Conditional recoding of variables for Data Analysis

Getting analytical variables from original/raw data for epidemiological studies

Lindsay Trujillo, PhD, MPH

Email: lindttruj@gmail.com | LinkedIn | Google Scholar

Introduction

This is the first document I created to demonstrate conditional recoding of analytical variables from original data. This arises from SAS programming practices I have done for local and national health surveillance data processing such as the National HIV Behavioral Surveillance, the Youth Risk Behavioral Surveillance, as well as analyses focusing on disease surveillance data and the Behavioral Risk Factor Surveillance system (BRFSS).

Known as filter questions, skip patterns are known to reduce the amount of time spent for each participant based on prior responses. As they help set up conditional denominators for calculating descriptive statistics, they are essential tools for survey methodology. For example, BRFSS asks questions on whether they have done an action in their life, such as smoking.[1] For those who never smoked, they may skip past detailed questions regarding smoking behaviors.

There are occasions that systems take into account skip patterns, such as imputing values, collapsing categories, or using metadata to preserve the survey logic.[2] Just as it is practical to have skip patterns programmed during data collection, it is also vital to consider for analysis. In public health and epidemiology, skip patterns reflect a type of eligibility criteria only considered for survey's purpose, and must be refined to fit the eligibility criteria that is appropriate for study. When they are ignored, prevalence estimates can be inflated due to misleading denominators and could be susceptible to misclassification bias. When skip patterns are considered during the first stage of your study design, researchers can improve precision, transparency, and improved interpretability for stakeholders.

This document will serve as the first example of demonstrating the impact of this type of data manipulation for R users. A modified version of the BRFSS 2008 data will be used, containing 11,046 observations with 46 variables.

For this document, there will be one aim: to create a dichotomous variable called Current Smoker derived from two variables of interest.

Data

```r
library(dplyr)
library(ggplot2)
library(tidyverse)
library(knitr)
library(kableExtra)
library(scales)

#Bringing in data
load("brfss08_samp.RData")
```

Two variables of interest to create one

For this request, I was asked to consider two variables of interest to create an analytical variable for current smoker:

Table 1: Survey Questions and Response Options		
Variable	**Question**	**Values**
SMOKE100	Have you smoked at least 100 cigarettes[1] in your entire life?	1 = Yes 2 = No 7 = Don't know/Not sure 9 = Refused
SMOKDAY2	Do you now smoke cigarettes every day, some days, or not at all?	1 = Every day 2 = Some days 3 = Not at all 7 = Don't know/Not sure 9 = Refused

[1]Note: 5 packs = 100 cigarettes

Let's look at one-way frequencies of each variable.

Table 2: Frequency of SMOKE100

Code	Ever Smk 100 cigs	Count
1	Yes	4,068
2	No	6,940
7	DK/NS	37
9	Missing	1
Total	Total Sample	11,046

From table 2, we can see that 4,068 participants had smoke at least 100 cigarettes in their entire lives, 6,940 participants did not smoke at least 100 cigarettes in their entire lives, 37 did not know or weren't sure, and 1 person refused to answer the question.

