Package 'HDSinRdata'

May 7, 2026

Type Package

Title Data for the 'Mastering Health Data Science Using R' Online Textbook

Version 0.3.0

Maintainer Alice Paul <alice_paul@brown.edu>

Description
Contains ten datasets used in the chapters and exercises of Paul, Alice (2023) ``Health Data Science in R" <https://alicepaul.github.io/health-data-science-using-r/>.

License CC BY 4.0

Encoding UTF-8

LazyData true

LazyDataCompression xz

Depends R (>= 2.10)

RoxygenNote 7.3.2

NeedsCompilation no

Author Alice Paul [aut, cre],
Hannah Eglinton [aut],
Jialin Liu [ctb],
Joanna Walsh [aut],
Xinbei Yu [ctb]

Repository CRAN

Date/Publication 2025-12-17 17:00:02 UTC

Contents

breastcancer 2
covidcases 3
lockdowndates 4
mobility 5
NHANESsample 5
nyts 7

pain 11
tb_diagnosis 14
tb_diagnosis_raw 15
tex_itop 17

Index **20**

`breastcancer` *Data from the Original Wisconsin Diagnostic Breast Cancer Database*

Description

32 features of cell nuclei present in digitized images of fine needle aspirates of 212 malignant and 357 benign breast masses.

Usage

```
breastcancer
```

Format

A data frame with 569 rows and 32 variables. The first two variables are id and diagnosis, and then the mean, standard error, and "worst" or largest (mean of the three largest values) for each of ten features are reported as follows:

id ID number

diagnosis Diagnosis (M = malignant, B = benign)

radius_mean Mean of mean distances from center to points on the perimeter

texture_mean Mean of standard deviation of gray-scale values

perimeter_mean Mean of perimeter

area_mean Mean of area

smoothness_mean Mean of local variation in radius lengths

compactness_mean Mean of perimeter^2 / area - 1.0

concavity_mean Mean of severity of concave portions of the contour

concave_points_mean Mean of number of concave portions of the contour

symmetry_mean Mean of symmetry

fractal_dimension_mean Mean of "coastline approximation" - 1

radius_se Standard error of mean distances from center to points on the perimeter

texture_se Standard error of standard deviation of gray-scale values

perimeter_se Standard error of perimeter

area_se Standard error of area

smoothness_se Standard error of local variation in radius lengths

compactness_se Standard error of perimeter^2 / area - 1.0

concavity_se Standard error of severity of concave portions of the contour

concave_points_se Standard error of number of concave portions of the contour

symmetry_se Standard error of symmetry

fractal_dimension_se Standard error of "coastline approximation" - 1

radius_worst "Worst" or largest (mean of the three largest values) of mean distances from center to points on the perimeter

texture_worst "Worst" or largest (mean of the three largest values) of standard deviation of gray-scale values

perimeter_worst "Worst" or largest (mean of the three largest values) of perimeter

area_worst "Worst" or largest (mean of the three largest values) of area

smoothness_worst "Worst" or largest (mean of the three largest values) of local variation in radius lengths

compactness_worst "Worst" or largest (mean of the three largest values) of perimeter^2 / area - 1.0

concavity_worst "Worst" or largest (mean of the three largest values) of severity of concave portions of the contour

concave_points_worst "Worst" or largest (mean of the three largest values) of number of concave portions of the contour

symmetry_worst "Worst" or largest (mean of the three largest values) of symmetry

fractal_dimension_worst "Worst" or largest (mean of the three largest values) of "coastline approximation" - 1

All feature values are recoded with four significant digits.

Source

Wolberg,William. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z.

Obtained from the UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

---

covidcases *US Covid Data from the Covid-19 Data Hub*

Description

Weekly confirmed Covid-19 cases and deaths at the state and county level in 2020, downloaded from the COVID19 R package.

Usage

```
covidcases
```

Format

A data frame with 69,530 rows and 5 variables.

state State (administrative_area_level_2 from Covid-19 Data Hub)

county County (administrative_area_level_3 from Covid-19 Data Hub)

week Week of 2020

weekly_cases Weekly Covid-19 cases calculated from the Covid-19 Data Hub's cumulative counts of confirmed cases. Note that, according to the Data Hub, "some of these values are negative due to decreasing cumulative counts in the original data provider".

weekly_deaths Weekly Covid-19 deaths calculated from the Covid-19 Data Hub's cumulative counts of confirmed deaths. Again, note that "some of these values are negative due to decreasing cumulative counts in the original data provider".

Source

Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open Source Software 5(51):2376, doi:10.21105/joss.02376"

https://CRAN.R-project.org/package=COVID19

https://covid19datahub.io/index.html

`lockdowndates` *Lockdown dates from Ballotpedia*

Description

Start and end dates of statewide stay at home orders in response to the Covid-19 pandemic.

