Check for updates

SOFTWARE TOOL ARTICLE

rdhs: an R package to interact with The Demographic and Health Surveys (DHS) Program datasets [version 1; peer review: 1 approved, 1 approved with reservations]

Oliver J. Watson, Rich FitzJohn, Jeffrey W. Eaton

MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, W2 1PG, UK

v1 **First published:** 27 Jun 2019, **4**:103
https://doi.org/10.12688/wellcomeopenres.15311.1
Latest published: 27 Jun 2019, **4**:103
https://doi.org/10.12688/wellcomeopenres.15311.1

Abstract

Since 1985, the Demographic and Health Surveys (DHS) Program has conducted more than 400 surveys in over 90 countries. These surveys provide decision markers with key measures of population demographics, health and nutrition, which allow informed policy evaluation to be made. Though standard health indicators are routinely published in survey final reports, much of the value of DHS is derived from the ability to download and analyse standardised microdata datasets for subgroup analysis, pooled multi-country analysis, and extended research studies. We have developed an open-source freely available R package 'rdhs' to facilitate management and processing of DHS survey data. The package provides a suite of tools to (1) access standard survey indicators through the DHS Program API, (2) identify all survey datasets that include a particular topic or indicator relevant to a particular analysis, (3) directly download survey datasets from the DHS website, (4) load datasets and data dictionaries into R, and (5) extract variables and pool harmonised datasets for multi-survey analysis. We detail the core functionality of 'rdhs' by demonstrating how the package can be used to firstly compare trends in the prevalence of anaemia among women between countries before conducting secondary analysis to assess for the relationship between education and anemia.

Keywords

R, survey analysis, API, Demographic and Health Surveys, DHS

Open Peer Review

Approval Status ✓ ?

	1	2
version 1 27 Jun 2019	✓ view	? view

1. **Dana R. Thomson**, University of Southampton, Southampton, UK

2. **Mahmoud Elkasabi**, The Demographic and Health Surveys Program, ICF, Rockville, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Oliver J. Watson (o.watson15@imperial.ac.uk)

Author roles: Watson OJ: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **FitzJohn R**: Methodology, Software, Supervision, Writing – Review & Editing; **Eaton JW**: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was been supported by the Wellcome Trust [109312/Z/15/Z]. We also acknowledge joint Centre funding from the UK Medical Research Council and Department for International Development [MR/R015600/1]
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Watson OJ, FitzJohn R and Eaton JW. **rdhs: an R package to interact with The Demographic and Health Surveys (DHS) Program datasets [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2019, **4**:103 https://doi.org/10.12688/wellcomeopenres.15311.1

First published: 27 Jun 2019, **4**:103 https://doi.org/10.12688/wellcomeopenres.15311.1

Introduction

The Demographic and Health Surveys (DHS) Program has collected and disseminated population survey data from over 90 countries for more than 30 years[1]. In many countries, DHS provide the key data that mark progress towards targets such as the Sustainable Development Goals (SDGs) and inform health policy such as detailing trends in child mortality[2] and characterising the distribution of malaria control interventions since 2000[3]. The DHS program publishes statistics for over 3000 indicators of population, nutrition, health and HIV status. Measures are presented at the national and subnational level and can be further disaggregated according to a series of demographic characterisations such as urbanicity, wealth and education level. However, there are additional data that are not disseminated in these statistics, which can be accessed by analysing the raw microdata. Analysis of the raw survey data has been used to show that HIV prevention strategies should focus on both sexes in discordant couples[4] and to investigate the impact of non-malarial fevers on the selective pressure for diagnostic resistant malaria[5].

The analysis of microdata datasets, however, often requires the creation of a data set that contains harmonised information across multiple surveys. One of the main challenges when interacting with the raw DHS datasets is isolating the required data set variables across different countries. Since the DHS Program started, there have been 7 “phases” of questionnaires used since 1985. The data from each phase is then recoded for consistency and comparability across surveys. However, new questions are often included or amended between different phases of the DHS program, which results in variable names sometimes changing between different phases. As well as this, there are a number of country specific records that are not part of model questionnaires. As such, it can become increasingly difficult to identify which variables to use within your final curated data set.

