Estimation of Biomarker Responses in the Presence of Missing Biomarker Status

Shengyan Hong

MedImmune

One MedImmune way, Gaithersburg, MD 20878

Abstract

One of the objectives in early phase clinical trials studying a targeted therapy is to accurately estimate treatment response by the status of the biomarker of interest to determine whether the biomarker may be predictive or not. This is critical to inform further development strategy, such as whether the therapy should be developed in an enriched population. However, often the collection of baseline biomarker tissues for determining biomarker status is not mandatory in early phase clinical trials and/or some collected baseline biomarker tissue is not usable to determine biomarker status due to operational handling issues or assay sensitivity. A naïve method in presence of missing biomarker status is to estimate the response rate based only on patients whose biomarker status is known (observed data analysis), which is an unbiased estimate only if the missing biomarker status is missing completely at random (MCAR). In this paper, we improve this approach by utilizing observed response data for patients with unknown biomarker status by applying the EM algorithm under the assumption of missing at random (MAR). Simulations are conducted to compare these two methods and a case study using the methodology is provided.

Key words and phrases: Biomarker response, missing data, EM algorithm, estimation

1. Introduction

In recent years, clinical drug development has increasingly focused on development of so called targeted therapies that are devised to target specific gene or protein (biomarker) that is considered to cause or promote disease. As such, patients whose disease doesn't express or lowly expresses a specific disease-related biomarker would be expected to respond less (or not at all) to therapy that targets the biomarker than patients whose disease (highly) expresses the biomarker. Therefore, it is of great clinical interest to design clinical trials that can detect the treatment effect of a targeted therapy and identify a patient population who benefit most from such therapy more efficiently. There have been numerous publications about clinical trial designs and associated issues for targeted therapies (Fredlin and Simon [1]; Fredlin *et al*. [2]; Gosho *et al*. [3]; Jiang *et al*. [4]; Mandrekar and Sargent [5]; Simon [6]; Simon and Maitournam [7, 8]; Wang *et al.* [9]; etc). FDA recently published a guidance about enrichment strategies to be used in clinical trials for targeted therapy (FDA [10]). To this end, biomarker evaluation should start early in clinical drug development with aim to identify optimal biomarker expression level that can best separate responders from non-responders, and to accurately estimate clinical responses by biomarker status (positive vs negative). These are usually done in early phase all-comers clinical trials with biomarker tissues collected to enable retrospective analysis of biomarker responses.

The literature contains statistical methods that can be used for identifying optimal biomarker threshold to best separate responders from non-responders. One early publication in this area is Miller and Siegmund [11] in which the maximal chi-square method to identify optimal cut-off point from a continuous biomarker was proposed and its asymptotic property was studied. Halpren [12] studied small sample behaviors of this method. Boulesteix [13] studied the asymptotic property of the same method for ordinal biomarker variables. Once the biomarker status is identified, estimation of clinical responses by biomarker status is straightforward if all patients' biomarker status is known. However, in early stage drug development, especially in phase I clinical trials, the biomarker evaluation is often a secondary endpoint and as such, the collection of baseline biomarker tissue to determine biomarker status is often not mandatory, resulting in biomarker tissue availability in only a subset of patients. Furthermore, some of collected biomarker tissue may not be useable to determine biomarker status due to operational handling issues or assay sensitivity. Consequently, it is not uncommon that biomarker status may be unknown for quite a few patients.

A naïve method in the presence of missing biomarker status is to estimate the response rate based only on patients whose biomarker status is known (so called observed data analysis). As is well known, it is an unbiased estimate only if the missing biomarker status is missing completely at random (MCAR). Even if the MCAR assumption is plausible, the variability of the estimate based only on observed data is usually larger due to reduced information. In this paper, we propose to improve this approach by utilizing observed response data for patients with unknown biomarker status by applying the EM algorithm under the assumption of missing at random (MAR). In Section 2, we describe biomarker response estimates in the presence of missing biomarker status based on aforementioned naïve method and the proposed EM method. Simulations are provided in Section 3 to compare the two methods. A case study applying the proposed method will be presented in Section 4. Section 5 contains some discussions.