Table 3: Frequency of SMOKDAY2

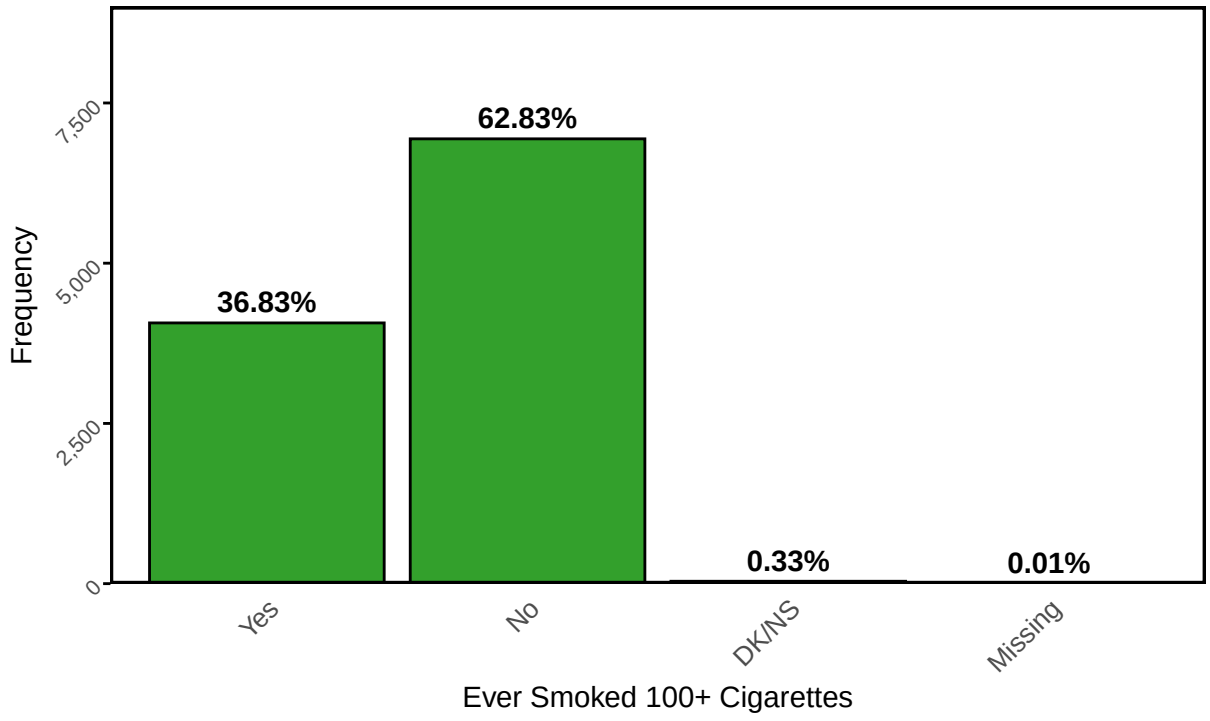
Code	Smoking Frequency	Count
1	Every day	1,097
2	Some days	395
3	Not at all	2,573
7	DK/NS	3
9	Refused	0
NA	NA	6,978
Total	Total Sample	11,046

In table 3, We can also see that 1,097 participants smoked cigarettes every day at the time of the survey, 395 smoked some days, 2,573 did not smoke at all, and 7 did not know or were not sure. We also see that we have 6,978 missing (NA) for this variable.

It seems intuitive to assume that those who did not smoke at least 100 cigarettes in their entire lives were not given the question. As well, it also seem intuitive to assume that those who did not know or refused to answer the question were not given the question either. However, a validation step is done to confirm the logic was truly followed.

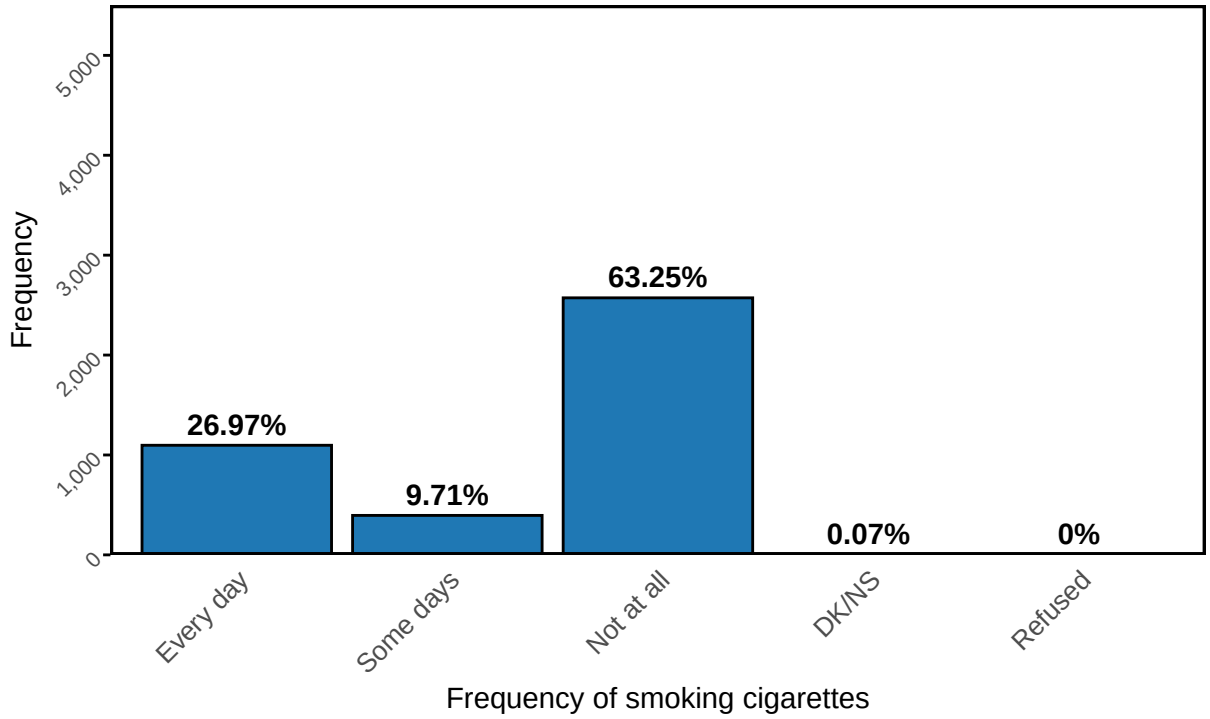
For fun, here's a way to visualize your raw distribution of both variables:

