Supplementary Material: Prognostic and predictive values and statistical interactions in the era of targeted treatment

Jaya M. Satagopan, Alexia Iasonos, and Qin Zhou

Memorial Sloan Kettering Cancer Center

February 27, 2015

A Theoretical value of AUC

Denoting X and Z as the two categorical risk factors of interest, for ease of exposition we shall write the logistic regression model of Equation (1) as:

$$\begin{aligned}\log\left\{\frac{\pi}{1-\pi}\right\} &= \mu + \sum_{j=1}^{L_1}\frac{I(X=j)-p_j}{\sqrt{p_x \times (1-p_x)}}\beta_j + \\ & \sum_{k=1}^{L_2}\frac{I(Z=k)-p_k}{\sqrt{p_k \times (1-p_k)}}\delta_k + \\ & \sum_{j=1}^{L_1}\sum_{k=1}^{L_2}\frac{I(X=j)-p_j}{\sqrt{p_x \times (1-p_x)}} \times \frac{I(Z=k)-p_k}{\sqrt{p_k \times (1-p_k)}}\gamma_{jk}\ . \qquad \text{(A.1)}\end{aligned}$$

Here $I(.)$ denotes the indicator function taking value 1 when the condition in the parentheses is true and taking value 0 otherwise, p_j is the prevalence of category j of X, and p_k is the prevalence of category k of Z.

To calculate the theoretical value of AUC, we postulate a log normal distribution for disease risk in the general population. Therefore $\log\{\pi\}$ is assumed to follow a normal distribution with mean m and variance σ^2. From Equation (A.1), given the parameters μ, $\boldsymbol{\beta} = \{\beta_j\}_{i=1}^{L_1}$, $\boldsymbol{\delta} = \{\delta_k\}_{k=1}^{L_2}$, and $\boldsymbol{\gamma} = \{\gamma_{jk}\}_{j,k=1}^{L_1,L_2}$, we can write the mean of $\log\{\pi\}$ under the rare disease assumption as $E(\log\{\pi\}) = \mu$. Its variance can be written as:

$$\sigma^2 \;=\; \begin{pmatrix}\boldsymbol{\beta}^T & \boldsymbol{\delta}^T & \boldsymbol{\gamma}\end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix}\boldsymbol{\beta}^T & \boldsymbol{\delta}^T & \boldsymbol{\gamma}\end{pmatrix}^T \;, \qquad \text{(A.2)}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the distribution of the risk factors. When the risk factors are independent and scaled to have variance 1, $\boldsymbol{\Sigma}$ is the identity matrix, and we can write:

$$\sigma^2 \;=\; \sum_{j=1}^{L_1} \beta_j^2 + \sum_{k=1}^{L_2} \delta_k^2 + \sum_{j=1}^{L_2} \sum_{k=1}^{L_2} \gamma_{jk}^2 \;. \qquad \text{(A.3)}$$

From the results of Begg [2002] and Pharoah et al. [2002], when risk has a log-normal distribution in the general population with mean μ and variance σ^2, the distribution of risk among the affected individuals is log-normal with mean $\mu + \sigma^2$ and variance σ^2. Note that AUC is the probability that disease risk among affected individuals is higher than that among the unaffected individuals. Denote R_1 and R_0 as the risk among affected and unaffected individuals. Then $\log\{R_1\}$ and $\log\{R_0\}$ are distributed independently as $N(\mu + \sigma^2, \sigma^2)$ and $N(\mu, \sigma^2)$, respectively. Therefore, the theoretical value of AUC is:

$$\begin{aligned} AUC &= P(R_1 > R_0) \\ &= P(\log\{R_1\} > \log\{R_0\}) \\ &= P\left(\frac{\log\{R_1\} - \log\{R_0\} - \sigma^2}{\sqrt{2\sigma^2}} > \frac{-\sigma^2}{\sqrt{2\sigma^2}}\right) \\ &= \Phi\left(\frac{\sigma}{\sqrt{2}}\right) \;, \qquad \text{(A.4)} \end{aligned}$$

where $\Phi(.)$ is the cumulative probability of the standard normal distribution.

