

linkrep: Data Linkage Quality Reports

Documentation

Authors: Elizabeth Stoughton
Barret Monchka

Table of Contents

| | | |
|-------|--|----|
| 1 | linkage_quality_report() | 2 |
| 2 | Data Inputs | 8 |
| 2.1 | main_data | 8 |
| 2.2 | missing_data_inidcaters | 9 |
| 2.3 | algorithm_summary_data | 11 |
| 2.4 | performance_measures_data | 11 |
| 2.5 | definitions | 12 |
| 2.6 | abbreviations | 12 |
| 2.7 | How to format data for report | 12 |
| 3 | Report Customizations | 14 |
| 3.1 | Quarto templates | 14 |
| 3.1.1 | Modify text elements | 14 |
| 3.1.2 | Add or remove tables & figures | 14 |
| 3.1.3 | Set metadata | 18 |
| 3.1.4 | Quarto tips: Understanding the report template | 19 |
| 3.2 | References | 22 |
| 3.2.1 | Modify reference list | 22 |
| 3.2.2 | Modify citation style | 24 |
| 3.3 | Word template | 25 |
| 3.3.1 | Modify font | 25 |
| 3.3.2 | Modify footers & headers | 27 |
| 3.3.3 | Additional resources | 28 |
| 3.4 | Modify background images | 28 |
| 3.5 | Table customizations | 30 |

1 linkage_quality_report()

Description

The `linkage_quality_report` function takes in data from a linkage and outputs a data linkage quality report. The report contains information on data linkage and provides figures and tables to describe the data.

Arguments

| | |
|---|---|
| <code>main_data</code> | A data frame, a file path to an rds file that contains a data frame or a file path to a csv file. This data contains variables present in the left dataset of the linkage. |
| <code>report_title</code> | String indicating the title of the report. If <code>output_format = "docx"</code> , the title will only be used in the suggested citation. |
| <code>report_subtitle</code> | String indicating the subtitle of the report. If <code>output_format = "docx"</code> , the subtitle will only be used in the suggested citation. |
| <code>left_dataset_name</code> | String indicating the name of the left dataset. |
| <code>right_dataset_name</code> | String indicating the name of the right dataset. |
| <code>output_dir</code> | A path to a directory. All output files will be saved here. |
| <code>data_linker</code> | String indicating who performed the linkage. |
| <code>linkage_package</code> | String indicating the R package used to link the data. |
| <code>stratified_linkage_tbls_column_var</code> | A string of the name of a logical or binary variable present in <code>main_data</code> that indicates whether a record linked or not. Its values will be the columns in the linkage rate table and the linkaed data representativeness table. |
| <code>linked_data_representativeness_tbl_strata_vars</code> | A character vector of the names of the variables present in <code>main_data</code> to stratify the linked data representativeness table by. |
| <code>linkage_rate_tbl_strata_vars</code> | A character vector of the names of the variables present in <code>main_data</code> to stratify the linkage rate table by. |
| <code>linked_data_representativeness_tbl_footnotes</code> | A character vector of additional footnotes for the linked data representativeness table. Each element in the vector will be displayed on a new line. |

| | |
|--|--|
| <code>linkage_rate_tbl_footnotes</code> | A character vector of additional footnotes for the linkage rate table. Each element in the vector will be displayed on a new line. |
| <code>linkage_rate_tbl_display_total_col</code> | A logical indicating whether to display a third column in the linkage rate table displaying the overall totals for each value. |
| <code>stratified_linkage_tbls_continuous_stat</code> | A string indicating which statistic to use on continuous variables in the linkage rate table. Allowed values are "mean" or "median" (default). If "mean", mean \pm standard deviation will be output otherwise, median (Q1, Q3), where Q1 is the 25 th percentile, and Q3 is the 75 th percentile, will be output. |
| <code>stratified_linkage_tbls_output_to_csv</code> | A logical indicating whether to save the linkage rate table in a csv file. Default is FALSE. |
| <code>display_missingness_table</code> | A logical indicating whether to display the missingness table in the report. |
| <code>missing_data_indicators</code> | A data frame, a file path to an rds file that contains a data frame or a file path to a csv file. All variables in the data must be logical or binary, with 1 or TRUE representing a missing record for that variable. See Details section for more information on naming conventions and where this data is used in the report. |
| <code>missingness_tbl_footnotes</code> | A character vector of additional footnotes for the missingness table. Each element in the vector will be displayed on a new line. |
| <code>output_format</code> | String specifying the desired output format. Allowed values are "pdf" or "docx". |
| <code>linkage_package_version</code> | String indicating the version of the <code>linkage_package</code> used to link the data. You can obtain the package version via <code>packageVersion("package_name")</code> . |
| <code>linkrep_package_version</code> | String indicating the version of <code>linkrep</code> used to generate the report. You can obtain the package version via <code>packageVersion("linkrep")</code> . |
| <code>R_version</code> | String indicating the version of R used to generate the report. You can obtain the version of R via <code>R.version.string</code> . |
| <code>datastan_package_version</code> | String indicating the version of <code>datastan</code> used to preprocess the data. You can obtain the package version via <code>packageVersion("datastan")</code> . |

| | |
|---|--|
| <code>comprehensive_report</code> | A logical indicating whether to output a comprehensive report. A comprehensive report includes the Background and Methods sections. |
| <code>save_linkage_rate</code> | A logical indicating whether to save information on the linkage in an SQLite file. See Details section for details on what is saved in the file. |
| <code>project_id</code> | String indicating the project ID. |
| <code>num_records_right_dataset</code> | The number of records in the right dataset of the linkage. |
| <code>acquisition_year_var</code> | A string of the name of the numeric variable in <code>main_data</code> that represents the acquisition year. This must be provided for the datasets date range to be output on the title page of a PDF output. |
| <code>acquisition_month_var</code> | A string of the name of the numeric variable in <code>main_data</code> that represents the acquisition month. |
| <code>algorithm_summary_data</code> | A data frame, a file path to an rds file that contains a data frame or a file path to a csv file. This data contains information on the linkage algorithm (ex. linking variables on each pass). |
| <code>algorithm_summary_tbl_footnotes</code> | A character vector of additional footnotes for the algorithm summary table. Each element in the vector will be displayed on a new line. |
| <code>performance_measures_data</code> | A data frame, a file path to an rds file that contains a data frame or a file path to a csv file. This data contains performance measures (ex. classification metrics such as sensitivity/recall). Data must be in percentages with values between 0 and 100 when using the default quarto template. |
| <code>performance_measures_tbl_footnotes</code> | A character vector of additional footnotes for the performance measures table. Each element in the vector will be displayed on a new line. |
| <code>ground_truth</code> | String indicating the ground truth used to produce the performance measures. Must be provided with <code>performance_measures_data</code> . |
| <code>num_pairs_non_missing_ground_truth</code> | The number of record pairs with non-missing ground truth. |
| <code>num_record_pairs</code> | The total number of record pairs (cartesian product). |
| <code>definitions</code> | A data frame, a file path to an rds file that contains a data frame or a file path to a csv file. Data must contain two columns: the list of terms in the first and their definitions in the second. |

