Import AQI data

Michael Andreae

Friday, June 19, 2015

Insurance status predicts antiemetic use

We investigate the Hypothesis that socioeconomic status (**SES**) predicts anesthesia quality.

We use either insurance status or median income in patient address zip code as markers of socioeconomic status and antiemetic use as marker of anesthesia quailty. We use the population in the Public Use File *NACOR*, the National Anesthesia Clinical Outcome Registry, of the Anesthesia Quality Institute *AQI* with electronic anesthesia records recording antiemetic use.

Import Data and generate clean data files

we load the original dataset and save as it as *PUF_Q4_2013.Rdata*

Original Data

```
# run only once
PUF_Q4_2013 <- read.csv("Analysis/Data/PUF_Q4_2013_Antimetic.csv")
save(PUF_Q4_2013, file="Analysis/Data/PUF_Q4_2013.Rdata")
```

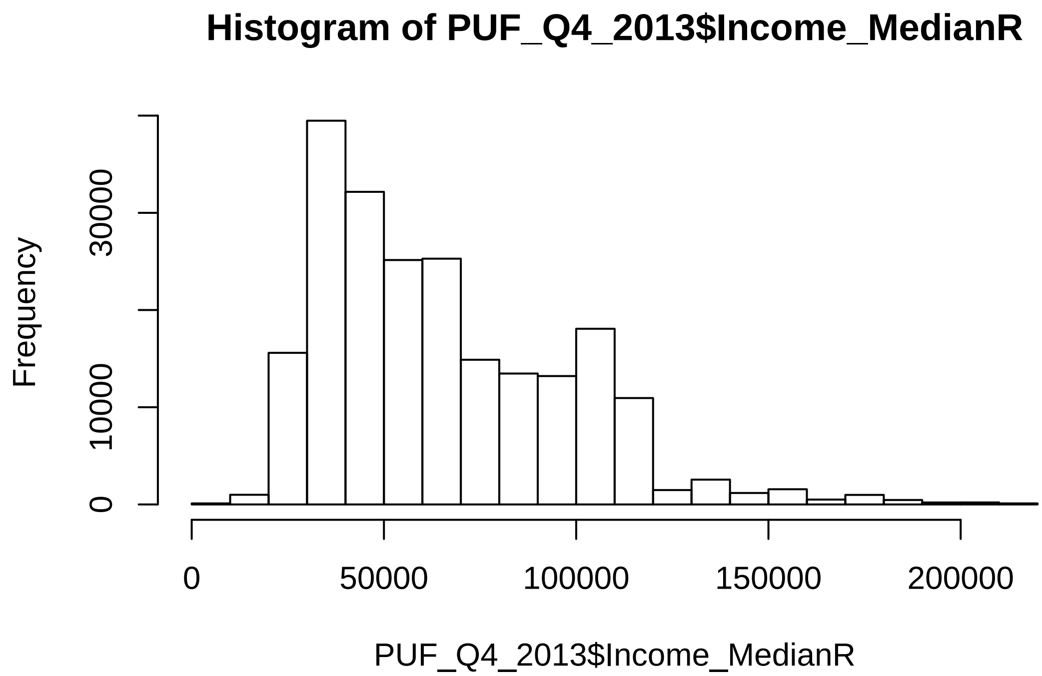
Load Rdata AQI raw dataset *PUF_Q4_2013.Rdata*

```
rm(list = ls())
load("Analysis/Data/PUF_Q4_2013.Rdata")
```

