$r = \frac{n_{B,F}}{n_{L,F}} = \frac{754}{924} = 0.816.$$

We combined this value of r with $k = 1$ and $\rho = 0.66$ to obtain

$$I_F = [Var(\ln(\widehat{HRR}_F))]^{-1} = \frac{k\, n_{L,F}}{(k+1)^2} \frac{r}{(1+r-2\rho\sqrt{r})} = 302.271$$

and $I_S = 0.3 \times I_F = 90.681$.

We then carried out the following steps for each of the 10000 simulated BICR sample datasets.

1) Calculate $\widehat{HR}_{L,S}$, $\widehat{HR}_{BICR,S}$ and, hence, $\widehat{HRR}_S$ for the BICR sample.

2) Calculate $Z_S = \ln(\widehat{HRR}_S)\sqrt{I_S}$ and, following (2), reject $H_0: \widehat{HRR}_F \geq 0.944$ if

$$Z_S < \ln(0.944)\sqrt{I_S} - \Phi^{-1}(0.9)\sqrt{\frac{I_F - I_S}{I_F}} = -1.624 = Acceptance\ Threshold$$

Out of 10000 simulated BICR samples, 8985 (89.9%) led to acceptance of H_0, in close agreement with the theoretical sensitivity of 90%. The above exercise was repeated using sampling rates of 20%, 40% and 50%, with 10000 replicates in each case. Again, correlation was assumed to be 0.66 so I_F remained the same but I_S varied with the value of n_{sample}. Table 3Table 3 shows the acceptance threshold for $\widehat{HRR}_S$, i.e., *AT* from equation (3), and the percentage of cases out of 10000 simulations in which H_0 was accepted. All these estimates of sensitivity are close to 90%.

Table 3 Acceptance Thresholds for $\widehat{HRR}_S$ and estimated sensitivity for tests of H_o: $\widehat{HRR}_F \geq 0.944$

% Sampling (Number of patients)	Acceptance Threshold, AT, for sensitivity 0.9 under H_0	Estimated sensitivity from 10000 simulations
20% (286)	0.814	89.8%
30% (428)	0.843	89.8%
40% (570)	0.862	90.2%
50% (712)	0.877	89.6%

The above results concern a single point in the distribution of Z_S. We can go further and compare the full distribution of the simulated values of Z_S against the theoretical density of Z_S given $\widehat{HRR}_F = 0.944$ or, equivalently, $Z_F = \ln(\widehat{HRR}_F)\sqrt{I_F} = \ln(0.944) \times \sqrt{302.271} = -1.002$. This conditional distribution of Z_S is given by (1) with $\tilde{Z}_F = -1.002$, $I_F = 302.271$ and, for 30% sampling, $I_S = 0.3 \times I_F = 90.681$. Figure 6 shows a smoothed kernel density estimate based on the simulated values of Z_S for the case of 30% sampling plotted with the conditional density of $Z_S \mid Z_F = -1.002$ given by equation (1). The critical value $Z_S = -1.624$, below which H_0 is rejected, is indicated in the figure. Figure 6 also compares results in terms of $\widehat{HRR}_S$, showing the smoothed kernel density estimate based on simulated values of $\widehat{HRR}_S$ and the theoretical conditional density of $\widehat{HRR}_S$ given $\widehat{HRR}_F = 0.944$. In this case the critical value for $\widehat{HRR}_S$, below which H_0 is rejected, is $\exp\left(\frac{-1.624}{\sqrt{90.681}}\right) = 0.843$, and this is also the value AT obtained from (3). The results in Figure 6 demonstrate excellent agreement between the distribution of the simulated data and the theoretical null distribution.