| | |
|---|--|
| <code>definitions_display_header</code> | A logical indicating whether to display the column headers in the definitions table. Only applied when <code>output_format = "docx"</code> . |
| <code>abbreviations</code> | A data frame, a file path to an rds file that contains a data frame or a file path to a csv file. Data must contain two columns: the list of abbreviations in the first and their meaning in the second. |
| <code>abbreviations_display_header</code> | A logical indicating whether to display the column headers in the abbreviation table. Only applied when <code>output_format = "docx"</code> . |
| <code>thousands_separator</code> | A string specifying the style of the thousand's separator in all numeric values. Default is <code>" , "</code> . |
| <code>decimal_mark</code> | A string specifying the style of the decimal mark in all numeric values. Default is <code>" . "</code> . |
| <code>num_decimal_places</code> | A number specifying the number of digits to output after the decimal mark of all necessary numeric values. Default is 1. |
| <code>display_percent_symbol</code> | A logical indicating whether to display a percent symbol next to percentages in the linkage rate and missingness tables. Default is <code>FALSE</code> . |
| <code>text_font_size</code> | A number specifying the font size for the inline text. Default is 12. |
| <code>table_font_size</code> | A number specifying the font size for the table text. Default is 12. |
| <code>font_style</code> | A string specifying the font style. Must be present in <code>system_fonts()\$name</code> or <code>system_fonts()\$family</code> . See system_fonts for more details. |
| <code>cover_page</code> | A file path to a png, pdf or jpg file that contains the desired cover page. Default is https://github.com/CHIMB/linkrep/blob/main/inst/background_images/cover_page.pdf . |
| <code>content_portrait_page</code> | A file path to a png, pdf or jpg file that contains the desired content portrait page. Default is https://github.com/CHIMB/linkrep/blob/main/inst/background_images/content_portrait_page.pdf . |
| <code>content_landscape_page</code> | A file path to a png, pdf or jpg file that contains the desired content landscape page. Default is https://github.com/CHIMB/linkrep/blob/main/inst/background_images/content_landscape_page.pdf . |
| <code>display_back_cover_page</code> | A logical indicating whether to display the back cover page in the output. |

| | |
|--|---|
| <code>back_cover_page</code> | A file path to a png, pdf or jpg file that contains the desired back cover page. Default is https://github.com/CHIMB/linkrep/blob/main/inst/background_images/back_cover_page.pdf . |
| <code>blank_background</code> | A logical indicating whether to display the report on blank white background. Default is <code>FALSE</code> . |
| <code>temp_data_output_dir</code> | A path to a directory. All complex data (ex. tables) must be passed into the quarto report through an rds file therefore, all temporarily generated rds files containing the report elements will be stored in this directory. Default is <code>tempdir(check = TRUE)</code> |
| <code>quarto_report_template</code> | A file path to a quarto (qmd) file that renders the report. Use this parameter to apply additional customization to the report output. Default is https://github.com/CHIMB/linkrep/blob/main/inst/templates/base_quarto_report_template.qmd . |
| <code>extra_textual_content_quarto_template</code> | A file path to a quarto (qmd) file that contains the Background and Methods sections of the report. Default is <code>system.file("templates", "base_quarto_report_template.qmd", package = "linkrep")</code> . |
| <code>references</code> | A file path to a BibTex (bib) file that contains the references used in the report. For references to be displayed in the References section of the report, they must be cited in the quarto document. Default is https://github.com/CHIMB/linkrep/blob/main/inst/templates/references.bib . |
| <code>word_template</code> | A file path a to a word document that specifies the output styles for a word report. Default is https://github.com/CHIMB/linkrep/blob/main/inst/templates/word_template.docx . |
| <code>set_background_images_template</code> | A file path to a LaTeX file that specifies how the background images are placed onto a PDF report. Default is https://github.com/CHIMB/linkrep/blob/main/inst/templates/set_background_images.tex . |
| <code>citation_style</code> | A file path to a csl file containing the citation style. To find different styles visit https://github.com/citation-style-language/styles/blob/master/american-medical-association.csl . If the location of the citation must change in the text you must modify its location in the quarto report template. |

Details

All tables display the variable labels in their headings before reverting to the raw variable name. To label variables in your data before calling this function use `label`.

Information on `missing_data_indicators`:

- Used in the linkage rate table, linked data representativeness table and the missingness table. All variables present in the data will be displayed in both tables.

Linkage rate and linked data representativeness table:

- Variables associated with those in `main_data` must either have the same variable name suffixed by `"_missing"` or have the same label for it to be displayed in the linkage rate table as a value of that variable. In this case, the variable will be relabelled `"Missing"` and tabbed under the header of the variable it's associated with. If the variable is not associated with one in `main_data` it will be relabelled with its label or variable name prefixed by `"Missing "`.

Missingness table:

- No label changes are made in this table. It is made entirely up of the data from this dataset.

Information on `save_linkage_rate`:

If `save_linkage_rate = TRUE`, an SQLite file (`linkage_rate.sqlite`) will be saved in the `output_dir`.

The file will include the following:

- The report generation date
- The report generation year
- The name of the data linker (provided by `data_linker`)
- The name of the left dataset (provided by `left_dataset_name`)
- The name of the right dataset (provided by `right_dataset_name`)
- The overall linkage rate
- The acquisition dates of the left dataset
- The project ID

2 Data Inputs

2.1 main_data

main_data includes the de-identified data from the left dataset as well as a binary or logical variable indicating whether the record linked or not. Missing values in the dataset should be represented by NA. See the [How to Format Data for Report](#) section for details on formatting data for desired output.

E.g.,

| | gender Gender | birth_year Birth Year | linked |
|----------|-------------------------|---------------------------------|---------------|
| 1 | M | 1963 | 1 |
| 2 | F | 2003 | 0 |
| 3 | F | 1965 | 1 |

main_data is used in the creation of the following elements:

Tables:

- Linked data representativeness table
- Linkage rate table

Figures:

- Linkage rate over time bar plot

Values:

- Linkage rate
- Number of records in the left dataset

Why isn't the linkage rates over time plot displaying?

The plot will only display if acquisition year and acquisition month variables are present in main_data and passed to **linkage_quality_report** through acquisition_year_var and acquisition_month_var

E.g.,

| | gender Gender | birth_year Birth Year | linked | acq_year Acquisition Year | acq_month Acquisition Month |
|---|-------------------------|---------------------------------|---------------|-------------------------------------|---------------------------------------|
| 1 | M | 1963 | 1 | 2010 | 10 |
| 2 | F | 2003 | 0 | 2022 | 3 |
| 3 | F | 1965 | 1 | 2015 | 1 |

If the plot is still not displaying, the date range is not wide enough to display a meaningful plot (e.g., if `acq_year` = 2024 and `acq_month` = 4 for all records, then only one bar can be made in the plot hence, there's no distribution to look at therefore, the plot will never display).

