# Supplementary Material for "Group Testing Can Improve the Cost-Efficiency of Prospective-Retrospective Biomarker Studies"

Wei Zhang[1], Zhiwei Zhang[2,*], Julia Krushkal[2] and Aiyi Liu[1]

[1]Biostatistics and Bioinformatics Branch, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA

[2]Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

[*]zhiwei.zhang@nih.gov

## Evaluating a Prognostic Biomarker

In general, a measure of association between $X$ and $Y$, say $\delta g(p_1, p_0) = g(p_1) - g(p_0)$, can be estimated by substituting estimates of $(p_1, p_0)$. If $\delta g(p_1, p_0)$ is the log-odds ratio, $\log[p_1(1-p_0)/\{p_0(1-p_1)\}]$, it can also be expressed in terms of $(q_1, q_0)$ as $\log[q_1(1-q_0)/\{q_0(1-q_1)\}]$ (e.g., Agresti, 2013, Chapter 2), and thus can be estimated by substituting estimates of $(q_1, q_0)$. For a different measure of association, $\delta g(p_1, p_0)$ is not a function of $(q_1, q_0)$; however, estimates of $(q_1, q_0)$ may still be useful for estimating $(p_1, p_0)$ because, by Bayes' theorem,

$$\begin{aligned} p_1 &= \frac{\lambda q_1}{\lambda q_1 + (1-\lambda) q_0}, \\ p_0 &= \frac{\lambda(1-q_1)}{\lambda(1-q_1) + (1-\lambda)(1-q_0)}, \end{aligned} \tag{S.1}$$

where $\lambda = \mathrm{P}(Y = 1)$.

The "full data"can be represented as $(X_i, Y_i)$, $i = 1, \ldots, n$, where the subscript $i$ denotes the $i$th subject in the trial. Under the standard design, the full data are fully observed, and it is straightforward to estimate $p_x$ as a sample proportion:

$$\widehat{p}_x^S = \frac{\sum_{i=1}^n I(X_i = x) Y_i}{\sum_{i=1}^n I(X_i = x)}, \qquad x = 0, 1,$$

where $I(\cdot)$ is the indicator function and the superscript $S$ denotes the standard design. The resulting estimate of $\delta g(p_1, p_0)$ is simply $\delta g(\widehat{p}_1^S, \widehat{p}_0^S)$.

Under the RS design, the $X_i$'s are incompletely observed. Let $R_i = 1$ if $X_i$ is observed; 0 otherwise. The RS design implies that

$$\mathrm{P}(R_i = 1|X_i, Y_i) = \mathrm{P}(R_i = 1|Y_i),$$

so $X_i$ is missing at random in the sense of Rubin (1976). This further implies that

$$\mathrm{P}(X_i = 1|Y_i = y, R_i = 1) = \mathrm{P}(X_i = 1|Y_i = y) = q_y, \qquad y = 0, 1,$$

which motivates the following estimates:

$$\widehat{q}_y^{RS} = \frac{\sum_{i=1}^n I(R_i = 1, Y_i = y, X_i = 1)}{\sum_{i=1}^n I(R_i = 1, Y_i = y)}, \qquad y = 0, 1.$$

As noted earlier, if $\delta g(p_1, p_0)$ is the log-odds ratio, it can be estimated as $\delta g(\widehat{q}_1^{RS}, \widehat{q}_0^{RS})$. For other measures of association, we can invoke (S.1) and estimate $(p_1, p_0)$ as

$$\widehat{p}_1^{RS} = \frac{\widehat{\lambda}\widehat{q}_1^{RS}}{\widehat{\lambda}\widehat{q}_1^{RS} + (1 - \widehat{\lambda})\widehat{q}_0^{RS}},$$
$$\widehat{p}_0^{RS} = \frac{\widehat{\lambda}(1 - \widehat{q}_1^{RS})}{\widehat{\lambda}(1 - \widehat{q}_1^{RS}) + (1 - \widehat{\lambda})(1 - \widehat{q}_0^{RS})},$$

where $\widehat{\lambda} = n^{-1} \sum_{i=1}^n Y_i$. The resulting estimate of $\delta g(p_1, p_0)$ is $\delta g(\widehat{p}_1^{RS}, \widehat{p}_0^{RS})$.