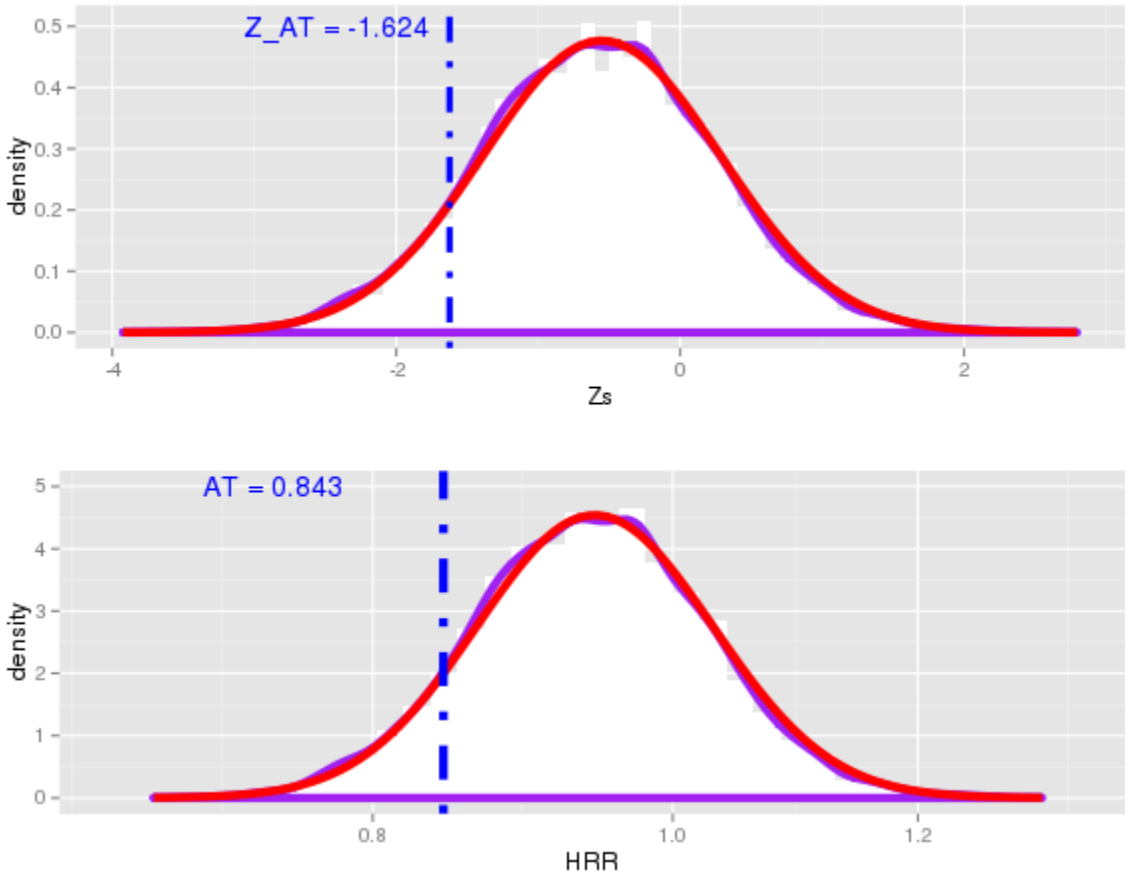


Figure 6 Observed density of Z_S and $\widehat{HRR}_S$ (purple) versus density given $\widehat{HRR}_F = 0.944$ (red)

Demonstrating Specificity for a given Null Distribution

We now use the same example to confirm that the theoretically derived value for specificity is observed in practice. To this end, suppose it is desired to test H_0: $\widehat{HRR}_F \geq HRR_U = 1.25$ with $\alpha = 0.1$. For the data set we are considering, $\widehat{HRR}_F = 0.944$, and $I_F = 302.271$. With k% sampling, $I_S = \left(\frac{k}{100}\right) \times I_F$ and equation (4) gives the specificity under $\widehat{HRR}_F = 0.944$ as

$$\Phi\left[(\ln(1.25) - \ln(0.944))\sqrt{\frac{I_S I_F}{(I_F - I_S)}} - \Phi^{-1}(0.9)\right] = \Phi\left[\ln(1.324)\sqrt{\frac{k \times 302.271}{(100-k)}} - \Phi^{-1}(0.9)\right] \quad (7)$$

Table 4 compares the estimated specificity, based on 10000 simulated BICR samples, with the values given by (7) for 20%, 30%, 40% and 50% sampling. We see that in each case the estimated specificity is a little higher than the theoretical value. While the differences are greater than might be explained by the sampling error in 10000 replications, they are still small and do not give any serious cause for concern.