2.2 missing_data_indicators

`missing_data_indicators` includes binary or logical missingness indicators for the variables you wish to be displayed in the linked data representativeness table, the linkage rate table and the missingness table.

E.g.,

| | first_name_missing Missing Given Name | last_name_missing Missing Surname | postal_code_missing Missing Postal Code |
|---|---|---|---|
| 1 | TRUE | FALSE | TRUE |
| 2 | FALSE | FALSE | FALSE |
| 3 | FALSE | FALSE | TRUE |

Naming conventions:

If you have variables listed in your `strata_vars` that also appear in `missing_data_indicators`, use one of the following naming conventions for them to be recognized as the same variable:

1. Give them the same variable name suffixed by “_missing”

E.g., `main_data:`

| | gender | birth_year | linked |
|---|---------------|-------------------|---------------|
| 1 | NA | 1963 | 1 |
| 2 | F | 2003 | 0 |
| 3 | F | 1965 | 1 |

`missing_data_indicators:`

| | gender_missing | first_name_missing |
|---|-----------------------|---------------------------|
| 1 | 1 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 0 |

2. Label them the same

E.g., `main_data`:

| | gender | birth_year | linked |
|---|--------|------------|--------|
| | Gender | | |
| 1 | NA | 1963 | 1 |
| 2 | F | 2003 | 0 |
| 3 | F | 1965 | 1 |

`missing_data_indicators`:

| | mis_gender | first_name_missing |
|---|------------|--------------------|
| | Gender | |
| 1 | 1 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 0 |

Linkage rate table with naming conventions:

| | Linked (N = 2, 66.7%) | Unlinked (N = 1, 33.3%) |
|---------------------------------|--------------------------|----------------------------|
| Gender | | |
| F | 1 (50.0) | 1 (50.0) |
| Missing | 1 (100.0) | 0 (0.0) |
| birth_year | | |
| 1963 | 1 (100.0) | 0 (0.0) |
| 1965 | 0 (0.0) | 1 (100.0) |
| 2003 | 1 (100.0) | 0 (0.0) |
| Missing first_name | 0 (0.0) | 1 (100.0) |
| Data are presented as n (row %) | | |

Linkage rate table without naming conventions:

Shows up twice

| | Linked (N = 2, 66.7%) | Unlinked (N = 1, 33.3%) |
|---------------------------------|--------------------------|----------------------------|
| Gender | | |
| F | 1 (50.0) | 1 (50.0) |
| Missing | 1 (100.0) | 0 (0.0) |
| birth_year | | |
| 1963 | 1 (100.0) | 0 (0.0) |
| 1965 | 0 (0.0) | 1 (100.0) |
| 2003 | 1 (100.0) | 0 (0.0) |
| Missing mis_gen | 1 (100.0) | 0 (0.0) |
| Missing first_name | 0 (0.0) | 1 (100.0) |
| Data are presented as n (row %) | | |

Missingness is pulled from the NA values of that variable therefore, the only reason to include the same variable in both **main_data** and **missing_data_indicators** is to display its missingness in the missingness table. To ensure the missingness table is displayed, set the `display_missingness_table` parameter to `TRUE`.

For the other variables that do not appear in `main_data`, “**Missing**” will be appended to the front of either the label or variable name (e.g., label = “Surname”, output = “Missing Surname”, label = “Missing name”, output = “Missing Missing name”).

missing_data_indicators is used in the creation of the following elements:

Tables:

- Linked data representativeness table
- Linkage rate table
- Missingness table

2.3 algorithm_summary_data

algorithm_summary_data is used in the creation of the following element:

- Algorithm summary table

The data is output exactly as it is received, so format it according to your desired output.

E.g.,

| | blocking_vars Blocking Schema | matching_vars Matching Criteria | linkage_rate Linkage Rate |
|---|---|---|-------------------------------------|
| 1 | given name, surname | birth year | 72.3 |
| 2 | given name, birth year, birth month | surame, sex | 5.5 |

2.4 performance_measures_data

performance_measures_data is used in the creation of the following elements:

Tables:

- Performance measures table

Figures:

- Performance measures radar chart

The data is output exactly as it is received, so format it according to your desired output.

E.g.,

| | ppv Positive Predictive Value | npv Negative Predictive Value | sensitivity Sensitivity | specificity Specificity |
|---|---|---|-----------------------------------|-----------------------------------|
| 1 | 80.7 | 99 | 92.4 | 100 |

2.5 definitions

definitions includes term definitions for important terms used throughout the report. It should be formatted as a two-column dataset: Term and Definition.

E.g.,

| Term | Definition |
|---------------------------------|---|
| Positive predictive value (PPV) | Proportion of predicted positive matches that are truly positive |
| Negative predictive value (NPV) | Proportion of predicted negative matches that are truly negative |
| Sensitivity | Proportion of positive matches the algorithm correctly identified |
| Specificity | Proportion of negative matches the algorithm correctly identified |

2.6 abbreviations

abbreviations includes abbreviations used throughout the report. It should be formatted as a two-column dataset: Abbreviation and Definition.

E.g.,

| Abbreviation | Definition |
|--------------|---|
| CHI | George & Fay Yee Centre for Healthcare Innovation |
| MCHP | Manitoba Centre for Healthcare Policy |
| MH | Manitoba Health |
| PHIN | Personal Health Identification Number |

2.7 How to format data for report

Data elements are used as received, so the more formatted they are, the cleaner the output will be.

Variable Names:

To modify the output of the variable names in the tables do one of the following:

1. Label the variable names using `label()` in the `Hmisc` package (Harrell Jr F (2024). `Hmisc`: Harrell Miscellaneous. R package version 5.1-3, <https://CRAN.R-project.org/package=Hmisc>).

E.g.,

```
label(data$birth_year) <- "Birth Year"
label(data$gender) <- "Gender"
```

| | gender | birth_year | linked |
|---|--------|------------|--------|
| | Gender | Birth Year | |
| 1 | M | 1963 | 1 |
| 2 | F | 2003 | 0 |
| 3 | F | 1965 | 1 |

Example linkage rate table:

Labelled

Not labelled

| | Linked (N = 2, 66.7%) | Unlinked (N = 1, 33.3%) |
|---------------------------------|--------------------------|----------------------------|
| Gender | | |
| F | 1 (50.0) | 1 (50.0) |
| Missing | 1 (100.0) | 0 (0.0) |
| birth_year | | |
| 1963 | 1 (100.0) | 0 (0.0) |
| 1965 | 0 (0.0) | 1 (100.0) |
| 2003 | 1 (100.0) | 0 (0.0) |
| Missing first_name | 0 (0.0) | 1 (100.0) |
| Data are presented as n (row %) | | |

2. Name the original variables as desired

E.g.,

| | Gender | Birth Year | linked |
|---|--------|------------|--------|
| 1 | M | 1963 | 1 |
| 2 | F | 2003 | 0 |
| 3 | F | 1965 | 1 |

Variable values:

Values output in the stratified tables are output as received.

If you wish to change their values or categorize them, you must do so before passing the data to the function through **main_data**.

