00_Preprocessing_MSD_refactored

main	00_helper_product_data.R	00_a4d_patient_tracker_extract_helper.R

main

- store output (not used) path_output <- "xx"
- tracker excel codebook_data_file <- "xx"
- Get data tracker_data_file <- "xx"
- Main function to extract product data reading_a4d_products_from_tracker <- function(tracker_data_file, codebook_data_file)
- Get column names and cleaned names to match to columns_synonyms <-
- Extracting info sheet_list <- *from tracker_data_file* month_list <- *from sheet_list* year <- *from month_list*
- **Looping through each month**

CurrSheet in month_list

- read in current sheet as data frame tracker_data <- ...
- Get country code cc_codes <-
- Save country and clinic code country_code <- *from cc_codes* clinic_code <- *from cc_codes*
- sheet contains string "product" or "Description of Support"
- product_df <-
- check that product_df not empty
- Get cleaned column names product_df <-
- Remove rows where <=1 entry in a column is not NA del_rows <- *rows to delete* product_df <- remove del_rows
- if 2021
- Fill product info down and **remove last row** of each product product_df <- ...
- Split product cells with multiple products product_df <- .
- Add columns missing according to codebook_data_file and set to NA missing_cols <.. missing_cols_names <- ... product_df[missing_cols_names]
- Add index column
- Recode Excel dates product_df <-
- If not done already, fill products and **sort by name, and date**
- All NA to 0 product_df <-
- Cleaning column product_received_from product_df <-
- Cleaning column product_units_received product_df <-
- All NA to 0 product_df <-
- Compute balance part1 product_df <-
- Compute balance part2 product_df <-
- Compute balance part3 product_df <-
- Add info to table product_df <- add info to columns: country, hospital, month, year, sheet name
- Adjust classes product_df <-
- Remove index product_df <- ...
- Add to final if not already exists **df_final** <- rbind(df_final, product_df)
- return(df_final)

00_helper_product_data.R

- Saves the data from the excel for product data read_column_synonyms_product(codebook_data_file)
- Extracted df with 2 columns: clean column name & column name to match to
- Extracting all product data from sheet extract_product_data <- function(tracker_data)
- Extract start and end of table **Loop through df**
  - Extracting all product data from sheet start <- get_msd_start(...) end <- get_msd_end(..)
    - helper_is_msd_start_row <- function(df, i) helper_is_msd_end_row <- function(df, i)
  - by looping through each row, script checks for specific strings, if they exists, saves the index
- Extracting product table product_data_df <- monthly_tracker_df[...]
- Get correct headers product_data_df <- set_second_row_as_headers_and_remove_first_row(
- Get correct headers harmonize_input_data_columns <- function(product_df, columns_synonyms)
- Remove empty columns /headers
- Get cleaned names via sanitize_column_name and then match
- important step, because stuff later makes problems
- extract_product_multiple <- function(product_df)
- Split rows which have multiple products in "product" column
- Add column "product_units_notes"
- Generate product_units_received and product_units_received if not existing and fill with NA
- Add column "product_units_notes " and fill with units extracted from product name
- *Loop through table and recode units received / released if stuff true??*
- format_date(product_df)
  - recode depending on format format_date_excelnum <- function(product_df) format_date_exceldate <- function(product_df)
- date and balance don't match
- NA to 0 for columns product_units_* recode_unitcolumnstozero (product_df)
- Check for START in product_units_received and put product_received_from to NA in rows with numeric & non NA entry in product_received_from clean_receivedfrom(product_df)
- Put character values to = when having START|END|BALANCE in column product_units_received clean_unitsreceived(product_df)
- NA to 0 for columns product_units_* Rerun again in case NA generated before recode_unitcolumnstozero (product_df)
- delete rows where all product info NA compute_balance_cleanrows (product_df)
- identify start and end of each product and define product_balance_status: start, end, change compute_balance_status (product_df)
- start/end wrong because of error above (2021 data)
- compute_balance (product_df, year)
- put balance columns to numeric
- start balance: take units_received value if NA
- if year < 2021
- loop through each row and calculate balance= balance of last date - released current data + received current data
- if year >= 2021
- loop through each row and calculate balance= balance of last date - released current data + received current data, **BUT "end" last row will be kept because it should be end balance (but it's not)** !!!
- columns names which should be character, date, numeric adjust_column_classes(product_df)

00_a4d_patient_tracker_extract_helper.R

- Get country code extract_country_clinic_code (tracker_data)
  - wrong code for 2021 data
- takes 2nd column: extracts if match with string "country" or "clinic" otherwise first string containing "_" is processed