Table 4 Acceptance Threshold for testing H_0: HRR ≥ 1.25, Estimated Specificity based on 10000 Simulations and Theoretical Specificity from Equation (7)

% Sampling (Number of patients)	Acceptance Threshold *AT* for testing H_o: $\widehat{HRR}_F \geq 1.25$	Estimated specificity from simulations	Theoretical Specificity
20% (286)	1.079	88.9%	87.7%
30% (428)	1.117	97.8%	97.2%
40% (570)	1.142	99.76%	99.66%
50% (712)	1.161	100%	99.98%

"Real life" Properties of the Method

In the preceding calculations and simulations regarding sensitivity and specificity, we have used the full study information to define $r = n_{B,F}/n_{L,F}$, to compute I_F, and to find a bootstrap estimate of ρ. In practice, this complete information would not be known at the time of carrying out a sample BICR. Instead, we would use the information in the sample, directly calculating I_S from the estimated variances of the log hazard ratios for the sample BICR and sample LE data returned by standard software packages and a separate bootstrap estimate of ρ from each sample. I_F could then be calculated as $\gamma^{-1}I_S$, where γ is the sampling fraction used.

In order to assess the proposed procedure as it would be used in practice, we analysed the same sets of simulated samples from the previous section in this way. The percentages of simulations in which H_0: $\widehat{HRR}_F \geq 0.944$ was not rejected are given in Table 5.

Table 5 Estimated sensitivity for tests of H_o: $\widehat{HRR}_F \geq 0.944$ when I_S, I_F and ρ are estimated from information in the sample data only

% Sampling (Number of patients)	Estimated Sensitivity from 10000 simulations
20% (286)	92.23%
30% (428)	92.15%
40% (570)	92.23%
50% (712)	91.72%

The observed sensitivities are close to the intended value of 90% and perhaps slightly conservative, i.e., with an error rate under H_0 below $\alpha = 0.1$.

We have repeated the calculations of sensitivity and specificity in simulated sample data sets from a smaller study of 196 gliobastoma patients who were randomised in a ratio of 2 to 1 between

experimental and control treatments. We found similarly agreement between theoretical and empirical properties of the proposed procedure, including the "real-life" case where values for r, I_F and ρ based on the sample data sets themselves. (See Appendices.)

Discussion: Practical Considerations and Potential Applications

Methods

We have presented a method whereby a sample of centrally reviewed cases can be used to decide if a full review of local assessments of progression free survival is needed. This method is simple to apply and effective in reducing the volume of BICR when the audit of a sample of patients supports use of the hazard ratio from local evaluation in determining the study conclusion. The method's theoretical statistical properties have been confirmed in examples of historical data from Phase III trials of metastatic colorectal cancer and gliobastoma.

In the proposed method, we define a null hypothesis under which the level of bias in local evaluations is unacceptable. If the audit sample leads to rejection of this null hypothesis, we conclude that local evaluations are sufficiently close to independent reviews (or biased against the experimental treatment) and a full BICR is unnecessary. The approach is in keeping with the idea that a full study BICR is appropriate unless there is evidence to demonstrate this is not necessary.

If the audit sample triggers a full study BICR and the hazard ratio ratio observed in the full-study data indicates a difference between the LE and BICR estimates of hazard ratio, then both these estimates may be subject to bias. The LE sample may indicate progression that BICR does not confirm, while limited availability of post-progression scans causes informative censoring for the BICR estimate. Methods have been proposed for such a BICR situation, for example to include an event at the visit subsequent to the LE progression [16]. Another possibility in this situation could be a multiple imputation approach [17].

