 # S2 File. The R script used to download data from CTRI, process them and store them in an SQLite database.

```
#loading libraries
library(DBI)
library(XML)
library(tidyr)
library(stringr)
library(readr)

#set range for searching valid trial webpages
ids <- c(1:60000)
counter = 0

#-------------------------------define and connect the db----------------
mydb <- dbConnect(RSQLite::SQLite(), "ctri-ecsite.sqlite")

#-------------------------------writing records to the db----------------

for (i in ids) {
  #load and check if url is valid
  myurl <- paste("http://ctri.nic.in/Clinicaltrials/pmaindet2.php?trialid=",i)
  myurl = gsub(" ","",myurl)
  page <- readHTMLTable(myurl)
  page <- page$`NULL`
  if (is.null(page)) {
    next
  }
  #preliminary adjustments for efficient data processing
  out<-NULL
  out$col1 <- page$V1
  out$col2 <- page$V2
  out$col3 <- page$V3
  if(length(page)==4)
  {
    out$col4 <- page$V4
  }
  write.csv(out,  file = "output.csv", row.names = FALSE)
  tmp <- read.csv("output.csv", na.strings = c(""))
  tmp$col1=gsub("Â","",tmp$col1)
  tmp$col2=gsub("Â","",tmp$col2)
  tmp$col3=gsub("Â","",tmp$col3)
  if(length(page)==4)
  {
    tmp$col4=gsub("Â","",tmp$col4)
  }
  write.csv(tmp, file = "raw.csv", row.names = FALSE)
  tmp$col1 <- trimws(tmp$col1)
  tmp$col2 <- trimws(tmp$col2)
  tmp$col3 <- trimws(tmp$col3)
  if(length(page)==4)
  {
    tmp$col4 <- trimws(tmp$col4)
```

```
  }


  result<-NULL #will save desired information here

  #-----------------------------urls table--------------

  result$ctri_number <- gsub("\\[.*","",tmp$col2[which(tmp$col1 == "CTRI Number")])
  result$registered_on <- gsub("\\
.*","",str_trim(gsub("]","",gsub(".*:","",tmp$col2[which(tmp$col1 == "CTRI Number")]) )))
  result$reg_type <- gsub(".*\\Trial Registered ", "", tmp$col2[which(tmp$col1 == "CTRI
Number")])

  urls <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    ctri_url <- toString(myurl)
  )
  dbWriteTable(mydb, "urls", urls, append = TRUE )
  write.table(urls, "urls.csv", sep = ",",row.names = FALSE, col.names = !file.exists("urls.csv"),
append = T)

  #-----------------------------reg_details table--------------

  reg_details <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    registered_on <- toString(paste("'",result$registered_on,"'")),
    reg_type <- toString(result$reg_type)
  )
  dbWriteTable(mydb, "reg_details", reg_details, append = TRUE )
  write.table(reg_details, "reg_details.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("reg_details.csv"), append = T)

  #-----------------------------study_details table--------------

  result$type_of_trial <- sub("...", "", toString(tmp$col2[which(grepl("Type of Trial", tmp$col1,
fixed=TRUE))]))
  result$type_of_study <- sub("...", "", toString(tmp$col2[which(grepl("Type of Study", tmp$col1,
fixed=TRUE))]))
  result$study_design <- sub("...", "", toString(tmp$col2[which(grepl("Study Design", tmp$col1,
fixed=TRUE))]))
  result$public_title <- sub("...", "", toString(tmp$col2[which(grepl("Public Title of Study",
tmp$col1, fixed=TRUE))]))
  result$scientific_title <- sub("...", "", toString(tmp$col2[which(grepl("Scientific Title of Study",
tmp$col1, fixed=TRUE))]))
  result$acronym <- sub("...", "", toString(tmp$col2[which(grepl("Trial Acronym:", tmp$col1,
fixed=TRUE))]))
  result$phase <- sub("...", "", toString(tmp$col2[which(grepl("Phase of Trial", tmp$col1,
fixed=TRUE))]))
  result$pgt <- sub("...", "", toString(tmp$col2[which(grepl("Post Graduate Thesis", tmp$col1,
fixed=TRUE))]))
```

```
  study_details <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    type_of_trial <- toString(result$type_of_trial),
    type_of_study <- toString(result$type_of_study),
    study_design <- toString(result$study_design),
    phase <- toString(result$phase),
    post_graduate_thesis <- result$pgt,
    acronym <- result$acronym
  )
  dbWriteTable(mydb, "study_details", study_details, append = TRUE )
  write.table(study_details, "study_details.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("study_details.csv"), append = T)

  #-----------------------------study_name table--------------

  study_name <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    public_title <- result$public_title,
    scientific_title <- result$scientific_title
  )
  dbWriteTable(mydb, "study_name", study_name, append = TRUE )
  write.table(study_name, "study_name.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("study_name.csv"), append = T)

  #-----------------------------dates table--------------

  result$lmo <- sub("...", "", toString(tmp$col2[which(grepl("Last Modified On:", tmp$col1,
fixed=TRUE))]))
  result$dfei <- sub("...", "", toString(tmp$col2[which(grepl("Date of First Enrollment (India)",
tmp$col1, fixed=TRUE))]))
  result$dfeg <- sub("...", "", toString(tmp$col2[which(grepl("Date of First Enrollment (Global)",
tmp$col1, fixed=TRUE))]))
  result$dsci <- sub("...", "", toString(tmp$col2[which(grepl("Date of Study Completion (India)",
tmp$col1, fixed=TRUE))]))
  result$dscg <- sub("...", "", toString(tmp$col2[which(grepl("Date of Study Completion (Global)",