B Resampling approach to test $H_0 : \Delta AUC = 0$ in relation to interactions

When a new biomarker is included in the model, a resampling procedure can be used to test the null hypothesis $H_0 : \Delta AUC = 0$ [Seshan et al. 2013]. This approach retains the outcome and all the other risk factors of each individual, and permutes the biomarker value to calculate the null distribution of ΔAUC. Although the concept of permutation is applicable in our setting, the method of Seshan et al. [2013] cannot be used directly. This is because the definition of interaction depends upon the definition of a main effect [Finney 1948; Wang et al. 2010; Satagopan and Elston 2013]. In any regression setting, the estimates of the main effects are weighted averages of the outcomes (for example, weighted averages of log odds in logistic regression). The weights depend upon whether or not interaction terms are included in the model. Therefore, under a naive permutation of the interaction column of the design matrix, the main and interaction effects will no longer be interpretable in this canonical sense. Therefore, we propose a novel resampling procedure for evaluating the null distribution of ΔAUC when the new risk factor of interest is an interaction term. We develop this approach when the interaction is between two categorical risk factors.

For binary disease traits, interactions have a unique interpretation in terms of the odds ratios of association between the risk factors calculated separately in the affected and unaffected individuals. Therefore, we will generate data under the null such that there is no interaction between the two risk factors, but the association between the two risk factors in the unaffected individuals will be the same as that in the observed data.

For each individual, denote Y as the binary disease trait taking value 1 when the person is affected (i.e., has the event of interest) and taking value 0 otherwise. Let X and Z denote the two categorical risk factors having L_1 and L_2 levels, respectively. The association between the two risk factors in the affected and unaffected individuals can be measured via the conditional probability $P(X=j|Z,Y)$. Setting $Y=1$ (or 0) provides the association in affected (or unaffected) individuals. It is easy to see that:

$$\frac{P(X=j|Z,Y=1)}{P(X=j|Z,Y=0)} = \frac{P(Y=1|X=j,Z)}{P(Y=0|X=j,Z)} \times \frac{P(Y=0|Z)}{P(Y=1|Z)} \; . \tag{B.1}$$

From Equation (1) of the paper, it follows that the first term on the right hand side of Equation (B.1) is equal to

$$\exp\left\{\mu + \beta_j + \sum_{k=1}^{L_2} \delta_k I(Z=k) + \sum_{k=0}^{L_2} \gamma_{jk} I(Z=k)\right\} \; .$$

We posit a logistic regression model for Y given Z, and write the second term on the right hand side of Equation (B.1) as: $\exp\left\{-\left(\tilde{\mu} + \sum_{k=0}^{L_2} \tilde{\delta}_k I(Z=k)\right)\right\}$, where $\tilde{\mu}$ and $\tilde{\delta}_k$ are the parameters of this model. Therefore, the right hand side of Equation (B.1) can be written as:

$$\frac{P(X=j|Z,Y=1)}{P(X=j|Z,Y=0)} = \exp\left\{\mu^* + \beta_j + \sum_{k=1}^{L_2} \delta_k^* I(Z=k) + \sum_{k=1}^{L_2} \gamma_{jk} I(Z=k)\right\} \; .$$

When the risk factor X is held at its baseline level of 0, we have:

$$\frac{P(X=0|Z,Y=1)}{P(X=0|Z,Y=0)} = \exp\left\{\mu^* + \sum_{k=1}^{L_2} \delta_k^* I(Z=k)\right\} \; .$$

The odds that $X=j$ among the affected individuals relative to the odds among the unaffected individuals can be written using the above equations as:

$$\frac{P(X=j|Z,Y=1)}{P(X=0|Z,Y=1)} \times \frac{P(X=0|Z,Y=0)}{P(X=j|Z,Y=0)} = \exp\left\{\beta_j + \sum_{k=1}^{L_2-1} \gamma_{jk} I(Z=k)\right\} \; . \tag{B.2}$$