2. Biomarker Response Estimation in the Presence of Missing Biomarker Status

Suppose that a total of N patients are enrolled in a clinical trial. The efficacy endpoint of interest is whether patients respond to the treatment under study. Each subject can be categorized to one of two biomarker statuses: biomarker positive (BM+) or biomarker negative (BM-). At the end of study, while all patients' treatment responses are known, some patients' biomarker statuses are unknown, resulting in observed data in the form the following frequency table

{Insert Table 1 here}

The task is to estimate the response rates p_+ for BM+ and p_- for BM-, respectively. Also of interest is to estimate prevalence rate p_0 for BM+.

Without missing data, the parameters p_i can simply be estimated by MLEs

$$\hat{p}_0 = \frac{N_{k+} + N_{u+}}{N}, \qquad \hat{p}_i = \frac{R_{ki} + R_{ui}}{N_{ki} + N_{ui}}, \qquad i = +, -$$

where, as indicated in Table 1, N_{k+} and N_{k-} are number of patients with BM+ and BM-, respectively, in known BM status group, N_{u+} and N_{u-} are the number of patients with BM+ and BM-, respectively, in unknown BM status group, and R_{ji} are the number of responders out of N_{ji} for j=k, u and i=+, -.

In the presence of missing biomarker status, N_{ui} and R_{ui} are unknown, hence the above MLEs are unobtainable. The naïve estimates are MLEs based only on non-missing data

$$\hat{p}_0 = \frac{N_{k+}}{N_{k+} + N_{k-}}, \qquad \hat{p}_i = \frac{R_{ki}}{N_{ki}}, \qquad i = +, -$$

These naïve estimates are unbiased under MCAR, but would be biased under MAR or MNAR (missing not at random). Under the assumption of MAR, that is, the missing biomarker status is dependent only on observed data, then the observed response data for patients with missing biomarker status can be utilized in the estimation of the parameters p_i by applying the following EM algorithm (Dempster et al. [14]). Denote

$$X_{ij} = \begin{cases} 1 & if\ response \\ 0 & otherwise \end{cases}, \qquad Z_{ij} = \begin{cases} 1 & if\ BM + \\ 0 & if\ BM - \end{cases},$$

$$i = 1\ (\text{known BM status}), 2\ (\text{unknown BM status});\ j = 1, \dots, N_i$$

The observed data are $Y = \{X_{1j},\ Z_{1j},\ j = 1, \dots, N_1;\ X_{2j},\ j = 1, \dots, N_2\}$ and the unobserved data are $Z_2 = \{Z_{2j},\ j = 1, \dots, N_2\}$. Then the likelihood function for given the parameter vector $p = \{p_0,\ p_+,\ p_-\}$ is

$$L(Y,\ Z_2 \mid p) = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left[p_0 {p_+}^{X_{ij}} (1 - p_+)^{(1 - X_{ij})}\right]^{Z_{ij}} \left[(1 - p_0) {p_-}^{X_{ij}} (1 - p_-)^{(1 - X_{ij})}\right]^{1 - Z_{ij}}$$

The EM algorithm is an iterative process consisting of two steps at each iteration to estimate the parameter p. Start with an initial guesstimate $p^{(0)}$. A good choice would be above naïve estimate. Let $p^{(t)}$ denote the estimate of p obtained at the t-th iteration, then the estimate will be updated to $p^{(t+1)}$ obtained at the (t+1)-th iteration of the following two steps:
E-Step (Expectation Step):

$$Q\left(p \middle| p^{(t)}\right) = E_{L(Z_2 | Y, p^{(t)})}[logL(Y, Z_2 | p)]$$

M-Step (Maximization Step):

$$p^{(t+1)} = \arg\max_{p} Q\left(p \middle| p^{(t)}\right)$$

The EM iterations stop when the estimates between last two iterations are within a pre-specified margin, such as 10^{-4}, and the estimate from the last iteration is chosen as the estimate for p.

