**READ ME coding notes**

# Aim

* To convert csv files downloaded from trials registries into RIS files, readable in EPPI
* To include text in the abstract to say whether the trial is conducted in an OECD country

**lineRead\_csv function will need modifying to refer to wherever you're picking up the csv file from.**

**The test csv below can be found here:** [L:\IR\\_IS\Learning & Dev\Coding group\Current projects\CTdotgov to RIS](file:///L:/IR/_IS/Learning%20&%20Dev/Coding%20group/Current%20projects/CTdotgov%20to%20RIS)

# 30/06/2023 (updated after ct.gov column name changes)

library(tidyverse)

# How to run script:

# (1) Insert name of CSV file to convert into line 10 (between speech marks)

# (2) Select all in this window (CTRL+A) and click "Run"

# (3) RIS file (clinicaltrialsris.txt) will appear in files pane location (bottom right window)

# (4) If using Posit Cloud ris file can be exported by clicking on the cog symbol

df <- read\_csv("menopause trials.csv")

# Basic method here is to combine relevant fields to make an abstract;

# create new columns for field tags then unite these with data from the CSV.

df <- df |>

select(`Study Title`, `Study Status`, `Study Design`, Conditions, Age, Enrollment, Interventions, `Primary Outcome Measures`, `Study URL`, `First Posted`, `NCT Number`, `Primary Completion Date`, Locations) |>

mutate(`Study Design` = str\_replace\_all(`Study Design`, "\\|", "\n\t"),

`Primary Outcome Measures` = str\_replace\_all(`Primary Outcome Measures`, "\\|", "\n\t"),

Conditions = str\_replace\_all(Conditions, "\\|", "\n\t"),

Interventions = str\_replace\_all(Interventions, "\\|", "\n\t"),

UK = str\_detect(Locations, "United Kingdom"),

UK = case\_when(UK == TRUE ~ "yes",

UK == FALSE ~ "no",

is.na(UK) ~ "not stated"),

OECD = str\_detect(Locations, "Australia|Austria|Belgium|Canada|Chile|Colombia|Costa Rica|Czech Republic|Czechia|Denmark|Estonia|Finland|France|Germany|Greece|Hungary|Iceland|Ireland|Israel|Italy|Japan|Korea|Latvia|Lithuania|Luxembourg|Mexico|Netherlands|New Zealand|Norway|Poland|Portugal|Slovak Republic|Slovakia|Slovenia|Spain|Sweden|Switzerland|Türkiye|Turkey|United Kingdom|United States"),

OECD = case\_when(OECD == TRUE ~ "yes",

OECD == FALSE ~ "none",

is.na(OECD) ~ "not stated"),

Age = str\_replace\_all(Age, "...\\(.\*", ""),

`Study Status` = paste0("THIS IS A TRIALS REGISTY RECORD. The trial status is currently recorded as ", str\_to\_lower(`Study Status`), ". This may not be reflective of the actual state of the trial. The completion date for the primary outcome is given as: ", `Primary Completion Date`, "\n"),

`Study Design` = paste0("STUDY DESIGN:\n\t", `Study Design`, "\n\tNumber of participants: ", Enrollment, "\n\tAge: ", Age, "\n\tUK trial centre(s): ", UK, "\n"),

Conditions = paste0("CONDITION(S):\n\t", Conditions, "\n"),

Interventions = paste0("INTERVENTION(S):\n\t", Interventions, "\n"),

`Primary Outcome Measures` = paste0("OUTCOME(S):\n\t", `Primary Outcome Measures`),

# First Posted is a non-controlled string so have just

# truncated to leave 4-digit year code rather than use lubridate.

# This field becomes the year in the EPPI record. Needs to be consistent for deduping,

# hence use of first posted date rather than completion date.

`First Posted` = str\_trunc(`First Posted`, 4, side = "left", ellipsis = ""),

NCT2 = paste0(`NCT Number`, " - ")

) |>

# Create new columns to represent RIS field tags

mutate(TY = "TY - DBASE", TI = "TI - ", AB = "AB - ", Y1 = "Y1 - ",

UR = "UR - ", JF = "JF - clinicaltrials.gov", SP = "SP - 1", ER = "ER - ", VL = "VL - ") |>

unite(TI, c(TI, NCT2, `Study Title`), sep = "") |>

unite(AB, c(AB, `Study Status`, `Study Design`, Conditions, Interventions, `Primary Outcome Measures`), sep = "") |>

unite(UR, c(UR, `Study URL`), sep = "") |>

unite(Y1, c(Y1, `First Posted`), sep = "") |>

unite(VL, c(VL, `NCT Number`), sep = "")