In the GT design, we allow pools in the same stratum to have different sizes for full generality. Suppose the subjects in the $Y = y$ stratum are randomly grouped into $m_y$ pools of sizes $k_{jy}$, $j = 1, \ldots, m_y$. The marker status of the $j$th pool in the $Y = y$ stratum is given by $X_{jy}^* = \max_{1 \le i \le k_{jy}} X_{ijy}$, where $X_{ijy}$ is the marker status of the $i$th subject in the same pool. It follows that

$$\mathrm{P}(X_{jy}^* = 1) = 1 - (1 - q_y)^{k_{jy}},$$

and the likelihood for $q_y$ is

$$\prod_{j=1}^{m_y} \left\{1 - (1 - q_y)^{k_{jy}}\right\}^{X_{jy}^*} \left\{(1 - q_y)^{k_{jy}}\right\}^{1 - X_{jy}^*},$$

which can be maximized to estimate $q_y$. The resulting maximum likelihood estimates of $(q_1, q_0)$ can be used to estimate $\delta g(p_1, p_0)$ in the same manner as in the RS design.

# Evaluating a Predictive Biomarker

In general, the interaction coefficient $\beta_{TX}$ can be estimated by substituting estimates of the $p_{tx}$'s into equation (2) in the main text. For the logit link,

$$\beta_{TX} = \log\left\{\frac{p_{11}(1-p_{10})(1-p_{01})p_{00}}{(1-p_{11})p_{10}p_{01}(1-p_{00})}\right\}$$

can be alternatively expressed as

$$\beta_{TX} = \log\left\{\frac{q_{11}(1-q_{10})(1-q_{01})q_{00}}{(1-q_{11})q_{10}q_{01}(1-q_{00})}\right\}; \tag{S.2}$$

see, for example, Liu et al. (2012, Supplementary Materials). Thus, in this case, $\beta_{TX}$ can also be estimated by substituting estimates of the $q_{ty}$'s. For a different link function, $\beta_{TX}$ is not a function of the $q_{ty}$'s but its estimation can be helped by estimation of the $q_{ty}$'s, as Bayes' theorem implies that

$$\begin{aligned} p_{t1} &= \frac{\lambda_t q_{t1}}{\lambda_t q_{t1} + (1-\lambda_t)q_{t0}}, \\ p_{t0} &= \frac{\lambda_t(1-q_{t1})}{\lambda_t(1-q_{t1}) + (1-\lambda_t)(1-q_{t0})}, \end{aligned} \tag{S.3}$$

where $\lambda_t = \mathrm{P}(Y=1|T=t)$, $t=0,1$.

In this setting, the full data can be represented as $(X_i, T_i, Y_i)$, $i=1,\ldots,n$, where the subscript $i$ denotes the $i$th subject in the trial. Under the standard design, where all variables are fully observed, each $p_{tx}$ can be estimated as a sample proportion:

$$\widehat{p}^{S}_{tx} = \frac{\sum_{i=1}^{n} I(T_i=t, X_i=x)Y_i}{\sum_{i=1}^{n} I(T_i=t, X_i=x)},$$

which can then be substituted into equation (2) to estimate $\beta_{TX}$.

Under the RS design, the $X_i$'s are incompletely observed. Let $R_i=1$ if $X_i$ is observed; 0 otherwise. The RS design implies that

$$\mathrm{P}(R_i=1|X_i,T_i,Y_i) = \mathrm{P}(R_i=1|T_i,Y_i),$$

or equivalently,

$$\mathrm{P}(X_i=1|T_i,Y_i,R_i=1) = \mathrm{P}(X_i=1|T_i,Y_i).$$

Therefore, we can estimate each $q_{ty}$ with

$$\widehat{q}_{ty}^{RS} = \frac{\sum_{i=1}^{n} I(R_i = 1, T_i = t, Y_i = y, X_i = 1)}{\sum_{i=1}^{n} I(R_i = 1, T_i = t, Y_i = y)}.$$

These estimates can be substituted into equation (S.2) to estimate $\beta_{TX}$ under the logit link. For other link functions, equation (S.3) suggests that each $p_{tx}$ can be estimated as

$$\widehat{p}_{t1}^{RS} = \frac{\widehat{\lambda}_t \widehat{q}_{t1}^{RS}}{\widehat{\lambda}_t \widehat{q}_{t1}^{RS} + (1 - \widehat{\lambda}_t)\widehat{q}_{t0}^{RS}},$$
$$\widehat{p}_{t0}^{RS} = \frac{\widehat{\lambda}_t (1 - \widehat{q}_{t1}^{RS})}{\widehat{\lambda}_t (1 - \widehat{q}_{t1}^{RS}) + (1 - \widehat{\lambda}_t)(1 - \widehat{q}_{t0}^{RS})},$$

where $\widehat{\lambda}_t = \sum_{i=1}^{n} I(T_i = t) Y_i / \sum_{i=1}^{n} I(T_i = t)$, $t = 0, 1$. The $\widehat{p}_{tx}^{RS}$'s can be substituted into equation (2) to estimate $\beta_{TX}$.