Note that

$$Q\left(p\middle|p^{(t)}\right) =$$

$$\sum_{j=1}^{N_1}\left[\begin{array}{c} Z_{1j}\left(logp_0 + X_{1j}logp_+ + \left(1-X_{1j}\right)\log(1-p_+)\right) \\ +\left(1-Z_{1j}\right)\left(log(1-p_0) + X_{1j}logp_- + \left(1-X_{1j}\right)\log(1-p_-)\right)\end{array}\right]+$$

$$\sum_{j=1}^{N_2}\left[\begin{array}{c} E\left(Z_{2j}|X_{2j},\, p^{(t)}\right)\left(logp_0 + X_{2j}logp_+ + \left(1-X_{2j}\right)\log(1-p_+)\right) \\ +\left(1-E\left(Z_{2j}|X_{2j},\, p^{(t)}\right)\right)\left(log(1-p_0) + X_{2j}logp_- + \left(1-X_{2j}\right)\log(1-p_-)\right)\end{array}\right]$$

The expectation of unobserved variable Z_{2j} conditional on observed data X_{2j} and given $p^{(t)}$ is

$$Z_{2j}{}^{(t)} \triangleq E\left(Z_{2j}|X_{2j},\, p^{(t)}\right) = P\left(Z_{2j} = 1|X_{2j},\, p^{(t)}\right)$$

$$= \frac{P\left(Z_{2j} = 1|\, p^{(t)}\right)P\left(X_{2j}|\, Z_{2j} = 1,\, p^{(t)}\right)}{P\left(X_{2j}|\, p^{(t)}\right)}$$

$$= \frac{p_0{}^{(t)}(p_+{}^{(t)})^{X_{2j}}(1-p_+{}^{(t)})^{1-X_{2j}}}{p_0{}^{(t)}(p_+{}^{(t)})^{X_{2j}}(1-p_+{}^{(t)})^{1-X_{2j}} + (1-p_0{}^{(t)})(p_-{}^{(t)})^{X_{2j}}(1-p_-{}^{(t)})^{1-X_{2j}}}$$

Hence, the $p^{(t+1)}$ of M-Step are obtained as follows by solving $\left.\frac{\partial Q}{\partial p}\right|_{p=p^{(t+1)}} = 0$

$$p_0{}^{(t+1)} = \frac{1}{N}\left(\sum_{j=1}^{N_1} Z_{1j} + \sum_{j=1}^{N_2} Z_{2j}{}^{(t)}\right)$$

$$p_+{}^{(t+1)} = \frac{1}{Np_0{}^{(t+1)}}\left(\sum_{j=1}^{N_1} X_{1j}Z_{1j} + \sum_{j=1}^{N_2} X_{2j}Z_{2j}{}^{(t)}\right)$$

$$p_-{}^{(t+1)} = \frac{1}{N(1-p_0{}^{(t+1)})}\left(\sum_{j=1}^{N_1} X_{1j}(1-Z_{1j}) + \sum_{j=1}^{N_2} X_{2j}(1-Z_{2j}{}^{(t)})\right)$$

The variance of $p^{(t+1)}$ are approximately as follows obtained by solving the inverse of $-\frac{\partial^2 Q}{\partial p \partial p'}$

$$Var\left(p_0{}^{(t+1)}\right) \cong \frac{p_0{}^{(t+1)}\left(1-p_0{}^{(t+1)}\right)}{N},$$

$$Var\left(p_+{}^{(t+1)}\right) \cong \frac{p_+{}^{(t+1)}\left(1-p_+{}^{(t+1)}\right)}{Np_0{}^{(t+1)}},$$

$$Var\left(p_-{}^{(t+1)}\right) \cong \frac{p_-{}^{(t+1)}\left(1-p_-{}^{(t+1)}\right)}{N(1-p_0{}^{(t+1)})}$$