E.g.,

```

birth_year_cut <- c(-Inf, 1970, 1980, 1990, 2000, Inf)
birth_year_labels <- c("<1970", "1970-1979", "1980-1989",
  "1990-1999", "2000+")
data <- mutate(data,
  birth_year = cut(data$birth_year,
    birth_year_cut,
    birth_year_labels,
    right = FALSE))
label(data$sex) <- "Sex"
label(data$birth_year) <- "Birth Year"

```

Linked data representativeness table:

| | Linked (N = 489) | Total (N = 1,000) |
|------------------------------------|---------------------|----------------------|
| Sex | | |
| Female | 265 (54.2) | 518 (51.8) |
| Male | 224 (45.8) | 482 (48.2) |
| Birth Year | | |
| <1970 | 113 (23.1) | 228 (22.8) |
| 1970-1979 | 103 (21.1) | 221 (22.1) |
| 1980-1989 | 92 (18.8) | 199 (19.9) |
| 1990-1999 | 126 (25.8) | 225 (22.5) |
| 2000+ | 55 (11.2) | 127 (12.7) |
| Data are presented as n (column %) | | |

3 Report Customizations

3.1 Quarto templates

To perform any of the following, you will need to modify at least one of the Quarto templates. If you have not done so already, download the default templates by visiting https://github.com/CHIMB/linkrep/blob/main/inst/templates/base_quarto_report_template.qmd and https://github.com/CHIMB/linkrep/blob/main/inst/templates/extra_textual_content.qmd and selecting “Download raw file” on each. Open the files in RStudio to begin editing.

If you need guidance on understanding the template files, please refer to the [Quarto tips: Understanding the report template](#) section.

3.1.1 Modify text elements

Step 1: Locate the text elements you wish to modify

Step 2: Modify the text:

- For more information on writing in Markdown visit <https://www.markdownguide.org/basic-syntax/>
- If you are incorporating dynamic elements (e.g., table reference) into the text, refer to the syntax used in the template, or for more guidance visit <https://quarto.org/docs/visual-editor/technical.html>
- To modify the citations (e.g., [[@fellegisunter](#)]), see the [References](#) section

Step 3: Save the files you modified and pass their paths to the **linkage_quality_report** function through the appropriate parameter: `base_quarto_report_template.qmd` → `quarto_report_template`; `extra_textual_content.qmd` → `extra_textual_content_quarto_template`.

3.1.2 Add or remove tables & figures

Note: The datasets passed into **linkage_quality_report** are not accessible in the Quarto report template therefore, they must be provided separately using one of the strategies below.

For each new element, you must create at least one new R chunk in the Quarto file template. Create the R chunk where you wish the element to output.

Option 1: Code the data and element directly into the Quarto file

Create the data directly in the R chunk:

```
{r}
data <- data.frame(a = c("apple", "apricot"), b = c("banana", "blueberry"))
```

Or import the data:

```
{r}
data <- read.csv("C:/Users/stoughte/Downloads/linked_data.csv")
```

Search “import __ file in R” where __ is a file type (e.g., xlsx) for guidance on how to import other file types

Then create the new element below your data:

```
{r generate new table, message=FALSE}
#| label: tbl-new_table
#| tbl-cap: This is a table about fruits
library(flextable)
data <- read.csv("C:/Users/stoughte/Downloads/linked_data.csv")
flextable(data)
```

Table caption and label

New table

Option 2: Source the data into the Quarto file and create the element directly in it

Open a new R file and create your data in it:

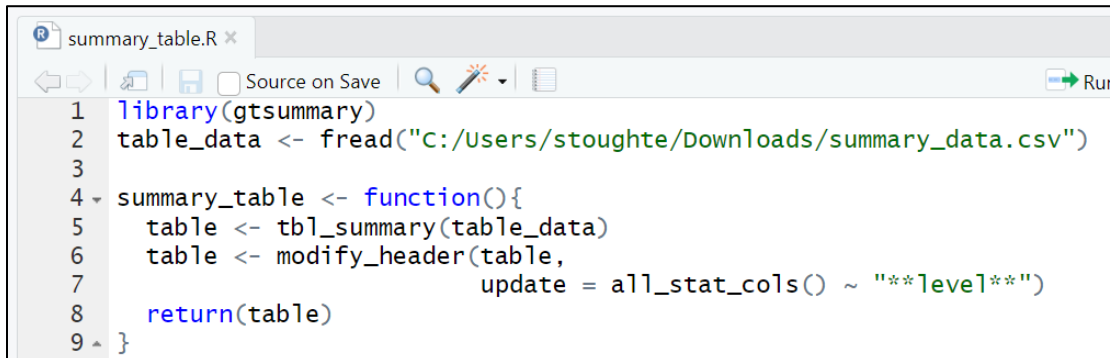
```
plot_data.R x
Source on Save
1 library(dplyr)
2 data <- read.csv("C:/Users/stoughte/Downloads/linked_data.csv")
3 data <- mutate(
4   data,
5   age = as_factor(age),
6   linkage_rate = ifelse(gender == "M", -linkage_rate, linkage_rate)
7 )
```


Save the file and source it into the Quarto file:

```
~~~~{r}
source("C:/Users/stoughte/Documents/linkrep/R/plot_data.R")
~~~~
```

Option 3: Source the data and a function to generate the element into the Quarto file

Open a new R file and create your data and a function to generate the new element:



The screenshot shows an R script editor window titled 'summary_table.R'. The script contains the following code:

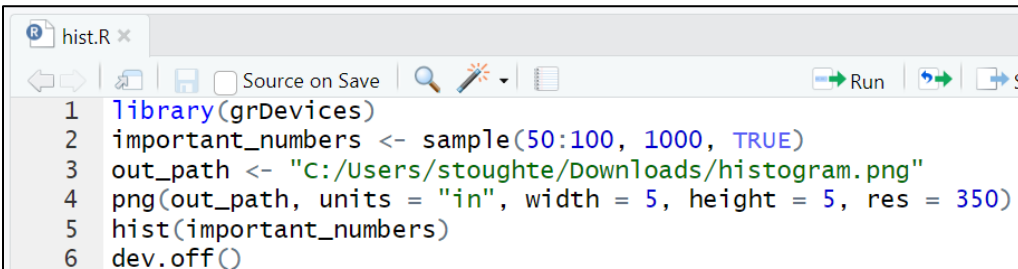
```
1 library(gtsummary)
2 table_data <- fread("C:/Users/stoughte/Downloads/summary_data.csv")
3
4 summary_table <- function(){
5   table <- tbl_summary(table_data)
6   table <- modify_header(table,
7                           update = all_stat_cols() ~ "***level***")
8   return(table)
9 }
```

Source the file into the Quarto document and call the function:

```
~~~~{r}
#| label: tbl-summary
#| tbl-cap: This is a summary table
source("C:/Users/stoughte/Documents/linkrep/R/summary_table.R")
summary_table()
~~~~
```

Option 4: Save the element as an image and import it into the Quarto file

Generate the element and save it as an image:



The screenshot shows an R script editor window titled 'hist.R'. The script contains the following code:

```
1 library(grDevices)
2 important_numbers <- sample(50:100, 1000, TRUE)
3 out_path <- "C:/Users/stoughte/Downloads/histogram.png"
4 png(out_path, units = "in", width = 5, height = 5, res = 350)
5 hist(important_numbers)
6 dev.off()
```

Import the image where you want it to be displayed in the report:

```
! [Important numbers.] (C:/Users/stoughte/Downloads/histogram.png) {#fig-imp_nums}
```

Save the modified Quarto file and pass its file path to the **linkage_quality_report** function through the **quarto_report_template** parameter.