For the GT design, suppose the subjects in the $(T = t, Y = y)$ stratum are randomly grouped into $m_{ty}$ pools of sizes $k_{jty}$, $j = 1, \ldots, m_{ty}$. The marker status of the $j$th pool in the $(T = t, Y = y)$ stratum is given by $X_{jty}^{*} = \max_{1 \le i \le k_{jty}} X_{ijty}$, where $X_{ijty}$ is the marker status of the $i$th subject in the same pool. It follows that

$$\mathrm{P}(X_{jty}^{*} = 1) = 1 - (1 - q_{ty})^{k_{jty}},$$

and the likelihood for $q_{ty}$ is

$$\prod_{j=1}^{m_{ty}} \left\{1 - (1 - q_{ty})^{k_{jty}}\right\}^{X_{jty}^{*}} \left\{(1 - q_{ty})^{k_{jty}}\right\}^{1 - X_{jty}^{*}}.$$

Maximum likelihood estimates of the $q_{ty}$'s can be used to estimate $\beta_{TX}$ in the same manner as in the RS design.

# Choosing a Pool Size

When planning the retrospective part of a P-R biomarker study with GT, the relevant variance to minimize is the conditional variance of an estimator given observed data from the prospective part of the study. To fix ideas, consider a predictive biomarker study aiming to estimate $\beta_{TX}$ for an arbitrary (but specified) link function $g$. Given $\mathcal{O} = \{(T_i, Y_i) : i = 1, \ldots, n\}$, the conditional variance of $\widehat{\beta}_{TX}^{GT}$ is a monotone function of the conditional variance

of $\widehat{\boldsymbol{q}}^{GT} = (\widehat{q}_{11}^{GT}, \widehat{q}_{10}^{GT}, \widehat{q}_{01}^{GT}, \widehat{q}_{00}^{GT})'$, the vector of maximum likelihood estimates of the $q_{ty}$'s. Specifically, $\text{var}(\widehat{\beta}_{TX}^{GT}|\mathcal{O})$ decreases when $\text{var}(\widehat{\boldsymbol{q}}^{GT}|\mathcal{O})$ becomes smaller in the sense of non-negative definiteness. Because $\text{var}(\widehat{\boldsymbol{q}}^{GT}|\mathcal{O})$ is a diagonal matrix, $\text{var}(\widehat{\beta}_{TX}^{GT}|\mathcal{O})$ is monotone in $\text{var}(\widehat{q}_{ty}^{GT}|\mathcal{O})$ for each $(t, y)$ pair. Now, consider a fixed $(t, y)$ pair, and assume that the $m_{ty}$ pools in the $(T = t, Y = y)$ stratum have the same size, say $k$. If $m_{ty}$ is reasonably large, $\text{var}(\widehat{q}_{ty}^{GT}|\mathcal{O})$ is approximately the inverse of the Fisher information about $q_{ty}$ in $\{X_{jty}^*, j = 1, \ldots, m_{ty}\}$, which is easily found to be $m_{ty} I_k(q_{ty})$, where

$$I_k(q_{ty}) = \frac{k^2(1 - q_{ty})^{2(k-1)}}{(1 - q_{ty})^k\{1 - (1 - q_{ty})^k\}}$$

is the Fisher information about $q_{ty}$ in a single pooled assay result $X_{jty}^*$. If $m_{ty}$ is fixed and $n_{ty}$ is large enough, then the optimal value of $k$ is the one that maximizes $I_k(q_{ty})$. Although this argument is made for a predictive biomarker, it can be applied to a prognostic biomarker with minor modifications.

# References


Agresti A. *Categorical Data Analysis*, 3rd ed. 2013. John Wiley and Sons: Hoboken, NJ.

Liu A, Liu C, Zhang Z, Albert PS. Optimality of group testing in the presence of misclassification. *Biometrika* 2012; 99:245–251.

Rubin DB. Inference and missing data. *Biometrika* 1976; 63:581–592.