3. Simulations

In this section we conduct simulations to study the performance under MCAR and MAR of the estimates by the naïve method and the proposed EM method. The sample size N is set at 40, 60 and 80, respectively, typical size of early phase trial for preliminary efficacy assessment. The true prevalence rate p_0 for BM+ is set at 30%, the true response rate p_- for BM- is set at 15%, and the true response rate p_+ for BM+ is set at 40% and 20%, representing large and small difference in response rate between BM+ and BM-, respectively. The response data and missing BM status process are as follows:

For MCAR case, each subject's BM status is drawn from $B(1, p_0)$ and whether or not it's known is drawn from Bernoulli $B(1, \pi)$ with π=50%. Meanwhile, the subject's response status is drawn from $B(1, p_+)$ for BM+ and from $B(1, p_-)$ for BM-.

For MAR case, draw a latent biomarker variable $X \sim U(0,1)$ for each subject, where BM+ is defined as $X>0.7$ so that the true prevalence rate for BM+ is $p_0=P(X>0.7)=30\%$. Each subject's response R is drawn as $R\sim B(1, p_+)$ for BM+ and $R\sim B(1, p_-)$ for BM-, and the subject's BM status K , where K=1 indicates that the BM status is known, is drawn according to the observed response R as follows: $K\sim B(1, p_{\text{BM}}(R))$ where $p_{BM}(R) = \exp(0.5 + R)/(1 + \exp(0.5 + R))$.

The simulation is run 100000 times for each case. The margin 10^{-4} is used to stop EM iteration process. The results are presented in Tables 2 and 3.

{Insert Table 2 here}

{Insert Table 3 here}

For MCAR case, while both naïve and EM estimates are unbiased and very close to the MLEs based on all data (pretending all BM statuses are known), the EM method has smaller SD, and hence better precision, than the naïve method. It is expected as the naïve method only relies on observed data and hence essentially throws away response information on patients with unknown biomarker status. It is a defacto reduction in sample size. The EM method utilizes this information and hence results in less variability.

For MAR case, the naïve estimates for p_+ and p_- are clearly biased, while the EM estimates are unbiased and very close to the MLEs based on all data (pretending all BM status are known), and the SD is also smaller. Again, it is expected as the naïve method essentially assumes that the subset of patients with known BM status is a random sample of overall population (MCAR), and would be biased if this assumption is not true. The EM method lessens this assumption.

4. A Case Study

A phase I/II trial was conducted to study an experimental anti-cancer therapy targeting a certain biomarker. The baseline fresh tissue samples were collected but not mandatory to determine the biomarker status categorized as BM positive or BM negative. Of clinical interest is to estimate tumor response rate by biomarker status to inform future development. However, among 169 patients with a certain tumor type enrolled to the study and evaluable for response assessments, the biomarker status for about 35% of patients could not be determined, mostly due to unavailable tissue samples. Table 4 displays the observed data and Table 5 displays the percentage of patients with known BM status in overall population as well as in two complementary subpopulations (represented by A and B for convenience) of interest, and the estimates and standard errors of prevalence rate p_0, the tumor response rate p_+ for BM+ and p_- for BM- in each (sub)population, based only on known BM status data and by EM algorithm, respectively.

{Insert Table 4 here}

{Insert Table 5 here}

Both methods produced similar results in overall population as well as in subpopulation A. However, the estimates for p_1 between two methods in subpopulation B are quite different, 17.6% vs 25%. As a result, if the estimates based only on known BM status data are used for decision making, then one may conclude that since the response rate for BM+ in subpopulation A is more than double that in subpopulation B, this substantial difference seems to support an enriched development strategy only in subpopulation A. However, the estimates by EM algorithm did not show such a drastic difference in response rate for BM+ between the two subpopulations, and hence may lead to a different conclusion/decision.