Figure 1: Raw distribution of ever smoked 100 cigarettes question



Abbreviation: DK/NS = Don't Know/Not Sure

Figure 2: Raw distribution of Frequency of smoking cigarettes



Abbreviation: DK/NS = Don't Know/Not Sure

Checking skip patterns

To confirm the skip patterns of the variables were followed, here is a two-way frequency table of SMOKE100 and SMOKDAY2:

Table 4: Cross-tabulation of SMOKE100 and SMOKDAY2 Response

Ever Smoked 100 Cigarettes	Smoking Frequency	Count
Yes	Every day	1,097
Yes	Some days	395
Yes	Not at all	2,573
Yes	DK/NS	3
Yes	Refused	0
No	Every day	0
No	Some days	0
No	Not at all	0
No	DK/NS	0
No	Refused	0
DK/NS	Every day	0
DK/NS	Some days	0
DK/NS	Not at all	0
DK/NS	DK/NS	0
DK/NS	Refused	0
Missing	Every day	0

(continued)

Ever Smoked 100 Cigarettes	Smoking Frequency	Count
Missing	Some days	0
Missing	Not at all	0
Missing	DK/NS	0
Missing	Refused	0
No	NA	6,940
DK/NS	NA	37
Missing	NA	1
Total	Total	11,046

Here, we safely see that respondents who said No (SMOKE100 = 2), Don't know or not sure (SMOKE100 = 7), or refused (SMOKE100 = 9) were marked as missing within the SMOKDAY2 variable. Here is a formal method to demonstrate this:

Table 4 demonstrates a suggested structure to export and "map" out all possible combinations to consider when creating your analytical variable for current smoker. In practice, this is exported out as a table (Excel) for reference during this stage of data manipulation.

Creating the analytical variable

Let's say that your workgroup decides that current smoker only include those who responded Every day or some days to the SMOKDAY2 question and those who responded No to SMOK100 and responded Not at all to SMOKDAY2 were defined as not being a current smoker. For re-coding, those deemed to be current smokers were to be coded as '1' for the new analytical variable (current_smk) and '0' for those who were not current smokers with anyone else marked as Unknown (NA_character_).

Here's an example of assuming that the question skip patterns were followed and not conducting any checks:

```
#Creating analytical variable
currsmk_wrong <- brfss08_samp %>%
  mutate(
    current_smk = case_when(
      # Unknown/refused
      SMOKDAY2 %in% c(7, 9)                    ~ NA_character_,
      # Smoke every day or some days
      SMOKDAY2 %in% c(1, 2)                              ~ "1",
      # Smoked not at all
      SMOKDAY2 == 3                                      ~ "0",
      TRUE ~ NA_character_  # default to missing (.X)
    )
  )

#Demonstrating issue:
currsmk_wrong %>%
count(SMOKE100, SMOKDAY2, current_smk, .drop = FALSE) %>%
  mutate(SMOKE100 = as.character(SMOKE100),
```

```
        SMOKDAY2 = as.character(SMOKDAY2),
        current_smk = as.character(current_smk)) %>%
  complete(SMOKE100 = all_codes1,
           SMOKDAY2 = all_codes2,
           current_smk = c("0", "1", NA_character_),
           fill = list(n = 0)) %>%
  bind_rows(
    tibble(SMOKE100 = "Total",
           SMOKDAY2 = "Total",
           current_smk = "Total",
           n = sum(.$n, na.rm = TRUE))
  ) %>%
  select(current_smk, SMOKE100, SMOKDAY2, n) -> crosstab1smk
```