Note: Once data has been imported into the Quarto file it can be used anywhere below therefore, if you wish to add multiple elements using the same data, you only need to import the data once. By ‘import’ I’m referring to any of the strategies used above.

For more information on including figures in Quarto see <https://quarto.org/docs/authoring/figures.html>.

For more information on including tables in Quarto see <https://quarto.org/docs/authoring/tables.html>

Remove elements:

Option 1: Utilize the function parameters

Linkage rates over time plot:

Leave the following parameters out of the **linkage_quality_report** function call:

- acquisition_year_var
- acquisition_month_var

Algorithm summary table:

Leave the following parameters out of the **linkage_quality_report** function call:

- algorithm_summary_data
- algorithm_summary_tbl_footnotes

Performance measures table and figure:

- Leave the following parameters out of the **linkage_quality_report** function call:
- performance_measures_data
- performance_measures_tbl_footnotes
- classification_metrics_used
- ground_truth
- ground_truth_missing_var

Missingness table:

Leave the following parameters out of the **linkage_quality_report** function call:

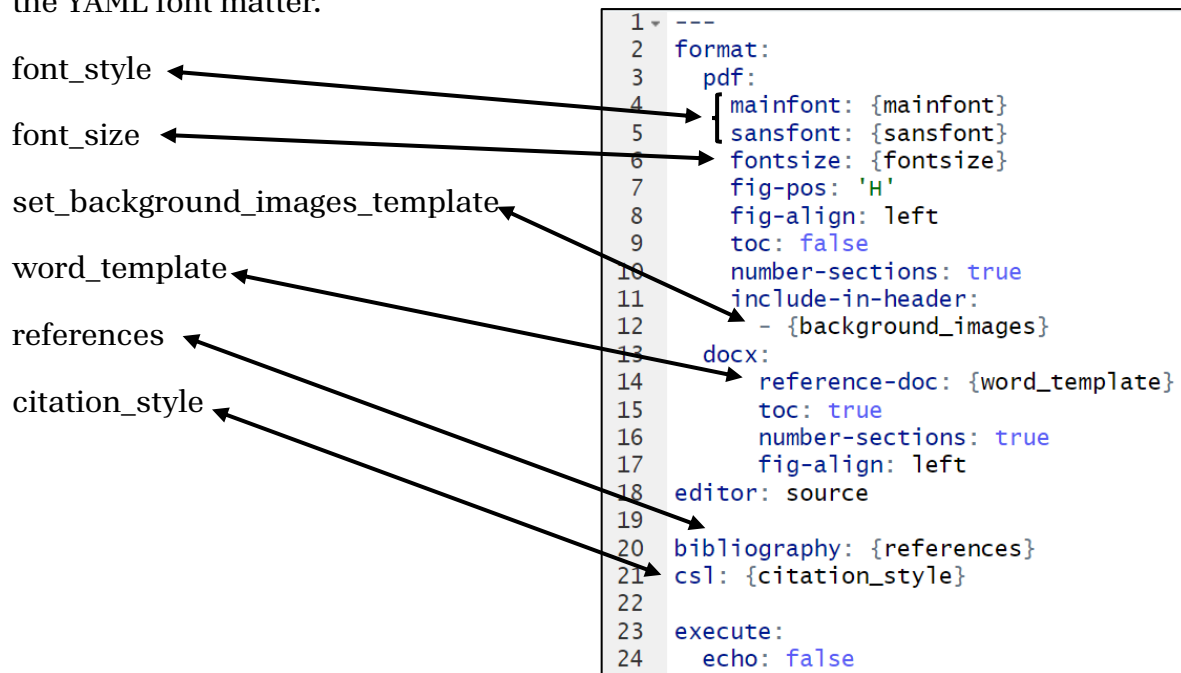
- missing_data_indicators
- missingness_tbl_footnotes

Option 2: Permanently remove the elements from the Quarto file

Suppose you wish to remove elements that cannot be removed using the parameter inputs, such as removing the performance measures table but keeping the figure. In that case, you will need to remove them from the Quarto template file. To do so, delete the R chunks associated with that element and any of its references throughout the report (e.g., @tbl-performance_measures_tbl). Save the modified Quarto file and pass its file path to the **linkage_quality_report** function through the **quarto_report_template** parameter.

3.1.3 Set metadata

There are a couple of parameters in the **linkage_quality_report** that can be set directly in the Quarto template file if you wish. They are located at the top of the Quarto file in the YAML front matter.



Example:

```
1 ---
2 format:
3   pdf:
4     mainfont: "Arial"
5     sansfont: "Arial"
6     fontsize: 10pt
7     fig-pos: 'H'
8     fig-align: left
9     toc: false
10    number-sections: true
11    include-in-header:
12      - "C:/Users/stoughte/Downloads/set_bg_images.tex"
13  docx:
14    reference-doc:
15      "C:/Users/stoughte/Downloads/word_template.docx"
16    toc: true
17    number-sections: true
18    fig-align: left
19  editor: source
20 bibliography: "C:/Users/stoughte/Downloads/references.bib"
21 csl: "C:/Users/stoughte/Downloads/apa.csl"
22
23 execute:
24   echo: false
```

If you choose to do this, you do not need to input those parameters when you call the function.

Save the modified Quarto file and pass its file path to the **linkage_quality_report** function through the **quarto_report_template** parameter.

3.1.4 Quarto tips: Understanding the report template

YAML font matter: contains metadata and configuration options for the report

```
1 ---
2 format:
3   pdf:
4     mainfont: {mainfont}
5     sansfont: {sansfont}
6     fontsize: {fontsize}
7     fig-pos: 'H'
8     fig-align: left
9     toc: false
10    number-sections: true
11    include-in-header:
12      - {background_images}
13  docx:
14    reference-doc: {word_template}
15    toc: true
16    number-sections: true
17    fig-align: left
18  editor: source
19
```

⋮

R code chunk: performs tasks in R

```
{r load packages, include=FALSE, message=FALSE}
library(flextable)
library(xfun)
```

LaTeX code chunk: performs tasks in LaTeX

```
{=tex}
\section*{List of Abbreviations}
\addcontentsline{toc}{section}{List of Abbreviations}
```

Conditional content: outputs the contents between the div (::: :::) if the condition evaluates to true.

