**The description of the processing of data starts** with a SAS dataset **covariate\_file\_020922.sas7bdat** that contains variables of interest defined based on read codes for conditions observed 1-10 years and 10-32 years before the index date (diagnosis or selection). Variable names incorporate position in hierarchy, definition sensitivity (1 versus 2+ codes within time frame) and timing of observation (1-10 versus 10-32 years).

**Descriptions of R code used to process and analyze data**

**DATA PROCESSING FILES**

**GOLD Data processing files.**

1. **FirstProcessData.R**
   1. **Purpose:** Process GOLD population SAS file
   2. **Input:** A SAS datafile
   3. **Output**: R file used in future analyses (FullDSAgeProper.Feb2022.RData )
   4. **Variables contained within SAS file include**
      1. Demographic and lifestyle information.
      2. Registration information.
      3. Selection or Diagnosis date.
      4. **Medical condition flags.** These flags indicating the presence of particular diseases were named to accommodate the timing and specificity of definition. The name also contains information regarding the hierarchy. See supplemental Table 1 for hierarchy of condition and associated variable names. SAS variable names could only be 32 characters long and that impacted the variable name.
         1. Sensitive condition definition flag = 1+ observation of condition within time frame (1-10 years prior to diagnosis (**primary analysis**))
            1. Variable name = “CONDITION NAME AS WRITTEN ON EXCEL SHEET IN SUPPLEMENTAL INFORMATION (Supplementary Material S1\_DiscoveryGOLDCodes.xlsx)” + “\_lt10”.
            2. Date of the first diagnosis in 1-10 year before index date. Date variable name = “CONDITION NAME AS WRITTEN ON EXCEL SHEET IN SUPPLEMENTAL INFORMATION (Supplementary Material S1\_DiscoveryGOLDCodes.xlsx)” + “\_d”.
         2. 10-32 years prior to diagnosis (**secondary analyses**)
            1. Variable name = “CONDITION NAME AS WRITTEN ON EXCEL SHEET IN SUPPLEMENTAL INFORMATION (Supplementary Material S1\_DiscoveryGOLDCodes.xlsx)” + “gt10”
            2. Date of the first diagnosis in 1-10 year before index date. Date variable name = Date variable name = “CONDITION NAME AS WRITTEN ON EXCEL SHEET IN SUPPLEMENTAL INFORMATION (Supplementary Material S1\_DiscoveryGOLDCodes.xlsx)” + “\_d”.
         3. Alternate exposure definitions (\_alt\_lt10, \_alt\_gt10) (not used for this project):
2. **Medication use flags.** These were named to accommodate timing and specificity of definition. SAS names could only be 32 characters long and that impacted the naming of variables. There were no hierarchies related to this definition
   * 1. Sensitive definition flag (1+ observation of condition within time frame)
        1. 1-10 years prior to diagnosis (primary analysis)
           1. “MEDICATION NAME AS WRITTEN ON EXCEL SHEET IN SUPPLEMENTAL INFORMATION” (Supplementary Material S3\_MedicationsGOLD.xlsx)
           2. Date of the first observed medication in 1-10 year before index date. Date variable name = “CONDITION NAME AS WRITTEN ON EXCEL SHEET IN SUPPLEMENTAL INFORMATION” (Supplementary Material S3\_MedicationsGOLD.xlsx)” + “\_date”.
        2. 10-32 years prior to diagnosis (secondary analyses)
     2. Alternate exposure definitions (\_alt\_lt10, \_alt\_gt10) (not used for this project):
3. **CleanHierarchy.R**

**Purpose**: Used to create file that can be used by run\_proc\_glimmix\_general.sas to perform hierarchical analysis

**Input**: Processed GOLD dataset generated in FirstProcessData.R

**Output**: Excel file that can be used in hierarchical analyses run\_proc\_glimmix\_general.sas

**Aurum (Validation) data processing file**

1. ***NSLC\_ProcessAurumData.R***

**Purpose:** Process Aurum population SAS file

**Input:** A SAS datafile (*covariate\_file\_051022.sas7bdat*) created by IMS.

**Outputs:** R file used for rest of analyses (*AurumDeDupped.May2024.RData*)

**NOTE:** Variable naming conventions are similar to those described in GOLD Data processing files. See (Supplementary Material S5\_ValidationAurumCodes.xlsx; Supplementary Material S4\_MedicationsAurum[55].xlsx). Used (\_lt10, \_date, \_gt10) to identify meaning of variable.

**Additional NOTE:** Processing this file is a little more complicated than that of GOLD because we first need to identify all individuals who did not appear in GOLD. This was determined from a text file that CPRD sent us.

**ANALYSIS CODE FILES**

**Analytic files**

1. *ProcessConditionalLogitApr2021.R* (Contains functions that iterate through a dataset to estimate odds ratios using conditional logistic regression). It is used for several of the R files described below

**GOLD-specific**

1. *GOLDPrimaryAnalyses.R.*
   1. **Purpose:** Performs conditional logistic regression associations
   2. **Input**: R file created in *FirstProcessData*.R (FullDSAgeProper.Feb2022.RData)
   3. **Input**: functions in *ProcessConditionalLogitApr2021.R*
   4. **Output**: Text files that form the basis for **Table 2**, a few supplementary tables
2. runII is a SAS program that calls the macro\_proc\_glimmix\_general.sas
   1. **Purpose**: Performs hierarchical analysis on group that was significant (infections and inflammation in this case)
   2. **Input:** formatted to run the hierarchical analysis in SAS. See **CleanHierarchy.R.**
   3. **Output:** Excel document with hierarchy results that form basis of **Table 3** and supplemental tables

**Aurum-specific**

1. *AurumPrimaryAnalyses.R:*
   1. **Purpose**: Performs conditional logistic regression analyses in validation population.
   2. **Input**: R file (*AurumDeDupped.May2024.RData*) created in the R code described above in Aurum Data processing section (*NSLC\_ProcessAurumData.R)*
   3. **Output**: Text files that form basis of **Table 4** and supplemental tables X which contain associations between subtypes of some conditions that are heterogenous (anemia, gastritis) and LCINS

**SENSITIVITY ANALYSIS CODE FILES**

1. *NSLC\_GOLDMedAnalyses.R*
   1. **Purpose**: Performs conditional logistic regression in the GOLD population to estimate the association between the medical condition and LCINS while statistically adjusting for medications indicated for the condition
   2. **Input**: R file (*FullDSAgeProper.Feb2022.RData*) created in the R code described above in processing section (*FirstProcessData*.R*)*
   3. **Output**: Text files that form basis of **Table 5** and supplemental tables
2. *NCLC\_AurumMedAnalyses.R*
   1. **Purpose**: Performs conditional logistic regression in the Aurum population to estimate the association between the medical condition and LCINS while statistically adjusting for medications indicated for the condition
   2. **Input**: R file (*AurumDeDupped.May2024.RData*) created in the R code described above in Aurum Data processing section (*NSLC\_ProcessAurumData.R)*
   3. **Output**: Text files that form basis of **Table 5** and supplemental tables
3. *Associations\_BMISES.R.* 
   1. **Purpose**: Performs conditional logistic regression analyses in both populations and mutually adjusts for SES (in deciles) and BMI (categories: underweight, normal, overweight, obese, missing).
   2. **Input**: R files (*AurumDeDupped.May2024.RData,* *FullDSAgeProper.Feb2022.RData)* created about in data processing code (*NSLC\_ProcessAurumData.R, FirstProcessData*.R*)*
   3. **Output**: Text files that form basis of **Tables X** and supplemental tables