{cleanepi}

0.0.2

Stage of development:

Experimental

Authors:

Karim Mané
Bankolé Ahadzie
Bubacarr Bah
Abdoelnaser Degoot
Nuredin Mohammed

Clean, curate, and standardize epidemiological data

MRC Unit The Gambia

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

https://epiverse-trace.github.io/cleanepi

epiverse-trace/cleanepi

functionalities

- scan_data()
- standardize_date()
- check_subject_ids()
- calculate_age()
- check_date_sequence()
- find_duplicates()
- remove_duplicates()
- convert_to_numeric()
- clean_using_dictionary()
- clean_data()

scan_data(data = test_data)

study_id	event	code	country	date admitted	DOB	pcr_date	sex
PS001P2	day 0	2	Gambia	01/12/2020	06/01/1972	Dec 01, 2020	1
PS002P2	day 0	2	Gambia	28/01/2021	02/20/1952	Jan 01, 2021	1
PS004P2-1	day 0	2	Gambia	15/02/2021	06/15/1961	Feb 11, 2021	-99
PS003P2	day 0	2	Gambia	11/02/2021	11/11/1947	Feb 01, 2021	1
P0005P2	day 0	2	Gambia	17/02/2021	09/26/2000	Feb 16, 2021	2
PS006P2	day 0	2	Gambia	17/02/2021	-99	May 02, 2021	2

Determine data types

- numeric, character, and date
- % of missings in each column

data_type	study_id	event	code	country	date admitted	DOB	pcr_date	sex
missing	0	0	0	0	0	0.0	0	0
numeric	0	0	1	0	0	0.1	0	1
date	0	0	0	0	1	0.9	1	0
character	1	1	0	1	0	0.0	0	0
logical	0	0	0	0	0	0.0	0	0

standardize_date()

- Convert date columns into %Y-%m-%d format
- Convert character columns into date

date admitted	DOB	pcr date
01/12/2020	06/01/1972	"Dec 01, 2020"
28/01/2021	02/20/1952	"Jan 01, 2021"
15/02/2021	06/15/1961	"Feb 11, 2021"
11/02/2021	11/11/1947	"Feb 01, 2021"
17/02/2021	09/26/2000	"Feb 16, 2021"
17/02/2021	-99	"May 02, 2021"

date_admitted	DOB	pcr_date
2020-12-01	01-06-1972	2020-12-01
2021-01-28	1952-02-20	2021-01-01
2021-02-15	1961-06-15	2021-02-11
2021-02-11	1947-11-11	2021-02-01
2021-02-17	2000-09-26	2021-02-16
2021-02-17	NA	2021-05-02

check_subject_ids()

Detect and remove incorrect subject IDs

study_id
PS001P2
PS002P2
PS004P2-1
PS003P2
P0005P2
PB500P2

study_id
PS001P2
PS002P2
PS004P2-1
PS003P2
P0005P2
PB500P2

calculate_age()

- In years / months / weeks / days

study_id	DOB
PS001P2	1-Jun-1972
PS002P2	20-Feb-1952
PS004P2	06/15/1961
PS003P2	11/11/1947
P0005P2	26/09/2000
PS006P2	-99

study_id	DOB	age_months
PS001P2	1972-06-01	615
PS002P2	1952-02-20	859
PS004P2	1961-06-15	747
PS003P2	1947-11-11	910
P0005P2	2000-09-26	276
PS006P2	NA	NA

check_date_sequence()

-e.g., DOB <= date_admitted

date admitted	DOB
06/01/1972	01/12/2020
28/01/2021	02/20/1952
15/02/2021	06/15/1961
11/02/2021	11/11/1947
09/26/2000	17/02/2021
17/02/2021	13/03/2022

DOB	date_admitted
01/12/2020	06/01/1972
02/20/1952	28/01/2021
06/15/1961	15/02/2021
11/11/1947	11/02/2021
17/02/2021	09/26/2000
13/03/2022	17/02/2021

find_duplicates()

group_id	dt_onset	dt_report	sex	outcome
1	2015-05-21	2015-06-03	M	Alive
3	2015-05-31	2015-06-02	M	Dead
2	2015-05-30	2015-06-06	M	Alive
1	2015-05-21	2015-06-03	M	Alive
2	2015-05-30	2015-06-06	M	Alive

remove_duplicates()

group_id	dt_onset	dt_report	sex	outcome
1	2015-05-21	2015-06-03	M	Alive
3	2015-05-31	2015-06-02	M	Dead
2	2015-05-30	2015-06-06	M	Alive

convert_to_numeric()

x <- c(2, "ten", NA, 2, "twenty", 6)

x <- c(2, 10, NA, 2, 20, 6)

clean_using_dictionary()

The data dictionary

options	values	grp	orders
1	male	sex	1
2	female	sex	2

Add_to_dictionary()

options	values	grp	orders
1	male	sex	1
2	female	sex	2
-99	unknown	sex	3

sex
1
1
-99
1
2
2

sex
male
female
unknown
male
female
female

clean_data()

{

- scan_data()
- standardize_date()
- check_subject_ids()
- calculate_age()
- check_date_sequence()
- find_duplicates()
- remove_duplicates()
- convert_to_numeric()

}

Contributors




Reviewers: Hugo, Pratik, Andree

Our colleagues in statistics and bioinformatics, data management, and app development teams at MRCG.