Note: content within the div will be run even if the condition evaluates to false. Wrap conditional code in an if statement within the conditional output to ensure the program doesn't halt.

```
::: {.content-visible when-format="pdf"}
  {=tex}
  \section*{Acknowledgements}
  ...
:::
::: {.content-visible when-format="docx"}
# Acknowledgements {.unnumbered}
:::
```

Conditional content based on the output format. For more information see <https://quarto.org/docs/authoring/conditional.html>.

```
{r if(!is.null(linkage_rates_over_time_plot)) "::::
{.content-visible}" else ":::: {.content-hidden}"

{r linkage rates over time plot}
#| label: fig-linkage_rate_dist
#| fig-cap: !expr paste0("Distribution of linkage rates over
acquisition dates for records in ", params$left_dataset_name, ".")
if (!is.null(linkage_rates_over_time_plot)){
  linkage_rates_over_time_plot
}
}
::::
```

Conditional content based on R objects

Don't forget the ending
::::

Headings: unnumbered headings will be unnumbered on the report and in the table of contents; numbered headings will be numbered on the report and in the table of contents.

Unnumbered headings in table of contents:

```
%%{=tex}
\section*{List of Abbreviations}
\addcontentsline{toc}{section}{List of Abbreviations}
```

```
# List of Abbreviations {.unnumbered}
\addcontentsline{toc}{section}{List of Abbreviations}
```

Unnumbered headings not in table of contents:

```
%%{=tex}
\section*{Acknowledgements}
```

```
# Acknowledgements {.unnumbered}
```

Numbered heading in table of contents:

```
# Results
```

```
## Linkage Rate Summary
```

Table & Figure labels and captions: labels are needed to reference the tables and figures. For tables, labels must begin with **tbl-**; for figures, labels must begin with **fig-**. Captions can be made dynamic with the use of R objects.

Note: Must place **!expr** before caption using `paste0()` to incorporate dynamic elements. You do not need it if you are using simple text.

```
%%{r linkage_rate_table, ft.align="left"}
#| label: tbl-linkage_rates
#| tbl-cap: !expr paste0("Stratified linkage rates for records in ",
  params$left_dataset_name, " that linked to the records in ",
  params$right_dataset_name, " (N = ", params$num_records_left_dataset,
  ifelse(params$data_time_period == "", "", ", ", params$data_time_period, ").")

linkage_rate_table
```

Parameters to the Quarto file.

Images:

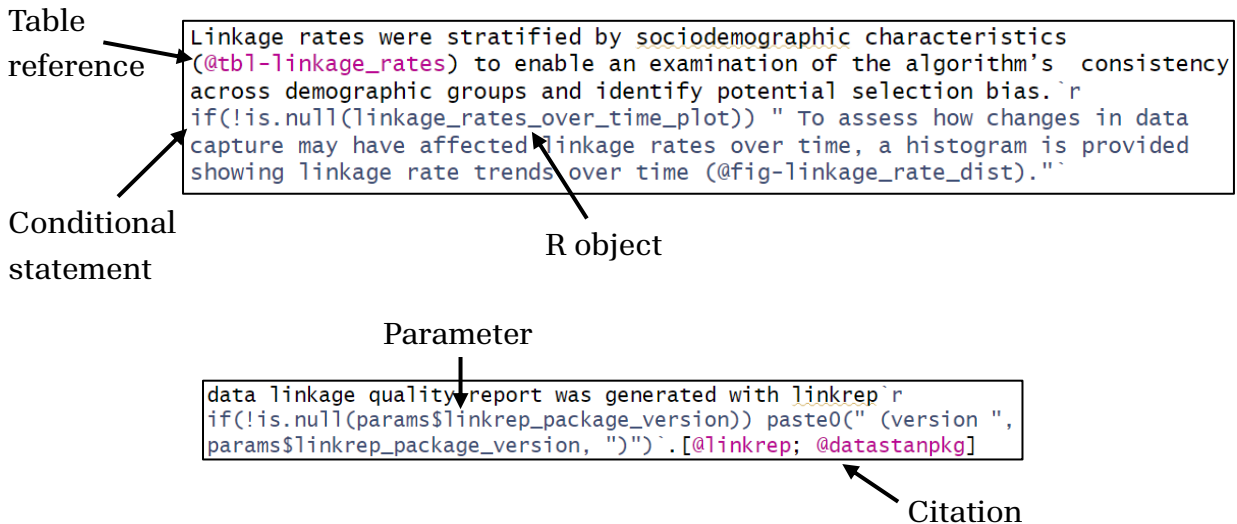
```
![Visualization of the performance metrics provided in @tbl-performance_measures_tbl.](r params$performance_measures_plot_path){#fig-performance_measures_chart}
```

Figure caption. Includes a table reference

File path to the image.
In this case the parameter contains the file path

Figure label

Written portions: written portions can incorporate R objects, parameters, table/figure references, citations and much more. Refer to [Technical Writing – Quarto](#) for more information.



3.2 References

To perform either of the following, you will need to modify the Quarto template and the references file. If you have not done so already, download the default Quarto template file by visiting

https://github.com/CHIMB/linkrep/blob/main/inst/templates/base_quarto_report_template.qmd and selecting “Download raw file.” Download the default references file by visiting <https://github.com/CHIMB/linkrep/blob/main/inst/templates/references.bib> and selecting “Download raw file.”

Open both files in RStudio to begin editing.

3.2.1 Modify reference list

Add references:

Step 1: Add new references to the default references file (.bib file)

Method 1: Manually enter new references using BibTeX syntax

- For more guidance on writing references in BibTeX see <https://web.mit.edu/rsi/www/pdfs/bibtex-format.pdf>

Method 2: Use Google Scholar's citation generator

Step 1: Go to Google Scholar (search “Google Scholar in Google”)

Step 2: Search for the paper you wish to cite and locate it in the displayed list:

A theory for record linkage

IP Fellegi, AB Sunter - Journal of the American Statistical ..., 1969 - Taylor & Francis

... -oriented **record linkage** operations have already been reported in the literature ([4], [5], [6], [7], [8], [11], [12], [13]) as well as at least two attempts to develop a **theory for record linkage** ([1...

☆ Save **Cite** Cited by 3745 Related articles All 10 versions

Step 3: Select ‘Cite’

Step 4: Select BibTeX

×

Cite

MLA

Fellegi, Ivan P., and Alan B. Sunter. "A theory for record linkage." *Journal of the American Statistical Association* 64.328 (1969): 1183-1210.

APA

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

Chicago

Fellegi, Ivan P., and Alan B. Sunter. "A theory for record linkage." *Journal of the American Statistical Association* 64, no. 328 (1969): 1183-1210.

Harvard

Fellegi, I.P. and Sunter, A.B., 1969. A theory for record linkage. *Journal of the American Statistical Association*, 64(328), pp.1183-1210.

Vancouver

Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association*. 1969 Dec 1;64(328):1183-210.