tmp$col1, fixed=TRUE))]))


  dates <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    registered_on <- toString(paste("",result$registered_on,"") ),
    last_modified_on <- toString(paste("",result$lmo,"") ),
    date_of_first_enrollment_india <- toString(paste("",result$dfei,"")),
    date_of_first_enrollment_global <- toString(paste("",result$dfeg,"")),
    date_of_study_completion_india <- toString(paste("",result$dsci,"")),
    date_of_study_completion_global <- toString(paste("",result$dscg,""))
  )
  dbWriteTable(mydb, "dates", dates, append = TRUE )
```

```
  write.table(dates, "dates.csv", sep = ",",row.names = FALSE, col.names = !file.exists("dates.csv"),
append = T)

  #-----------------------------brief_summary table--------------
  result$bs <- sub("...", "", toString(tmp$col2[which(grepl("Brief Summary", tmp$col1,
fixed=TRUE))]))

  brief_summary <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    brief_summary <- toString(result$bs)
  )
  dbWriteTable(mydb, "brief_summary", brief_summary, append = TRUE )
  write.table(brief_summary, "brief_summary.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("brief_summary.csv"), append = T)

  #-----------------------------publication table--------------

  result$publication <- sub("...", "", toString(tmp$col2[which(grepl("Publication Details", tmp$col1,
fixed=TRUE))]))

  publication <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    publication <- toString(result$publication)
  )
  dbWriteTable(mydb, "publication", publication, append = TRUE )
  write.table(publication, "publication.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("publication.csv"), append = T)

  #-----------------------------recruitment table--------------


  result$rsti <- sub("...", "", toString(tmp$col2[which(grepl("Recruitment Status of Trial (India)",
tmp$col1, fixed=TRUE))]))
  result$rstg <- sub("...", "", toString(tmp$col2[which(grepl("Recruitment Status of Trial (Global)",
tmp$col1, fixed=TRUE))]))

  recruitment <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    recruitment_status_india <- toString(result$rsti),
    recruitment_status_global <- toString(result$rstg)
  )
  dbWriteTable(mydb, "recruitment", recruitment, append = TRUE )
  write.table(recruitment, "recruitment.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("recruitment.csv"), append = T)

  #-----------------------------method table--------------

  result$method_gsr <- sub("...", "", toString(tmp$col2[which(grepl("Method of Generating
Random Sequence", tmp$col1, fixed=TRUE))]))
```

```
  result$method_concealment <- sub("...", "", toString(tmp$col2[which(grepl("Method of
Concealment", tmp$col1, fixed=TRUE))]))
  result$bm <- sub("...", "", toString(tmp$col2[which(grepl("Blinding/Masking", tmp$col1,
fixed=TRUE))]))

  method <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    method_of_generating_random_sequence <- result$method_gsr,
    method_concealment <- result$method_concealment,
    blinding_masking <- result$bm
  )
  dbWriteTable(mydb, "method", method, append = TRUE )
  write.table(method, "method.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("method.csv"), append = T)


  #----------------------------Inclusion table--------------

  x <- sub("...", "", toString(which(grepl("Inclusion Criteria", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))

  y <- sub("...", "", toString(which(grepl("ExclusionCriteria", tmp$col1, fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))

  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Age From" )){
      result$inclusion_age_from <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Age To" )){
      result$inclusion_age_to <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Gender" )){
      result$inclusion_gender <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Details" )){
      result$inclusion_details <- toString(tmp$col2[[j]])
    }
  }

  inclusion <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    inclusion_age_from <- result$inclusion_age_from,
    inclusion_age_to <- result$inclusion_age_to,
    inclusion_gender <- result$inclusion_gender,
    inclusion_details <- result$inclusion_details
  )

  dbWriteTable(mydb, "inclusion", inclusion, append = TRUE )
  write.table(inclusion, "inclusion.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("inclusion.csv"), append = T)
```

```
  #----------------------------Exclusion table--------------

  x <- sub("...", "", toString(which(grepl("ExclusionCriteria", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Method of Generating Random Sequence", tmp$col1,
fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))

  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Details" )){
      result$inclusion_details <- toString(tmp$col2[[j]])
    }
  }
  exclusion <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    exclusion_details <- result$inclusion_details
  )

  dbWriteTable(mydb, "exclusion", exclusion, append = TRUE )
  write.table(exclusion, "exclusion.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("exclusion.csv"), append = T)