This motivates a polytomous logistic regression model for the j-th level of risk factor X

relative to its baseline level, given by:

$$\log\left\{\frac{P(X=j|Z,Y)}{P(X=0|Z,Y)}\right\} = m_j + \beta_j Y + \sum_{k=1}^{L_2-1} d_{jk} I(Z=k) + \sum_{k=1}^{L_2-1} \gamma_{jk} Y I(Z=k)\ , \tag{B.3}$$

where m_j measures the frequency of the j-th level of the risk factor X among the unaffected individuals when Z is held at its baseline level, β_j is the association between the j-th level of X and disease when Z is held at its baseline level, d_{jk} is the log odds ratio for the association between the j-th level of X and the k-th level of Z in the unaffected individuals, and $d_{jk}+\gamma_{jk}$ denotes the association between the j-th level of X and the k-th level of Z in the affected individuals. Under the null hypothesis of no interaction, we have $\gamma_{jk} = 0$ for all j and k. This model motivates a resampling procedure for estimating the null distribution of ΔAUC. Our model is derived for obtaining the null distribution of ΔAUC under the requirements that: (i) the association between the risk factors among the unaffected individuals must remain the same as that in the observed data; and (ii) their association among affected individuals must be the same as that in the unaffected individuals. Equation (B.3) has parallels to the conditional distribution of genetic factors given environmental factors and disease status, described by Han et al [Han et al. 2012], which was developed to obtain the null distribution of a test statistic for interactions under monotonicity constraints.

Our proposed resampling procedure based on Equation (B.3) proceeds as follows.

1. First, calculate ΔAUC for the observed data. Denote this as ΔAUC_{obs}.

2. Now begin the resampling procedure as follows. Fit the polytomous logistic regression model of Equation (B.3) under the null hypothesis (i.e., by setting $\gamma_{jk} = 0$ for all j and k) using the observed data.

3. Get the estimated parameters of the model.

4. For each individual, retain their observed values of Y and Z, and calculate the conditional probability that $X = j$ by plugging the estimated parameters into the right hand side of Equation (B.3).

5. Use this probability to sample X for each individual.

6. Fit 2 models to this null data set. The first model will be Equation (1) of the paper that includes interactions. The second model will be an additive logistic regression model i.e., Equation (1) but with $\gamma_{jk} = 0$ for all j and k.

7. Calculate the AUCs of these 2 models and obtain ΔAUC.

8. Repeat Steps 5 to 8 for a total of B times (we used B = 1000).

9. The resulting vector of ΔAUCs provides the required null distribution.

10. Calculate the p-value as the proportion of the null ΔAUCs that are greater than ΔAUC_{obs}.

C Simulation setup

We simulated N independent individuals (N_1 affected and $N_0 = N - N_1$ unaffected individuals), each having two categorical risk factors: X taking values $1, 2, \cdots, L_1$ and Z taking values $1, 2, \cdots, L_2$. We assumed that X and Z conferred disease risk when they exceed some threshold i.e., when $X \geq C_1$ and $Z \geq C_2$. This choice is motivated by threshold models that are commonly postulated for complex diseases such as cancer, whereby disease risk increases considerably when an underlying risk factor (for example, RNA or protein expression, body mass index, or cholesterol level) exceeds a certain limit.

Setting $P(X \geq C_1) = p_x$, we assumed $P(X = j) = p_x/(L_1 - C_1 + 1)$ for $j = C_1, C_1 + 1, \cdots, L_1$. Further, we assumed that $P(X = j) = (1 - p_x)/(C_1 - 1)$ for $j = 1, \cdots, C_1 - 1$. Similarly, setting $P(X \geq C_2) = p_z$, we assumed $P(Z = k) = p_z/(L_2 - C_2 + 1)$ for $k = C_2, \cdots, L_2$, and $P(Z = k) = (1 - p_z)/(C_2 - 1)$ for $k = 1, \cdots, C_2 - 1$.