5. Discussions

Biomarker response evaluation is critical in early phase clinical trials of targeted therapies to inform further development strategy. Because of small sample size to begin with for early phase clinical trials and the desire to make the biomarker response estimate as accurate as possible, the proposed EM method offers clear advantage over the naïve method in terms of providing unbiased estimation under less stringent MAR assumption as well as having smaller variability due to using data from all patients. Of course if the biomarker status is missing not at random (MNAR), then both methods would be biased. Because it is usually difficult to verify the missingness pattern, it is important to avoid missingness as much as possible by trying to collect baseline biomarker samples from all enrolled patients and to minimize unusable samples. If there is still considerably high percentage of missing biomarker status despite of all the efforts to prevent this, then the proposed EM estimation method is recommended over the naïve method, unless there is clear evidence of MNAR, in which case both methods are not reliable and alternative solutions need to be sought.

References

[1] Fredlin B., and Simon R. 2005. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research.* **11**:7872–7878.

[2] Fredlin B., McShane L.M., and Korn E.L. 2010. Randomized clinical trials with biomarkers: Design issues, *Journal of National Cancer Institutes*. **102**:152-160

[3] Gosho M., Nagashima K., and Sato Y. 2012. Statistical design and statistical analysis for biomarker research. *Sensors.* **12**:8966-8986

[4] Jiang W., Freidlin B., and Simon R. 2007. Biomarker adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *J. Natl. Cancer Inst.* **99**:1036–1043.

[5] Mandrekar S.J., and Sargent D.J. 2011. All-comers versus enrichment design strategy in phase II trials. *Journal of thoracic oncology*. **6**:658-660

[6] Simon R. 2008. Development and validation of biomarker classifiers for treatment selection. *J. Statist Plan Infer*. **138**:308-320

[7] Simon R., and Maitournam A. 2004. Evaluating the efficiency of targeted designs for randomized clinical Trials. Clinical Cancer Research. **10**:6759-6763

[8] Simon R., and Maitournam A. 2006. Correction & Supplement: Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research. **12**:3229

[9] Wang S., O'Neill R.T., and Hung J.M.J. 2007. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics.* **6**:227–244.

[10] FDA Guidance for Industry. 2012. Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products.

[11] Miller R., and Siegmund D. 1982. Maximally selected Chi square statistics. *Biometrics*. **38**:1011-1016.

[12] Halpren J. 1982. Maximally selected Chi square statistics for small samples. *Biometrics*. **38**:1017-1023.

[13] Boulesteix A.L. 2006. Maximally selected Chi square statistics for ordinal variables. *Biomedical Journal*. **48**:451-462.

[14] Dempster A.P., Laird N.M., and Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser B*. **39**: 1-38.

Table 1. Response data in the presence of missing BM status

	BM+	BM-	BM unknown	Total
Response	R_{k+}	R_{k-}	$R_u = R_{u+} + R_{u2}$	R
Non-Response	NR_{k+}	NR_{k-}	$NR_u = NR_{u+} + NR_{u-}$	NR
Total	N_{k+}	N_{k-}	$N_u = N_{u+} + N_{u-}$	N

Table 2. Simulation results for MCAR case

N	% of known BM status	True Parameter			Mean estimates/SD pretending BM status known for all patients			Mean estimates/SD based only on known BM status			Mean estimates/SD by EM algorithm		
		p_0	p_+	p_-	p_0	p_+	p_-	p_0	p_+	p_-	p_0	p_+	p_-
40	50%	30%	40%	15%	30.0% 0.072	40.1% 0.147	15.0% 0.069	30.0% 0.104	40.1% 0.218	15.0% 0.098	29.9% 0.104	40.4% 0.210	14.9% 0.089
40	50%	30%	20%	15%	30.0% 0.072	20.3% 0.121	15.0% 0.069	30.0% 0.104	20.2% 0.182	15.0% 0.098	29.9% 0.105	20.0% 0.176	15.1% 0.085
60	50%	30%	40%	15%	30.0% 0.059	40.1% 0.119	15.0% 0.056	30.0% 0.085	40.1% 0.174	15.0% 0.079	30.0% 0.084	40.4% 0.162	14.9% 0.070
60	50%	30%	20%	15%	30.0% 0.059	20.2% 0.098	15.0% 0.056	30.0% 0.085	20.1% 0.143	15.0% 0.079	29.9% 0.085	20.0% 0.137	15.0% 0.067
80	50%	30%	40%	15%	30.0% 0.051	40.1% 0.102	15.0% 0.048	30.0% 0.073	40.1% 0.147	14.9% 0.068	30.0% 0.072	40.4% 0.135	14.9% 0.060
80	50%	30%	20%	15%	30.0% 0.051	20.1% 0.084	15.0% 0.048	30.0% 0.073	20.1% 0.121	14.9% 0.068	29.9% 0.073	20.1% 0.115	14.9% 0.058