  #----------------------------Intervention / Comparator Agent table--------------

  x <- sub("...", "", toString(which(grepl("Intervention / Comparator Agent", tmp$col1,
fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Inclusion Criteria", tmp$col1, fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))

  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Comparator Agent" )){
      result$comparator_name <- toString(tmp$col2[[j]])
      result$comparator_details <- toString(tmp$col3[[j]])
    }

    if (identical(toString(tmp$col1[[j]]) , "Intervention" )){
      result$intervention_name <- toString(tmp$col2[[j]])
      result$intervention_details <- toString(tmp$col3[[j]])
    }

  }

  intervention_comparator = NULL

  intervention_comparator$trial_id <- toString(i)
  intervention_comparator$ctri_number <- toString(result$ctri_number)
  intervention_comparator$comparator_name <- result$comparator_name
  intervention_comparator$comparator_details <-result$comparator_details
  intervention_comparator$intervention_name <- result$intervention_name
```

```
  intervention_comparator$intervention_details <- result$intervention_details

  intervention_comparator <- as.data.frame(intervention_comparator)
  dbWriteTable(mydb, "intervention_comparator", intervention_comparator, append = TRUE )
  write.table(intervention_comparator, "intervention_comparator.csv", sep = ",",row.names =
FALSE, col.names = !file.exists("intervention_comparator.csv"), append = T)

  #----------------------------Secondary IDs table----------

  x <- sub("...", "", toString(which(grepl("Secondary IDs if Any", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Details of Principal Investigator", tmp$col1,
fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))
  x <- as.numeric(x)
  y <- as.numeric(y)

  if((y-x) != 1){

    for (item in (x:y)) {
      if (identical(toString(tmp$col1[[item]]) , "Secondary ID" )){
        xa <- item+1
      }

    }
    y <- y-1


    for (j in (xa:y)) {

      result$sec_id <- toString(tmp$col1[[j]])
      result$sec_identifier <- toString(tmp$col2[[j]])

      secondary_id <-NULL

      secondary_id$trial_id <- toString(i)
      secondary_id$ctri_number <- toString(result$ctri_number)
      secondary_id$id <- paste("'",result$sec_id,"'")
      secondary_id$identifier <- result$sec_identifier

      secondary_id <- as.data.frame(secondary_id)

      dbWriteTable(mydb, "secondary_id", secondary_id, append = TRUE )
      write.table(secondary_id, "secondary_id.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("secondary_id.csv"), append = T)
    }
  }else {
    secondary_id <-NULL

    secondary_id$trial_id <- toString(i)
    secondary_id$ctri_number <- toString(result$ctri_number)
    secondary_id$id <- "Not Available"
```

```
    secondary_id$identifier <- "Not Available"

    secondary_id <- as.data.frame(secondary_id)

    dbWriteTable(mydb, "secondary_id", secondary_id, append = TRUE )
    write.table(secondary_id, "secondary_id.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("secondary_id.csv"), append = T)
  }

  #-----------------------------Details of Principal Investigator or overall Trial Coordinator (multi-
center study)  table----------


  x <- sub("...", "", toString(which(grepl("Details of Principal Investigator", tmp$col1,
fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Scientific Query", tmp$col1, fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))

  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Name" )){
      result$pi_name <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Designation" )){
      result$pi_designation <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Affiliation" )){
      result$pi_affiliation <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Address" )){
      result$pi_address <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Phone" )){
      result$pi_phone <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Fax" )){
      result$pi_fax <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Email" )){
      result$pi_email <- toString(tmp$col2[[j]])
    }

  }
  pi <- NULL

  pi$trial_id <- toString(i)
  pi$ctri_number <- toString(result$ctri_number)
  pi$name <- result$pi_name
  pi$designation <- result$pi_designation
  pi$affiliation <- result$pi_affiliation
  pi$address <- paste("",result$pi_address,"")
  pi$phone <- paste("",result$pi_phone,"")
```

```
  pi$fax <- paste("'",result$pi_fax,"'")
  pi$email <- paste("'", result$pi_email,"'")

  pi <- as.data.frame(pi)
  dbWriteTable(mydb, "pi", pi, append = TRUE )
  write.table(pi, "pi.csv", sep = ",",row.names = FALSE, col.names = !file.exists("pi.csv"), append =
T)

  #----------------------------Details of Contact Person Scientific Query  table----------

  x <- sub("...", "", toString(which(grepl("Scientific Query", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Public Query", tmp$col1, fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))


  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Name" )){
