Extract Clinical Data – RISK Study

v1.1: September 25, 2018

Introduction

extract_clinical_data.R script reads RISK study clinical data tables downloaded from IBD Plexus in csv or txt format from a specified directory, filters tables, extracts specified columns and values, calculates wPCDAI score, orders columns by dictionary order and outputs a summary table as an excel workbook.

Requirements

RISK study patient data downloaded from IBD Plexus with the following tables are expected in the input directory: Encounter, Demographics, Diagnosis, Family_History_Diagnosis, Family_History_Demographics, Labs, Master_Patient, Observations, Omics_Patient, Patient_History, Patient_Problem, Prescriptions, Procedures, Vaccines. Multiple tables of the same type are combined by the script.

IBD Plexus data dictionary excel workbook in similar format as "IBD Plexus Data Dictionary_20180716.xlsx" provided for the RISK study.

R version 3.5.0 or above
R libraries dplyr, reshape2, data.table, and openxlsx. The libraries can be installed by running:
install.packages("dplyr")

install.packages("reshape2")

install.packages("data.table")

install.packages("openxlsx")

Assumptions

The clinical data is in a similar format to the RISK study patient data downloaded from IBD Plexus. The following tables are being prefiltered or transformed as described in this document. Additional filtering can be supplied as input into the extract_clinical_data() function.

Omics_Patient:

DEIDENTIFIED_PATIENT_ID column is renamed to DEIDENTIFIED_MASTER_PATIENT_ID

Diagnosis:

DIAG_CONCEPT_NAME “Extra-Intestinal Manifestations Follow-up” is combined with "Extra-Intestinal Manifestations".

For DIAG_STATUS_CONCEPT_NAME the answer to the leading question “Yes/Unknown” is filtered out except for "Extra-Intestinal Manifestations".

DIAGNOSIS_DATE is for IBD diagnosis at enrollment.

For “Ankylosing Spondylitis”, the answer to the family history leading question is filtered out.

Encounter:

"VISITENC_ID" column is renamed to "VISIT_ENCOUNTER_ID".

Filter out unknown VISIT_ENCOUNTER_START_DATE.

Filter TYPE_OF_ENCOUNTER to "Enrollment Visit", "6-Month Follow-up Visit", "12-Month Follow-up Visit”, "18-Month Follow-up Visit", "24-Month Follow-up Visit", "30-Month Follow-up Visit”, "36-Month Follow-up Visit".

Remove " 00:00:00.0" time stamp from VISIT_ENCOUNTER_START_DATE.

Prescriptions:

MEDICATION_ADMINISTRATED is filtered for "Yes", "Current", "No", "Not Current".

MED_ACTION_CONCEPT_NAME is filtered for "Started Since Last Review", "Ongoing Treatment", "Received".

Remove " 00:00:00.0" timestamp from MED_START_DATE.

Prescriptions are split into two columns:

1. MED_ACTION_CONCEPT_NAME is "Started Since Last Review" or "Received".
2. MED_ACTION_CONCEPT_NAME is "Ongoing Treatment".

Methotrexate medication is split into two columns based on the route of administration.

Observations:

Endoscopic observations are split into two columns based on ANA_SITE_CONCEPT_NAME "Rectum" and "Ileum".

Run Script

1. source("extract_clinical_data.R")
2. Run function with default parameters:

```
extract_clinical_data(dir = "10141031/",
                      filename = "RISK Summary.xlsx",
                      dictionary = "IBD Plexus Data Dictionary_20180716.xlsx")
```

3. To change selected columns and values edit select.col and select.val parameters. To change filtered columns and values edit filter.col and filter.val parameters:

```
extract_clinical_data (dir          = "10141031/",
                       dictionary = "IBD Plexus Data Dictionary_20180716.xlsx",
                       select.col  = c("DIAG_CONCEPT_NAME",
"DIAGNOSIS_DATE", "GENDER", "TYPE_OF_ENCOUNTER",
"VISIT_ENCOUNTER_START_DATE", "AGE_AT_ENCOUNTER", "BIRTH_YEAR",
"LAB_TEST_CONCEPT_NAME", "ASSAY.NAME", "MEDICATION_NAME",
"OBS_TEST_CONCEPT_NAME", "ANA_SITE_CONCEPT_NAME"),
                       select.val   = c("DIAG_STATUS_CONCEPT_NAME",
"DIAGNOSIS_DATE", "GENDER", "TYPE_OF_ENCOUNTER",
"VISIT_ENCOUNTER_START_DATE", "AGE_AT_ENCOUNTER", "BIRTH_YEAR",
"TEST_RESULT_NUMERIC", "RAW.DATA.FILE.NAME", "MED_START_DATE",
"TEST_RESULT_NUMERIC", "TEST_RESULT_NUMERIC"),
                       filter.col   = c("DIAG_CONCEPT_NAME",
"OBS_TEST_CONCEPT_NAME"),
                       filter.val   = c("IBD - Family History", "Disease Location",
"Endoscopic Assessment - Deep Ulceration", "Endoscopic Assessment - Superficial
Ulceration", "Endoscopic Assessment - Amount of Surface Ulcerated",  "Endoscopic
```