We assumed the two risk factors to have correlation ρ, given by

$$\rho \; = \; \frac{p_x \times \{P\left(Z \geq C_2 | X \geq C_1\right) - p_z\}}{\sqrt{p_x(1 - p_x) \;\; p_z(1 - p_z)}} \; . \tag{C.1}$$

Let Y be the binary disease status, with $Y = 1$ and 0 denoting affected and unaffected statuses, respectively. Given the two risk factors and the thresholds, disease risk was assumed to follow a logistic regression model given by:

$$\begin{aligned} \log\left\{\frac{P(Y = 1|X, Z)}{P(Y = 0|X, Z)}\right\} \; &= \; \log\left\{\frac{\pi}{1 - \pi}\right\} + \beta \times \frac{\{I(X \geq C_1) - p_x)\}}{\sqrt{p_x(1 - p_x)}} + \delta \times \frac{\{I(Z \geq C_1) - p_z)\}}{\sqrt{p_z(1 - p_z)}} + \\ & \gamma \times \frac{\{I(X \geq C_1) - p_x)\} \times \{I(Z \geq C_2) - p_z)\}}{\sqrt{p_x(1 - p_x) \; p_z(1 - p_z)}} \; , \end{aligned} \tag{C.2}$$

where π denotes baseline disease risk when the frequencies of X and Z exceeding the threshold are at their average levels; $I(.)$ is the indicator function taking value 1 when the condition within parentheses is true and taking value 0 otherwise; β and δ are interpreted as the main effects of the two risk factors; and γ is interpreted as the interaction effect.

We set $\pi = 0.05$, $p_x = 0.50$, $p_z = 0.70$, $L_1 = 3$, $C_1 = 2$, $L_2 = 2$, $C_2 = 1$, and simulated 100 data-sets under the following configurations for each data-set.

Sample Size: $N_1 = N_0 = 500$ and 200.

Correlation: $\rho = \{0, 0.25, 0.50\}$.

Effects under the null: We set $\delta = K_1 \times \beta$ and $\gamma = K_2 \times \beta$, with $K_1 = 0.50$. To examine type I errors of the three tests based on the likelihood ratio statistic, ΔAUC, and RERI, we simulated data under the null by setting $K_2 = 0$ (i.e., $\gamma = 0$). Note that when the risk factors are independent, the AUC based on Equation (C.2) is

$\Phi\left\{\frac{\sqrt{\beta^2+\delta^2+\gamma^2}}{\sqrt{2}}\right\} = \Phi\left\{\frac{\beta \times \sqrt{1+K_1^2+K_2^2}}{\sqrt{2}}\right\}$. We simulated β such that the AUCs under the null (i.e., AUC_0) were 0.55, 0.60, and 0.65.

Non-null effects: To examine the power of the three tests, we generated non-null data by setting K_2 such that $\Delta AUC = \{0.05, 0.10, 0.15\}$.

Table S1 shows various parametric configurations used in our simulations.

References

Begg C. B. 2002. On the use of familial aggregation in population-based case probands for calculating penetrance. *JNCI* 94:1221–1226.

Finney D. J. 1948. Main effects and interactions. *JASA* 43:566–571.

Han S. S, Rosenberg P. S, and Chatterjee N. 2012. Testing for gene-environment and gene-gene interactions under monotonicity constraints. *JASA* 107:1441–1452.

Pharoah P. D. P, Antoniou A, Bobrow M, Zimmern R. L, Easton D. F, and Ponder B. A. 2002. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* 31:33–36.

Satagopan J. M and Elston R. C. 2013. Evaluation of removable statistical interaction for binary traits. *Stat Med* 32:1164–1190.

Seshan V. E, Gonen M, and Begg C. B. 2013. Comparing roc curves derived from regression models. *Stat Med* 32:1483–1493.