Let's look at the first 12 combinations from our cross-walk (crosstab1smk).

```
bind_rows(slice_head(crosstab1smk, n = 12))
```

```
## # A tibble: 12 x 4
##    current_smk SMOKE100 SMOKDAY2     n
##    <chr>       <chr>    <chr>    <int>
##  1 0           1        1            0
##  2 1           1        1         1097
##  3 <NA>        1        1            0
##  4 0           1        2            0
##  5 1           1        2          395
##  6 <NA>        1        2            0
##  7 0           1        3         2573
##  8 1           1        3            0
##  9 <NA>        1        3            0
## 10 0           1        7            0
## 11 1           1        7            0
## 12 <NA>        1        7            3
```

We see that the records that were coded as '1' for our analytical variable were all those who reported smoking every day (SMOKE100 = 1 & SMOKDAY2 = 1, n = 1,097) and those who reported smoking some days (SMOKE100 = 1 & SMOKDAY2 = 2, n = 395). We also see that those who were coded as not current smokers seem to be coded appropriately (SMOKE100 = 1 & SMOKDAY2 = 2, n = 2,573) as well as our missing (SMOKE100 = 1 & SMOKDAY2 = 7, n = 3). So far, we see 4,068 participants were recoded, but what about the remaining 6,978?

```
bind_rows(slice_tail(crosstab1smk, n = 4))
```

```
## # A tibble: 4 x 4
##   current_smk SMOKE100 SMOKDAY2     n
```

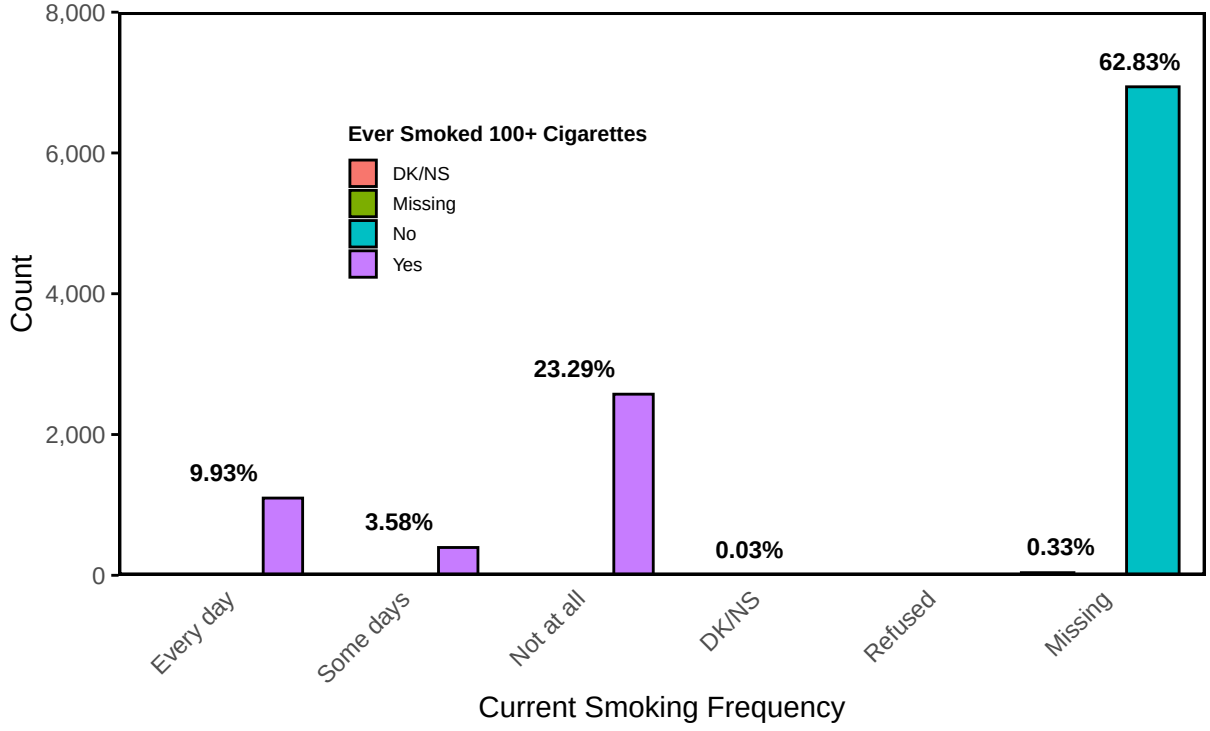
```
##   <chr>       <chr>    <chr>    <int>
## 1 <NA>        2        <NA>      6940
## 2 <NA>        7        <NA>        37
## 3 <NA>        9        <NA>         1
## 4 Total       Total    Total    11046
```