The `rdhs` package was designed to address these needs and facilitate the management and processing of DHS survey data in the R statistical software environment[6]. This occurs through both functioning as an application programming interface (API) client, allowing access to all data provided within the DHS API, and helping to download the standardised recoded microdatasets from the DHS website and read them into conventional R data structures. Lastly, the package caches data dictionaries associated with the survey datasets, enabling fast querying for survey variables of interest across multiple surveys.

Methods

Implementation

`rdhs` is designed to ease the process by which the user can identify and create a curated data set for their research and analysis purposes. To help guide the user to the datasets that contain the desired information, `rdhs` provides a wrapper to the DHS program application programming interface (API). Each of the twelve API endpoints is wrapped, which allows the user to query the API for survey metadata that can identify the relevant datasets. API results are by default cached for the user, which serves to both increase the speed with which repeated API requests can be returned and ensures previous requests can be accessed without an internet connection.

The identified datasets from the API are used by `rdhs` to specify which files are to be downloaded. Downloaded files are first read in and parsed to create a data dictionary for each survey. Both the dictionary and survey data are saved locally as R objects, which enables efficient querying of downloaded datasets for survey questions of interest. The responses to the identified survey questions are then used to build a curated data set that also includes geographical metadata of the survey locations.

A complete overview of the functionality within `rdhs` is described within the package’s manual or through the documented package website.

Operation

The package can be used with R version 3.3.0 or later on Linux, Mac and Windows. `rdhs` can be installed from within R via the command `install.packages("rdhs")`.

The main functionality of `rdhs` can be summarised as follows:

1. An Application Programming Interface (API) client for The DHS Program API.

2. Download, parse and cache micro datasets from The DHS Program to enable pooling of harmonised datasets for multi-survey analysis.

We will explore both these areas through an example use case in which we will compare estimates of the prevalence of any anemia among women from Demographic and Health Surveys conducted in Burundi, Ethiopia, Rwanda and Uganda since 2005.

Use case

API client

Anemia is a common cause of fatigue, and women of reproductive age are at particularly high risk for anemia. Anemia prevalence among women of reproductive age is reported as a core indicator through the DHS STATcompiler. These indicators can be accessed directly from R using the API client provided by `rdhs`, which provides a wrapper for each of the twelve endpoints of the DHS program API. Each of the twelve API endpoints can be accessed using their respective API functions, which all start with the prefix `"dhs_"`. For example, we can query the **countries** endpoint to see which countries are included in the DHS Program and their 2-letter DHS Country Code:

```
# query country names and 2-letter codes
> dhs_countries(
returnFields = c("CountryName", "DHS_CountryCode")
)

   DHS_CountryCode  CountryName
1:              AF  Afghanistan
2:              AL      Albania
3:              AO       Angola
. . .
```

The 2-letter country codes are used extensively within both the DHS API and the naming conventions of the survey files. However, the 2-letter DHS Country Codes are unique to the DHS Program. For example, the DHS country code for Burundi (BU) differs from the International Organization for Standardization (ISO) 2-letter code (BI). Once we are familiar with the country codes we need to identify the indicators related to anemia. The API includes statistics for thousands of survey indicators, which we can query directly using the **indicators** endpoint. Alternatively, the DHS program has created a system of tags that group indicators within topics of interest. We can query the **tags** endpoint to identify the DHS tag that corresponds to our topic of interest, which in this case would be related to anemia. Once identified, the identified tag can be used to query for the indicators relevant to our request.

```
# look up the DHS tags
> dhs_tags()
   TagType         TagName TagID TagOrder
1:       2  DHS Quickstats     0        0
2:       2      DHS Mobile    77        1
...
11:      0          Anemia    13      180
...

anemia_indicators <- dhs_indicators (tagIds = 13)
```

In the anemia tag identified, there are multiple indicators that consider the severity range of anemia for both children and adults. For our purposes, we want the percentage of women classified as having any anemia (<12.0 g/dl for non-pregnant women and <11.0 g/dl for pregnant women), which is given by the indicator ID “AN_ANEM_W_ANY”. Querying the **data** endpoint for the identified indicator ID will retrieve measures of the prevalence of any anemia in women in Burundi, Ethiopia, Rwanda and Uganda since 2005 (Figure 1).