Wang X, Elston R. C, and Zhu X. 2010. The meaning of interaction. *Hum Hered* 70:269–277.

Table S1: Values of model parameters used for simulations

β	δ	γ	AUC0	AUC1
Quantitative Interactions				
0.15	0.1	0.3	0.55	0.60
0.15	0.1	0.7	0.55	0.70
0.35	0.1	0.6	0.60	0.69
0.55	0.1	0.5	0.65	0.70
0.55	0.1	1.1	0.65	0.80
0.75	0.1	0.9	0.70	0.80
0.95	0.1	0.7	0.75	0.80
Qualitative Interactions				
0.15	0.1	-0.3	0.55	0.60
0.15	0.1	-0.7	0.55	0.70
0.35	0.1	-0.7	0.60	0.71
0.55	0.1	-0.5	0.65	0.70
0.55	0.1	-1.05	0.65	0.80
0.75	0.1	-0.9	0.70	0.80
0.75	0.6	-0.75	0.75	0.81

Table S2: Theoretical, estimated and attained AUCs and ΔAUC shown for simulations with 200 affected and 200 unaffected individuals. The first column shows the true values of β, δ, and γ used for generating disease risk from Equation (C.2).

		$\rho = 0$			$\rho = 0.5$		
(β, δ, γ)	AUC	AUC0	AUC1	ΔAUC	AUC0	AUC1	ΔAUC
Quantitative Interactions							
(0.15, 0.10, 0.30)	theoretical	0.551	0.598	0.05	0.561	0.583	0.02
	attainable	0.568 (0.028)	0.603 (0.023)	0.04	0.561 (0.025)	0.574 (0.025)	0.01
	estimated	0.57 (0.023)	0.608 (0.026)	0.04	0.568 (0.023)	0.582 (0.024)	0.01
(0.15, 0.10, 0.70)	theoretical	0.551	0.696	0.15	0.561	0.663	0.10
	attainable	0.554 (0.029)	0.672 (0.023)	0.12	0.58 (0.022)	0.606 (0.02)	0.03
	estimated	0.569 (0.026)	0.67 (0.025)	0.10	0.594 (0.027)	0.622 (0.024)	0.03
(0.55, 0.10, 0.50)	theoretical	0.654	0.702	0.05	0.666	0.693	0.03
	attainable	0.652 (0.025)	0.687 (0.023)	0.04	0.641 (0.023)	0.658 (0.023)	0.02
	estimated	0.652 (0.026)	0.693 (0.024)	0.04	0.644 (0.022)	0.662 (0.024)	0.02
(0.75, 0.10, 0.90)	theoretical	0.704	0.798	0.09	0.715	0.778	0.06
	attainable	0.69 (0.032)	0.753 (0.024)	0.06	0.668 (0.025)	0.698 (0.027)	0.03
	estimated	0.688 (0.029)	0.753 (0.024)	0.07	0.666 (0.03)	0.695 (0.025)	0.03
Quantitative Interactions							
(0.15, 0.10, -0.30)	theoretical	0.551	0.598	0.05	0.561	0.605	0.04
	attainable	0.542 (0.023)	0.582 (0.023)	0.04	0.566 (0.022)	0.587 (0.021)	0.02
	estimated	0.547 (0.026)	0.589 (0.022)	0.04	0.57 (0.027)	0.594 (0.024)	0.02
(0.15, 0.10, -0.70)	theoretical	0.551	0.695	0.14	0.561	0.687	0.13
	attainable	0.545 (0.028)	0.665 (0.026)	0.12	0.569 (0.032)	0.661 (0.026)	0.09
	estimated	0.556 (0.031)	0.669 (0.025)	0.11	0.577 (0.041)	0.664 (0.026)	0.09
(0.55, 0.10, -0.50)	theoretical	0.654	0.702	0.05	0.666	0.708	0.04
	attainable	0.639 (0.03)	0.67 (0.026)	0.03	0.618 (0.026)	0.647 (0.021)	0.03
	estimated	0.639 (0.029)	0.669 (0.024)	0.03	0.63 (0.024)	0.656 (0.022)	0.03
(0.75, 0.10, -0.90)	theoretical	0.703	0.793	0.09	0.715	0.793	0.08
	attainable	0.718 (0.032)	0.747 (0.026)	0.03	0.638 (0.029)	0.7 (0.019)	0.06
	estimated	0.714 (0.025)	0.747 (0.023)	0.03	0.637 (0.027)	0.7 (0.022)	0.06