BibTeX

EndNote

RefMan

RefWorks

Step 5: Copy and paste the generated citation into the references.bib file

```
@article{fellegi1969theory,
  title={A theory for record linkage},
  author={Fellegi, Ivan P and Sunter, Alan B},
  journal={Journal of the American Statistical Association},
  volume={64},
  number={328},
  pages={1183--1210},
  year={1969},
  publisher={Taylor & Francis}
}
```

Step 2: Cite all new references in the Quarto template file using their citation keys

Reference:

Citation key

```
@article{fellegisunter,  
  title={A theory for record linkage},  
  author={Fellegi, Ivan P and Sunter, Alan B},  
  journal={Journal of the American Statistical Association},  
  volume={64},  
  number={328},  
  pages={1183--1210},  
  year={1969},  
  publisher={Taylor & Francis}  
}
```

```
citation @fellegisunter style 1  
citation [@fellegisunter] style 2  
multiple citations [@fellegisunter; @jarowinkler; @bohensky2011]
```

Citation:

Step 3: Save the modified files and pass their file paths to the **linkage_quality_report** function through the **quarto_report_template** and **references** parameters.

Remove references:

To remove references, you need to remove their citations in the Quarto template file.

Step 1: Locate the references you wish to remove in the references.bib file and identify their citation key.

Step 2: Locate all citations for those references in the Quarto file (e.g., `[@fellegisunter]`).

Step 3: Delete all citations in the Quarto file

Step 4 (optional): Delete the references in the references.bib file. The above three steps will remove the references from the References section of the report therefore, this step is not necessary.

Step 5: Save the modified files and pass their file paths to the **linkage_quality_report** function through the **quarto_report_template** and **references** parameters.

3.2.2 Modify citation style

To modify the citation style, you will need to pass a file path to a csl file containing the desired citation style to the **linkage_quality_report** function through the **citation_style** parameter.

To modify the citation style, you will need to utilize the **citation_style** parameter in the **linkage_quality_report**.

Step 1: Download your desired citation style.

- Download a file from [GitHub - citation-style-language/styles: Official repository for Citation Style Language \(CSL\) citation styles](#). or,
- Search “{style} citation in csl” (e.g., “apa citation in csl”)

Step 2 (optional): If your citation style requires a location change within the text, you will need to manually change their locations. To do so, open the Quarto template file and modify the locations of all the citations.

Step 3: Save the modified files and pass their file paths to the **linkage_quality_report** function through the **quarto_report_template** and **citation_style** parameters.

3.3 Word template

The Word template defines the font, font size, footers, headers, captions, page numbers, and much more for Word reports. To modify the default output styles, download the default Word template by visiting https://github.com/CHIMB/linkrep/blob/main/inst/templates/word_template.docx and selecting “Download raw file.”

Open the file in Word to begin editing.

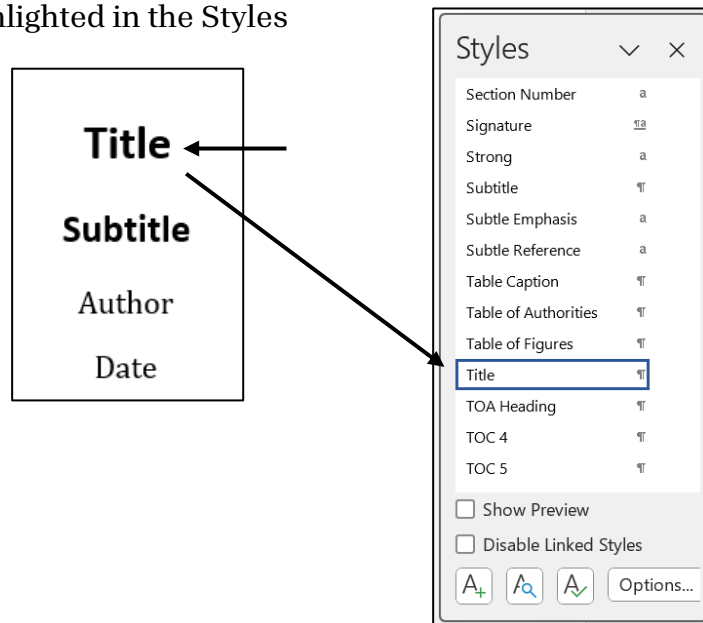
3.3.1 Modify font

To modify the font style, size, colour, etc., perform the following steps.

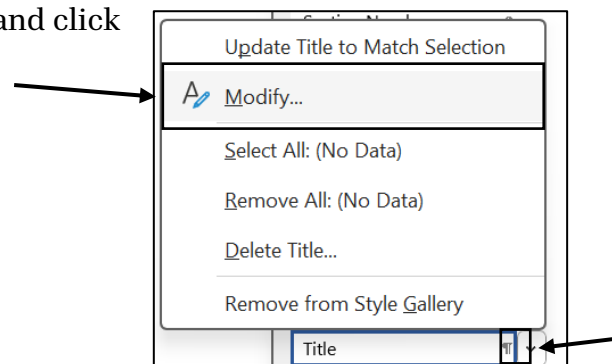
Step 1: Select the expander in the **Styles** pane on the **Home** tab.



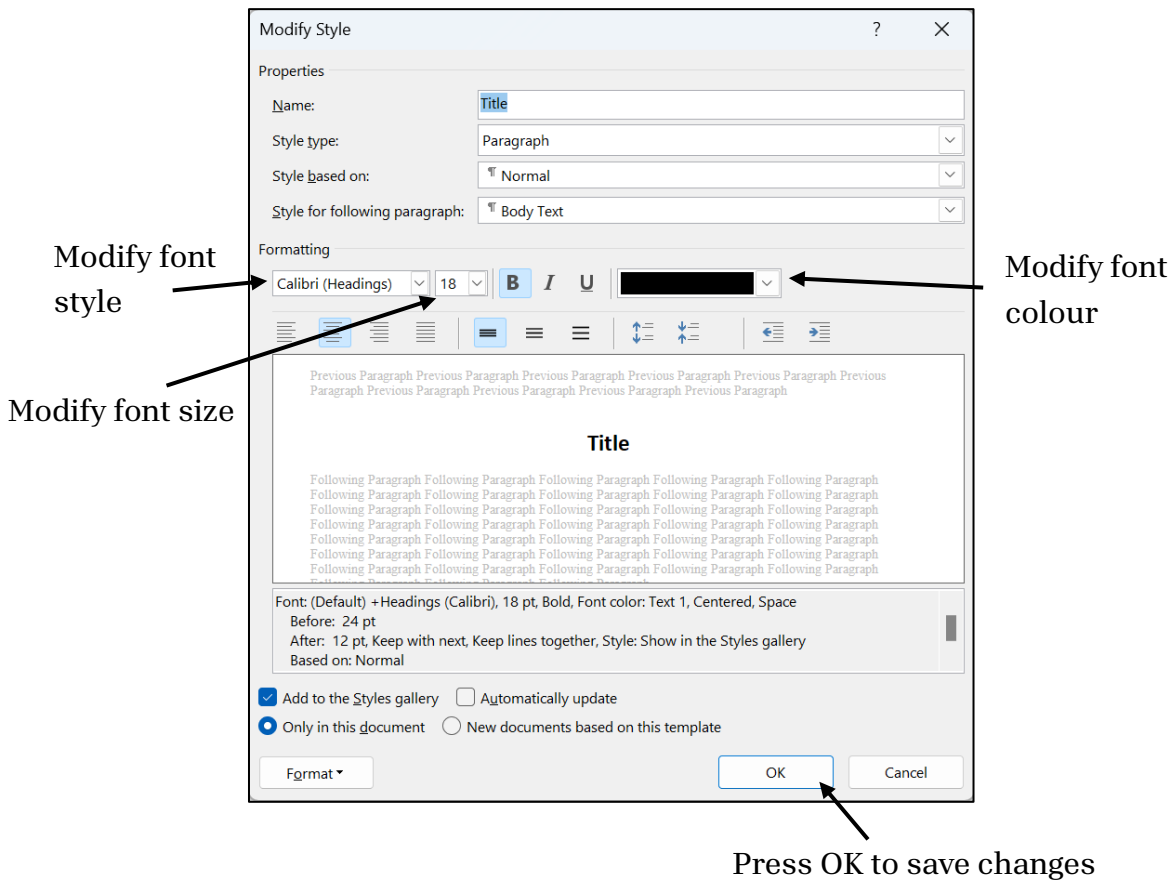
Step 2: Click on the text in the Word document you wish to modify, and its style will become highlighted in the Styles window.