Median income versus insurance status as predictors of socioeconomic status

Median income

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   40000   58000   65880   88000  220000  223065
```



A predictor of socioeconomic status could be the median income in the zip code of the patients home address, coded *Income_MedicanR* , a integer with 223065 NA values. We will generate a separate file with median income as predictor at the bottom.

Insurance status

The predictor insurance status (*Payment*) is coded in 4 levels as Commercial, MEDICAID, Medicare, SELF, and has 265311 cases with without insurance information, *NA*

Missing data on insurance status and median income

	Observed Income	NA
Observed Insurance	115757	60577
NA	102823	162488

There is considerable overlap in the missingness of information on *Income_MedicanR* and insurance status coded as *Payment*. In other words, data are missing in different subpopulatons for *Income_MedianR* and *Payment*. This makes it feasible to test our hypothesis in two different subsets of the NACOR database

- The population with complete data on median income at patient address, using *Income_MedianR* as predictor for poor healthcare quality.
- The population with complete data on insurance status, using insurance status coded as *Payment* as predictor for poor healthcare quality.

Summary: We generate several clean data files. The first set of files uses insurance status as predictor; the second set of files uses median income (in patient address zip code) as predictor.

Predictor: insurance status

We generate the first set of files; We limit the analysis on the complete cases with *Payment* information on insurance status.

The predictor insurance status (*Payment*) is coded in 4 levels as Commercial, MEDICAID, Medicare, SELF, after we removed 265311 cases without insurance information, (orginally coded as “”), with 176334 unique cases remaining.

Outcome variables: antiemetic administration

We focus on the antiemetics *ondansetron*, *dexamethason* and *droperidol*, the only agents with convincing evidence for effect.

Table 2: Cases with Ondansetron versus Dexamethason

	no Dex	Dex
no Ondan	79842	4873
Ondan	61191	30428

Table 3: Cases with Ondansetron versus Droperidol

	no Drope	Drope
no Ondan	84398	317
Ondan	89067	2552

Table 4: Cases with Dexamethason versus Droperidol

	no Drope	Drope
no Dex	140010	1023
Dex	33455	1846

The antiemetics *ondansetron* and *dexamethason* were sometimes administered together. This is coded in *ondan_dex_either*

Potential confounders and other variables

practice ID versus facility ID

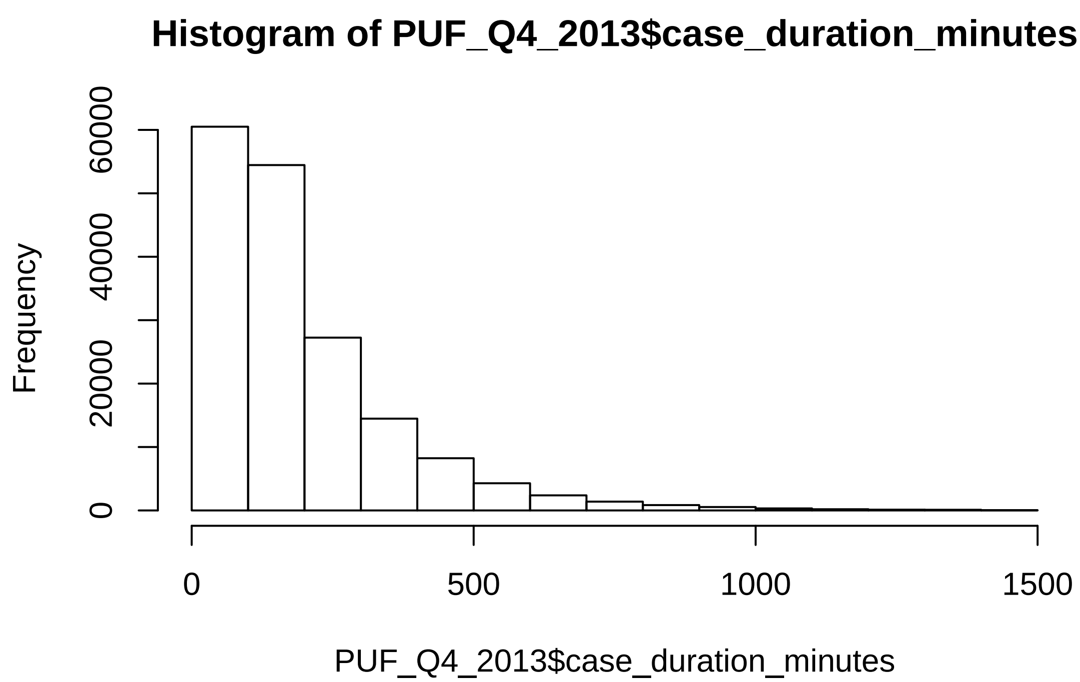
	193055	691419	5013437	5610264
136085	0	23241	0	0
1116623	36127	0	0	0
9485541	0	0	0	94080
23100212	0	0	79	0
46100453	0	0	15040	0
53228659	0	0	1	0
71100339	0	0	7766	0

The table of facility ID versus practice ID suggests that five practices have only one facility ID and one practice (=5013437) has three (sub) facilities. We will simplify by using practice ID, which has no NA.

We recode the practices to A through D to prevent reidentification.

case_duration_minutes