We see that the remaining records would be recoded as missing in our analytical variables. Those that reported don't know (SMOKE100 = 7 & SMOKDAY2 = NA, n = 37) and those that refused to answer the smoking question (SMOKE100 = 9 & SMOKDAY2 = NA, n = 1) are ok. But what about those that reported to not have smoked (SMOKE100 = 2 & SMOKDAY2 = NA, n = 6,940)? That's over 6,900 (63.2%) missing!

This is because skip patterns introduce a structural missingness; a respondent skips a question and is marked 'NA' but not due to nonresponse.[3] Those participants were omitted from this question because they have never smoked, or precisely smoked at least 100 cigarettes in their lives. From an epidemiological perspective, the deliberate skip of this question from those participants is based on survey logic, not error. Therefore, why would they be removed when the lack of exposure was documented?

Here's a visual to demonstrate this phenomenon.

Figure 3: Smoking Frequency by Smoking History



Abbreviation: DK/NS = Don't Know/Not Sure

From this example, we can see that we have a overwhelming amount of responses who reported "Yes" to Ever smoked 100+ cigarettes answering the Current Smoking Frequency question. However, we see all those who reported to never have smoked 100+ cigarettes (in teal blue) marked as missing.

This is theoretically correct based on survey questionnaire design. Why would those who never smoked be given this question? It would be a waste of time for the participant.

This is why it is critical to conduct these checks!

Therefore, with conditional recoding, this would be the appropriate method to create the current smoker analytical variable:

```
#Creating analytical variable
currsmk_right <- brfss08_samp %>%
  mutate(
    current_smk = case_when(
      # Unknown/refused
      SMOKE100 %in% c(7, 9)                  ~ NA_character_,
      SMOKE100 == 2                                  ~ "0",
      # Current smoker
      SMOKE100 == 1 & SMOKDAY2 %in% c(1, 2)          ~ "1",
      # Non-smoker
      SMOKE100 == 1 & SMOKDAY2 == 3                  ~ "0",
      # Unknown/refused
      SMOKE100 == 1 & SMOKDAY2 %in% c(7, 9) ~ NA_character_,
      TRUE ~ NA_character_  # default to missing (.X)
    )
  )

currsmk_right %>%
  count(SMOKE100, SMOKDAY2, current_smk, .drop = FALSE) %>%
  mutate(SMOKE100 = as.character(SMOKE100),
         SMOKDAY2 = as.character(SMOKDAY2),
         current_smk = as.character(current_smk)) %>%
  complete(SMOKE100 = all_codes1,
           SMOKDAY2 = all_codes2,
           current_smk = c("0", "1", NA_character_),
           fill = list(n = 0)) %>%
  bind_rows(
    tibble(SMOKE100 = "Total",
           SMOKDAY2 = "Total",
           current_smk = "Total",
           n = sum(.$n, na.rm = TRUE))
  ) %>%
  select(current_smk, SMOKE100, SMOKDAY2, n) -> crosstab2smk

#Getting first 12 combinations
bind_rows(slice_head(crosstab2smk, n = 12))
```

```
## # A tibble: 12 x 4
##    current_smk SMOKE100 SMOKDAY2     n
##    <chr>       <chr>    <chr>    <int>
##  1 0           1        1            0
```

```
##  2 1           1        1         1097
##  3 <NA>        1        1            0
##  4 0           1        2            0
##  5 1           1        2          395
##  6 <NA>        1        2            0
##  7 0           1        3         2573
##  8 1           1        3            0
##  9 <NA>        1        3            0
## 10 0           1        7            0
## 11 1           1        7            0
## 12 <NA>        1        7            3
```

```
#Getting last 3 combinations
bind_rows(slice_tail(crosstab2smk, n = 4))
```

```
## # A tibble: 4 x 4
##   current_smk SMOKE100 SMOKDAY2     n
##   <chr>       <chr>    <chr>    <int>
## 1 0           2        <NA>      6940
## 2 <NA>        7        <NA>        37
## 3 <NA>        9        <NA>         1
## 4 Total       Total    Total    11046
```