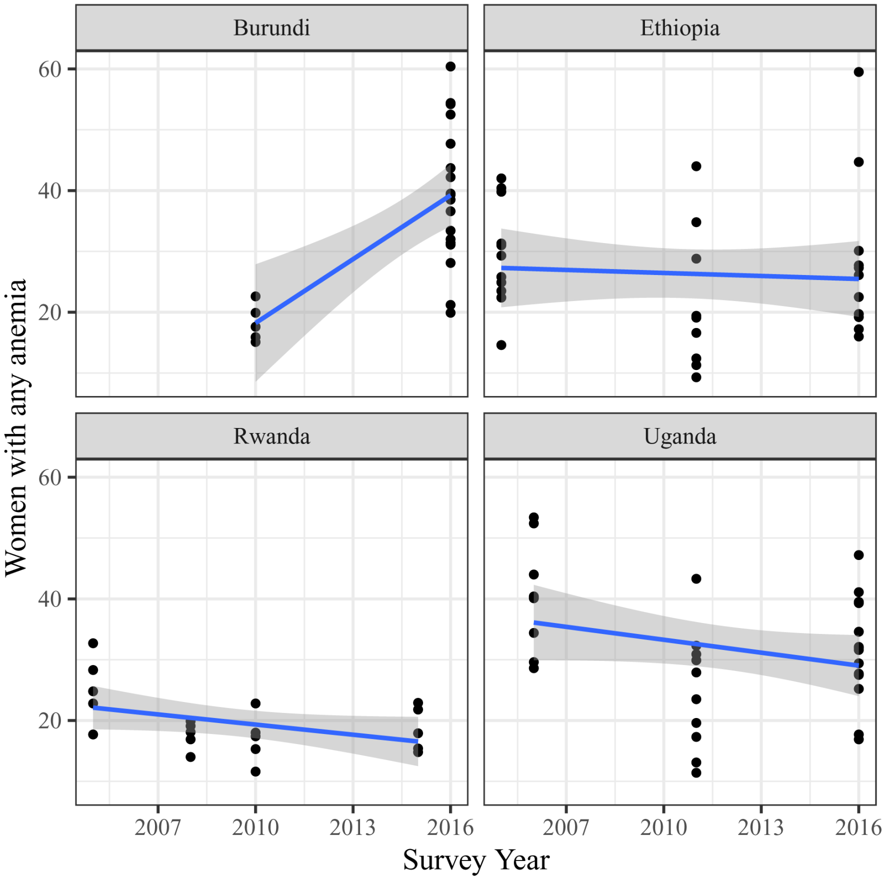

Figure 1. Trends in women with any level of anemia since 2005 for Burundi, Ethiopia, Rwanda and Uganda. Each point represents a subnational administrative unit and a linear regression is shown in blue with the 95% confidence interval.

```
# request data on any anemia in women
resp <- dhs_data(indicatorIds = "AN_ANEM_W_ANY",
                 countryIds = c("BU","ET","RW","UG"),
                 surveyYearStart = 2005,
                 breakdown = "subnational")

# and plot the results
library(ggplot2)
ggplot(resp, aes(x = SurveyYear, y = Value)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw() +
  ylab(unique(resp$Indicator)) +
  facet_wrap(CountryName, ncol = 2)
```

Lastly, the DHS API contains metadata about the characteristics, dates, and sample sizes for all DHS surveys conducted. This eases the identification of survey datasets that are relevant to anemia. We first query the **survey characteristics** endpoint to identify the survey characteristic ID that indicates that the survey includes questions related to anemia testing.

```
# look up survey characteristics
> surveychar <- dhs_survey_characteristics()

   SurveyCharacteristicID  SurveyCharacteristicName
1:                     16                  Abortion
2:                     33       Alcohol consumption
3:                     15          Anemia questions
4:                     41            Anemia testing
...
```