Usage

```
lockdowndates
```

Format

A data frame with 50 rows and 3 variables:

State State

Lockdown_Start Start date of the statewide order in YYYY-MM-DD format

Lockdown_End End date of the statewide order in YYYY-MM-DD format

Source

Raifman, J., Nocka, K., Jones, D., Bor, J., Lipson, S., Jay, J., Cole, M., Krawczyk, N., Benfer, E. A., Chan, P., Galea, S. (2022). COVID-19 US State Policy Database. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2022-03-30. https://doi.org/10.3886/E119446V143

https://www.openicpsr.org/openicpsr/project/119446/version/V143/view

mobility *Data for Mobility Changes in Response to COVID-19*

Description

2020 mobility statistics at the state level from Descartes Labs.

Usage

```
mobility
```

Format

A data frame with 9,333 rows and 5 variables:

state State (originally admin1)

date Date in YYYY-MM-DD format

samples The number of samples observed in the state on that date (summed across counties)

m50 The median of the max-distance mobility (representing the distance a typical member of a given population moves in a day) for all samples in a county, averaged across counties.

m50_index The percent of normal m50 in the region, with normal m50 defined during 2020-02-17 to 2020-03-07, averaged across counties.

Note from the data website: "Data for 2020-04-20, 2020-05-29, 2020-10-08, 2020-12-11 through 2020-12-18, 2021-01-08 through 2021-01-14, 2021-04-07, 2021-04-12 and 2021-04-21 to present did not meet quality control standards, and was not released."

Source

Data was obtained from Descartes Labs

Warren, Michael S. & Skillman, Samuel W. "Mobility Changes in Response to COVID-19". arXiv:2003.14228 [cs.SI], Mar. 2020. arxiv.org/abs/2003.14228

https://github.com/dlarchives/DL-COVID-19

NHANESsample *A sample of data from the National Health and Nutrition Examination Survey (NHANES)*

Description

Lead, blood pressure, and demographic variables from NHANES 1999-2018, downloaded from the nhanesA package. Data was filtered to adults 20 years of age or older with nonmissing blood lead level, blood pressure, and demographic information.

Usage

```
NHANESsample
```

Format

A data frame with 31,265 rows and 15 variables:

ID Respondent sequence number ("SEQN" in NHANES)

AGE Age ("RIDAGEYR" in NHANES: Best age in years of the sample person at time of HH screening. Individuals 85 and over are topcoded at 85 years of age up to 2006 and individuals 80 and over are topcoded at 80 years of age after 2006.)

SEX Gender ("RIAGENDR" in NHANES)

RACE Race and ethnicity ("RIDRETH1" in NHANES)

EDUCATION Education Level ("DMDEDUC2" in NHANES: What is the highest grade or level of school you have completed or the highest degree you have received?)

INCOME Poverty income ratio (PIR): a ratio of family income to poverty threshold ("INDFMPIR" in NHANES)

SMOKE Smoking status (Combination of SMQ020 (Have you smoked at least 100 cigarettes in your entire life?) and SMQ040 (Do you now smoke cigarettes?) in NHANES: equal to "Still Smoke" if respondent answered "Yes" to SMQ020 and either "Every day" or "Some days" to SMQ040, equal to "Quit Smoke" if respondent answered "Yes" to SMQ020 and "Not at all" to SMQ040, and equal to "Never Smoke" otherwise.)

YEAR Year of the Study (Equal to the first year of the two year interval in which the response was recorded - NHANES surveys are grouped in two-year intervals)

LEAD Lead (ug/dL): "LBXBPB" in NHANES unless the reported level of lead was less than the lower limit of detection (llod), as defined by the paper cited above, for the relevant year, in which case "LBXBPB" was replaced by llod/sqrt(2))

BMI_CAT Body Mass Index Category (kg/m^2): Based on "BMXBMI" in NHANES

LEAD_QUANTILE Quantile membership for blood lead levels based on the distribution of lead levels in the data

HYP Hypertension Status: Based on "BPQ020" (Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?) and "BPQ040A" (Because of your high blood pressure/hypertension, have you ever been told to take prescribed medicine?) in NHANES. Equal to 1 if the respondent answered "Yes" to either of these questions, or, if data on either of these questions isn't answered, if SBP >= 130 or DBP >= 80, and equal to 0 otherwise.

ALC Alcohol Use: Based on "ALQ120Q" (In the past 12 months, how often did you drink any type of alcoholic beverage?) up to 2016 and "ALQ121" (the same question, but used after 2016) in NHANES. Equal to "Yes" if the respondent's answer to either of these questions was > 0 and equal to "No" otherwise.

DBP1 First Diastolic Blood Pressure (mmHg) reading: "BPXDI1" in NHANES.

DBP2 Second Diastolic Blood Pressure (mmHg) reading: "BPXDI2" in NHANES.

DBP3 Third Diastolic Blood Pressure (mmHg) reading: "BPXDI3" in NHANES.

DBP4 Fourth Diastolic Blood Pressure (mmHg) reading: "BPXDI4" in NHANES.

SBP1 First Systolic Blood Pressure (mmHg) reading: "BPXSY1" in NHANES.