We have presented a situation with a single value of maximum acceptable HRR (1.25) to illustrate the proposed method. In practical application, we propose that a graded approach be taken, such that the limit varies depending on the observed LE HR. It would seem logical to have greater tolerance for possible bias (higher HRR, >1) in the presence of a strong treatment effect according to the LE HR, and smaller tolerance (lower HRR, closer to 1) in the case of a weaker LE treatment effect. A possible graded approach to satisfy this requirement would be to set the HRR threshold such that it preserves the majority of the observed LE HR. Table 6 for example, illustrates the HRR which would result from preserving 2/3^{rds} of the observed full study LE HR for a range of LE hazard ratios. Using this approach it is suggested the sample is designed to have sufficient specificity against the HRR that corresponds to the minimally clinically important LE HR.

Table 6 Graded Approach to Choice of HRR Threshold

Full Study Local Evaluation Hazard Ratio	HRR Threshold to preserve 2/3rd of LE HR
0.3	1.78
0.5	1.33
0.7	1.14
0.9	1.04

399

400 We have proposed using an alpha of 10% instead of the typical 2.5%. Given the prior data consistently
401 demonstrating the concordance in treatment effects estimated by the BICR and LE we feel this is
402 appropriate in most situations.

403 Application of the proposed method requires an initial estimate of the correlation between LE and BICR
404 hazard ratio estimates. In principle, the correlation observed in the first part of an audit sample could be
405 used to re-calculate the necessary sample size. However, for simplicity of application, it may well be
406 preferable to adopt a conservative approach and assume a low value for the correlation, since this will
407 lead to a specificity above the target value as long as the estimated correlation is below the true value.

408 For studies with long durations, or known operational changes during conduct, a stratified approach
409 could be followed (e.g., early/late, before/after) where correlations, and HRR estimates, are allowed to
410 vary between levels of the stratification factor. The BICR sample would select proportionately from
411 each level, so that the overall HRR estimate is representative of the whole study. Alternatively, if there
412 was concern about the potential for bias in certain subgroups of patients, such as those with non-
413 measurable disease, these patients could be enriched in the sample. In this case, the HRR in each
414 subgroup would need re-weighting to provide an estimate of the HRR in the overall population.

415 **Practical Implementation**

416 The practical considerations for performing a BICR can be challenging. To benefit most from the
417 proposed approach, the study should be sufficiently large that an audit sample size can be chosen which
418 is big enough to determine whether a full BICR is necessary, yet small enough that carrying out BICR only
419 for this audit sample represents a worthwhile saving. Our experience indicates that $1-$1.25m could be
420 saved if 50% of a trial were sampled, with the costs for the process being equally split between
421 collecting and reading the scans. Plans should be in place to collect and store scans from all patients,
422 with which there is an associated cost.

423 There are two potential options for implementation. The sample BICR may be initiated prior to database
424 lock to allow a rapid decision on whether full study review is required, so that such a review can be
425 conducted without a major impact on reporting timelines. However, the review process cannot be
426 started too early since it must not have any impact on the study conduct. The maximum acceptable HRR
427 (graded by observed HR(LE)), the sample selection process, and a mechanism for collecting the

appropriate scans promptly should all be prepared at the start of the trial. One possibility is that an Independent Data Monitoring Committee, or other independent party, could be supplied, close to database lock, with the random scheme and PFS data for the LE and BICR in the sample. They could then indicate to the sponsor whether the sample had been accepted without revealing either treatment effect. Alternatively, the decision to initiate the sample BICR could be taken after database lock and the primary LE analysis results are available to the sponsor. Clearly, if there is no significant treatment effect according to the LE, the sample BICR would not be required.

If a sample were pre-specified at the trial design stage, a BICR could feasibly be conducted in real time. Real-time BICR results can be used for improving data quality, and ensuring independent verification prior to treatment crossover on progression if permitted, during a trial. However, our method identifies the need for a full BICR based on the observed treatment effect, and not, for example, on observed quality of data collected. Assessing sample quality based on emerging treatment effect data is beyond the scope of this paper.