      result$cpsq_name <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Designation" )){
      result$cpsq_designation <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Affiliation" )){
      result$cpsq_affiliation <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Address" )){
      result$cpsq_address <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Phone" )){
      result$cpsq_phone <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Fax" )){
      result$cpsq_fax <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Email" )){
      result$cpsq_email <- toString(tmp$col2[[j]])
    }

  }

  cpsq <- NULL

  cpsq$trial_id <- toString(i)
  cpsq$ctri_number <- toString(result$ctri_number)
  cpsq$name <- result$cpsq_name
  cpsq$designation <- result$cpsq_designation
  cpsq$affiliation <- result$cpsq_affiliation
  cpsq$address <- paste("'",result$cpsq_address,"'")
  cpsq$phone <- paste("'",result$cpsq_phone,"'")
  cpsq$fax <- paste("'",result$cpsq_fax,"'")
  cpsq$email <- paste("'", result$cpsq_email,"'")
```

```
  cpsq <- as.data.frame(cpsq)
  dbWriteTable(mydb, "cpsq", cpsq, append = TRUE )
  write.table(cpsq, "contact_person_scientific_query.csv", sep = ",",row.names = FALSE, col.names
= !file.exists("contact_person_scientific_query.csv"), append = T)

  #-----------------------------Details of Contact Person Public Query  table--------------


  x <- sub("...", "", toString(which(grepl("Public Query", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Source of Monetary or Material Support", tmp$col1,
fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))


  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Name" )){
      result$cppq_name <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Designation" )){
      result$cppq_designation <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Affiliation" )){
      result$cppq_affiliation <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Address" )){
      result$cppq_address <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Phone" )){
      result$cppq_phone <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Fax" )){
      result$cppq_fax <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Email" )){
      result$cppq_email <- toString(tmp$col2[[j]])
    }

  }

  cppq <- NULL

  cppq$trial_id <- toString(i)
  cppq$ctri_number <- toString(result$ctri_number)
  cppq$name <- result$cppq_name
  cppq$designation <- result$cppq_designation
  cppq$affiliation <- result$cppq_affiliation
  cppq$address <- paste("",result$cppq_address,"")
  cppq$phone <- paste("",result$cppq_phone,"")
  cppq$fax <- paste("",result$cppq_fax,"")
  cppq$email <- paste("", result$cppq_email,"")
```

```
  cppq <- as.data.frame(cppq)
  dbWriteTable(mydb, "cppq", cppq, append = TRUE )
  write.table(cppq, "contact_person_public_query.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("contact_person_public_query.csv"), append = T)



  #-----------------------------Source of Monetary or Material Support table--------------


  result$source_of_monetory_material_support <- sub("...", "",
toString(tmp$col2[which(grepl("Source of Monetary or Material Support", tmp$col1,
fixed=TRUE))]))

  smms <- NULL
  smms$trial_id <- toString(i)
  smms$ctri_number <- toString(result$ctri_number)
  smms$source_of_monetory_material_support <-
toString(result$source_of_monetory_material_support)

  smms <- as.data.frame(smms)
  dbWriteTable(mydb, "smms", smms, append = TRUE )
  write.table(smms, "source_of_monetory_material_support.csv", sep = ",",row.names = FALSE,
col.names = !file.exists("source_of_monetory_material_support.csv.csv"), append = T)

  #-----------------------------Primary Sponsor table--------------


  x <- sub("...", "", toString(which(grepl("Primary Sponsor", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Details of Secondary Sponsor", tmp$col1,
fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))


  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Name" )){
      result$ps_name <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Address" )){
      result$ps_address <- toString(tmp$col2[[j]])
    }
    if (identical(toString(tmp$col1[[j]]) , "Type of Sponsor" )){
      result$ps_typeofsponsor <- toString(tmp$col2[[j]])
    }


  }

  ps <- NULL
```

```
  ps$trial_id <- toString(i)
  ps$ctri_number <- toString(result$ctri_number)
  ps$name <- result$ps_name
  ps$address <- paste("",result$ps_address,"")
  ps$type_of_sponsor <- paste("",result$ps_typeofsponsor,"")

  ps <- as.data.frame(ps)
  dbWriteTable(mydb, "ps", ps, append = TRUE )
  write.table(ps, "primary_sponsor.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("primary_sponsor.csv"), append = T)

  #----------------------------Details of Secondary Sponsor table--------------

  x <- sub("...", "", toString(which(grepl("Details of Secondary Sponsor", tmp$col1,
fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Countries of Recruitment", tmp$col1, fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))


  for(j in (x:y)){