```
PUF_Q4_2013$case_duration_minutes [PUF_Q4_2013$case_duration_minutes == -1] <- NA
missing <- sum(is.na(PUF_Q4_2013$case_duration_minutes))
hist(PUF_Q4_2013$case_duration_minutes)
```

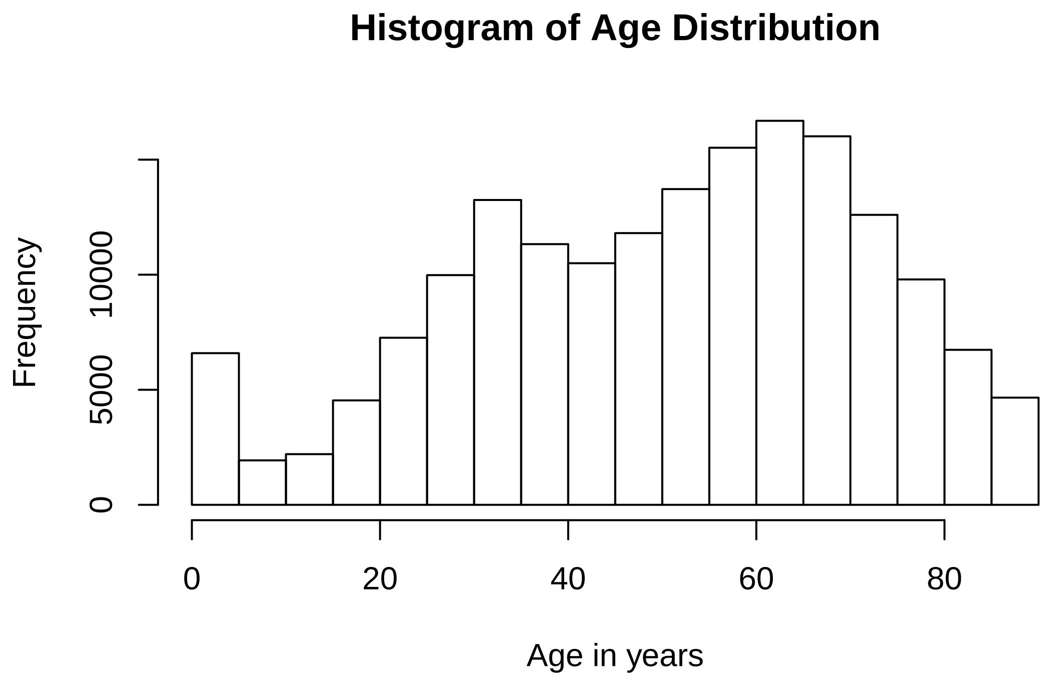



```
PUF_Q4_2013 <- PUF_Q4_2013 [complete.cases(PUF_Q4_2013$case_duration_minutes),]
```

Case duration in minutes (*case_duration_minutes*) is an integer and has 1211 missing values coded as -1, which we recoded as NA and removed from the dataset, leaving 175123 unique cases.

patient age

```
PUF_Q4_2013$patient_age[PUF_Q4_2013$patient_age==-1] <- NA
hist(PUF_Q4_2013$patient_age,
     main = "Histogram of Age Distribution",
     xlab = "Age in years")
```




```
missing <- sum(patient_age==-1)
PUF_Q4_2013 <- PUF_Q4_2013[complete.cases(PUF_Q4_2013$patient_age),]
```

Patient age (*patient_age*) is an integer with a distribution above and has 8 missing values coded as -1, which we recoded as NA and removed from the dataset, leaving 175115 unique cases.

patient_age_group

