The Rand index compares two clustering results [1]. However, to control for randomness and compare clustering scores from clusters of different sizes, the adjusted Rand index (ARI) is a more suitable metric [1]. The ARI is 1 for a perfect clustering that matches the true cluster labels.

On the other hand, a score close to 0 indicates a poor clustering. Based on the assumption of tumor characteristics in patients with low risk in HPV- and patients with intermediate risk in HPV-, we simulated patients change in outcomes using normal distribution as illustrated in Table 1. The change in outcome is .

|  |  |  |
| --- | --- | --- |
| Radiographic Signal | HPV- Low Risk | HPV- Intermediate Risk |
| Volume change | Normal (0.92, 0.05) | Normal (0.80, 0.05) |
| PET SUV max | Normal (0.55, 0.03) | Normal (0.45, 0.03) |
| Diffusion | Normal (1.18, 0.02) | Normal (1.12, 0.02) |

Table 1: Simulation parameters for change in outcomes in HPV- patients.

To evaluate our operating characteristics of the pilot study, we performed 10,000 different simulations. We used the ARI to evaluate clusters of the simulated data where the true cluster labels are known. We assumed the patients that enroll in the study have 50% chance of either being HPV- low risk and HPV- intermediate risk. To study the power of the design, we simulated the outcomes using parameters in Table 1 with 20 patients. We used k-means clustering to cluster the data into 2 groups. We compare the simulated cluster results with the patient assignment of risk. If the ARI exceeds 0.75, we assume the data were well clustered. Replicating the experiment 10,000 times, we obtain a power of 0.878. Similarly, to test type-I error, we enroll 20 patients with 50% chance of being assigned to the low and intermediate risk groups. However, all 20 patients were simulated using HPV- low risk parameters and used k-means clustering to cluster the data into two groups. We compare the clustered data with patient’s assignment of treatment. If the ARI exceeds 0.25, we assumed the data is clustered reasonably well to the patient assignment. The type-I error rate obtained from the simulation of 10,000 different simulations is 0.013.

1. *Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), pp.846-850.*