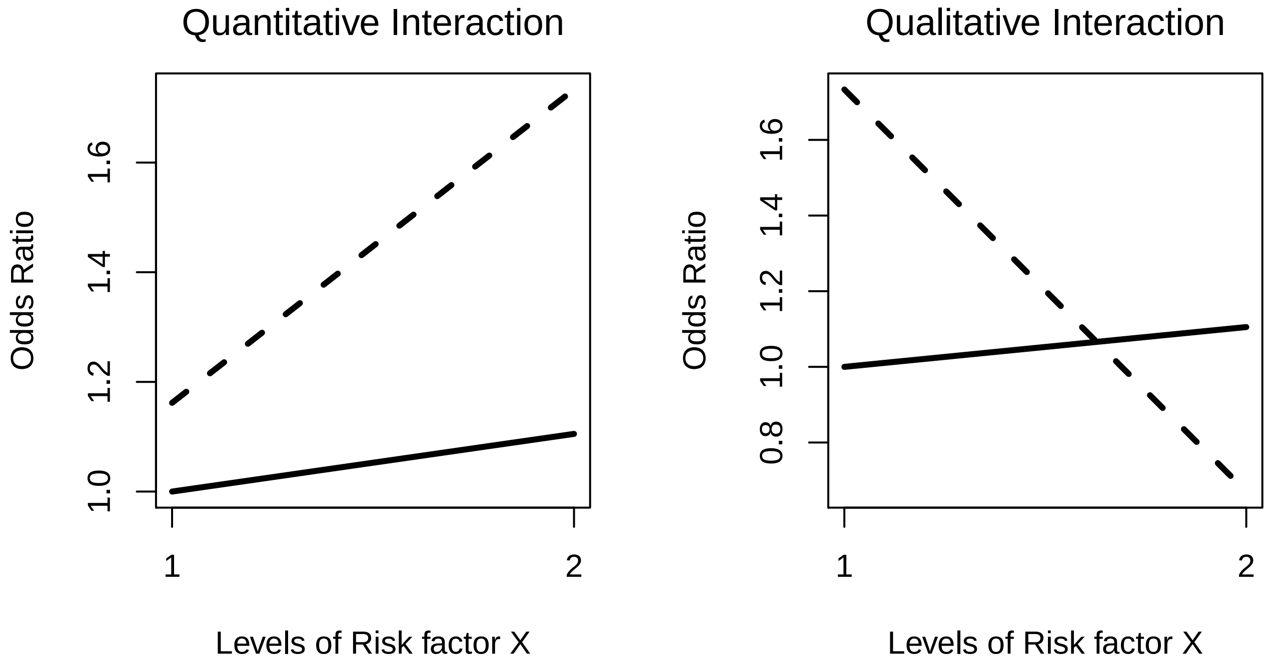

Figure S1: A visual representation of quantitative and qualitative interactions.