Next we query the **surveys** endpoint in the API to identify the surveys that have this characteristic ID and were conducted in our countries of interest. Finally, after identifying the relevant surveys, we can query the **datasets** endpoint to identify the individual recode (IR) survey datasets for each of these surveys. The DHS program provides datasets as SAS, Stata, SPSS, flat or Hierarchical formats. We recommend users to download the flat file format, which are both faster to load and have more complete data dictionaries available.

```
# find surveys including questions related to anemia testing
surveys <- dhs_surveys(surveyCharacteristicIds = 41,
                       countryIds = c("BU","ET","RW","UG"))

# return the individual recode datasets from these surveys
datasets <- dhs_datasets(surveyIds = surveys$SurveyId,
                         fileType = "IR",
                         fileFormat = "flat")
```

Downloading datasets

Downloading DHS survey datasets requires registration for a user login to the DHS website, provide a brief description of the project for which data are required, and request permission to access the required datasets. Instructions for this process are found on the DHS website. Once we have created an account and obtained permission to download datasets (typically within 1–2 days of making a request), our credentials must be provided to `rdhs` using the function `set_rdhs_config()`. This function requires the user to provide the email and project name used to create their account, before prompting the user to enter a password securely. The user is also asked whether they provide permission for `rdhs` to cache downloaded datasets outside of a temporary directory. If permission is granted, API queries and downloaded datasets will be saved on the computer hard drive for future use without requiring downloading again. If not granted, API queries and datasets will only be available in the present R session. After creating our DHS config, we download the identified datasets using `get_datasets()`.

```
## set up your credentials
set_rdhs_config(email = "rdhs.tester@gmail.com" ,
                project = "AnemiaInvestigations")

# download datasets
downloads <- get_datasets(datasets = datasets)
```

The default behaviour provided will download the requested datasets, read them in and save the resultant object as a *.rds* object within a cache directory. It also creates and caches the data dictionary within the cache directory. The cache directory is used when we interact with the DHS API; however, unless the user has provided permission for datasets and API calls to be cached outside of a session temporary directory these will be lost between R sessions. By default, the cache directory is established within the user's machine specific cache director. Alternatively, the user can specify a different directory for datasets and API calls to be cached within using `set_dhs_config()` or `update_dhs_config()`. `rdhs` will also check for any recent updates to the DHS API and will flush all potentially out of date API responses.

By caching both the dataset and the associated data dictionary we can quickly identify survey variables required for our analysis on anemia. The function `search_variable_labels()` facilitates regular expression (regex) searching of the cached data dictionaries to identify all questions in the surveys which contain data elements related to our analysis on hemoglobin levels in pregnant women. We can also search for variables related to the survey design as well as potentially interesting covariates such as education and urban/rural residence.

```
# search for relevant variable descriptions
questions <- search_variable_labels(
  dataset_filenames = datasets,
  search_terms = "hemoglobin|pregnant|education|residence"
)
```

Searching within the variable labels will likely identify the questions we are interested in as well as many other questions that are not immediately relevant for our analysis. Consequently, we can also search by variable names using the function `search_variables()`. In this case we could request variables for the cluster number (v001), sample weight (v005), region (v024), urban/rural residence (v025), education level (v106), pregnancy status (v454) and hemoglobin level after adjusting for altitude and smoking (v456).

```
# search for specific variables
questions <- search_variables(
  dataset_filenames = datasets,
  variables = c("v001", "v005", "v024", "v025",
                "v106", "v454", "v456")
)
```

The curated list of survey questions is used to extract the raw data with the function `extract_dhs()`.

```
# search for specific variables
extract <- extract_dhs(questions)
```