Step 3: Select the arrow next to the style and click “Modify...”



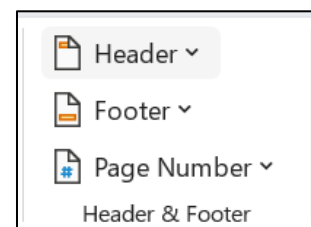
Step 4: Modify the font



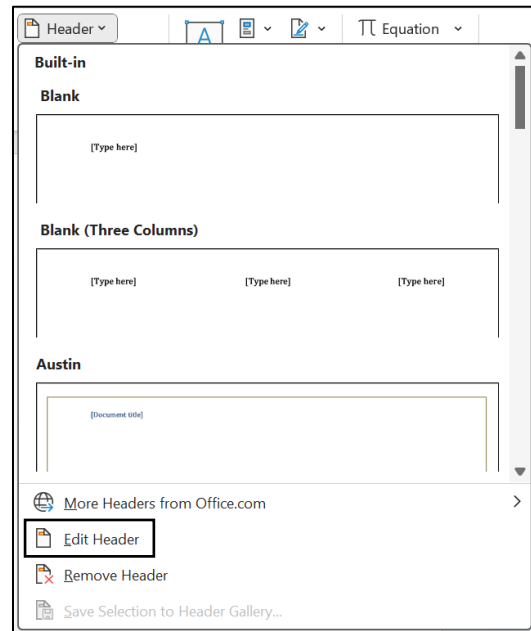
Step 5: Save the modified Word template and pass its file path to the **linkage_quality_report** function through the **word_template** parameter.

3.3.2 Modify footers & headers

Step 1: Navigate to the **Header & Footer** pane in the **Insert** tab and select the section you wish to modify



Step 2: Select either a Built-in option or the edit option. The edit option allows more flexibility, for example, you can add an image (e.g., logo) to the header through that option.



Step 3: Close the header & footer pane.

Step 4: Save the modified Word template and pass its file path to the **linkage_quality_report** function through the **word_template** parameter.

3.3.3 Additional resources

For more information, refer to <https://quarto.org/docs/output-formats/ms-word-templates.html> or utilize search engines (ex. Search: How to remove page numbers from word document) for more guidance.

3.4 Modify background images

To modify the background of a PDF report, use the image parameters provided in the **linkage_quality_report** function. Images should be letter size (8.5in x 11in) to fit the entire pages width.

Not all parameters need to be used, only the pages you wish to change. Pass the file path of your new background to the function.

Image Parameters:

- cover_page

- content_portrait_page
- content_landscape_page
- back_cover_page
- display_back_cover_page
- blank_background

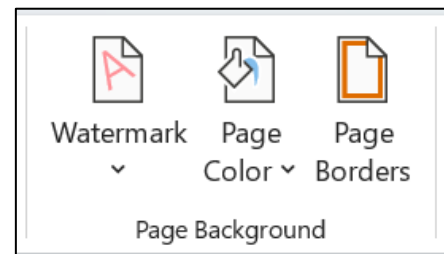
Word Report:

To modify the background of a Word report you can make use of the Word template, however, this will cause the full report to have the same background. If you wish to have more than one background you will need to change each page manually.

Change background in Word template:

Step 1: If you have not done so already, download the default Word template by visiting https://github.com/CHIMB/linkrep/blob/main/inst/templates/word_template.docx and selecting “Download raw file.”

Step 2: Navigate to the **Page Background** pane in the **Design** tab



Step 3: Select one of the options

- **Watermark:** If you wish to add a background image select Watermark, Custom Watermark, Picture Watermark, and select your image. If you don't want the image to washout, unselect the Washout box.
- **Page Color:** Select a colour for your background
- **Page Borders:** Select a border for your background

Step 4 (optional): Modify the headers and footnotes to contain logos or additional images. See the [Modify footers & headers](#) section for guidance.

Step 5: Save the modified Word template and pass its file path to the `linkage_quality_report` function through the `word_template` parameter.

3.5 Table customizations

Utilize the parameters to modify the table output:

E.g.,

```
table_font_size = 10,  
display_percent_symbol = TRUE,  
decimal_mark = ",",  
linkage_rate_tbl_footnotes = "Some other important information with a special character: \u00B1")
```

| | Linked (N = 489, 48.9%) | Unlinked (N = 511, 51.1%) |
|---------------------------------|----------------------------|------------------------------|
| Sex | | |
| Female | 265 (51.2) | 253 (48.8) |
| Male | 224 (46.5) | 258 (53.5) |
| Birth Year | | |
| <1970 | 113 (49.6) | 115 (50.4) |
| 1970-1979 | 103 (46.6) | 118 (53.4) |
| 1980-1989 | 92 (46.2) | 107 (53.8) |
| 1990-1999 | 126 (56.0) | 99 (44.0) |
| 2000+ | 55 (43.3) | 72 (56.7) |
| Data are presented as n (row %) | | |



| | Linked (N = 489, 48.9%) | Unlinked (N = 511, 51.1%) |
|--|----------------------------|------------------------------|
| Sex | | |
| Female | 265 (51,2%) | 253 (48,8%) |
| Male | 224 (46,5%) | 258 (53,5%) |
| Birth Year | | |
| <1970 | 113 (49,6%) | 115 (50,4%) |
| 1970-1979 | 103 (46,6%) | 118 (53,4%) |
| 1980-1989 | 92 (46,2%) | 107 (53,8%) |
| 1990-1999 | 126 (56,0%) | 99 (44,0%) |
| 2000+ | 55 (43,3%) | 72 (56,7%) |
| Some other important information with a special character: ± | | |
| Data are presented as n (row %) | | |