```
levels(PUF_Q4_2013$patient_age_group)[2] <- "1-18"
levels(PUF_Q4_2013$patient_age_group)[1] <- NA
PUF_Q4_2013$patient_age_group <- factor(PUF_Q4_2013$patient_age_group,
                levels(PUF_Q4_2013$patient_age_group)[c(2:6,1)])
missing <- sum(is.na(PUF_Q4_2013$patient_age_group))
PUF_Q4_2013 <- PUF_Q4_2013[complete.cases(PUF_Q4_2013$patient_age_group),]
```

Patient age group (*patient_age_group*) is a factor with 6 levels: 19 - 49, 50 - 64, 65 - 79, 80+, Under 1, 1-18; it has 0 missing values, leaving 175115 unique cases.

(Missing values were initially coded as -1, which we recoded as NA and removed as a level; we corrected the miscoding from “18-Jan”" to “1-18”).

patient_sex

Patient gender (*patient_sex*) is recoded as factor with the two levels female, male and 18 NAs, which are removed from the dataset, leaving 175097 unique cases.

in_or_out_patient

in- or outpatient status (*in_or_out_patient*) is recoded as a factor with the two levels Outpatient, Inpatient and 62432 NAs, which are too numerous to exclude.

surgical_cpt code

We considered to control with a random effect for *surgical_ctp* code but 58916 cases do not have a *surgical_ctp* code defined, which are too many to exclude.

combined_cpt code

We considered to control with a random effect for *combined_cpt* code but 71552 cases do not have a *combined_cpt* code defined, which are too many to exclude.

reported_anesthesia_code

We considered to control with a random effect for *reported_anesthesia_code* code but 174153 cases do not have a *reported_anesthesia_code* code defined, which are too many to exclude.

primary_anesthesia_type

primary_anesthesia-type is recoded as a factor with 7 levels [General, Neuroaxial, Regional, MAC, Sedation, Local, Other]. We considered to control with a fixed or a effect for *primary_anesthesia_type* code but 1914 cases do not have a *primary_anesthesia_type* code defined, which may be too many to exclude.

We did exclude NA leaving us with 173183 unique cases.

procedure_status

It would make sense to try to control for *procedure_status*, (which indicates if the case was Emergency or Elective); but 139046 of the remaining cases do not have a *procedure_status* code defined, which obviously are too many to exclude.

case_type

It would make sense to try to control for *case_type*, (which indicates if the case was Non - OR, OB/GYN NON Surgical, OB/GYN Surgical, OR, OTHER . . . , but 57743 of the remaining cases do not have a *case_type* code defined, which obviously are too many to exclude.

asaps_imputed

asaps

It would make sense to try to control for *asaps* or *asaps_imputed*, (ASA Status, which indicates how sick a patient was, and only 50 of the remaining cases do not have an ASA status recorded; so we exclude them, leaving us with 173133 unique cases with also *asaps* as a predictor.

prov1

It would be great to control for individual provider behavior, to show variability among providers in their propensity to admister antiemetics contingent on insurance status. There are 720 different *prov1* levels, I believe they are coding for individual providers. 26254 of the remaining cases do not have the *prov1* recorded; if we exclude them, it leaves us with 146879 unique cases with provider coded as *prov1* as predictor.

Save cleaned datasets

Larger dataset without provider information in *fullAQI_4_14*

a clean dataframe without provider info but more unique cases is saved as *fullAQI_4_14.Rdata*