Summary

The possibility of requiring expert review of outcomes arises across a range of therapy area indications. Cardiovascular outcome trials often have evidence of stroke or myocardial infarction centrally reviewed by an independent cardiologist. X-rays used to assess scale of bone deterioration in rheumatoid arthritis patients are also frequently independently reviewed. The proposed methodology has the potential for more general use. Indeed, its application is likely to be more straightforward when the primary outcome is measured at a single time point and issues of repeated assessment and possible informative censoring issue do not arise.

In summary we propose a sampling method that is simple to implement and reliable that would enable conclusions about bias to be assessed at reduced cost.

References

[1] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (version 1.1). *European Journal of Cancer* 2009; **45**: 228-247. DOI:10.1016/j.ejca.2008.10.026.

[2] Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, Munden RF. Interobserver and Intraobserver Variability in Measurement of Non-Small-Cell Carcinoma Lung Lesions: Implications for Assessment of Tumor Response. *Journal of Clinical Oncology* 2003;**21**:2574-2582.

[3] FDA: FDA Briefing Document Oncology Drug Advisory Committee Meeting December 5, 2007: BLA STN 125085/91.018 Avastin (bevacizumab), http://www.fda.gov/ohrms/dockets/ac/07/briefing/2007-4332b1-01-FDA.pdf *Accessed 2nd August 2014.*

[4] Dodd LE, Korn EL, Freidlin B, Jaffe CC, Rubinstein LV, Dancey J, Mooney MM . Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense? *Journal of Clinical Oncology* 2008; **26**:3791-3796.

[5] Amit O, Mannino F, Stone AM, Bushnell W, Denne J, Helterbrand J, Burger HU. Blinded Independent CEntral Review of Progression in Cancer Clinical Trials: Results from a meta analysis. *European Journal of Cancer* 2011;**47**:1772-1778.

[6] Zhang JJ, Chen H, He K, Tang S, Justice R, Keegan P, Pazdur R, Sridhara R. Evaluation of Blinded Independent Central Review of Tumor Progression in Oncology Clinical Trials: A Meta-analysis. *Drug Information Journal; Therapeutic Innovation & Regulatory Science* 2013;**47**:167-174.

[7] FDA. *Oncologic Drug Advisory Committee - Evaluation of Radiologic Review of PFS in Non-Hematologic Malignancies. July 24, 2012 Meeting.* http://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/drugs/oncologicdrugsadvisorycommittee/ucm311141.pdf

Accessed 2nd August 2014.

[8] Dodd L, Korn EL, Freidlin B, Gray R, Bhattacharya S. An Audit Strategy for Progression-Free Survival. *Biometrics* 2011;**67**:1092-1099.

[9] Mannino FV, Amit O, Lahiri S. Evaluation of Discordance Measures in Oncology Studies with Blinded Independent Central Review of Progression-Free Survival Using an Observational Error Model. *Biopharmaceutical Statistics* 2013;**23**:971-985.

[10] Amit O. A Sample Based Approach for Independent Review of PFS using Differential Discordance http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM315082.pdf

Accessed 19th January 2015.

[11] Zhang JJ, Zhang L, Chen H, Murgo AJ, Dodd LE, Pazdur R, Sridhara R. Assessment of Audit Methodologies for Bias Evaluation of Tumor Progression in Oncology Clinical Trials. *Clinical Cancer Research* 2013;**19**:2637-2645.

[12] Jennison C, Turnbull B. "10.4 A Parameter Free Approach." In *Group Sequential Methods with Applications to Clinical Trials*, by Turnbull B Jennison C, 213-214. Chapman & Hall/CRC, 2000.