    if (identical(toString(tmp$col1[[j]]) , "Name" )){
      result$ss_name <- toString(tmp$col1[[j+1]])
      result$ss_address <- toString(tmp$col2[[j+1]])
    }
  }

  ss <- NULL

  ss$trial_id <- toString(i)
  ss$ctri_number <- toString(result$ctri_number)
  ss$name <- result$ss_name
  ss$address <- paste("",result$ss_address,"")

  ss <- as.data.frame(ss)
  dbWriteTable(mydb, "ss", ss, append = TRUE )
  write.table(ss, "secondary_sponsor.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("secondary_sponsor.csv"), append = T)

  #----------------------------Countries of Recruitment table--------------


  result$cor <- sub("...", "", toString(tmp$col2[which(grepl("Countries of Recruitment", tmp$col1,
fixed=TRUE))]))

  cor <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    countries_of_recruitment <- toString(result$cor)
  )

  cor <- as.data.frame(cor)
```

```
  dbWriteTable(mydb, "cor", cor, append = TRUE )
  write.table(cor, "countries_of_recruitment.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("countries_of_recruitment.csv"), append = T)

  #----------------------------Sites of Study table--------------

  x <- sub("...", "", toString(which(grepl("Sites of Study", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Details of Ethics Committee", tmp$col1, fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))
  x <- as.numeric(x)
  y <- as.numeric(y)

  result$sos_noofsites <- gsub(".*\\No of Sites =", "", sub("...", "",
toString(tmp$col1[which(grepl("No of Sites =", tmp$col1, fixed=TRUE))])))

  for (item in (x:y)) {
    if (identical(toString(tmp$col1[[item]]) , "Name of Principal\nInvestigator" )){
      xa <- item+1
    }

  }

  y <- y-1

  for (j in (xa:y)) {
    result$sos_pi_name <- toString(tmp$col1[[j]])
    result$sos_sitename <- toString(tmp$col2[[j]])
    result$sos_siteaddress <- toString(tmp$col3[[j]])
    if(length(page)==4){
      result$sos_phonefax <- toString(tmp$col4[[j]])
    }
    sos <- NULL

    sos$trial_id <- toString(i)
    sos$ctri_number <- toString(result$ctri_number)
    sos$noofsites <- result$sos_noofsites
    sos$pi <- result$sos_pi_name
    sos$name <- result$sos_sitename
    sos$address <- paste("",result$sos_siteaddress,"")
    if(length(page)==4){
      sos$phonefaxemail <- paste("",result$sos_phonefax,"")
    }

    sos <- as.data.frame(sos)
    dbWriteTable(mydb, "sos", sos, append = TRUE )
    write.table(sos, "sites_of_study.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("sites_of_study.csv"), append = T)

  }

  #----------------------------Ethical Committee table--------------
```

```
  x <- sub("...", "", toString(which(grepl("Details of Ethics Committee", tmp$col1, fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Regulatory Clearance Status from DCGI", tmp$col1,
fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))
  x <- as.numeric(x)
  y <- as.numeric(y)

  result$ec_noofcom<- gsub(".*\\No of Ethics Committees=", "", sub("...", "",
toString(tmp$col1[which(grepl("No of Ethics Committees=", tmp$col1, fixed=TRUE))])))


  for (item in (x:y)) {
    if (identical(toString(tmp$col1[[item]]) , "Name of Committee" )){
      xa <- item+1
    }

  }

  y <- y-1

  for (j in (xa:y)) {
    result$ec_name <- toString(tmp$col1[[j]])
    result$ec_approval_status <- toString(tmp$col2[[j]])

    ec <- NULL

    ec$trial_id <- toString(i)
    ec$ctri_number <- toString(result$ctri_number)
    ec$noofcom <- result$ec_noofcom
    ec$name <- result$ec_name
    ec$approval_status <- result$ec_approval_status

    ec <- as.data.frame(ec)
    dbWriteTable(mydb, "ec", ec, append = TRUE )
    write.table(ec, "ethics_committee.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("ethics_committee.csv"), append = T)

  }

  #-----------------------------DCGI status table--------------

  x <- sub("...", "", toString(which(grepl("Regulatory Clearance Status from DCGI", tmp$col1,
fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Health Condition / Problems Studied", tmp$col1,
fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))
  x <- as.numeric(x)
  y <- as.numeric(y)
```

```
  for (item in (x:y)) {
    if (identical(toString(tmp$col1[[item]]) , "Status" )){
      xa <- item+1
    }

  }

  result$dcgi_status <- toString(tmp$col1[[xa]])
  dcgi <- NULL

  dcgi$trial_id <- toString(i)
  dcgi$ctri_number <- toString(result$ctri_number)
  dcgi$dcgi_status <- result$dcgi_status

  dcgi <- as.data.frame(dcgi)
  dbWriteTable(mydb, "dcgi", dcgi, append = TRUE )
  write.table(dcgi, "dcgi_status.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("dcgi_status.csv"), append = T)


  #----------------------------Health Condition table----------

  x <- sub("...", "", toString(which(grepl("Health Condition / Problems Studied", tmp$col1,
fixed=TRUE))))
  x <- gsub(",","",gsub("\\,.*",",",x))
  y <- sub("...", "", toString(which(grepl("Intervention / Comparator Agent", tmp$col1,
fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))
  x <- as.numeric(x)
  y <- as.numeric(y)


  for (item in (x:y)) {
    if (identical(toString(tmp$col1[[item]]) , "Health Type" )){