SBP2 Second Systolic Blood Pressure (mmHg) reading: "BPXSY2" in NHANES.

SBP3 Third Systolic Blood Pressure (mmHg) reading: "BPXSY3" in NHANES.

SBP4 Fourth Systolic Blood Pressure (mmHg) reading: "BPXSY4" in NHANES.

Source

Data was obtained from the nhanesA package https://CRAN.R-project.org/package=nhanesA.

Variable selection and feature engineering were conducted in an effort to replicate the analyses conducted by

Huang, Z. (2022). Association Between Blood Lead Level With High Blood Pressure in US (NHANES 1999-2018). Frontiers in Public Health, 892.

https://www.frontiersin.org/articles/10.3389/fpubh.2022.836357/full.

---

nyts *Data from the 2021 National Youth Tobacco Survey*

Description

Variables relating to demographic information, frequency of tobacco (e-cigs, cigarettes, and cigars) use, and methods of obtaining said tobacco as reported by students on the 2021 NYTS.

Usage

```
nyts
```

Format

A data frame with 20,413 rows and 35 variables:

location Survey Setting: Answer to the question "Where are you currently taking the survey?"

age Age: Answer to QN1: "How old are you?"

sex Sex: Answer to QN2: "What is your sex?"

grade Grade: Answer to QN3: "What grade are you in?"

race_and_ethnicity Race and Ethnicity: Equal to "Hispanic" if any of QN4B ("Are you Hispanic, Latino, Latina, or of Spanish origin?" (Yes, Mexican, Mexican American, Chicano, or Chicana)), QN4C ("Are you Hispanic, Latino, Latina, or of Spanish origin?" (Yes, Puerto Rican)), QN4D ("Are you Hispanic, Latino, Latina, or of Spanish origin?" (Yes, Cuban)), or QN4E ("Are you Hispanic, Latino, Latina, or of Spanish origin?" (Yes, Another Hispanic, Latino, Latina, or Spanish origin)) are selected. Otherwise, equal to "non-Hispanic Black" if QN5C ("What race or races do you consider yourself to be?" (Black or African American)) is selected, equal to "non-Hispanic White" if QN5E ("What race or races do you consider yourself to be?" (White)) is selected, and equal to "non-Hispanic other race" if QN5A ("What race or races do you consider yourself to be?" (American Indian or Alaska Native)), QN5B ("What race or races do you consider yourself to be?" (Asian)), or QN5D ("What race or races do you consider yourself to be?" (Native Hawaiian or Other Pacific Islander)) is selected.

otherlang Speaks Language other than English at Home: Answer to QN154: "Do you speak a language other than English at home?"

grades_in_past_year Grades in the Past Year: Answer to QN165: "During the past 12 months, how would you describe your grades in school?"

LGBT LGBT Status: Equal to "Yes" if respondent answered QN155 ("Which of the following best describes you") with "Gay or Lesbian" or "Bisexual" or if respondent answered QN156 ("Some people describe themselves as transgender when their sex at birth does not match the way they think or feel about their gender. Are you transgender?") with "Yes, I am transgender". Equal to "Not Sure" if respondent answered QN155 with "Not Sure" or answered QN156 with "I am not sure if I am transgender". Equal to "No" if respondent answered QN155 with "Heterosexual (straight)" and answered QN156 with "No, I am not transgender".

psych_distress Psychological Distress: As defined in the online supplement for the linked paper: "Psychological distress was assessed with the Patient Health Questionnaire for Depression and Anxiety (PHQ-4), a composite scale made up of four questions: "During the past two weeks, how often have you been bothered by any of the following problems?": QN157A: Little interest or pleasure in doing things; QN157B: Feeling down, depressed, or hopeless; QN157C: Feeling nervous, anxious, or on edge; QN157D: Not being able or stop or control worrying. Response options were provided with a numeric value of 0 for "not at all," 1 for "several days," 2 for "more than half of the days," and 3 for "nearly every day". Responses were summed (range: 0 – 12) and categorized as none (0–2), mild (3–5), moderate (6–8) and severe (9–12)."

family_affluence Family Affluence: As defined in the online supplement for the linked paper: "Family affluence was assessed with the Family Affluence Scale (FAS), a composite scale made up of four questions. Numeric values were assigned to each response and summed across responses: QN161: "Does your family own a vehicle (such as a car, van, or truck)? (No=0; Yes, one=1; Yes, two or more=2); QN162: "Do you have your own bedroom?" (No=0; Yes=1); QN163: "How many computers (including laptops and tablets, not including game consoles and smartphones) does your family own?" (None=0; One=1; Two=2; More than two=3); and QN164: "During the past 12 months, how many times did you travel on vacation with your family? (Not at all=0; Once=1; Twice=2; More than twice=3). Summed responses (range: 0–9) were categorized into low (0–5), medium (6–7), and high (8–9)."