[13] Jennison C, Turnbull B. Repeated Confidence Intervals for Group Sequential Clinical Trials, *Controlled Clinical Trials,* 1984;**5:**33-45

[14] Stone A, Bushnell W, Denne J, Sargent DJ, Amit O, Chen C, Bailey-Iacona R, Helterbrand J, Williams G. Research outcomes and recommendations for the assessment of progression in cancer clinical trials from a PhRMA working group' European Journal of Cancer 2011;47:1763-1771

[15] Schmol HJ, Cunningham D, Sobrero A, Karapetis CS, Rougier P, Koski SL, Kocakova I, Bondarenko I, Bodoky G, Mainwaring P, Salazar S, Barker P, Mookerjee B, Robertson J, Van Cutsem E.Cediranib With mFOLFOX6 Versus Bevacizumab With mFOLFOX6 As First-Line Treatment for Patients With Advanced Colorectal Cancer: A Double-Blind, Randomized Phase III Study (HORIZON III). *J Clin Oncol* 2012;**30**:3588-3595.

[16] Fleischer F, Gaschler-Markefski B, Bluhmki E. How is retrospective independent review influenced by investigator-introduced informative censoring: a quantitative approach. *Stat Med.* 2011;**30**:3373-3386.

[17] Hsu CH, Taylor JM, Murray S, Commenges D. Multiple imputation for interval censored data with auxiliary variables. *Stat Med.* 2007;**26**:769-781.

[18] Batchelor TT, Mulholland P, Neyns B, Nabors LB, Campone M, Wick A, Mason W, Mikkelsen T, Phuphanich S, Ashby LS, Degroot J, Gattamaneni R, Cher L, Rosenthal M, Payer F, Jürgensmeier JM, Jain RK, Sorensen AG, Xu J, Liu Q, van den Bent M. Phase III randomized trial comparing the efficacy of cediranib as monotherapy, and in combination with lomustine, versus lomustine alone in patients with recurrent glioblastoma. *J Clin Oncol* 2013;**31**:3212-3218 DOI: 10.1200/JCO.2012.47.2464.

Appendices

Consistency in Bootstrap Correlation Estimate with Sample Size

For each of 10000 samples generated for a given sample size (20% - 50% of the whole study [14]), 1000 bootstrap samples were generated in order to estimate the correlation between $\ln(HR_{LE})$ and $\ln(HR_{BICR})$. The distribution of these correlation estimates is summarised in Table 7. This suggests that the statistical properties of the correlation estimates do not vary much with the % sampling of the whole study.

Table 7 Summary of Bootstrap Correlation Estimates by size of BICR sample

Sample BICR as % of total study		Summary of Correlation between HR_{LE} and HR_{BICR}				
	Min	1st Quartile	Median	Mean	3rd Quartile	Max
20%	0.356	0.551	0.586	0.584	0.619	0.743
30%	0.409	0.559	0.588	0.587	0.617	0.722

40%	0.429	0.564	0.590	0.589	0.613	0.709
50%	0.473	0.569	0.591	0.590	0.613	0.704

Sensitivity of the method in a trial in Gliobastoma [17]

Table 8 Local Evaluation of PFS in a study of glioblastoma

Randomised Treatment Arm	Number of Patients (Number of Progression Events)	Hazard Ratio	95% Confidence Interval For Hazard Ratio
Active Treatment	131 (107)	0.837	(0.59,1.18)
Control Treatment	65 (47)		

Table 9 BICR Evaluation of PFS in a study of glioblastoma

Randomised Treatment Arm	Number of Patients (Number of Progression Events)	Hazard Ratio	95% Confidence Interval for Hazard Ratio
Active Treatment	131 (109)	1.015	(0.71, 1.48)
Control Treatment	65 (44)		

Study HRR = 1.212

Theoretical sensitivity

10000 simulations using 50% sampling were run (to ensure a reasonable number of progression events within BICR sample time to event analysis). N_B/N_{LE} = 153/154 was assumed as fixed for n_B/n_{LE} and ρ was set to 0.67 (the mean and median correlation observed using the bootstrap approach in 1000 earlier BICR simulations). Figure 7 shows the close concordance between the distribution of the simulated sample BICRs and the expected distribution. Approximate 90% sensitivity is demonstrated in Table 9.