The resultant extract is a list, with an item for each dataset that we have extracted. In addition the survey ID is added to the dataset by default, which allows us to identify the survey source when we combine our datasets for further analysis. The function `rbind_labelled()` was developed for this purpose, with the additional argument `labels`, which describes how to combine variable levels for all datasets. For example, the following ensures that we keep all labels for the region variable (v024) while providing a consistent set of value labels to be used for v454 (currently pregnant) across all datasets.

```
# combine the list of extracted data
dat <- rbind_labelled(
  extract,
  labels = list(v024 = "concatenate",
                  v454 = c("no/don't know" = 0L,
                           "yes" = 1L,
                           "missing" = 9L))
)
```

We now have a pooled labelled dataset for 154,439 individuals, which we can use to test for relationships between potential risk factors and the presence of anemia. To prepare for analysis of the dataset, we first need to recode a new binary variable indicating the presence of anemia, defined as hemoglobin level <12.0 g/dl for non-pregnant women and <11.0 g/dl for pregnant women. We also re-scale the sample weight variable by dividing by 1,000,000 as the weight is an eight-digit variable with six implied decimal places.

```
dat$anemia <- as.integer(
  dat$v456  < ifelse(dat$v454 == "yes", 110, 120)
)
dat$weight <- dat$v005 / 1e6
```

Finally, we use our data set to investigate the relationship between anemia prevalence and education level (v106) using logistic regression, adjusting for urban/rural residence (v025) and including fixed effects for each survey. To account for the survey design we will use the R package `survey`[7].

```
library (survey)

# create our survey design
des <- svydesign (
  ids = v001 + SurveyId, data = dat, weights = weights
)
# logistic regression
> summary(svyglm(anemia ~ SurveyId + v025 + v106,
                 design = des, family = "binomial"))
...
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.880623   0.090569 -20.765  < 2e-16 ***
SurveyIdBU2016DHS  1.068473   0.065034  16.429  < 2e-16 ***
SurveyIdET2005DHS  0.265482   0.074498   3.564 0.000391 ***
SurveyIdET2011DHS -0.147038   0.071638  -2.053 0.040492 *
SurveyIdET2016DHS  0.274877   0.071375   3.851 0.000128 ***
SurveyIdRW2005DHS  0.427045   0.070342   6.071 2.09e-09 ***
SurveyIdRW2008DHS  0.034822   0.075505   0.461 0.644806
SurveyIdRW2010DHS  0.007708   0.069819   0.110 0.912129
SurveyIdRW2015DHS  0.188178   0.069159   2.721 0.006672 **
SurveyIdUG2006DHS  1.238592   0.076239  16.246  < 2e-16 ***
SurveyIdUG2011DHS  0.418362   0.088090   4.749 2.48e-06 ***
SurveyIdUG2016DHS  0.903546   0.067845  13.318  < 2e-16 ***
v025               0.339089   0.037586   9.022  < 2e-16 ***
v106              -0.228964   0.015720 -14.565  < 2e-16 ***
---
Significance: 0 "***" 0.001 "**" 0.01 "*" 0.05
```

The results suggest that anemia prevalence is lower among women with higher education (v106) after adjusting for differences between rural and urban residences (v025). We can double check that this is the correct way to interpret the regression coefficients associated with these variables by looking at the labels encoded in these variables. For example, we can confirm that individuals in rural residences have an increased risk of anemia and that the education variable increases with increased levels of education:

```
> attr (dat$v025, "labels")
urban rural
    1     2

> attr (dat$v106, "labels")
no education   primary  secondary higher
           0         1           2      3
```

Discussion

Between 1987 and March 2019 the DHS Program conducted and published data from 315 surveys, which represents over 12,000 dataset files that can be freely downloaded for further analysis. The published datasets are available in formats that can be analysed using proprietary software including Stata, SAS and SPSS. In this paper we introduced the `rdhs` package, a freely available software package for the R software environment that extends the available toolkit for analysing DHS datasets.