num_e_cigs Days of E-cig Use in the Past 30 days: Answer to QN9: "During the past 30 days, on how many days did you use e-cigarettes?". Equal to 0 if respondent answered QN6 ("Have you ever used an e-cigarette, even once or twice") with "No"

num_cigarettes Days of Cigarette Use in the Past 30 days: Answer to QN38: "During the past 30 days, on how many days did you smoke cigarettes?". Equal to 0 if respondent answered QN35 ("Have you ever smoked a cigarette, even one or two puffs") with "No"

num_cigars Days of Cigar Use in the Past 30 days: Answer to QN53: "During the past 30 days, on how many days did you smoke cigars, cigarillos, or little cigars?". Equal to 0 if respondent answered QN51 ("Have you ever smoked a cigar, cigarillo, or little cigar, even one or two puffs?") with "No"

perceived_cigarette_use Perceived Percentage of Students in Respondent's Grade who Smoke Cigarettes: Answer to QN125: "Out of every 10 students in your grade at school, how many do you think smoke cigarettes?" divided by 10

perceived_e_cig_use Perceived Percentage of Students in Respondent's Grade who Use e-cigarettes: Answer to QN126: "Out of every 10 students in your grade at school, how many do you think use e-cigarettes?" divided by 10

bought_myself "I bought them myself during the past 30 days": Equal to 1 if respondent selected any of QN20AA, QN20BA, QN20CA (During the past 30 days, how did you get your _____? (I bought them myself) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

had_someone_else_buy "I had someone else buy them for me during the past 30 days": Equal to 1 if respondent selected any of QN20AB, QN20BB, QN20CB (During the past 30 days, how did you get your _____? (I had someone else buy them for me) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

asked_someone_to_give_me_some "I asked someone to give me some during the past 30 days": Equal to 1 if respondent selected any of QN20AC, QN20BC, QN20CC (During the past 30 days, how did you get your _____? (I asked someone to give me some) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

someone_offered "Someone offered them to me during the past 30 days": Equal to 1 if respondent selected any of QN20AD, QN20BD, QN20CD (During the past 30 days, how did you get your _____? (Someone offered them to me) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

got_from_a_friend "I got them from a friend during the past 30 days": Equal to 1 if respondent selected any of QN20AE, QN20BE, QN20CE (During the past 30 days, how did you get your _____? (I got them from a friend) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

got_from_a_family_member "I got them from a family member during the past 30 days": Equal to 1 if respondent selected any of QN20AF, QN20BF, QN20CF (During the past 30 days, how did you get your _____? (I got them from a family member) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

took_them "I took them from a store or another person during the past 30 days": Equal to 1 if respondent selected any of QN20AG, QN20BG, QN20CG (During the past 30 days, how did you get your _____? (I took them from a store or another person) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

some_other_way "I got them in some other way during the past 30 days": Equal to 1 if respondent selected any of QN20AH, QN20BH, QN20CH (During the past 30 days, how did you get your _____? (I got them in some other way) for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

did_not_buy "I didn't buy tobacco products during the past 30 days": Equal to 1 if respondent selected all of QN21AA, QN21BA, QN21CA ("During the past 30 days, where did you buy your ____? (I did not buy ____ during the past 30 days)" for each tobacco product) or equal to 1 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_someone "I bought them from another person (a friend, family member, or someone else) during the past 30 days": Equal to 1 if respondent selected any of QN21AB, QN21BB, QN21CB ("During the past 30 days, where did you buy your ____? (I bought them from another person (a friend, family member, or someone else))" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_gas_station "I bought them from a gas station or convenience store during the past 30 days": Equal to 1 if respondent selected any of QN21AC, QN21BC, QN21CC ("During the past 30 days, where did you buy your ____? (A gas station or convenience store)" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_grocery_store "I bought them from a grocery store during the past 30 days": Equal to 1 if respondent selected any of QN21AD, QN21BD, QN21CD ("During the past 30 days, where did you buy your ____? (A grocery store)" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_drugstore "I bought them from a drugstore during the past 30 days": Equal to 1 if respondent selected any of QN21AE, QN21BE, QN21CE ("During the past 30 days, where did you buy your ____? (A drugstore)" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_mall "I bought them from a mall or shopping center kiosk/stand during the past 30 days": Equal to 1 if respondent selected any of QN21AF, QN21BF, QN21CF ("During the past 30 days, where did you buy your ____? (A mall or shopping center kiosk/stand)" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_vending_machine "I bought them from a vending machine during the past 30 days": Equal to 1 if respondent selected any of QN21AG, QN21BG, QN21CG ("During the past 30 days, where did you buy your ____? (A vending machine)" for each tobacco product. Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_internet "I bought them on the Internet (such as a product website or store website like eBay or Facebook Marketplace) during the past 30 days": Equal to 1 if respondent selected any of QN21AH, QN21BH, QN21CH ("During the past 30 days, where did you buy your ____? (On the Internet (such as a product website or store website like eBay or Facebook Marketplace))" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_through_mail "I bought them through the mail during the past 30 days": Equal to 1 if respondent selected any of QN21AI, QN21BI, QN21CI ("During the past 30 days, where did you buy your ____? (through the mail)" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_through_delivery "I bought them through a delivery service (such as DoorDash or Postmates) during the past 30 days": Equal to 1 if respondent selected any of QN21AJ, QN21BJ, QN21CJ ("During the past 30 days, where did you buy your ____? (through a delivery service (such as DoorDash or Postmates))" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_from_smoke_shop "I bought them from a vape shop or tobacco shop during the past 30 days": Equal to 1 if respondent selected any of QN21AK, QN21BK, QN21CK ("During the past 30 days, where did you buy your ____? (a vape shop or tobacco shop)" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