```
## 'data.frame':    173133 obs. of  12 variables:
##  $ ond      : Factor w/ 2 levels "no Ondan","Ondan": 2 1 2 1 1 2 2 2 1 2 ...
##  $ dex      : Factor w/ 2 levels "no Dex","Dex": 1 1 1 1 1 1 1 1 1 1 ...
##  $ drop     : Factor w/ 2 levels "no Drope","Drope": 1 1 1 1 1 1 1 1 1 1 ...
##  $ any      : Factor w/ 2 levels "neither","either": 2 1 2 1 1 2 2 2 1 2 ...
##  $ pay      : Factor w/ 4 levels "Commercial","MEDICAID",..: 1 1 1 3 1 3 2 1 3 1 ...
##  $ age      : int  50 53 58 73 64 73 19 27 85 59 ...
##  $ age_group: Factor w/ 6 levels "19 - 49","50 - 64",..: 2 2 2 3 2 3 1 1 4 2 ...
##  $ sex      : Factor w/ 2 levels "female","male": 2 2 1 2 2 1 1 1 2 2 ...
##  $ ASA      : Factor w/ 6 levels "1","2","3","4",..: 2 3 3 2 3 3 2 2 3 2 ...
##  $ duration : int  59 43 190 56 37 116 93 108 70 93 ...
##  $ anes_type: Factor w/ 7 levels "General","Neuroaxial",..: 1 4 1 2 3 1 2 1 1 1 ...
##  $ practice : Factor w/ 4 levels "A","B","C","D": 2 2 2 2 2 2 2 2 2 2 ...
```

Smaller dataset with provider information in *prov1_AQI_4_14*

a more limited dataset with the individual provider as predictor is saved (after removing cases with *prov1* NA) in *prov1_AQI_4_14.Rdata*

```
## 'data.frame':    146879 obs. of  13 variables:
##  $ ond      : Factor w/ 2 levels "no Ondan","Ondan": 2 1 2 2 2 2 1 1 2 2 ...
##  $ dex      : Factor w/ 2 levels "no Dex","Dex": 1 1 1 1 1 2 1 1 1 2 ...
##  $ drop     : Factor w/ 2 levels "no Drope","Drope": 1 1 1 1 1 1 1 1 1 1 ...
##  $ any      : Factor w/ 2 levels "neither","either": 2 1 2 2 2 2 1 1 2 2 ...
##  $ pay      : Factor w/ 4 levels "Commercial","MEDICAID",..: 3 2 1 1 3 1 1 1 3 1 ...
##  $ age      : int  73 31 56 59 64 49 36 45 62 26 ...
##  $ age_group: Factor w/ 6 levels "19 - 49","50 - 64",..: 3 1 2 2 2 1 1 1 2 1 ...
##  $ sex      : Factor w/ 2 levels "female","male": 1 1 2 2 1 2 2 1 1 2 ...
##  $ ASA      : Factor w/ 6 levels "1","2","3","4",..: 4 3 2 2 3 1 2 2 3 2 ...
##  $ duration : int  172 80 172 122 133 83 136 81 112 110 ...
##  $ anes_type: Factor w/ 7 levels "General","Neuroaxial",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ provider : Factor w/ 720 levels "5622","5623",..: 161 155 152 163 156 156 161 161 156 153 ...
##  $ practice : Factor w/ 4 levels "A","B","C","D": 2 2 2 2 2 2 2 2 2 2 ...
```

Predictor: median income

We generate the second set of files; We limit the analysis on the complete cases with *Median_incomeR* information on median income in the patient's home address zip code.

Load Rdata AQI raw dataset *PUF_Q4_2013.Rdata*

```
rm(list = ls())
load("Analysis/Data/PUF_Q4_2013.Rdata")
summary(PUF_Q4_2013$Income_MedianR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   40000   58000   65880   88000  220000  223065
```