`rdhs` enables the DHS API to be accessed directly from within R using a series of functions that wrap each of the twelve API endpoints. API requests made in this way are cached and will be automatically updated in response to any future updates in the API. These functions assist users in both searching for published health and demographic statistic and identifying datasets needed for further analysis. Once identified `rdhs` enables users to easily download datasets and converts them into standard R data structures and caches the associated data dictionaries, enabling quick searching across multiple datasets for survey questions of interest. We demonstrated in this paper how a user can create a curated data set containing harmonised responses for 154,439 individuals in fewer than ten lines of R code.

The DHS program periodically updates published datasets. However, each update causes a change in the file name that will also be updated in the API. This feature ensures that updated datasets will both be detected by `rdhs` via the API and consequently the newer dataset files will be downloaded. For example, in our use case the downloaded dataset from the survey conducted in Uganda 2016 has the file name "UGIR7AFL.ZIP". The DHS file naming conventions indicate that this file is the second released version. As such, the code used in this example will ensure that the second released version of the survey will be downloaded, which helps to ensure reproducible research that can be easily shared between individuals. In this way, it is hoped that `rdhs` will enable more researchers to interact with the wealth of data published by the DHS Program through providing a tool that enables information rich datasets to be easily accessed and reproducibly produced.

Software availability

The software (v0.6.3) is available at https://CRAN.R-project.org/package=rdhs

Development version (v0.6.3) available from: https://github.com/ropensci/rdhs

Archived source code (v0.6.3) as at time of publication: http://doi.org/10.5281/zenodo.2598510[8]

Licence: MIT

Grant information

This work was been supported by the Wellcome Trust [109312/Z/15/Z]. We also acknowledge joint Centre funding from the UK Medical Research Council and Department for International Development [MR/R015600/1].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to thank rOpenSci and Anna Krystalli for helping to review `rdhs`, with specific thanks to Lucy McGowan and Duncan Gillespie for their individual reviews of `rdhs`. We would also like to thank Trevor Croft and Phil Satlof for their help interacting with the DHS API.

References

1. ICF: **The DHS Program. Funded by USAID.**
Reference Source

2. Silva R: **Child mortality estimation: consistency of under-five mortality rate estimates using full birth histories and summary birth histories.** *PLoS Med.* 2012; **9**(8): e1001296.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Bhatt S, Weiss DJ, Cameron E, *et al.*: **The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015.** *Nature.* 2015; **526**(7572): 207–11.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Eyawo O, de Walque D, Ford N, *et al.*: **HIV status in discordant couples in sub-Saharan Africa: a systematic review and meta-analysis.** *Lancet Infect Dis.* 2010; **10**(11): 770–777.
PubMed Abstract | Publisher Full Text

5. Watson OJ, Slater HC, Verity R, *et al.*: **Modelling the drivers of the spread of *Plasmodium falciparum hrp2* gene deletions in sub-Saharan Africa.** *eLife.* 2017; **6**: pii: e25008.
PubMed Abstract | Publisher Full Text | Free Full Text

6. R Core Team: **R: A Language and Environment for Statistical Computing.** 2019.
Reference Source

7. Lumley T: **Analysis of complex survey samples.** *J Stat Softw.* 2004; **9**(1): 1–19.
Publisher Full Text

8. Watson OJ, Eaton J, FitzJohn R, *et al.*: **ropensci/rdhs v0.6.3 (Version v0.6.3).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2598510

Open Peer Review

Current Peer Review Status: ✓ ?

Version 1

Reviewer Report 30 July 2019

https://doi.org/10.21956/wellcomeopenres.16712.r36009

?

Mahmoud Elkasabi
The Demographic and Health Surveys Program, ICF, Rockville, MD, USA

This is a well-written short paper describing the rdhs R package for facilitating management and processing of DHS survey data. The package should be a very useful tool for R users who are interested in searching for published indicators or analysing the raw microdata of DHS surveys.

Overall the paper looks good. I just have found minor editorial errors in the text or in the R code. These errors need to be corrected. I also have few suggestions. Some of the suggestions are strongly recommended. Finally I found some problems with the R codes that need to be addressed.

