Supplementary Material

S1: Statistical Tests for Individual Variables

Given a dataset with variables from all sites, the first step is to conduct appropriate statistical tests, for each variable, to check if there is any site that is different from others. Here, we focus on two published statistical models, namely likelihood ratio test (LRT) and random effects model, for count and continuous variables, respectively.

- **Test based on LRT for Count Variables**

For a binary variable denoting whether a patient may have a serious adverse event (SAE), a non-serious adverse event (NSAE), a protocol deviation, or death, the data from all subjects within a site can be summarized by a count variable (Var_j) as shown in Table S1, where the first column represents the site ID, the second column represents the # of patients with at least one event of Var_j, in each of the sites, and the third column represents the total # of patients in the corresponding site.

Table S1. Data structure for a count variable

Site ID	# of patients with Var_j	Total # of patients
1	n_{1j}	$n_{1\cdot}$
⋮	⋮	⋮
i	n_{ij}	$n_{i\cdot}$
⋮	⋮	⋮
I	n_{Ij}	$n_{I\cdot}$
Total	$n_{\cdot j}$	$n_{\cdot\cdot}$

With data described in the above count table, the likelihood ratio test (LRT) method developed by Huang et al.[3] can be applied to identify the sites that are different from the other sites for Var_j.

We assume that $n_{ij} \sim^{ind} Pois(n_{i\cdot} p_{ij})$, and $n_{\cdot j} - n_{ij} \sim^{ind} Pois((n_{\cdot\cdot} - n_{i\cdot})q_{ij})$ for all sites i =1,...,I, where p_{ij} is the parameter representing the risk/rate of site i for Var_j and q_{ij} is the parameter representing the risk/rate of all the other sites combined, except site i. For the jth count variable (i.e., Var_j), we test the global null hypothesis $H_0 : p_{ij} = q_{ij} = p_0$ for all $i = 1, \ldots, I$, where p_0 is unknown common value, against the alternative hypothesis $H_a : p_{ij} > q_{ij}$ for at least one i.

The likelihood ratio for site i, fixed at Var_j, is $LR_{ij} = \frac{L_{H_a}}{L_{H_0}} = \frac{(\hat{p}_{ij})^{n_{ij}} (\hat{q}_{ij})^{n_{\cdot j} - n_{ij}}}{(\hat{p}_0)^{n_{\cdot j}}} = \frac{(\frac{n_{ij}}{n_{i\cdot}})^{n_{ij}} (\frac{n_{\cdot j} - n_{ij}}{n_{\cdot\cdot} - n_{i\cdot}})^{n_{\cdot j} - n_{ij}}}{(\frac{n_{\cdot j}}{n_{\cdot\cdot}})^{n_{\cdot j}}}$ and the log likelihood ratio is $LLR_{ij} = n_{ij} \log(\hat{p}_{ij}) + (n_{\cdot j} - n_{ij}) \log(\hat{q}_{ij}) - n_{\cdot j} \log(\hat{p}_0)$.

The likelihood ratio test statistic for testing H_0 versus H_a is the maximum likelihood ratio ($MLLR_j$) across all sites, $i = 1, ..., I$, where $\hat{p}_{ij} > \hat{q}_{ij}$, i.e. $MLLR_j = \max_i (LLR_i) I(\hat{p}_{ij} > \hat{q}_{ij})$. A Monte Carlo (MC) simulation is used to obtain the empirical distribution of $MLLR_j$. The p-values for each site is determined by ranking the observed LLR_{ij} in the empirical distribution of $MLLR_j$.

- **Test based on Random Effects Model for Continuous Variables**

For continuous variables, such as BMI, blood pressure, baseline continuous biomarker, and medical device success rate, etc., the random effects model developed by Desmet et al.[5] can be applied to identify sites with significantly different values of mean when compared with other

sites. Using a random effects model, for a continuous variable, the outcome for subject s in site i is modeled as $y_{is} = \mu + \gamma_i + \varepsilon_{is}$, with $\gamma_i \sim N(0, \tau^2), \varepsilon_{is} \sim N(0, \sigma^2)$, where μ is the fixed effect, γ_i are the site-level random effects, and ε_{is} are the random errors. The sample mean for site i is then $y_{i\cdot} = \frac{1}{n_{i\cdot}} \sum_{s=1}^{n_{i\cdot}} y_{is} \sim N(\mu, \tau^2 + \frac{\sigma^2}{n_{i\cdot}})$, so that $y_{i\cdot} - \mu \sim N(0, \tau^2 + \frac{\sigma^2}{n_{i\cdot}})$. Let the maximum likelihood estimates (MLEs) of μ, τ^2 and σ^2 be $\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2$. The p-value of one-sided test (for $H_0 : \mu_i = \mu_0$ vs. $H_a : \mu_i > \mu_0$) for site i is calculated as $p_i = P(\mathrm{Z} \geq \frac{y_{i\cdot} - \hat{\mu}}{\sqrt{\hat{\tau}^2 + \frac{\hat{\sigma}^2}{n_{i\cdot}}}})$; and the p-value of two-sided test ($H_0 : \mu_i = \mu_0$ vs. $H_a : \mu_i \neq \mu_0$) for site i is calculated as $p_i = 2 * P(\mathrm{Z} \geq abs(\frac{y_{i\cdot} - \hat{\mu}}{\sqrt{\hat{\tau}^2 + \frac{\hat{\sigma}^2}{n_{i\cdot}}}}))$.

In situations where the subject-level data are not available and only site-level summary data are provided, i.e., when for each site i, only the site size n_i, sample mean $y_{i\cdot}$ and sample variance s_i^2 are available, other parameter estimation methods such as ANOVA approach or DerSimonian-Laird (DL) approach[6], can be used to estimate the reference distribution, and then the corresponding statistical tests for each site and variable pair can be carried out. Lastly, to control the false discovery rate (FDR) across the sites/rows in the presence of multiple comparisons (I), "Benjamini Hochberg (BH)" method[8] is performed, for each variable, to adjust the raw p-values.