bought_elsewhere "I bought them from some other place not listed here during the past 30 days": Equal to 1 if respondent selected any of QN21AL, QN21BL, QN21CL ("During the past 30 days, where did you buy your ____? (some other place not listed here)" for each tobacco product). Equal to 0 if days used in the past 30 days is equal to 0 for all three tobacco products.

Source

Data was downloaded from the CDC's website at the following link:

https://www.cdc.gov/tobacco/about-data/surveys/national-youth-tobacco-survey.html.

Variables were selected and defined in a similar manner to those in

Park-Lee, E., Gentzke, A. S., Ren, C., Cooper, M., Sawdey, M. D., Hu, S. S., & Cullen, K. A. (2023). Impact of Survey Setting on Current Tobacco Product Use: National Youth Tobacco Survey, 2021. Journal of Adolescent Health, 72(3), 365-374.

https://pubmed.ncbi.nlm.nih.gov/36470692/

`pain` *Data from Alter et al. (2021)'s Study on Patient-Reported Pain*

Description

Information from patient-reported pain assessments using the Collaborative Health Outcomes Information Registry (CHOIR) at baseline and at a 3-month follow-up.

Usage

```
pain
```

Format

A data frame with 21,659 rows and 92 variables. Data and variable descriptions were downloaded from the "S1 Dataset".

PATIENT_NUM Deidentified study identification number

X101 Body Region Selected = 1; not selected = 0

X102 Body Region Selected = 1; not selected = 0

X103 Body Region Selected = 1; not selected = 0

X104 Body Region Selected = 1; not selected = 0

X105 Body Region Selected = 1; not selected = 0

X106 Body Region Selected = 1; not selected = 0

X107 Body Region Selected = 1; not selected = 0

X108 Body Region Selected = 1; not selected = 0

X109 Body Region Selected = 1; not selected = 0

X110 Body Region Selected = 1; not selected = 0

X111 Body Region Selected = 1; not selected = 0

X112 Body Region Selected = 1; not selected = 0

X113 Body Region Selected = 1; not selected = 0

X114 Body Region Selected = 1; not selected = 0

X115 Body Region Selected = 1; not selected = 0

X116 Body Region Selected = 1; not selected = 0

X117 Body Region Selected = 1; not selected = 0

X118 Body Region Selected = 1; not selected = 0

X119 Body Region Selected = 1; not selected = 0
X120 Body Region Selected = 1; not selected = 0
X121 Body Region Selected = 1; not selected = 0
X122 Body Region Selected = 1; not selected = 0
X123 Body Region Selected = 1; not selected = 0
X124 Body Region Selected = 1; not selected = 0
X125 Body Region Selected = 1; not selected = 0
X126 Body Region Selected = 1; not selected = 0
X127 Body Region Selected = 1; not selected = 0
X128 Body Region Selected = 1; not selected = 0
X129 Body Region Selected = 1; not selected = 0
X130 Body Region Selected = 1; not selected = 0
X131 Body Region Selected = 1; not selected = 0
X132 Body Region Selected = 1; not selected = 0
X133 Body Region Selected = 1; not selected = 0
X134 Body Region Selected = 1; not selected = 0
X135 Body Region Selected = 1; not selected = 0
X136 Body Region Selected = 1; not selected = 0
X201 Body Region Selected = 1; not selected = 0
X202 Body Region Selected = 1; not selected = 0
X203 Body Region Selected = 1; not selected = 0
X204 Body Region Selected = 1; not selected = 0
X205 Body Region Selected = 1; not selected = 0
X206 Body Region Selected = 1; not selected = 0
X207 Body Region Selected = 1; not selected = 0
X208 Body Region Selected = 1; not selected = 0
X209 Body Region Selected = 1; not selected = 0
X210 Body Region Selected = 1; not selected = 0
X211 Body Region Selected = 1; not selected = 0
X212 Body Region Selected = 1; not selected = 0
X213 Body Region Selected = 1; not selected = 0
X214 Body Region Selected = 1; not selected = 0
X215 Body Region Selected = 1; not selected = 0
X216 Body Region Selected = 1; not selected = 0
X217 Body Region Selected = 1; not selected = 0
X218 Body Region Selected = 1; not selected = 0
X219 Body Region Selected = 1; not selected = 0