```
n_Income_MedianRNA <-sum(is.na(PUF_Q4_2013$Income_MedianR))
PUF_Q4_2013 <- PUF_Q4_2013[complete.cases(PUF_Q4_2013$Income_MedianR),]
```

A predictor of socioeconomic status could be the median income in the zip code of the patients home address, coded *Income_MedicanR* , a integer with 223065 NA values. We will generate a separate file with median income as predictor at the bottom, after we removed 223065 cases without information on median income, with 218580 unique cases remaining.

Outcome variables: antiemetic administration

We focus on the antiemetics *ondansetron*, *dexamethason* and *droperidol*, the only agents with convincing evidence for effect.

Table 6: Cases with Ondansetron versus Dexamethason

	no Dex	Dex
no Ondan	113700	5400
Ondan	53227	46253

Table 7: Cases with Ondansetron versus Droperidol

	no Drope	Drope
no Ondan	118675	425
Ondan	95219	4261

Table 8: Cases with Dexamethason versus Droperidol

	no Drope	Drope
no Dex	165721	1206
Dex	48173	3480

The antiemetics *ondansetron* and *dexamethason* were sometimes administered together. This is coded in *ondan_dex_either*

Potential confounders and other variables

practice ID versus facility ID

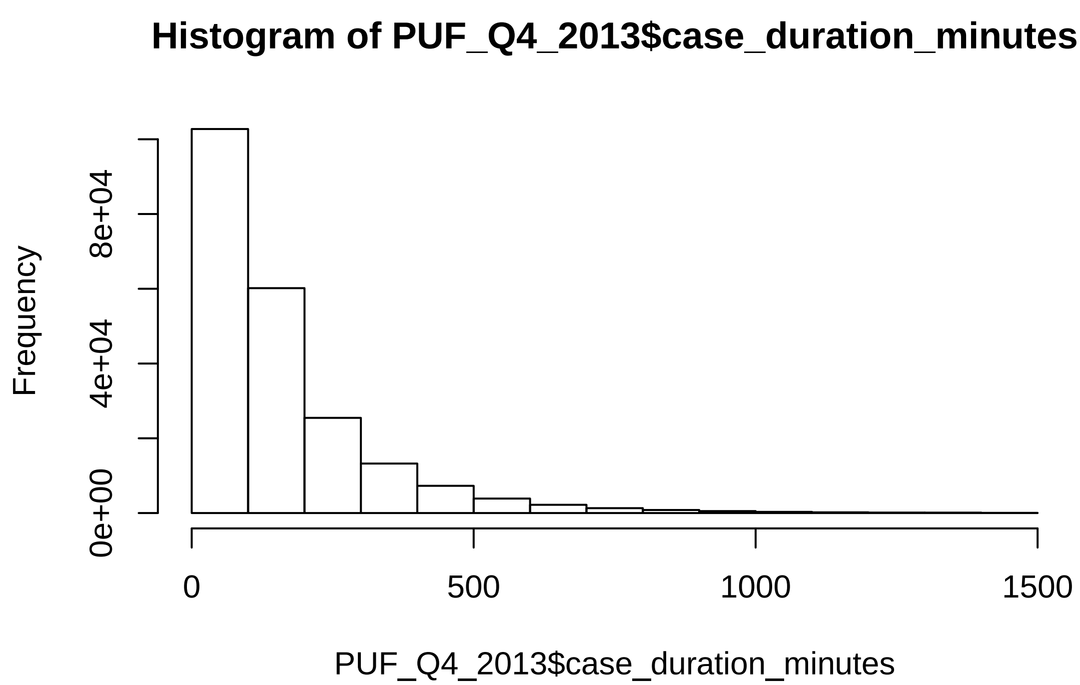
	5013437	5610264
9485541	0	179347
23100212	16721	0
46100453	14823	0
53228659	1	0
71100339	7688	0

The table of facility ID versus practice ID suggests that only two practices contribute data to this subset.

We recode the practices to A through B to prevent reidentification.

case_duration_minutes

