# Introduction to the Datasets

Ashley I Naimi

April 2023

## Contents

# 1 Introduction to the Datasets

In this short course, we will have access to two datasets we will use to demonstrate the methods explored: 1) the NHEFS data, which we will use to demonstrate concepts and methods in the course; and 2) the Asthma data, which we will use for the homeworks and in class exercises.

Note that to analyze these data in the current course, we will rely on several simplifications to avoid getting bogged down by details that are important, but tangential to the concepts we need to cover to learn how to implement machine learning estimators for causal effects. These include the handling of missing data, confounder selection, trial emulation, information and selection bias, and other important topics. Indeed, there will be many ways to improve upon the analyses we conduct here, and I would encourage you to consider them.

# 2 NHANES Epidemiologic Follow-Up Study

We will be primarily relying on the NHANES Epidemiologic Follow-Up Study (NHEFS) to demonstrate the use of machine learning methods for estimating causal effects. These data were obtained from Hernán and Robins (Hernán and Robins, 2020). A subset of these data will be used to estimate the effect of quitting smoking on weight change (in kg) between 1971 and 1982.

```
a <- read_csv(url("https://tinyurl.com/2s432xv6")) %>%
    select(seqn, qsmk, sex, age, income,
        sbp, dbp, price71, tax71, race, wt82_71) %>%
    na.omit()

dim(a)
```

```
## [1] 1394   11
```

```
names(a)
```

```
##  [1] "seqn"    "qsmk"    "sex"     "age"     "income"  "sbp"     "dbp"
##  [8] "price71" "tax71"   "race"    "wt82_71"
```

```
head(a)
```

```
## # A tibble: 6 x 11
##    seqn  qsmk   sex   age income   sbp   dbp price71 tax71  race wt82_71
##   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>
## 1   233     0     0    42     19   175    96    2.18 1.10      1  -10.1 
## 2   235     0     0    36     18   123    80    2.35 1.36      0    2.60
## 3   244     0     1    56     15   115    75    1.57 0.551     1    9.41
## 4   245     0     0    68     15   148    78    1.51 0.525     1    4.99
## 5   252     0     0    40     18   118    77    2.35 1.36      0    4.99
## 6   257     0     1    43     11   141    83    2.21 1.15      1    4.42
```

```
write_csv(a, here("data", "nhefs.csv"))
```

Throughout this course, when using this dataset, our exposure of interest will be the indicator of whether the individual quit smoking (qsmk), our outcome will be weight change between 1971 and 1982, and our confounders of this relation will be all remaining variables (except the `seqn` participant ID).

It would be useful to become familiar with the variables selected to construct our analytic dataset. Information on these variables is available in the NHEFS codebook provided with the course materials.

## 3 Asthma Quality of Care Study

Most of the homework assignments will be completed in the Asthma dataset, which was used to evaluate the effect of being treated by different physician groups on perceived quality of asthma care in California (Huang et al., 2005).

```
# data taken from:
# www.biostat.jhsph.edu/~cfrangak/biostat_causal/asthma.txt

a <- read_delim(here("data", "asthma.txt"))

dim(a)
```

```
## [1] 276  10
```

```
names(a)
```

```
##  [1] "pg"       "age"      "sex"      "educ"     "insu"     "severity"
##  [7] "com"      "pcs"      "mcs"      "aqoc"
```

```
head(a)
```

```
## # A tibble: 6 x 10
##      pg   age   sex  educ  insu severity   com   pcs   mcs  aqoc
##   <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     0    36     0     5     1        2     4  50.0  49.6     1
## 2     0    37     1     6     2        3     3  40.9  56.1     1
## 3     0    43     1     5     1        4     3  50.3  57.1     1
## 4     0    39     1     6     1        4     1  49.3  49.7     1
## 5     0    46     1     1     2        4     0  37.4  53.0     0
## 6     0    40     1     3     1        3     2  40.6  50.7     0
```

In this dataset, the exposure is physician group (pg) , and the outcome is perceived asthma quality of care (aqoc), and our confounders of this relation will be all remaining variables in the dataset. Information on these variables is available in the asthma codebook provided with the course materials.

## References

M. A. Hernán and JM Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

I-Chan Huang, Constantine Frangakis, Francesca Dominici, Gregory B Diette, and Albert W Wu. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health services research*, 40(1):253–278, 2005.