**What’s inside (structure)**

* R/00\_setup\_packages.R — installs/loads everything.
* R/01\_config.R — set AOU\_PROJECT, AOU\_DATASET, and a writeable WORK\_SCHEMA.
* R/10\_build\_concepts.R — stubs for SOGI and cancer concept sets (replace with real concept\_ids).
* R/20\_extract\_cohort.R — runs the SQL to materialize SOGI flags, base cohort, and incident cancer tables.
* R/30\_covariates\_and\_ps.R — builds covariates table, computes PS, 1:1 matching, Love plot.
* R/40\_models\_primary\_secondary.R — Poisson/NB IRR and Cox HR; exports tables + crude incidence figure.
* R/50\_sensitivity\_robustness.R — overlap weighting (ATO) sensitivity IRR.
* R/60\_figures\_tables.R — formats tables; saves markdown and images.
* sql/20\_sogi\_exposure.sql, sql/21\_base\_cohort.sql, sql/22\_outcome\_incident\_cancer.sql, sql/30\_covariates.sql — BigQuery Standard SQL templates.
* report/analysis\_report.qmd — Quarto report that pulls results and renders a one-click methods appendix.
* outputs/ — where figures/tables land.
* README.md — quickstart.

**Runbook (Workbench RStudio)**

1. Open RStudio in AoU workspace.
2. Edit R/01\_config.R: set AOU\_PROJECT, AOU\_DATASET (the Controlled Tier CDR ID), and WORK\_SCHEMA (a writeable dataset).
3. Run once:
   1. source("R/00\_setup\_packages.R")
4. Populate concept IDs in R/10build\_concepts.R (see next section), then:
   1. source("R/10\_build\_concepts.R")
   2. source("R/20\_extract\_cohort.R")
   3. source("R/30\_covariates\_and\_ps.R")
   4. source("R/40\_models\_primary\_secondary.R")
   5. source("R/50\_sensitivity\_robustness.R")
   6. source("R/60\_figures\_tables.R")
   7. quarto::render("report/analysis\_report.qmd")
5. Still need to swap out placeholders (left explicit hooks to fill in real concept IDs when we pull them from AoU dataset creation
   1. **SOGI** (R/10\_build\_concepts.R + sql/20\_sogi\_exposure.sql):
      1. sogi\_orientation\_hetero, sogi\_orientation\_lgb
      2. sogi\_gender\_cis, sogi\_gender\_trans\_nb
      3. Use ATLAS/Athena to export concept sets; paste the integer concept\_ids into the vectors. Update the IN UNNEST([…]) lists in the SQL accordingly.
   2. **Cancer phenotype**:
      1. cancer\_condition\_concepts should be SNOMED malignant neoplasm concept\_ids (exclude non-melanoma skin if that’s our analytic rule).
      2. If we want “confirmatory evidence” logic (oncology drugs/procedures) beyond the simple first-event rule in 22\_outcome\_incident\_cancer.sql, extend the CTE by joining to drug\_exposure (ATC L01/L02) and procedure\_occurrence with our CPT/HCPCS lists.
   3. **Covariates** (sql/30\_covariates.sql):
      1. Replace stubs for preventive\_any, insurance, region, and cci with our exact OMOP concept logic.
      2. If we have ADI or RUCA/RUCC derivations, add them here.
6. Defaults (aligned to the SAP from 10.22)
   1. **Eligibility**: 2-year wash-in, cancer-free at baseline, at least 1 encounter in the year after index.
   2. **Exposure lock**: first affirmed SOGI date as index; treat as ITT.
   3. **Matching**: 1:1 nearest neighbor, logit-PS caliper = 0.2 SD, no replacement, common support enforced.
   4. **Inference**: Poisson w/ log(person-years) offset; auto-switch to NB if overdispersion >1.5; pair-clustered robust SEs.
   5. **Sensitivity**: overlap weighting (ATO) via WeightIt, quasipoisson with survey design.
   6. **Competing risk**: scaffold for Fine-Gray via cmprsk (add fstatus 1/2 in your data; model in a site-specific script if your Workbench image supports the package)
7. **Extensions needed:**
   1. **Site-specific analyses**: parameterize a vector of site concept sets and loop the model scripts to emit per-site IRRs/HRs with BH FDR q-values.
   2. **IPCW** for differential follow-up: add a small script to estimate censoring weights and refit.
   3. **Negative controls & E-values**: glue in EValue package and one or two negative-control outcomes to probe residual confounding.