Here, we see that with conditional recoding, we were able to capture the 6,978 records to be coded as '0' for the analytical variable for current smoker, where subsequential analysis could be done and reported accurately.

Table 5: Three-Way Cross-tabulation of Smoking Variables

Current Smoker	Ever Smoked 100 Cigarettes	Smoking Frequency	Count
1	1	1	1,097
1	1	2	395
1	1	3	0
1	1	7	0
1	1	9	0
1	2	1	0
1	2	2	0
1	2	3	0
1	2	7	0
1	2	9	0
1	7	1	0
1	7	2	0
1	7	3	0
1	7	7	0
1	7	9	0
1	9	1	0
1	9	2	0
1	9	3	0
1	9	7	0
1	9	9	0
0	1	1	0
0	1	2	0
0	1	3	2,573
0	1	7	0
0	1	9	0
0	2	1	0
0	2	2	0
0	2	3	0
0	2	7	0
0	2	9	0
0	2	NA	6,940
0	7	1	0
0	7	2	0
0	7	3	0
0	7	7	0
0	7	9	0
0	9	1	0
0	9	2	0
0	9	3	0
0	9	7	0
0	9	9	0
Missing	1	1	0
Missing	1	2	0
Missing	1	3	0
Missing	1	7	3
Missing	1	9	0
Missing	2	1	0
Missing	2	2	0
Missing	2	3	0
Missing	2	7	0
Missing	2	9	0
Missing	7	1	0
Missing	7	2	0
Missing	7	3	0
Missing	7	7	0
Missing	7	9	0
Missing	7	NA	37
Missing	9	1	0
Missing	9	2	0
Missing	9	3	0
Missing	9	7	0
Missing	9	9	0
Missing	9	NA	1
Total	Total	Total	11,046

Impact

Conditional recoding may seem redundant, but it plays a pivotal role in ensuring data integrity throughout your analysis stages and onward. By honoring survey logic and skip patterns, it prevents misclassification and bias and reduces the risk of making decisions based on assumptions instead of logic. Analysts will spend less time fixing broken logic or re-running models during the validation stage. Therefore, reducing time burden and labor costs.

Through this method, you enable transparency in how definitions are applied, fostering trust in your outputs and enhancing reproducibility. When stakeholders can trace logic back to its source, your data stories become not only credible but actionable.

Summary

This example is meant to demonstrate the importance of conditional formatting for data analysis. In the future, more documents will be developed with more complex combinations as well as tips for data management for documentation.

Suggested Citation:

Trujillo L. Conditional recoding of variables for Data Analysis. GitHub. Published August 26, 2025. https://github.com/lindst973404/Conditional-recording-analysis.

References

1. Centers for Disease Control and Prevention. Data Guides: Health Risks and Behaviors. Disability and Health Data System (DHDS). Published April 3, 2025. Accessed August 25, 2025. https://www.cdc.gov/dhds/data-guides/health-risks-and-behaviors.html
2. Dykema JD. Skip Pattern Coding in Survey Data: A Comparison of Methods [master's thesis]. Lawrence, KS: University of Kansas; 2012. Available from: https://hdl.handle.net/1808/11040
3. Zhang G, He Y, Cai B, Moriarity C, Shin HC, Parsons V, Irimata KE. Multiple imputation of missing data with skip-pattern covariates: a comparison of alternative strategies. J Stat Comput Simul. 2023;94(7):1543–1570. doi:10.1080/00949655.2023.2293124.