X220 Body Region Selected = 1; not selected = 0

X221 Body Region Selected = 1; not selected = 0

X222 Body Region Selected = 1; not selected = 0

X223 Body Region Selected = 1; not selected = 0

X224 Body Region Selected = 1; not selected = 0

X225 Body Region Selected = 1; not selected = 0

X226 Body Region Selected = 1; not selected = 0

X227 Body Region Selected = 1; not selected = 0

X228 Body Region Selected = 1; not selected = 0

X229 Body Region Selected = 1; not selected = 0

X230 Body Region Selected = 1; not selected = 0

X231 Body Region Selected = 1; not selected = 0

X232 Body Region Selected = 1; not selected = 0

X233 Body Region Selected = 1; not selected = 0

X234 Body Region Selected = 1; not selected = 0

X235 Body Region Selected = 1; not selected = 0

X236 Body Region Selected = 1; not selected = 0

X237 Body Region Selected = 1; not selected = 0

X238 Body Region Selected = 1; not selected = 0

PAIN_INTENSITY_AVERAGE Pain intensity NRS (0-10)

PROMIS_PHYSICAL_FUNCTION PROMIS physical function T-score, range 0-100

PROMIS_PAIN_BEHAVIOR PROMIS pain behavior T-score, range 0-100

PROMIS_DEPRESSION PROMIS depression T-score, range 0-100

PROMIS_ANXIETY PROMIS anxiety T-score, range 0-100

PROMIS_SLEEP_DISTURB_V1_0 PROMIS sleep disturbance T-score, range 0-100

PROMIS_PAIN_INTERFERENCE PROMIS pain interference, range 0-100

GH_MENTAL_SCORE PROMIS global mental health, range 0-100

GH_PHYSICAL_SCORE PROMIS global physical health, range 0-100

AGE_AT_CONTACT Age at baseline assessment extracted from EMR

BMI Body Mass Index at baseline extracted from EMR

CCI_TOTAL_SCORE Charlson Comorbidity Index extracted from EMR

PAIN_INTENSITY_AVERAGE.FOLLOW_UP Pain intensity NRS at follow up (range 0 - 10)

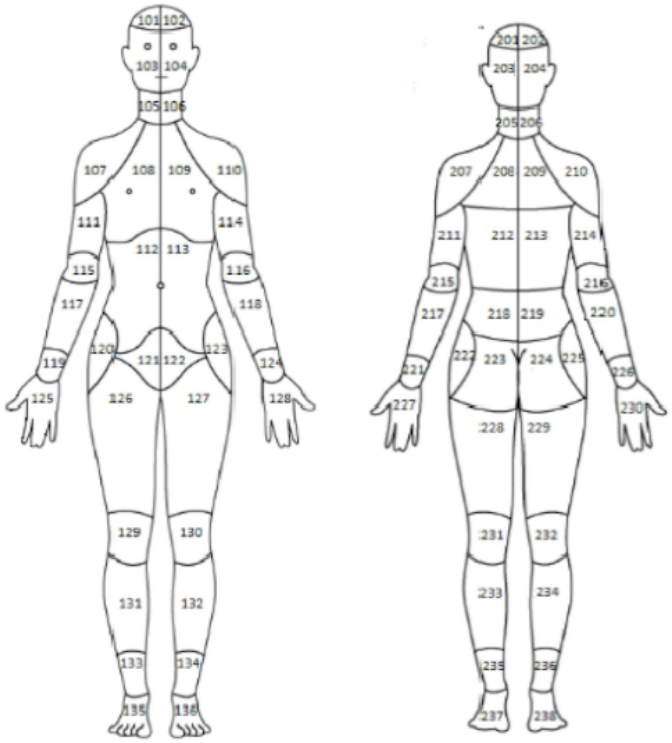
PAT_SEX Patient reported gender, "male" or "female", derived from EMR

PAT_RACE Patient reported race, 17 categories, EMR derived

CCI_BIN Binary Charlson Comorbidity Index: "No comorbidity" CCI score = 0; "Any comorbidity" CCI score > 0

MEDICAID_BIN Medicaid payor: "yes" or "no"

Here is a key for the coded body pain regions (S2 Fig from the linked paper):



Note that, as described in the paper, PROMIS is short for Patient-Reported Outcomes Measurement Information System: the source of the validated instruments for pain assessment used in the adaptive computerized test given to patients in accordance with the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT). EMR refers to the electronic medical record in the University of Pittsburgh's Patient Outcomes Repository for Treatment registry (PORT).

Source

Alter, B. J., Anderson, N. P., Gillman, A. G., Yin, Q., Jeong, J. H., & Wasan, A. D. (2021). Hierarchical clustering by patient-reported pain distribution alone identifies distinct chronic pain subgroups differing by pain intensity, quality, and clinical outcomes. PloS one, 16(8), e0254862.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0254862

`tb_diagnosis` *Data from TB studies in South Africa and Uganda*

Description

Demographic and health data collected from primary care clinic patients presenting with TB symptoms in rural South Africa (Kharitode study) and urban Uganda (STOMP study).