      xa <- item+1
    }

  }

  y <- y-1

  for (j in (xa:y)) {
    result$hc_type <- toString(tmp$col1[[j]])
    result$hc_condition <- toString(tmp$col2[[j]])

    hc <- NULL

    hc$trial_id <- toString(i)
    hc$ctri_number <- toString(result$ctri_number)
    hc$type <- result$hc_type
    hc$condition <- result$hc_condition
```

```
  hc <- as.data.frame(hc)

  dbWriteTable(mydb, "hc", hc, append = TRUE )

  write.table(hc, "health_condition.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("health_condition.csv"), append = T)

 }

 #----------------------------Primary Outcome table--------------

 x <- sub("...", "", toString(which(grepl("Primary Outcome", tmp$col1, fixed=TRUE))))
 x <- gsub(",","",gsub("\\,.*",",",x))
 y <- sub("...", "", toString(which(grepl("Secondary Outcome", tmp$col1, fixed=TRUE))))
 y <- gsub(",","",gsub("\\,.*",",",y))
 x <- as.numeric(x)
 y <- as.numeric(y)


 for (item in (x:y)) {
  if (identical(toString(tmp$col1[[item]]) , "Outcome" )){
    xa <- item+1
  }

 }

 y <- y-1

 for (j in (xa:y)) {
  result$po_outcome <- toString(tmp$col1[[j]])
  result$po_timepoint <- toString(tmp$col2[[j]])

  po <- NULL

  po$trial_id <- toString(i)
  po$ctri_number <- toString(result$ctri_number)
  po$outcome <- result$po_outcome
  po$timepoint <- result$po_timepoint

  po <- as.data.frame(po)
  dbWriteTable(mydb, "po", po, append = TRUE )

  write.table(po, "primary_outcome.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("primary_outcome.csv"), append = T)

 }

 #----------------------------Secondary Outcome table--------------

 x <- sub("...", "", toString(which(grepl("Secondary Outcome", tmp$col1, fixed=TRUE))))
 x <- gsub(",","",gsub("\\,.*",",",x))
```

```
  y <- sub("...", "", toString(which(grepl("Target Sample Size", tmp$col1, fixed=TRUE))))
  y <- gsub(",","",gsub("\\,.*",",",y))
  x <- as.numeric(x)
  y <- as.numeric(y)


  for (item in (x:y)) {
    if (identical(toString(tmp$col1[[item]]) , "Outcome" )){
      xa <- item+1
    }

  }

  y <- y-1

  for (j in (xa:y)) {
    result$so_outcome <- toString(tmp$col1[[j]])
    result$so_timepoint <- toString(tmp$col2[[j]])

    so <- NULL

    so$trial_id <- toString(i)
    so$ctri_number <- toString(result$ctri_number)
    so$outcome <- result$so_outcome
    so$timepoint <- result$so_timepoint

    so <- as.data.frame(so)
    dbWriteTable(mydb, "so", so, append = TRUE )

    write.table(so, "secondary_outcome.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("secondary_outcome.csv"), append = T)

  }

  #-----------------------------Target Sample Size table--------------

  result$tss <- sub("...", "", toString(tmp$col2[which(grepl("Target Sample Size", tmp$col1,
fixed=TRUE))]))

  tss <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    target_sample_size <- toString(result$tss)
  )

  tss <- as.data.frame(tss)
  dbWriteTable(mydb, "tss", tss, append = TRUE )

  write.table(tss, "target_sample_size.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("target_sample_size.csv"), append = T)

  #-----------------------------Estimated Duration of Trial table--------------
```

```
  result$edt <- sub("...", "", toString(tmp$col2[which(grepl("Estimated Duration of Trial", tmp$col1,
fixed=TRUE))]))

  edt <- data.frame(
    trial_id <- toString(i),
    ctri_number <- toString(result$ctri_number),
    estimated_duration <- toString(result$edt)
  )

  edt <- as.data.frame(edt)
  dbWriteTable(mydb, "edt", edt, append = TRUE )

  write.table(edt, "estimated_duration.csv", sep = ",",row.names = FALSE, col.names =
!file.exists("estimated_duration.csv"), append = T)

  #----------------------------CHECK--------------

  counter = counter + 1
  print(paste("Count = ", counter,"ID = ",i))
}

file.remove("output.csv")
file.remove("raw.csv")
dbDisconnect(mydb)
#code might require modifications based on new features in CTRI trial records webpages

# Further adjustments to the database created for increased usability
con <- dbConnect(RSQLite::SQLite(), "ctri-ecsite.sqlite")
fin <- dbConnect(RSQLite::SQLite(), "IC_CTRIdb.sqlite")

#---------------------urls table-----------------

urls <- dbGetQuery(con, 'select  * from urls')
names(urls)[names(urls) == "trial_id....toString.i."] <- "trial_id"
names(urls)[names(urls) == "ctri_number....toString.result.ctri_number."] <- "ctri_number"
names(urls)[names(urls) == "ctri_url....toString.myurl."] <- "ctri_url"
dbWriteTable(fin, "urls", urls)

#---------------------reg_details table-----------------

reg_details <- dbGetQuery(con, 'select  * from reg_details')
names(reg_details)[names(reg_details) == "trial_id....toString.i."] <- "trial_id"
names(reg_details)[names(reg_details) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
names(reg_details)[names(reg_details) ==
"registered_on....toString.paste.....result.registered_on......."] <- "registered_on"
names(reg_details)[names(reg_details) == "reg_type....toString.result.reg_type."] <- "reg_type"
dbWriteTable(fin, "reg_details", reg_details)

#---------------------study_details table-----------------
```

```