Table 3. Simulation results for MAR case

N	True Parameter			Mean %/SD of known BM status	Mean estimates/SD pretending BM status known for all patients			Mean estimates/SD based only on known BM status			Mean estimates/SD by EM algorithm		
	p_0	p_+	p_-		p_0	p_+	p_-	p_0	p_+	p_-	p_0	p_+	p_-
40	30%	40%	15%	66.6% 0.074	30.0% 0.072	40.2% 0.145	15.1% 0.069	31.6% 0.090	46.9% 0.180	19.0% 0.094	30.1% 0.088	41.0% 0.171	15.1% 0.074
40	30%	20%	15%	65.4% 0.074	30.0% 0.072	20.2% 0.120	15.1% 0.069	30.4% 0.090	24.8% 0.164	19.0% 0.094	30.1% 0.090	20.5% 0.142	15.2% 0.074
60	30%	40%	15%	66.6% 0.061	30.0% 0.059	40.2% 0.117	15.1% 0.056	31.7% 0.074	46.9% 0.144	18.9% 0.076	30.1% 0.072	40.7% 0.135	15.1% 0.060
60	30%	20%	15%	65.4% 0.061	30.0% 0.059	20.1% 0.096	15.1% 0.056	30.4% 0.074	24.7% 0.129	18.9% 0.076	30.1% 0.074	20.3% 0.110	15.1% 0.059
80	30%	40%	15%	66.6% 0.053	30.0% 0.051	40.0% 0.101	15.0% 0.048	31.6% 0.063	46.7% 0.123	18.9% 0.066	30.1% 0.062	40.4% 0.115	15.1% 0.052
80	30%	20%	15%	65.4% 0.053	30.0% 0.051	20.0% 0.083	15.0% 0.048	30.4% 0.063	24.7% 0.112	18.9% 0.066	30.1% 0.063	20.2% 0.094	15.1% 0.051

Table 4. Case study: Observed data

Evaluable Population		**BM+**	**BM-**	**BM Unknown**	**Total**
All Patients (N=169)	Response	10	7	10	27
	Non-Response	26	67	49	142
	Total	36	74	59	169
Subpopulation A (N=68)	Response	7	6	3	16
	Non-Response	12	30	10	52
	Total	19	36	13	68
Subpopulation B (N=101)	Response	3	1	7	11
	Non-Response	14	37	39	90
	Total	17	38	46	101

Table 5. Case study: Estimated response rates by biomarker status

Evaluable population	% of known BM status	Estimates (SE) based only on known BM status			Estimates (SE) by EM algorithm		
		p_0	p_+	p_-	p_0	p_+	p_-
All patients (N=169)	65%	32.7% (0.045)	27.8% (0.075)	9.5% (0.034)	32.9% (0.036)	28.6% (0.061)	9.8% (0.028)
Subpopulation A (N=68)	81%	34.5% (0.064)	36.8% (0.111)	16.7% (0.062)	34.5% (0.058)	36.7% (0.099)	16.6% (0.056)
Subpopulation B (N=101)	54%	30.9% (0.062)	17.6% (0.092)	2.6% (0.026)	32.6% (0.047)	25.0% (0.075)	4.1% (0.024)