Usage

```
tb_diagnosis
```

Format

A data frame with 1762 rows and 11 variables:

tb TB test result (1 = positive, 0 = negative)

age_group Age group

hiv_pos Answer to the question "What is your HIV status?" (1 = positive, 0 = negative)

diabetes Self-reported history of diabetes (1 = diabetes, 0 = no diabetes)

ever_smoke Answer to the question "Do you smoke tobacco?" (1 = "yes" or "not currently, but formally", 0 = "no, never")

past_tb Answer to the question "Have you ever been diagnosed with TB in the past?" (1 = yes, 0 = no)

male Sex (1 = male, 0 = female)

hs_less Answer to the question "What is the highest grade of education that you have attained?" (1 = Grade 12 or lower, 0 = Any postgraduate education or higher)

two_weeks_symp Answer to the question "How long had you had a TB symptom (cough, fever, night sweats, weight loss) before you came to clinic?" (1 = >2 weeks, 0 = <2 weeks)

num_symptoms Number of TB symptoms (cough, fever, night sweats, weight loss)

country Country in which data were collected (South Africa = Kharitode study, Uganda = STOMP study)

Source

Baik, Y., Rickman, H. M., Hanrahan, C. F., Mmolawa, L., Kitonsa, P. J., Sewelana, T., Nalutaaya, A., Kendall, E. A., Lebina, L., Martinson, N., Katamba, A., & Dowdy, D. W. (2020). A clinical score for identifying active tuberculosis while awaiting microbiological results: Development and validation of a multivariable prediction model in sub-Saharan Africa. PLoS medicine, 17(11), e1003420. doi:10.1371/journal.pmed.1003420

The data are held in the Johns Hopkins University Data Services database and available at doi:10.7281/T1/W2AG3A.

---

tb_diagnosis_raw *Data from TB study in South Africa (Kharitode Study)*

Description

Demographic and health data collected from primary care clinic patients presenting with TB symptoms in rural South Africa.

Usage

```
tb_diagnosis_raw
```

Format

A data frame with 1634 rows and 34 variables:

consent Did the individual consent to participate in the study? (1 = Yes)

new_suspect_fac Is the participant a patient recently tested for TB? (1 = Yes)

ic_adult_fac Has informed consent been provided by the participant if age 18 or older? (1 = Yes; 2 = No; 77 = Under 18)

ic_adol_fac Has parental consent and adolescent/child assent been provided if the participant is less than 18 years of age? (1 = Yes, 2 = No)

facility_crf_complete Complete? (2 = Yes)

xpert_status_fac Is the participant TB-negative or TB-positive? (1 = Positive; 2 = Negative)

age_group Age group

sex Sex (1 = Male; 2 = Female)

hiv_status_fac What is your HIV status? (1 = Positive, 2 = Negative, 3 = Unknown, 99 = Refused)

other_conditions_fac___1 Do you have any other medical conditions? – None

other_conditions_fac___3 Do you have any other medical conditions? – Diabetes

other_conditions_fac___88 Do you have any other medical conditions? – Refused

other_conditions_fac___99 Do you have any other medical conditions? – Missing

symp_fac___1 On the day of your clinic visit, which of the following symptoms did you have? – Cough

symp_fac___2 On the day of your clinic visit, which of the following symptoms did you have? – Fever

symp_fac___3 On the day of your clinic visit, which of the following symptoms did you have? – Weight loss

symp_fac___4 On the day of your clinic visit, which of the following symptoms did you have? – Drenching sweats at night

symp_fac___5 On the day of your clinic visit, which of the following symptoms did you have? – Pain in my chest

symp_fac___11 On the day of your clinic visit, which of the following symptoms did you have? – No symptoms

symp_fac___99 On the day of your clinic visit, which of the following symptoms did you have? – Missing

length_symp_unit_fac How long had you had that symptom before you came to clinic that day? Enter unit of response. (1 = Days, 2 = Weeks, 3 = Months, 4 = Years, 77 = Unknown)

length_symp_mnt_fac How long had you had that symptom before you came to clinic that day? Length of time in months.

length_symp_yr_fac How long had you had that symptom before you came to clinic that day? Length of time in years.

length_symp_days_fac How long had you had that symptom before you came to clinic that day? Length of time in days.

length_symp_wk_fac How long had you had that symptom before you came to clinic that day? Length of time in weeks.

educ_fac What is the highest grade of education that you have attained? (0 = None; 1 = Grade 1; 2 = Grade 2; 3 = Grade 3; 4 = Grade 4; 5 = Grade 5; 6 = Grade 6; 7 = Grade 7; 8 = Grade 8; 9 = Grade 9; 10 = Grade 10; 11 = Grade 11; 12 = Grade 12; 13 = Any postgraduate education; 14 = Attained postgraduate degree; 99 = Missing)

on_arvs_fac Are you currently taking antiretrovirals for your HIV? (1 = Yes; 2 = No; 77 = Don't know or not applicable)

smk_fac Do you smoke tobacco? (1 = Yes; 2 = Not currently, but formerly; 3 = No, never; 88 = Refused; 99 = Missing)

dx_tb_past_fac Have you ever been diagnosed with TB in the past? (1 = Yes, 2 = No, 77 = Don't know)

seek_care_unit_fac If you were to develop a mild cough, how long would it likely be before you saw a doctor or other healthcare professional for a diagnosis? Unit of response. (1 = Days; 2 = Weeks; 3 = Months; 4 = Years; 77 = Don't know)

seek_care_days_fac If you were to develop a mild cough, how long would it likely be before you saw a doctor or other healthcare professional for a diagnosis? Length of time in days.

