We have shared the site names and additional considerations with the client. This documents is for illustrative purposes only

Strategic site selection

The task

This report is a part of a larger exercise of identifying optimal clinical trial sites for a high-complexity early oncology trial targeting 7 cancer types in the dose expansion phase. We can't share details on why this study was crucial for our client but it was of utmost importance to select a small number of sites and ensure that they are able to carry all 7 arms. By leveraging Principal Component Analysis (PCA) and K-Means clustering, we synthesized high-dimensional performance data into a clear, actionable selection model

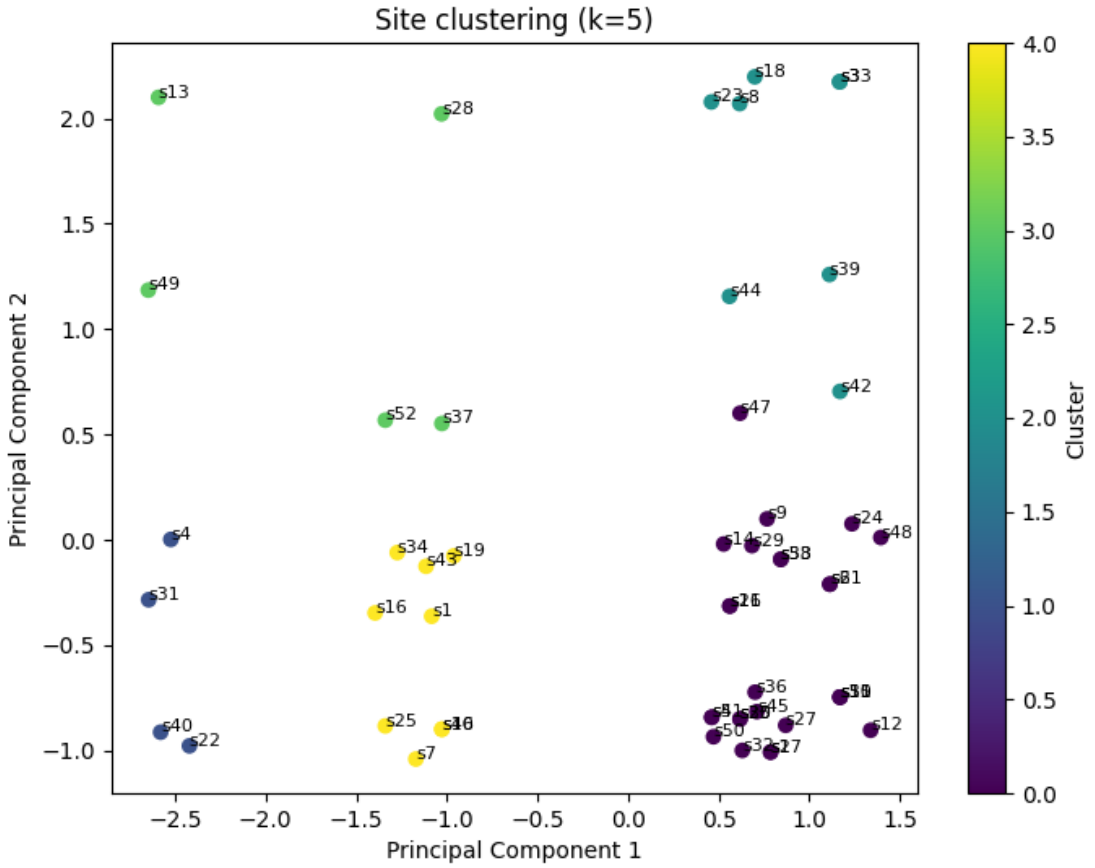
Methodology and Pre-Selection

Before the application of PCA, the initial pool of sites underwent a stringent pre-qualification process. All sites represented in this analysis were selected based on:

- **High Historical Enrollment:** A proven track record of patient recruitment in early-phase oncology
- **Capacity Analysis:** Comprehensive evaluation of site capacity and existing trial load to ensure the capability to support a multi-arm desig
- **Timelines assessment:** IRB type, contracting and startup timelines

After coming up with a short list of sites fitting all criteria, we used PCA to reduce the complexity of the dataset into two orthogonal vectors that capture the majority of the data's variance. This allows us to visualize "performance fingerprints" in a two-dimensional space.

- **Principal Component 1 (PC1 - "Overall" axis):** PC1 represents the primary driver of variance across the study. In this context, it aligns with what can be called universal capacity. Sites with higher PC1 values show a high level of engagement and recruitment across all the target indications. It is an excellent metric of the oncology enrolling power of the site
- **Principal Component 2 (PC2 - "Specialization" axis):** PC2 captures the secondary variance, which in the context of this dataset and study shows sites having a particulary strong performance in one or few indications and weak performance in the rest. A high PC2 value suggests a lack of balance we would recommend to exclude (if possible) from a small study with a very limited space for mistakes and inefficiencies



3. Visual interpretation of results

Utilizing K-Means clustering and after discussing with the client what is the sweet spot between detailed output and hypersegmentation, we have mathematically divided the pre-selected sites into 5 distinct operational profiles:

- **Cluster 0 (Deep Purple):** The "Universal Performers." High PC1 stability with controlled PC2 variance
- **Cluster 1 (Indigo):** Underperformers or "Low-Volume Specialty" sites on the far left.
- **Cluster 2 (Teal):** "High-Variance Specialists" located in the top-right; high output but potentially inconsistent across different cancer types
- **Cluster 3 and 4 (Green/Yellow):** Outliers and mid-range performers that lack the requisite density for core study inclusion

Recommendation: Prioritise sites from Cluster 0, keep 1-2 sites from Cluster 2 as a backup to supplement in case of enrollment lag in the smaller indications

- Cluster 0 consists of highly predictable sites
- They are an excellent fit for the main goal to run a highly complex trial in as few sites as possible
- This strategy mitigates the risk of an arm creating a bottle neck but the choice to pre-select 1-2 sites form Cluster 2 adds more flexibility without a significant overall cost increase