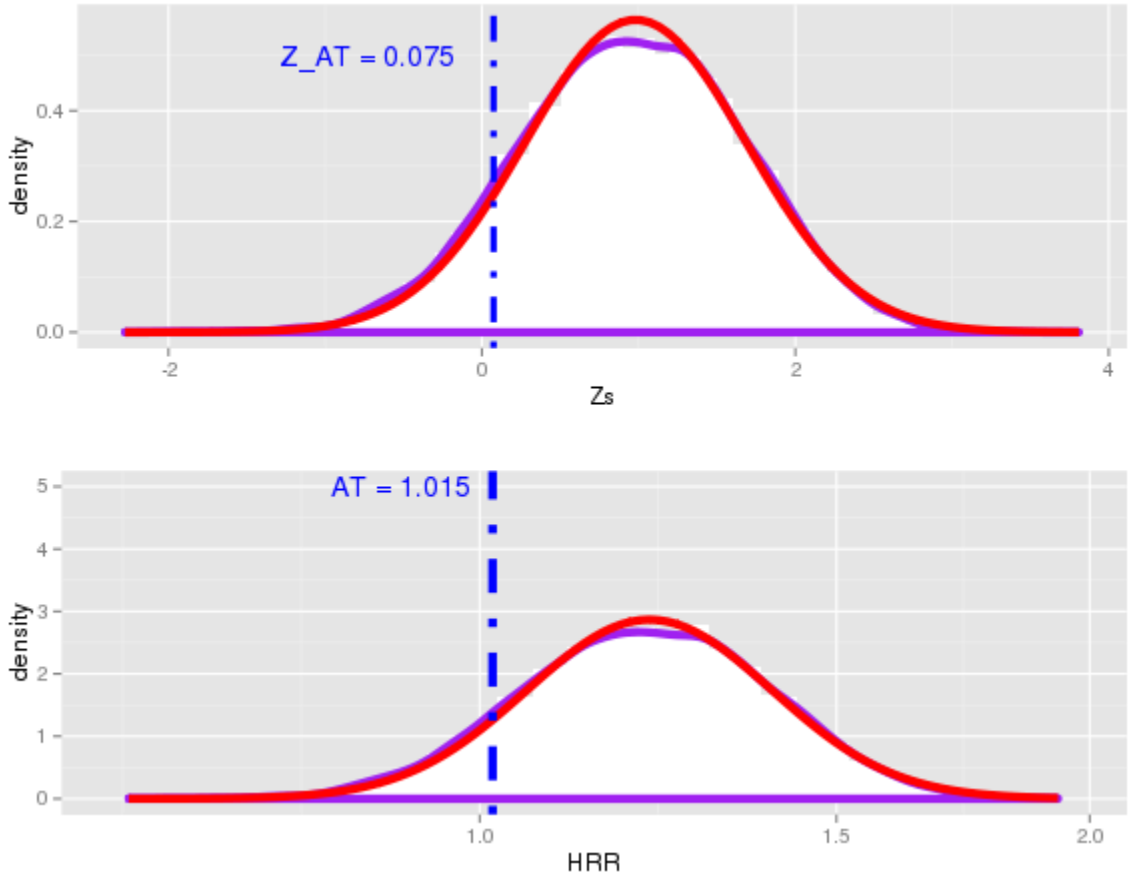


Figure 7 Observed density of Z_S and $\widehat{HRR}_S$ (purple) versus density given $\widehat{HRR}_F = 1.212$ (red) for the gliobastoma trial (50% sampling of 196 patients)

Table 10 Acceptance Thresholds for $\widehat{HRR}_S$ and estimated sensitivity for tests of H_o: $\widehat{HRR}_F \geq 1.212$

% Sampling (Number of patients)	Acceptance Threshold, AT, for sensitivity 0.9 under H_0	Estimated sensitivity from 10000 simulations
50% (99)	1.015	88.8%