seek_care_wks_fac If you were to develop a mild cough, how long would it likely be before you saw a doctor or other healthcare professional for a diagnosis? Length of time in weeks

seek_care_mth_fac If you were to develop a mild cough, how long would it likely be before you saw a doctor or other healthcare professional for a diagnosis? Length of time in months.

seek_care_yr_fac If you were to develop a mild cough, how long would it likely be before you saw a doctor or other healthcare professional for a diagnosis? Length of time in years.

Source

Baik, Y., Rickman, H. M., Hanrahan, C. F., Mmolawa, L., Kitonsa, P. J., Sewelana, T., Nalutaaya, A., Kendall, E. A., Lebina, L., Martinson, N., Katamba, A., & Dowdy, D. W. (2020). A clinical score for identifying active tuberculosis while awaiting microbiological results: Development and validation of a multivariable prediction model in sub-Saharan Africa. PLoS medicine, 17(11), e1003420. doi:10.1371/journal.pmed.1003420

The data are held in the Johns Hopkins University Data Services database and available at doi:10.7281/T1/W2AG3A.

`tex_itop` *2016-2021 Statistics on Induced Terminations of Pregnancy (ITOP) in Texas*

Description

Texas abortion counts and rates by race/ethnicity and county of residence from 2016 to 2021 from the Texas Department of State Health Services (up to June 2018) and the Health and Human Services Commission since then.

Usage

```
tex_itop
```

Format

A data frame with 1,524 rows and 18 variables:

county County of residence in Texas

total_itop Total number of abortions

asian_itop Total number of abortions among Asian women between the ages of 15 and 44

hispanic_itop Total number of abortions among Hispanic women between the ages of 15 and 44

white_itop Total number of abortions among White women between the ages of 15 and 44

black_itop Total number of abortions among Black women between the ages of 15 and 44

native_american_itop Total number of abortions among Native American women between the ages of 15 and 44

other_itop Total number of abortions among women of other races or ethnicities between the ages of 15 and 44

year year

urban Indicator for whether the county is 'rural' or 'urban' according to the Texas Department of Housing and Community Affairs

total_rate Abortion rate per 1000 women between the ages of 15 and 44

asian_rate Abortion rate per 1000 Asian women between the ages of 15 and 44

hispanic_rate Abortion rate per 1000 Hispanic women between the ages of 15 and 44

white_rate Abortion rate per 1000 White women between the ages of 15 and 44

black_rate Abortion rate per 1000 Black women between the ages of 15 and 44

native_american_rate Abortion rate per 1000 Native American women between the ages of 15 and 44

other_rate Abortion rate per 1000 women of other races or ethnicities between the ages of 15 and 44

county_type Indicator for whether the county is urban, suburban, or rural according to the RUCC (rural-urban continuum codes) from the U.S. Department of Agriculture in 2013. Counties with Rural-Urban Continuum codes of 1-3 were categorized as urban, counties with codes of 4-7 were categorized as suburban, and counties with codes of 8 or 9 were categorized as rural.

Note from the data website: for the year 2020, "Data do not include 82 reports submitted after statutory deadlines and that were not available when annual data were compiled."

Source

Abortion counts by county and race/ethnicity were obtained from Texas Health and Human Services ISTOP Statistics at the following link:

https://www.hhs.texas.gov/about/records-statistics/data-statistics/texas-induced-terminations-pregr

To calculate abortion rates, total female populations between the ages of 15 and 44 were retrieved using the tidycensus package in R:

https://CRAN.R-project.org/package=tidycensus

Census codes for females between the ages of 15 and 44 by each race/ethnicity were retrieved from the following website:

https://api.census.gov/data/2020/dec/dhc/variables.html.

Information on whether counties are categorized as rural or urban was obtained from the 2022 Index of Texas Counties from the Texas Department of Housing and Community Affairs.

The 2013 Rural-Urban Continuum Codes from the U.S. Department of Agriculture were obtained from the following site:

https://www.ers.usda.gov/data-products/rural-urban-continuum-codes

Index

∗ **datasets**
breastcancer, 2
covidcases, 3
lockdowndates, 4
mobility, 5
NHANESsample, 5
nyts, 7
pain, 11
tb_diagnosis, 14
tb_diagnosis_raw, 15
tex_itop, 17

breastcancer, 2

covidcases, 3

lockdowndates, 4

mobility, 5

NHANESsample, 5
nyts, 7

pain, 11

tb_diagnosis, 14
tb_diagnosis_raw, 15
tex_itop, 17