Formal Description of the GIST 2.0 Algorithm

Anando Sen

All the material described in this document is from our paper **"GIST 2.0: A Scalable Multi-Trait Metric for Quantifying Population Representativeness of Individual Clinical Studies"** published in the **Journal of Biomedical Informatics** in 2016. This document only provides the algorithm for computing GIST 2.0 scores. For greater understanding of the methodology and mathematical properties and validations of GIST 2.0, please consult the paper.

Sen, A., Chakrabarti, S., Goldstein, A., Wang, S., Ryan, P. B. and Weng, C. GIST 2.0: A scalable multi-trait metric for quantifying population representativeness of individual clinical studies. J. Biomed. Inform. 63, 325-336 (2016). (PubMed: 27600407 PMC: PMC5077682 DOI: 10.1016/j.jbi.2016.09.003)

The GIST 2.0 scores are always between zero and one with higher scores implying greater population representativeness. For a clinical study with n study traits this algorithm outputs $n + 1$ values: the sGIST of the n study traits and the mGIST of the study. An mGIST score of a trial estimates the fraction of diabetic patients that would be eligible for that trial. An sGIST score of a trait estimates the fraction of diabetic patients that would be satisfy the eligibility criterion for that trait in that clinical trial.

The GIST 2.0 algorithm uses a non-uniform significance (importance) for each eligibility trait. The clinical significance of each eligibility trait is different and may vary by trial. We use a trial-specific significance scale based on stringency where the traits with greater stringency had higher significance. For example, in a trial with HbA1C criterion greater than 6.5% the significance of HbA1C is lower than a trial requiring HbA1C between 9 and 11%. While the former criterion is aimed at ascertaining that a prospective patient is indeed diabetic, the latter plays the additional role of further restricting the diabetic patient population to a smaller subgroup, relevant to the objective of the trial. Formally, in a trial T and for trait f_j, let the fraction of patients in the EP satisfying its eligibility criteria be r_j. Then the significance of f_j is calculated as $s_j = 1 - r_j$.

In addition to the formal description of the GIST algorithm, we explain each step with a simple example. In this example we use only two study traits, HbA1C and glucose (referred to as f_1 and f_2) of a clinical trial (Clinical trial ID: NCT00570739). Figure 1 is used to visualize some steps of the algorithm. As continuous study traits have different ranges, we begin by standardizing them.

1. Standardize the summarized continuous traits of every patient using the mean and the standard deviation (SD). For any continuous trait f_j, let μ_j and σ_j denote its mean and SD over the EP patients. Then the standardized patient is $\tilde{P}_i = (\tilde{f}_1^i, \tilde{f}_2^i, \ldots, \tilde{f}_n^i)$, with $\tilde{f}_j^i = \frac{f_j^i - \mu_j}{\sigma_j}$ for continuous traits and categorical traits remaining unchanged ($\tilde{f}_j^i = f_j^i$).

   In the example - The means and standard deviations of f_1 and f_2 were calculated: $\mu_1 = 7.14$, $\sigma_1 = 1.65$, $\mu_2 = 165.81$, $\sigma_2 = 41.15$ and used for the standardization.

2. Calculate the significance s_j, of each trait as described above. For a patient with standardized traits $\tilde{P}_i$, compute its significance-scaled form by $\hat{P}_i = (\hat{f}_1^i, \ldots, \hat{f}_n^i) = (s_1\tilde{f}_1^i, \ldots, s_n\tilde{f}_n^i)$.

   In the example - The significance scales s_1 and s_2 were calculated to be 0.45 and 0.08 respectively. These were used to compute $\hat{f}_1$ and $\hat{f}_2$. The scatter plot between the standardized and significance-scaled patients ($\hat{f}_1$ and $\hat{f}_2$) is shown in Figure 1. For visual clarity only a random sample of 300 patients are displayed in this figure. Note that due to the lower significance of f_2, its range is much lower than that of f_1.

