

Analysis Plan Draft

Project Title: Genetic Variation in *ABC* and Longitudinal Fatigue in Breast Cancer Survivors

Authors: Dr. Lacey W. Heinsberg (Nursing, Human Genetics); Dr. Naomi Judge (Nursing); Dr. Tobias Vector (Biostatistics); Ms. Pepper Linear (Oncology); Dr. Quinn Marshall (Nursing)

Background: Fatigue is common among individuals undergoing or recovering from breast cancer treatment. We have longitudinal symptom data (fatigue measured 0-50 at 4 timepoints) and candidate SNPs in the *ABC* gene.

Purpose: Determine whether genetic variation in the *ABC* gene (rs1/rs2/rs3) is associated with fatigue over time.

Overview:

- Participants: N~500 breast cancer survivors
- Time: 4 repeated observations
 - T0 (baseline, pre-intervention)
 - T1 (6 months, intervention completion)
 - T2 (12 months)
 - T3 (18 months)
- Outcome: Fatigue (continuous; 0-50)
- Predictors: rs1, rs2, rs3 (coded additively, 0/1/2 based on number of copies of the minor allele)
- Covariates: age, race, education
- Missingness: likely MAR or MCAR; mild imbalance expected

Analysis plan: Linear mixed model with random intercepts:

$$\text{Fatigue} \sim \text{rs1} * \text{Time} + \text{age} + \text{race} + \text{education} + (1|\text{ID})$$

Rationale:

- Outcome is continuous
- Repeated within-subject observations
- Correlation among observations within a subject must be modeled
- LMM is interpretable, familiar, and robust to missingness under MAR

Assumptions:

- Residuals are approximately normal
- Missing data mechanism approximately MAR
- Random intercepts are sufficient to capture within-ID correlation
- Covariate measurement is unbiased

Alternative approaches we could consider:

- GEE using an exchangeable correlation structure
 - Not selected as primary approach because GEE estimates population-averaged effects, which are less aligned with our scientific question (inter-individual differences in trajectories). Also less convenient for plotting individual-level predicted curves, which are central to our interpretation.

Alternative approaches considered but not selected:

- Repeated-measures ANOVA: assumptions are quite restrictive, less flexible missing data handling
- Cross-sectional analyses at a single timepoint: would discard longitudinal richness, which is a shame because the sample size supports a more complex analysis

Planned sensitivity checks:

- Compare LMM vs GEE to evaluate consistency of direction/significance
- Evaluate alternative coding of genotypes (dominant/recessive)

Plan / deliverables:

- Develop R Markdown using synthetic data first and AI support
- Full reproducible code scaffold before real data is accessed
- Final analytic report to co-authors prior to manuscript results section drafting

Abbreviations:

Term, Meaning

ABC, Candidate gene of interest

ANOVA, Analysis of Variance

GEE, Generalized Estimating Equation

ID, Individual participant identifier

LMM, Linear Mixed Model

MAR, Missing At Random

MCAR, Missing Completely at Random

SNP, Single Nucleotide Polymorphism

Declaration of generative AI use: ChatGPT-5 was used to draft this document based on the author's scientific reasoning, and to support organization, editing, and formatting for clarity. The author reviewed and edited all content and takes full responsibility for the final plan.