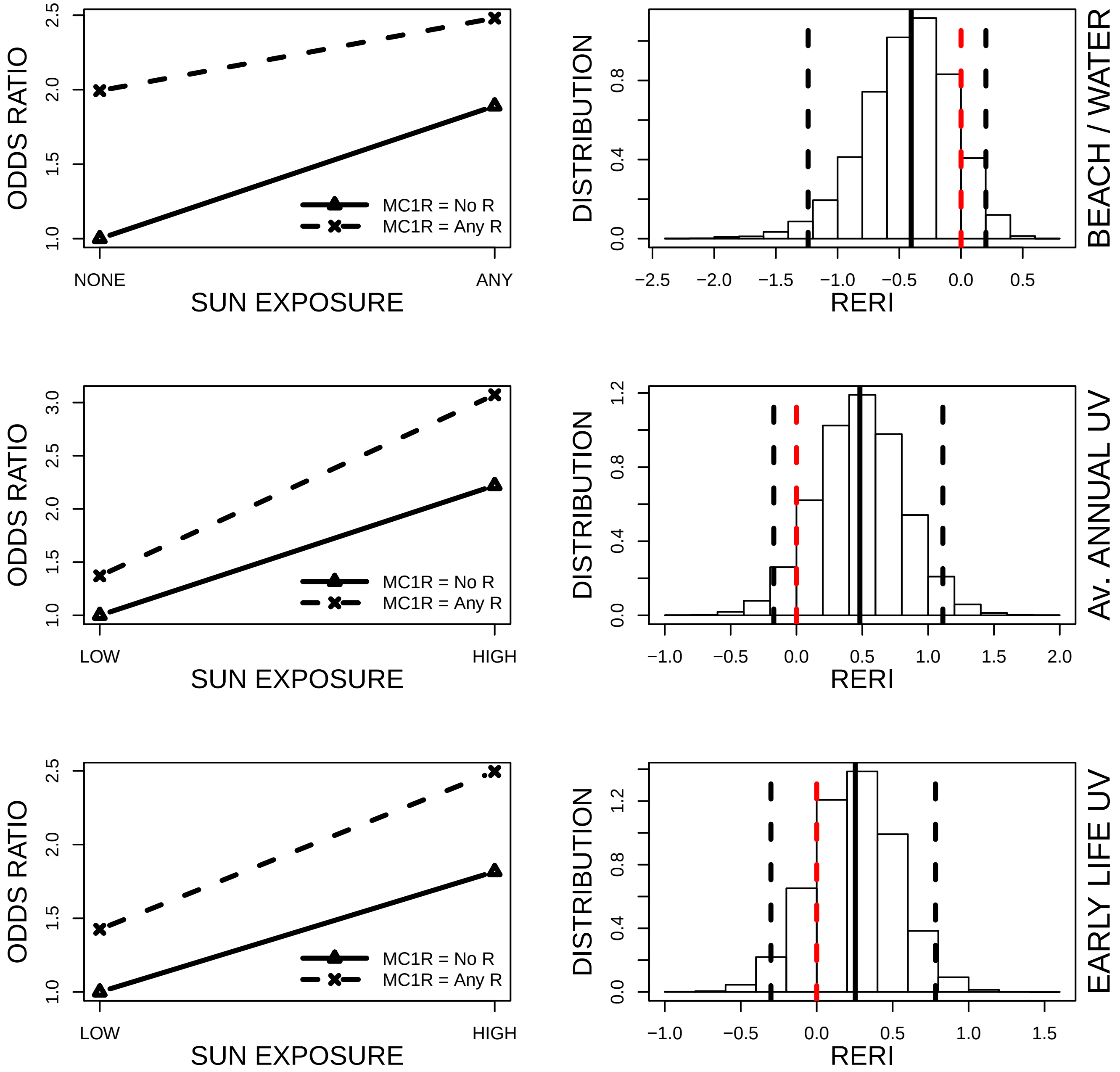

Figure S2: Odds ratios (left column) and empirical distribution of $RERI$ obtained via bootstrap (right column) for the three melanoma data applications. Rows 1, 2, and 3 show the results for sun exposure corresponding to beach and water activities from age 15, average annual lifetime ambient UV, and early life ambient UV, respectively. In the right column, the vertical bold black line shows the estimated $RERI$, the black dashed lines are the 95% confidence intervals, and the dashed red line denotes the benchmark value of $RERI = 0$.