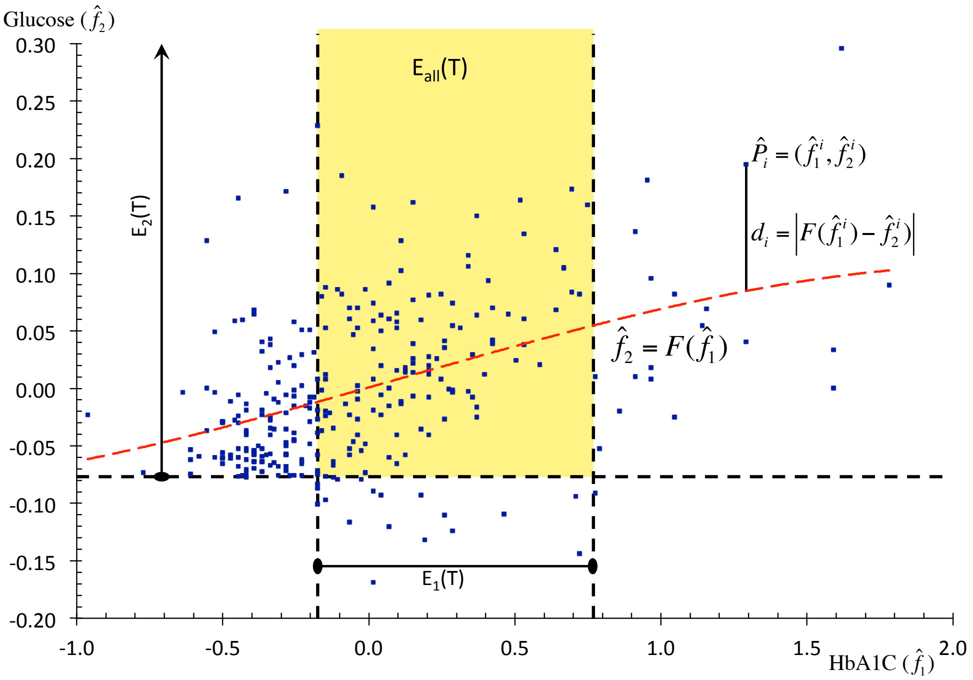

Figure 1: An example illustrating the GIST 2.0 methodology with two traits. The regression curve of dependency is shown by the red dashed curve. The eligibility criteria are marked by black dashed lines and points in the yellow region represent eligible patients.

3. Use the standardized and significance-scaled patients $\hat{P}_i$ to compute a non-linear regression hyper-surface F with one of the continuous traits (e.g. $\hat{f}_n$) designated as the dependent variable (and all other traits as independent variables), i.e. $\hat{f}_n = F(\hat{f}_1, \ldots, \hat{f}_{n-1})$.

   In the example - The hyper-surface $\hat{f}_2 = F(\hat{f}_1)$ was calculated in Matlab 2015b. It is shown by the red dashed curve in Figure 1.

4. For every patient (standardized and significance-scaled) $\hat{P}_i$, calculate its residual distance $d_i = |F(\hat{f}_1, \ldots, \hat{f}_{n-1}) - \hat{f}_n|$ from the hyper-surface F.

   In the example - In Figure 1, this distance is shown for one patient P_i. For this patient $d_i = 0.13$.

5. Assign a weight w_i to every patient P_i that is inversely proportional to the patient's distance d_i from F, i.e. $w_i = \frac{1}{1+d_i}$. (Note that $w_i = 1$ when $d_i = 0$).

   In the example - For the patient P_i in Figure 1 the corresponding d_i would be 0.88.

6. For a trial T, let the inclusion range (as given in the eligibility criteria) for trait f_j be denoted by the set $E_j(T)$. The notation $f_j^i \in E_j(T)$ implies that the jth trait of patient P_i falls within the corresponding inclusion range of the eligibility criteria. Now, the single-trait GIST score $sGIST_j$ for trait f_j is computed as,

$$sGIST_j(T) = \frac{\sum\limits_{f_j^i \in E_j(T)} w_i}{\sum\limits_{f_j^i} w_i} \tag{1}$$

   In the example - The eligibility criteria for HbA1C and glucose were $6.5 \leq f_1 \leq 10.5$ and $f_2 \geq 126$ respectively.

In the standardized and significance-scaled coordinates these transform to $-0.17 \leq \hat{f}_1 \leq 0.78$ and $\hat{f}_2 \geq -0.07$ respectively. These intervals form the sets $E_1(T)$ and $E_2(T)$. They are marked in Figure 1 with capped dashed lines. Elliptical caps imply bounded range and arrow-head caps imply unbounded range.

701 patients satisfied the criteria for f_1 and 1213 for f_2. The total weight of patients lying in E_1 was 642.75 and for those lying in E_2 was 1130.20. The total weight of all patients was 1231.71. Hence, $sGIST_1(T) = \frac{642.75}{1231.71} = 0.52$ and $sGIST_2(T) = \frac{1130.20}{1231.71} = 0.92$.

7. The multi-dimensional (overall) eligibility criteria of T is the logical combination of the eligibility criteria of the individual traits. Let the volume defined by the multi-dimensional eligibility criteria be denoted by $E_{all}(T)$ (e.g. $E_{all}(T) = E_1(T)\ AND\ [E_2(T)\ OR\ E_3(T)]$). A patient P_i satisfying the overall eligibility criteria for T is said to belong to this volume i.e. $P_i \in E_{all}(T)$. Now, the multiple-trait GIST score $mGIST$ of T is computed as,

$$mGIST(T) = \frac{\sum\limits_{P_i \in E_{all}(T)} w_i}{\sum\limits_{P_i \in EP} w_i} \tag{2}$$

In the example - The set $E_{all}(T)$ is the intersection of $E_1(T)$ and $E_2(T)$. 492 patients lie within $E_{all}(T)$ with a combined weight of 448.33. As the total weight of all patients was 1231.71, we get $mGIST(T) = \frac{448.33}{1231.71} = 0.36$.