Minor errors or inconsistencies

1. Information about number of surveys mentioned in the abstract (400 surveys conducted since 1985) and in the discussion section (data from 315 surveys published between 1987 and March 2019) are not consistent. Please check and correct if needed.

2. Some R codes start with “>” and some do not. Be consistent.

3. In page 5, the last line in the R code is missing “~”; it should read as below
facet_wrap(~ CountryName, ncol = 2)

4. In page 6, the last line in the R code has a wrong argument label; it should read as below
downloads <- get_datasets(dataset_filenames = datasets)

5. In the svydesign function, weights is declared as “weights”. it should be “weight” as you defined it in the weight line below
dat$weight <-- dat$v005 / 1e6

Suggestions/extensions

6. In the introduction section, there is a discussion about the harmonized information across multiple surveys. I strongly believe this topic cannot be discussed without mentioning the IPUMS-

DHS project.

7. Under the methods section, the paper could definitely benefit from the following extensions:
   - Before the “Implementation” subsection, readers might benefit from a paragraph about API and what does it mean and how it is used. Also, terms like “wrapper” and “endpoints” need to be defined.

   - I would recommend to add a flowchart to illustrate how the package work with the DHS API.

   - It might be helpful to list the most important functions under a separate subsection where reader can benefit from such introduction before the Use case section and without a need to check the package’s manual.

8. I found the example illustrated in Figure 1 very interesting and highlights the potential contributions of the package. In the illustration, trends for only one indicator (Anemia) were presented. I would suggest to add the trends of another indicator on the same graph; for example trends for U5MR and NNMR can be graphed to should how these two related indicators change over time. Perhaps you can think about another indicator related to anemia.

<u>Problems in the R code</u>

9. For some reason, the rbind_labelled function gives me the error below. Therefore, I was not able to check the next codes. I strongly believe this has to be checked/fixed.

```
> dat <- rbind_labelled(
+   extract,
+   labels = list(v024 = "concatenate",
+               v454 = c("no/don’t know" = 0L,
+                     "yes" = 1L,
+                     "missing" = 9L))
+ )
Error in `[.data.frame`(X[[i]], ...) : undefined columns selected
```

10. In the svydesign function, strata is not defined. Although v023 can be used as strata in most of the recent surveys, in older surveys this might not be the case. This might be tricky, since you need to check and identify stratification variable for each survey before you declare a unified variable for strata. I believe this issue should be highlighted and discussed in the paper.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets

and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Survey statistics, survey sampling, demography.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 22 July 2019

https://doi.org/10.21956/wellcomeopenres.16712.r35942

✓

Dana R. Thomson

Department of Social Statistics, University of Southampton, Southampton, UK

The authors describe a new package that leverages the DHS Program API to access and combine multiple DHS micro datasets in R. This R package and article stand to vastly improve the ease and reproducibility of DHS data analyses.

Clarification needed:

- In the first paragraph of the Methods: Implementation section, the phrase "the twelve API endpoints" is first mentioned. Please briefly explain what an API endpoint is, and what the 12 DHS Program API endpoints are.

Other suggestions:

- Is "any anemia" the correct wording? Should it be "any severity of anemia"?
- Consider "outdated" instead of "out of date".
- Both "dataset" and "data set" are used throughout. Some authors seem to use "dataset" to refer to a specific set of data, and "data set" as a general term - if this was the intent here, double check that your use of the terms are correct. Otherwise, choose one spelling throughout.
- Check that "DHS Program" and not "DHS program" is used throughout.

Grammar:

- Throughout, consider dropping or changing duplicate words used in the same sentence. E.g. "Once identified, the identified tag..."

- Several sentences throughout need additional commas for ease of understanding. E.g.:
  "... however, unless the user has provided permission for datasets and API calls to be cached outside of a session temporary directory[comma] these will be lost between R sessions."
  "By caching both the dataset and the associated data dictionary[comma] we can quickly identify survey variables required for our analysis on anemia."

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Household survey tool development, survey data analysis, public health.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.