study_details <- dbGetQuery(con, 'select  * from study_details')
names(study_details)[names(study_details) == "trial_id....toString.i."] <- "trial_id"
names(study_details)[names(study_details) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
names(study_details)[names(study_details) == "type_of_trial....toString.result.type_of_trial."] <-
"type_of_trial"
names(study_details)[names(study_details) == "type_of_study....toString.result.type_of_study."] <-
"type_of_study"
names(study_details)[names(study_details) == "study_design....toString.result.study_design."] <-
"study_design"
names(study_details)[names(study_details) == "phase....toString.result.phase."] <- "phase"
names(study_details)[names(study_details) == "post_graduate_thesis....result.pgt"] <-
"post_graduate_thesis"
names(study_details)[names(study_details) == "acronym....result.acronym"] <- "acronym"
dbWriteTable(fin, "study_details", study_details)

#----------------------study_name table-----------------

study_name <- dbGetQuery(con, 'select  * from study_name')
names(study_name)[names(study_name) == "trial_id....toString.i."] <- "trial_id"
names(study_name)[names(study_name) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
names(study_name)[names(study_name) == "public_title....result.public_title"] <- "public_title"
names(study_name)[names(study_name) == "scientific_title....result.scientific_title"] <-
"scientific_title"
dbWriteTable(fin, "study_name", study_name)

#----------------------dates table-----------------

dates <- dbGetQuery(con, 'select  * from dates')
names(dates)[names(dates) == "trial_id....toString.i."] <- "trial_id"
names(dates)[names(dates) == "ctri_number....toString.result.ctri_number."] <- "ctri_number"
names(dates)[names(dates) == "registered_on....toString.paste......result.registered_on......."] <-
"registered_on"
names(dates)[names(dates) == "last_modified_on....toString.paste......result.lmo......."] <-
"last_modified_on"
names(dates)[names(dates) == "date_of_first_enrollment_india....toString.paste......result.dfei.."] <-
"date_of_first_enrollment_india"
names(dates)[names(dates) == "date_of_first_enrollment_global....toString.paste......result.dfeg.."]
<- "date_of_first_enrollment_global"
names(dates)[names(dates) == "date_of_study_completion_india....toString.paste......result.dsci.."]
<- "date_of_study_completion_india"
names(dates)[names(dates) == "date_of_study_completion_global....toString.paste......result.dscg.."]
<- "date_of_study_completion_global"
dbWriteTable(fin, "dates", dates)

#----------------------inclusion table-----------------

inclusion <- dbGetQuery(con, 'select  * from inclusion')
names(inclusion)[names(inclusion) == "trial_id....toString.i."] <- "trial_id"
names(inclusion)[names(inclusion) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
```

```
names(inclusion)[names(inclusion) == "inclusion_age_from....result.inclusion_age_from"] <-
"inclusion_age_from"
names(inclusion)[names(inclusion) == "inclusion_age_to....result.inclusion_age_to"] <-
"inclusion_age_to"
names(inclusion)[names(inclusion) == "inclusion_gender....result.inclusion_gender"] <-
"inclusion_gender"
names(inclusion)[names(inclusion) == "inclusion_details....result.inclusion_details"] <-
"inclusion_details"
dbWriteTable(fin, "inclusion", inclusion)

#----------------------exclusion table-----------------

exclusion <- dbGetQuery(con, 'select  * from exclusion')
names(exclusion)[names(exclusion) == "trial_id....toString.i."] <- "trial_id"
names(exclusion)[names(exclusion) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
names(exclusion)[names(exclusion) == "exclusion_details....result.inclusion_details"] <-
"exclusion_details"
dbWriteTable(fin, "exclusion", exclusion)

#----------------------intervention_comparator table-----------------

intervention_comparator <- dbGetQuery(con, 'select  * from intervention_comparator')
dbWriteTable(fin, "intervention_comparator", intervention_comparator)

#----------------------method table-----------------

method <- dbGetQuery(con, 'select  * from method')
names(method)[names(method) == "trial_id....toString.i."] <- "trial_id"
names(method)[names(method) == "ctri_number....toString.result.ctri_number."] <- "ctri_number"
names(method)[names(method) ==
"method_of_generating_random_sequence....result.method_gsr"] <-
"method_of_generating_random_sequence"
names(method)[names(method) == "method_concealment....result.method_concealment"] <-
"method_concealment"
names(method)[names(method) == "blinding_masking....result.bm"] <- "blinding_masking"
dbWriteTable(fin, "method", method)

#----------------------secondary_id table-----------------

secondary_id <- dbGetQuery(con, 'select  * from secondary_id')
dbWriteTable(fin, "secondary_id", secondary_id)

#----------------------brief_summary table-----------------

brief_summary <- dbGetQuery(con, 'select  * from brief_summary')
names(brief_summary)[names(brief_summary) == "trial_id....toString.i."] <- "trial_id"
names(brief_summary)[names(brief_summary) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
names(brief_summary)[names(brief_summary) == "brief_summary....toString.result.bs."] <-