# Reordering columns for output. If I were reordering the rows that would go before the first comma.

df <- df[, c("TY", "TI", "Y1", "AB", "JF", "UR", "VL", "SP", "ER")]

# write.table used to create text files with new lines as separators and line ends.

write.table(df, file = "clinicaltrialsris.txt", quote = FALSE, sep = "\n", eol = "\n\n", col.names = FALSE, row.names = FALSE)

# Question on whether to add some sifting rules i.e. if primary completion

# >3 years in the future then ignore?

# 15/05/2023 Original code

library(tidyverse)

# How to run script:

# (1) Insert name of CSV file to convert into line 10 (between speech marks)

# (2) Select all in this window (CTRL+A) and click "Run"

# (3) RIS file (clinicaltrialsris.txt) will appear in files pane location (bottom right window)

# (4) If using Posit Cloud ris file can be exported by clicking on the cog symbol

df <- read\_csv("menopause trials.csv")

# Basic method here is to combine relevant fields to make an abstract;

# create new columns for field tags then unite these with data from the CSV.

df <- df |>

select(Title, Status, `Study Designs`, Conditions, Age, Enrollment, Interventions, `Outcome Measures`, URL, `First Posted`, `NCT Number`, `Primary Completion Date`, Locations) |>

mutate(`Study Designs` = str\_replace\_all(`Study Designs`, "\\|", "\n\t"),

`Outcome Measures` = str\_replace\_all(`Outcome Measures`, "\\|", "\n\t"),

Conditions = str\_replace\_all(Conditions, "\\|", "\n\t"),

Interventions = str\_replace\_all(Interventions, "\\|", "\n\t"),

UK = str\_detect(Locations, "United Kingdom"),

UK = case\_when(UK == TRUE ~ "yes",

UK == FALSE ~ "no",

is.na(UK) ~ "not stated"),

OECD = str\_detect(Locations, "Australia|Austria|Belgium|Canada|Chile|Colombia|Costa Rica|Czech Republic|Czechia|Denmark|Estonia|Finland|France|Germany|Greece|Hungary|Iceland|Ireland|Israel|Italy|Japan|Korea|Latvia|Lithuania|Luxembourg|Mexico|Netherlands|New Zealand|Norway|Poland|Portugal|Slovak Republic|Slovakia|Slovenia|Spain|Sweden|Switzerland|Türkiye|Turkey|United Kingdom|United States"),

OECD = case\_when(OECD == TRUE ~ "yes",

OECD == FALSE ~ "none",

is.na(OECD) ~ "not stated"),

Age = str\_replace\_all(Age, "...\\(.\*", ""),

Status = paste0("THIS IS A TRIALS REGISTY RECORD. The trial status is currently recorded as ", str\_to\_lower(Status), ". This may not be reflective of the actual state of the trial. The completion date for the primary outcome is given as: ", `Primary Completion Date`, "\n"),

`Study Designs` = paste0("STUDY DESIGN:\n\t", `Study Designs`, "\n\tNumber of participants: ", Enrollment, "\n\tAge: ", Age, "\n\tUK trial centre(s): ", UK, "\n"),

Conditions = paste0("CONDITION(S):\n\t", Conditions, "\n"),

Interventions = paste0("INTERVENTION(S):\n\t", Interventions, "\n"),

`Outcome Measures` = paste0("OUTCOME(S):\n\t", `Outcome Measures`),

# First Posted is a non-controlled string so have just

# truncated to leave 4-digit year code rather than use lubridate.

# This field becomes the year in the EPPI record. Needs to be consistent for deduping,

# hence use of first posted date rather than completion date.

`First Posted` = str\_trunc(`First Posted`, 4, side = "left", ellipsis = ""),

NCT2 = paste0(`NCT Number`, " - ")

) |>

# Create new columns to represent RIS field tags

mutate(TY = "TY - DBASE", TI = "TI - ", AB = "AB - ", Y1 = "Y1 - ",

UR = "UR - ", JF = "JF - clinicaltrials.gov", SP = "SP - 1", ER = "ER - ", VL = "VL - ") |>

unite(TI, c(TI, NCT2, Title), sep = "") |>

unite(AB, c(AB, Status, `Study Designs`, Conditions, Interventions, `Outcome Measures`), sep = "") |>

unite(UR, c(UR, URL), sep = "") |>

unite(Y1, c(Y1, `First Posted`), sep = "") |>

unite(VL, c(VL, `NCT Number`), sep = "")

