**Case-only analysis (early-onset vs late-onset cancers)**

The first thing to do now is a case-only analysis. Filter the dataset to include only participants with cancer (ever\_cancer == 1). Within this subset, the outcome variable becomes early\_onset\_proxy, which distinguishes early-onset cases (<50 years old) from later-onset cases (≥50 years old). The predictors should include standardized PHQ-9 (dpq\_total\_std), standardized CRP (log\_hscrp\_std), smoking status, sex, education level, race/ethnicity, and income (PIR).

What to look for here are the odds ratios for PHQ-9 and CRP. If the ORs are clearly greater than 1, this means that among cancer cases, early-onset patients are more stress- and inflammation-linked than late-onset patients. If the ORs are close to 1, it would suggest there is little difference in stress or CRP profiles between early- and late-onset cancers. This step is important because it directly asks whether early-onset cancers have distinct stress and inflammation characteristics compared with cancers diagnosed later in life.

**Sensitivity analysis: exclude possible acute inflammation**

Next, re-run the continuous model in the <50 group but exclude individuals with very high CRP values (hscrp\_mg\_l\_raw > 10). Extremely elevated CRP can be caused by acute infections rather than chronic inflammation, and keeping them may distort the signal.

Here the main thing to check is whether the odds ratio for CRP (log\_hscrp\_std) changes meaningfully compared to the original <50 model. If the CRP effect stays above 1, then the association is robust and not just driven by a few extreme cases. If the CRP effect drops substantially, then it suggests that outliers with very high CRP may have been driving the signal. Reporting this shows that you have tested the stability of the result and adds credibility to the findings.

**Younger cutoff analysis (<45 and <40)**

After that, repeat the continuous model within progressively younger age subsets, specifically <45 and <40. Start with the <50 dataset, then filter further by age to create these subsets. Run the same model each time with PHQ-9, CRP, smoking, sex, education, race/ethnicity, and income.

What you want to check are the odds ratios for PHQ-9 and CRP, as well as the AUC values. The hypothesis is that the associations will become stronger in the very young — that is, odds ratios larger than in the <50 group and AUC values slightly higher. If you see that, it supports the idea that stress and inflammation matter most for the youngest cancer patients. If the results are unstable with wide confidence intervals, that probably reflects small sample sizes, which is fine to note.

**Stratified analysis (effect differences across groups)**

Finally, run the <50 continuous model separately in subgroups to see if associations look different by sex or race/ethnicity. Start by splitting the <50 group into males and females, and run the model in each. If sample sizes permit, also run it within the major race/ethnicity categories: White, Black, Hispanic, and Asian/Pacific Islander.

The goal here is to compare the odds ratios for PHQ-9 and CRP across these groups. If one group shows much stronger effects, that suggests stress or CRP may play a more important role in that demographic. For example, if PHQ-9 is strongly predictive in women but not in men, or if CRP shows a clearer signal in one racial group than others, that is worth reporting. If the odds ratios are broadly similar across groups, the interpretation is that the effects are consistent regardless of sex or race. This adds depth to the analysis and helps address whether the observed patterns are generalizable or concentrated in certain populations.

**Share the code once you reach here and try to do the next steps**

**Propensity score weighting (robustness check for <50 group)**

As a more advanced step, implement propensity score weighting to make sure the stress and CRP signals are not just due to confounding factors like sex, race, socioeconomic status, or smoking. Define two exposures separately: high stress (dpq\_total >= 10) and high CRP (hscrp\_mg\_l\_raw > 3). For each exposure, fit a logistic regression where the exposure is the outcome and the predictors are sex, race/ethnicity, education level, income (PIR), and smoking status. This gives you a propensity score — the probability that someone would be exposed based on these confounders.

From this score, calculate inverse probability weights: if a person is exposed, their weight is 1/PS; if not exposed, their weight is 1/(1-PS). Then fit a weighted logistic regression with cancer (ever\_cancer) as the outcome and the exposure (high stress or high CRP) as the predictor.

What you are looking for is the weighted odds ratio for stress and CRP. If the weighted ORs remain similar to the unweighted ORs and still greater than 1, it means the associations are robust and not explained away by confounders. If the weighted ORs shrink toward 1, then confounding is likely responsible for some of the observed effect. This analysis is critical for making the findings credible in a publication, since reviewers will want to see that potential confounding has been addressed.