Extended data for “Non-AIDS defining malignancies in the combination ART era: immunological and socio-behavioral risk factors”, Ruffieux et al, F1000Research 2019

The data management and analysis were done with R. We have three sources of data:

1) the SHCS data,
2) results of the linkage to the cancer registry,
3) the nation-wide cancer rates (NICER)

Data management

pilot.R – run this to initialize the data preprocessing, for a given cancer (captured through the *tumor_restriction* variable, which can be, for example, “lung”, “anus”, “prostate”, “liver”). It executes the three programs described below:

preprocess_SHCS.R – loads the relevant SHCS data (patient characteristics, lab values, etc.).

preprocess_linkage.R – loads the linkage data (cases of cancer that could be linked to the SHCS). Cancers are identified based on the ICD codes (cancer registries) or custom categories (SHCS). Splits the data into two portions, those who develop the cancer (as specified by *tumor_restriction* in pilot.R) and those who don’t.

merge.R – appends the patient-level information, and creates a single data frame (*data_res_mrg)* with one line per patient.

risk_factors_monthly.R – creates a large dataset with one line per month, per patient. Time-updated variables are created here (CD4, RNA, etc). This is done independently of the above four files, and does not need to be repeated for every cancer.

write_res_df.R – sets up the data for the survival analysis for a specific cancer, using the ‘one line per month’ dataset created by risk_factors_monthly.R. Must run pilot.R (with the appropriate tumor restriction) first. Creates a *res_df* data frame which will be used for the survival analysis, for a specific cancer.

Analysis

cancer_counts.R – produces the counts (number of cancers per data source) reported in Table 1.

patient_characteristics.R – produces results in Table 2. Requires *res_df* object described above, for the appropriate cancer.

crude_incidence_rates.R – produces the incidence rates for the NADMs. Requires the *res_df* object described above, for the appropriate cancer.

SIR_NADCs.R – produces the SMRs for NADMs (Figure 2). Requires Excel files with Swiss cancer rates (www.nicer.org), and the *data_res_mrg* data frame for the appropriate tumor (created with pilot.R).

impute_smoking.R – multiple imputation of smoking variable, only requires running pilot.R to load in the patient-level data. Output is saved for use in the survival analysis.

Cox_model_selection.R – produces that AICs in Table 3. Must first run risk_factors_Cox.R (with the appropriate tumor restriction) once.

risk_factors_Cox.R – Cox regression to produce results in Tables 4 and 5. Requires the *res_df* object described above (for the appropriate cancer), and the result of the multiple imputations for smoking. Exposure variables (with amount of lag/moving average) to include must be specified via the *base_factors* variable. For example “cd4_ma12_lag24” means I include CD4, lagged 24 months, 12 month SMA, as a risk factor. For univariate regressions of exposure variables, adjusted for the ‘other’ variables (-> Table 4) set the *univariate_reg* variable to TRUE.