Merging data using dplyr

Using Administrative Data for Clinical and Health Services Research

Overview

- Merging in the SID
- Key variables
- Examples

Merging in the SID

- Most variables of interest available in the CORE files
- Exceptions:
  - Detailed charges
  - Hospital characteristics
  - External data (e.g., from Census Bureau)

Keys

- Variables used to merge data sets are referred to as *keys*
- Merging requires understanding of how data are organized and what variable(s) represent keys
- `KEY` in the SID makes this easy for discharge *records*

Examples

```
payer_info
```

```
## # A tibble: 3 × 2
##   VisitLink  PAY1
##       <dbl> <dbl>
## 1     46571     1
## 2     93576     3
## 3     69250     3
```

```
charge_info
```

```
## # A tibble: 3 × 2
##   VisitLink TOTCHG
##       <dbl>  <dbl>
## 1     46571  28004
## 2     93576   7140
## 3     73038  82450
```

Inner Join

- Combines observations with matches in the key variable(s)
    - Unmatched records are dropped

```
payer_and_charge_info <- inner_join(
  payer_info, charge_info, by = "VisitLink")
payer_and_charge_info
```

```
## # A tibble: 2 × 3
##   VisitLink  PAY1 TOTCHG
##       <dbl> <dbl>  <dbl>
## 1     46571     1  28004
## 2     93576     3   7140
```

- Or using %>%,

```
payer_and_charge_info <- payer_info %>%
  inner_join(charge_info, by = "VisitLink")
```

Left Join

- Combines observations with matches in the key variable(s)
    - Unmatched records in 'left' data set are retained, those in 'right' data set are dropped

```
payer_and_maybe_charge_info <- payer_info %>%
  left_join(charge_info, by = "VisitLink")
payer_and_maybe_charge_info
```

```
## # A tibble: 3 × 3
##   VisitLink  PAY1 TOTCHG
##       <dbl> <dbl>  <dbl>
## 1     46571     1  28004
## 2     93576     3   7140
## 3     69250     3     NA
```

- `right_join` works in an analogous way

Full join

- Combines observations with matches in the key variable(s)
    - Unmatched records in both data sets are retained

```
payer_or_charge_info <- payer_info %>%
  full_join(charge_info, by = "VisitLink")
payer_or_charge_info
```

```
## # A tibble: 4 × 3
##   VisitLink  PAY1 TOTCHG
##       <dbl> <dbl>  <dbl>
## 1     46571     1  28004
## 2     93576     3   7140
## 3     69250     3     NA
## 4     73038    NA  82450
```

Anti Join

- *Removes* observations with matches in the key variable(s)

```
missing_payer_info <- payer_or_charge_info %>%
  filter(is.na(PAY1))
missing_payer_info
```

```
## # A tibble: 1 × 3
##   VisitLink  PAY1 TOTCHG
##       <dbl> <dbl>  <dbl>
## 1     73038    NA  82450
```

```
charge_info_but_only_if_payer_info <- charge_info %>%
  anti_join(missing_payer_info, by = "VisitLink")
charge_info_but_only_if_payer_info
```

```
## # A tibble: 2 × 2
##   VisitLink TOTCHG
##       <dbl>  <dbl>
## 1     46571  28004
## 2     93576   7140
```

Merging Tip

- Make sure the only shared variables are the key variable(s)

```
payer_info2
```

```
## # A tibble: 3 × 3
##   VisitLink  PAY1   AGE
##       <dbl> <dbl> <dbl>
## 1     46571     1    25
## 2     93576     3    31
## 3     69250     3    44
```

```
charge_info2
```

```
## # A tibble: 3 × 3
##   VisitLink TOTCHG   AGE
##       <dbl>  <dbl> <dbl>
## 1     46571  28004    25
## 2     93576   7140    31
## 3     73038  82450    76
```

Merging Tip

- Make sure the only shared variables are the key variable(s)

```
payer_and_charge_info_with_duplicate_age <- payer_info2 %>%
  inner_join(charge_info2, by = "VisitLink")
payer_and_charge_info_with_duplicate_age
```

```
## # A tibble: 2 × 5
##   VisitLink  PAY1 AGE.x TOTCHG AGE.y
##       <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1     46571     1    25  28004    25
## 2     93576     3    31   7140    31
```

Merging Tip

- Make sure the only shared variables are the key variable(s)

```
payer_and_charge_info_without_duplicate_age <- payer_info2 %>%
  select(!AGE) %>%
  inner_join(charge_info2, by = "VisitLink")
payer_and_charge_info_without_duplicate_age
```

```
## # A tibble: 2 × 4
##   VisitLink  PAY1 TOTCHG   AGE
##       <dbl> <dbl>  <dbl> <dbl>
## 1     46571     1  28004    25
## 2     93576     3   7140    31
```

Merging Tip

- Alternatively, add shared variables to the key

```
payer_and_charge_info_without_duplicate_age <- payer_info2 %>%
  inner_join(charge_info2, by = c("VisitLink", "AGE"))
payer_and_charge_info_without_duplicate_age
```

```
## # A tibble: 2 × 4
##   VisitLink  PAY1   AGE TOTCHG
##       <dbl> <dbl> <dbl>  <dbl>
## 1     46571     1    25  28004
## 2     93576     3    31   7140
```