# Reordering columns for output. If I were reordering the rows that would go before the first comma.

df <- df[, c("TY", "TI", "Y1", "AB", "JF", "UR", "VL", "SP", "ER")]

# write.table used to create text files with new lines as separators and line ends.

write.table(df, file = "clinicaltrialsris.txt", quote = FALSE, sep = "\n", eol = "\n\n", col.names = FALSE, row.names = FALSE)

# Question on whether to add some sifting rules i.e. if primary completion

# >3 years in the future then ignore?

CT.gov code with in\_join

#Change Steve's converted files to csv format before starting

library(tidyverse)

df <- read\_csv("Clinicaltrialsgov Breast Cancer by last updated 20 06 23.csv")

# df <- read\_csv("22-12-20 Clinical Trials gov 2015-2022.csv")

df2 <- read\_csv("biological therapy for advanced breast cancer (CG81-1.3.12).csv")

df3 <- inner\_join(df, df2, by='NCT Number')

df4 <-

df3 %>%

rename("Title" = 3)%>%

rename("URL" = 27)

df4 <- df4 |>

select(Title, Status, `Study Designs`, Conditions, Age, Enrollment, Interventions, `Outcome Measures`, URL, `First Posted`, `NCT Number`, `Primary Completion Date`, Locations) |>

mutate(`Study Designs` = str\_replace\_all(`Study Designs`, "\\|", "\n\t"),

`Outcome Measures` = str\_replace\_all(`Outcome Measures`, "\\|", "\n\t"),

Conditions = str\_replace\_all(Conditions, "\\|", "\n\t"),

Interventions = str\_replace\_all(Interventions, "\\|", "\n\t"),

UK = str\_detect(Locations, "United Kingdom"),

UK = case\_when(UK == TRUE ~ "yes",

UK == FALSE ~ "no",

is.na(UK) ~ "not stated"),

OECD = str\_detect(Locations, "Australia|Austria|Belgium|Canada|Chile|Colombia|Costa Rica|Czech Republic|Czechia|Denmark|Estonia|Finland|France|Germany|Greece|Hungary|Iceland|Ireland|Israel|Italy|Japan|Korea|Latvia|Lithuania|Luxembourg|Mexico|Netherlands|New Zealand|Norway|Poland|Portugal|Slovak Republic|Slovakia|Slovenia|Spain|Sweden|Switzerland|Türkiye|Turkey|United Kingdom|United States"),

OECD = case\_when(OECD == TRUE ~ "yes",

OECD == FALSE ~ "none",

is.na(OECD) ~ "not stated"),

Age = str\_replace\_all(Age, "...\\(.\*", ""),

Status = paste0("THIS IS A TRIALS REGISTY RECORD. The trial status is currently recorded as ", str\_to\_lower(Status), ". This may not be reflective of the actual state of the trial. The completion date for the primary outcome is given as: ", `Primary Completion Date`, "\n"),

`Study Designs` = paste0("STUDY DESIGN:\n\t", `Study Designs`, "\n\tNumber of participants: ", Enrollment, "\n\tAge: ", Age, "\n\tUK trial centre(s): ", UK, "\n"),

Conditions = paste0("CONDITION(S):\n\t", Conditions, "\n"),

Interventions = paste0("INTERVENTION(S):\n\t", Interventions, "\n"),

`Outcome Measures` = paste0("OUTCOME(S):\n\t", `Outcome Measures`),

# First Posted is a non-controlled string so have just

# truncated to leave 4-digit year code rather than use lubridate.

# This field becomes the year in the EPPI record. Needs to be consistent for deduping,

# hence use of first posted date rather than completion date.

`First Posted` = str\_trunc(`First Posted`, 4, side = "left", ellipsis = ""),

NCT2 = paste0(`NCT Number`, " - ")

) |>

# Create new columns to represent RIS field tags

mutate(TY = "TY - DBASE", TI = "TI - ", AB = "AB - ", Y1 = "Y1 - ",

UR = "UR - ", JF = "JF - clinicaltrials.gov", SP = "SP - 1", ER = "ER - ", VL = "VL - ") |>

unite(TI, c(TI, NCT2, Title), sep = "") |>

unite(AB, c(AB, Status, `Study Designs`, Conditions, Interventions, `Outcome Measures`), sep = "") |>

unite(UR, c(UR, URL), sep = "") |>

unite(Y1, c(Y1, `First Posted`), sep = "") |>

unite(VL, c(VL, `NCT Number`), sep = "")

# Reordering columns for output. If I were reordering the rows that would go before the first comma.

df4 <- df4[, c("TY", "TI", "Y1", "AB", "JF", "UR", "VL", "SP", "ER")]