Assessment - Amount of Surface Involved", "Perianal Disease -", "EIM", "Disease Behavior - Stricturing/Fibrostenotic", "Disease Behavior - Internally Pentrating", "PCDAI"), filename = "RISK Summary.xlsx")

Default Parameters:

table	select.col	select.val	filter.col	filter.val	comment
Encounter	TYPE_OF_ENCOUNTER	TYPE_OF_ENCOUNTER			
Encounter	VISIT_ENCOUNTER_START_DATE	VISIT_ENCOUNTER_START_DATE			
Encounter	AGE_AT_ENCOUNTER	AGE_AT_ENCOUNTER			
Demographics	GENDER	GENDER			
Demographics	BIRTH_YEAR	BIRTH_YEAR			
Diagnosis	DIAG_CONCEPT_NAME	DIAG_STATUS_CONCEPT_NAME			
Diagnosis	DIAGNOSIS_DATE	DIAGNOSIS_DATE			IBD diagnosis at enrollment
Family History Diagnosis	DIAG_CONCEPT_NAME	DIAG_STATUS_CONCEPT_NAME	DIAG_CONCEPT_NAME	"IBD - Family History"	
Labs	LAB_TEST_CONCEPT_NAME	TEST_RESULT_NUMERIC			
Master Patient	GENDER	GENDER			
Master Patient	BIRTH_YEAR	BIRTH_YEAR			
Observations	OBS_TEST_CONCEPT_NAME	TEST_RESULT_NUMERIC	OBS_TEST_CONCEPT_NAME	"Disease Location", "Endoscopic Assessment - Deep Ulceration", "Endoscopic Assessment - Superficial Ulceration", "Endoscopic Assessment - Amount of Surface Ulcerated", "Endoscopic Assessment - Amount of Surface Involved", "Perianal Disease -", "EIM", "Disease Behavior - Stricturing/Fibrostenotic", "Disease Behavior - Internally Pentrating", "PCDAI"	select.val = DESCRIPTIVE_SYMP_TEST_RESULTS if TEST_RESULT_NUMERIC is not available
Observations	ANA_SITE_CONCEPT_NAME	TEST_RESULT_NUMERIC			select.val = DESCRIPTIVE_SYMP_TEST_RESULTS if TEST_RESULT_NUMERIC is not available
Omics Patient	ASSAY.NAME	RAW.DATA.FILE.NAME			
Prescriptions	MEDICATION_NAME	MED_START_DATE			select.val = MEDICATION_ADMINISTRATED if MED_START_DATE is not available

Parameters Description

dir — Path to directory of clinical files

dictionary — Path to IBD Plexus data dictionary worksheet

select.col — Columns to select

default = c("DIAG_CONCEPT_NAME", "DIAGNOSIS_DATE", "GENDER", "TYPE_OF_ENCOUNTER", "VISIT_ENCOUNTER_START_DATE", "AGE_AT_ENCOUNTER", "BIRTH_YEAR", "LAB_TEST_CONCEPT_NAME", "RAW.DATA.FILE.NAME", "MEDICATION_NAME","OBS_TEST_CONCEPT_NAME", "ANA_SITE_CONCEPT_NAME")

select.val — Columns with values for the selected columns. If select.val is the same as the select.col, then the whole column will be selected. If value is different than the select.col the selected column will be reshaped into a wide format with values from the select.val column.

default = c("DIAG_STATUS_CONCEPT_NAME", "DIAGNOSIS_DATE", "GENDER", "TYPE_OF_ENCOUNTER", "VISIT_ENCOUNTER_START_DATE", "AGE_AT_ENCOUNTER", "BIRTH_YEAR", "TEST_RESULT_NUMERIC", "RAW.DATA.FILE.NAME", "MED_START_DATE", "TEST_RESULT_NUMERIC", "TEST_RESULT_NUMERIC")

filter.col — Columns that need to be filtered prior to reshaping.

default = c("DIAG_CONCEPT_NAME", "OBS_TEST_CONCEPT_NAME")

filter.val — Values to filter on filter.col

default = c("IBD - Family History", "Disease Location", "Endoscopic Assessment - Deep Ulceration", "Endoscopic Assessment - Superficial Ulceration", "Endoscopic Assessment - Amount of Surface Ulcerated", "Endoscopic Assessment - Amount of Surface Involved", "Perianal Disease -", "EIM", "Disease Behavior - Stricturing/Fibrostenotic", "Disease Behavior - Internally Pentrating", "PCDAI")

filename — A filename to write the final output to.

Quality Control

Script was created by Rancho Bioscience. QC was performed on the final output, examining several columns from each of the original tables for correct entries. Code review was performed on the final script by another data scientist at Rancho Bioscience.