**Study Design**

1. **Background & Rationale**
   * Ovarian serous cystadenocarcinoma: incidence, mortality, need for molecular prognostic markers.
   * BRCA1 (and other homologous-recombination genes) known to influence outcome.
2. **Objectives & Questions**
   * *Exploratory:* Which genes’ expression levels are most strongly associated with overall survival in TCGA-OV?
   * *Clinical adjustment:* Which clinical features (stage, age, residual disease…) should be included in a prognostic model?
   * *Joint model:* Among top survival-associated genes + selected clinical covariates, which are independent risk (HR>1) or protective (HR<1) factors?
3. **Design**
   * Retrospective cohort analysis of TCGA-OV (n≈ 421), with right-censored OS data.
   * merged dataset for entire workflow.

* **Descriptive statistics & hypothesis tests** ✓

**Descriptive Statistics & Hypothesis Tests**

* 1. **Cohort overview**
     + Table 1: patient demographics and clinical features (n, %, mean±SD or median [IQR]).
     + Use χ²/Fisher’s exact for categorical comparisons (e.g. stage by BRCA1 high/low), t‐tests or Wilcoxon for continuous.
  2. **Survival summary**
     + Overall Kaplan–Meier curve for TCGA-OV.
     + Median OS (with 95% CI).
  3. **Exploratory gene expression**
     + Volcano plot of –log₁₀(p) vs HR for univariate Cox of all genes (highlight top 5).
     + KM plots (high vs low) for top 3 genes.
* **Survival models** ✓

**Modeling Techniques**

Must include at least one: Time-to-event models & Logistic Regression

**1.1 Survival Models**

* 1. **Step 1: Univariate Cox scan**
     + For each gene, fit Cox PH; adjust p’s via BH-FDR.
     + Select top 3–5 by q < 0.05 (or lowest q).
  2. **Step 2: Clinical feature selection**
     + Fit a **LASSO-penalized Cox** on clinical variables only (cross‐validated λ).
     + Identify nonzero coefficients → parsimonious adjustment set.
  3. **Step 3: Multivariable Cox**
     + Include selected gene(s) + LASSO‐chosen clinical covariates.
     + Report adjusted HRs (95% CI) and p-values.
* **Logistic regression** ✓

**2.1 Logistic Regression (secondary)**

For each top gene, define “high vs low expression” (median split).

Fit logistic model:

$logit[P(high gene)]=β0+β1(stage)+β2(CNV burden)+⋯$

Interpret ORs: what clinical/genomic factors predict high expression of a survival-relevant gene.