"brief_summary"
dbWriteTable(fin, "brief_summary", brief_summary)
```

```
#----------------------publication table-----------------

publication <- dbGetQuery(con, 'select  * from publication')
names(publication)[names(publication) == "trial_id....toString.i."] <- "trial_id"
names(publication)[names(publication) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
names(publication)[names(publication) == "publication....toString.result.publication."] <-
"publication"
dbWriteTable(fin, "publication", publication)

#----------------------recruitment table-----------------

recruitment <- dbGetQuery(con, 'select  * from recruitment')
names(recruitment)[names(recruitment) == "trial_id....toString.i."] <- "trial_id"
names(recruitment)[names(recruitment) == "ctri_number....toString.result.ctri_number."] <-
"ctri_number"
names(recruitment)[names(recruitment) == "recruitment_status_india....toString.result.rsti."] <-
"recruitment_status_india"
names(recruitment)[names(recruitment) == "recruitment_status_global....toString.result.rstg."] <-
"recruitment_status_global"
dbWriteTable(fin, "recruitment", recruitment)

#----------------------pi table-----------------

pi <- dbGetQuery(con, 'select  * from pi')
dbWriteTable(fin, "pi", pi)

#----------------------contact_person_scientific_query table-----------------

contact_person_scientific_query <- dbGetQuery(con, 'select  * from cpsq')
dbWriteTable(fin, "contact_person_scientific_query", contact_person_scientific_query)

#----------------------contact_person_public_query table-----------------

contact_person_public_query <- dbGetQuery(con, 'select  * from cppq')
dbWriteTable(fin, "contact_person_public_query", contact_person_public_query)

#----------------------source_of_monetory_material_support table-----------------

source_of_monetory_material_support <- dbGetQuery(con, 'select  * from smms')
dbWriteTable(fin, "source_of_monetory_material_support", source_of_monetory_material_support)

#----------------------primary_sponsor table-----------------

primary_sponsor <- dbGetQuery(con, 'select  * from ps')
dbWriteTable(fin, "primary_sponsor", primary_sponsor)

#----------------------secondary_sponsor table-----------------

secondary_sponsor <- dbGetQuery(con, 'select  * from ss')
dbWriteTable(fin, "secondary_sponsor", secondary_sponsor)
```

```
#----------countries_of_recruitment table----------

countries_of_recruitment <- dbGetQuery(con, 'select  * from cor')
names(countries_of_recruitment)[names(countries_of_recruitment) == "trial_id....toString.i."] <-
"trial_id"
names(countries_of_recruitment)[names(countries_of_recruitment) ==
"ctri_number....toString.result.ctri_number."] <- "ctri_number"
names(countries_of_recruitment)[names(countries_of_recruitment) ==
"countries_of_recruitment....toString.result.cor."] <- "countries_of_recruitment"
dbWriteTable(fin, "countries_of_recruitment", countries_of_recruitment)

#----------sites_of_study table----------

sites_of_study <- dbGetQuery(con, 'select  * from sos')
dbWriteTable(fin, "sites_of_study", sites_of_study)

#----------ethics_committee table----------

ethics_committee <- dbGetQuery(con, 'select  * from ec')
dbWriteTable(fin, "ethics_committee", ethics_committee)

#----------dcgi_status table----------

dcgi_status <- dbGetQuery(con, 'select  * from dcgi')
dbWriteTable(fin, "dcgi_status", dcgi_status)

#----------health_condition table----------

health_condition <- dbGetQuery(con, 'select  * from hc')
dbWriteTable(fin, "health_condition", health_condition)

#----------primary_outcome table----------

primary_outcome <- dbGetQuery(con, 'select  * from po')
dbWriteTable(fin, "primary_outcome", primary_outcome)

#----------secondary_outcome table----------

secondary_outcome <- dbGetQuery(con, 'select  * from so')
dbWriteTable(fin, "secondary_outcome", secondary_outcome)

#----------target_sample_size table----------

target_sample_size <- dbGetQuery(con, 'select  * from tss')
names(target_sample_size)[names(target_sample_size) == "trial_id....toString.i."] <- "trial_id"
names(target_sample_size)[names(target_sample_size) ==
"ctri_number....toString.result.ctri_number."] <- "ctri_number"
names(target_sample_size)[names(target_sample_size) ==
"target_sample_size....toString.result.tss."] <- "target_sample_size"
dbWriteTable(fin, "target_sample_size", target_sample_size)
```

```
#----------------------estimated_duration table-----------------

estimated_duration <- dbGetQuery(con, 'select  * from edt')
names(estimated_duration)[names(estimated_duration) == "trial_id....toString.i."] <- "trial_id"
names(estimated_duration)[names(estimated_duration) ==
"ctri_number....toString.result.ctri_number."] <- "ctri_number"
names(estimated_duration)[names(estimated_duration) ==
"estimated_duration....toString.result.edt."] <- "estimated_duration"
dbWriteTable(fin, "estimated_duration", estimated_duration)

dbDisconnect(fin)
dbDisconnect(con)

# database file ready for use after code completion - "IC_CTRIdb.sqlite".
# other  files may be deleted or kept for ease of backtracing.
```