# write.table used to create text files with new lines as separators and line ends.

#Change name to relevant BC file

write.table(df4, file = "biological therapy for advanced breast cancer (CG81-1.3.12).txt", quote = FALSE, sep = "\n", eol = "\n\n", col.names = FALSE, row.names = FALSE)

# Question on whether to add some sifting rules i.e. if primary completion

# >3 years in the future then ignore?

ISRCTN code

**Read\_csv function will need modifying to refer to wherever you're picking up the csv file from.**

**The test csv below can be found here:** [L:\IR\\_IS\Learning & Dev\Coding group\Current projects\CTdotgov to RIS](file:///L:/IR/_IS/Learning%20&%20Dev/Coding%20group/Current%20projects/CTdotgov%20to%20RIS)

# 15/05/2023

library(tidyverse)

library(readr)

df <- read\_csv("ISRCTN\_trials.csv")

#Basic method here is to combine relevant fields to make an abstract;

#create new columns for field tags then unite these with data from the CSV.

df2 <- df |>

select(Title, `Overall trial status`, `Primary study design`, Condition,

`Age group`, `Intervention type`, `Drug/device/biological/vaccine name(s)`, ISRCTN, Phase,

`Target number of participants`, `Total final enrolment`, `Country of recruitment`, `Overall trial end`, `ClinicalTrials.gov number`

) |>

unite(`Intervention type`, `Intervention type`, `Drug/device/biological/vaccine name(s)`, sep = ": ", na.rm=TRUE) |>

mutate(URL = ISRCTN) |>

mutate(URL = str\_replace(URL,"^", "<https://www.isrctn.com/>"))|>

filter(!grepl('NCT', `ClinicalTrials.gov number`))|>

mutate(OECD = `Country of recruitment`)|>

mutate(Status = paste0("THIS IS A TRIALS REGISTY RECORD", " \n",

"STATUS: ", `Overall trial status`, ". This may not be reflective of the actual state of the trial \n"),

`Primary study design` = paste0("STUDY DESIGN: ", `Primary study design`, " \n"),

Condition = paste0("CONDITION(S): ", Condition, " \n"),

`Age group` = paste0("AGE: ", `Age group`, " \n"),

`Intervention type` = paste0("INTERVENTION(S): ", `Intervention type`, " \n"),

Phase = paste0("PHASE: ", Phase, " \n"),

`Target number of participants`= paste0("EST PARTICIPANTS: ", `Target number of participants`, " \n"),

`Total final enrolment`= paste0("FINAL ENROLMENT: ", `Total final enrolment`, " \n"),

`Country of recruitment`= paste0("COUNTRY: ", `Country of recruitment`, " \n"),

OECD = str\_detect(OECD, "Australia|Austria|Belgium|Canada|Chile|Colombia|Costa Rica|Czech Republic|Czechia|Denmark|Estonia|Finland|France|Germany|Greece|Hungary|Iceland|Ireland|Israel|Italy|Japan|Korea|Latvia|Lithuania|Luxembourg|Mexico|Netherlands|New Zealand|Norway|Poland|Portugal|Slovak Republic|Slovakia|Slovenia|Spain|Sweden|Switzerland|Türkiye|Turkey|United Kingdom|United States"),

OECD = case\_when(OECD == TRUE ~ "Yes",

OECD == FALSE ~ "None",

is.na(OECD) ~ "Not stated"),

OECD = paste0("OECD: ", OECD, " \n"),

`Overall trial end` = paste0("TRIAL END DATE: ", `Overall trial end`)

) |>

# Create new columns to represent RIS field tags

mutate(TY = "TY - DBASE", TI = "TI - ", DO = "DO - <https://dx.doi.org/10.1186/>", AB = "AB - ", JF = "JF - ISRCTN", UR = "UR - ", ER = "ER - ") |>

unite(TI, c(TI, Title), sep = "") |>

unite(AB, c(AB, Status, `Primary study design`, Condition, `Intervention type`, Phase, `Target number of participants`, `Total final enrolment`, `Country of recruitment`, OECD, `Overall trial end`), sep = "") |>

unite(UR, c(UR, URL), sep = "")|>

unite(DO, c(DO, ISRCTN), sep = "")

#Reordering columns for output. If I were reordering the rows that would go before the comma.

df2 <- df2[, c("TY", "TI", "DO", "AB", "JF", "UR", "ER")]

#write.table can be used to create text files with new lines as separators and line ends.

write.table(df2, file = "ISRCTNris.txt", quote = FALSE, sep = "\n", eol = "\n\n", col.names = FALSE, row.names = FALSE)