```
PUF_Q4_2013$case_duration_minutes [PUF_Q4_2013$case_duration_minutes == -1] <- NA
missing <- sum(is.na(PUF_Q4_2013$case_duration_minutes))
hist(PUF_Q4_2013$case_duration_minutes)
```

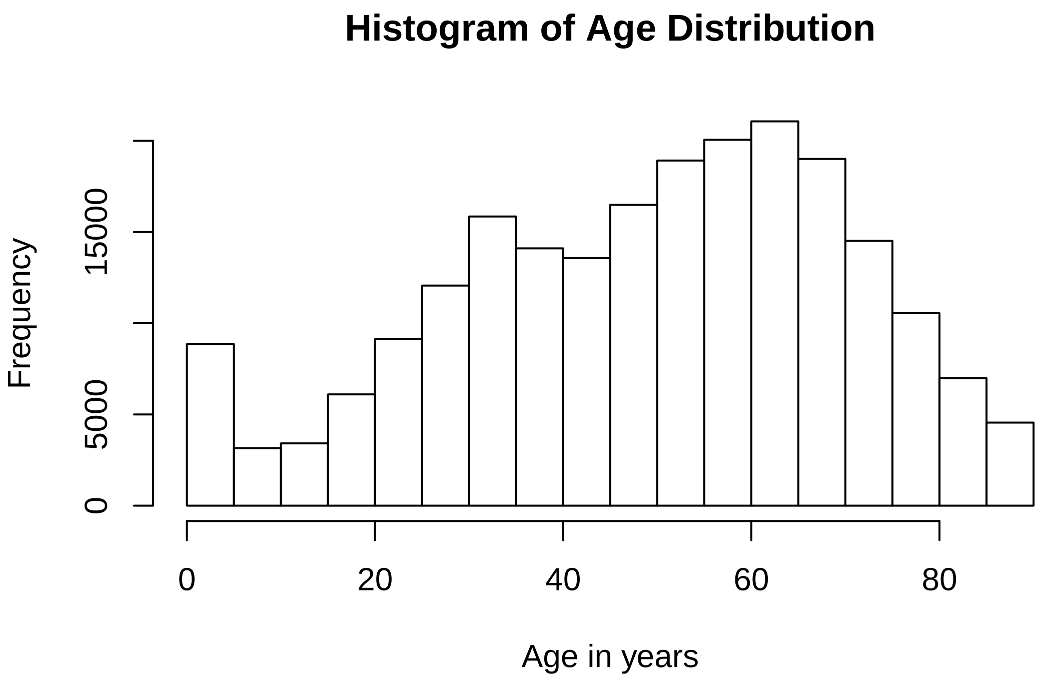



```
PUF_Q4_2013 <- PUF_Q4_2013 [complete.cases(PUF_Q4_2013$case_duration_minutes),]
```

Case duration in minutes (*case_duration_minutes*) is an integer and has 192 missing values coded as -1, which we recoded as NA and removed from the dataset, leaving 218388 unique cases.

patient age

```r
PUF_Q4_2013$patient_age[PUF_Q4_2013$patient_age==-1] <- NA
hist(PUF_Q4_2013$patient_age,
     main = "Histogram of Age Distribution",
     xlab = "Age in years")
```




```r
missing <- sum(patient_age==-1)
PUF_Q4_2013 <- PUF_Q4_2013[complete.cases(PUF_Q4_2013$patient_age),]
```

Patient age (*patient_age*) is an integer with a distribution above and has 7 missing values coded as -1, which we recoded as NA and removed from the dataset, leaving 218381 unique cases.

patient_age_group

```r
levels(PUF_Q4_2013$patient_age_group)[2] <- "1-18"
levels(PUF_Q4_2013$patient_age_group)[1] <- NA
PUF_Q4_2013$patient_age_group <- factor(PUF_Q4_2013$patient_age_group,
                levels(PUF_Q4_2013$patient_age_group)[c(2:6,1)])
missing <- sum(is.na(PUF_Q4_2013$patient_age_group))
PUF_Q4_2013 <- PUF_Q4_2013[complete.cases(PUF_Q4_2013$patient_age_group),]
```

Patient age group (*patient_age_group*) is a factor with 6 levels: 19 - 49, 50 - 64, 65 - 79, 80+, Under 1, 1-18; it has 0 missing values, leaving 218381 unique cases.

(Missing values were initially coded as -1, which we recoded as NA and removed as a level; we corrected the miscoding from “18-Jan”" to “1-18”).

patient_sex

Patient gender (*patient_sex*) is recoded as factor with the two levels female, male and 14 NAs, which are removed from the dataset, leaving 218367 unique cases.

in_or_out_patient

in- or outpatient status (*in_or_out_patient*) is recoded as a factor with the two levels Outpatient, Inpatient and 4091 NAs, which are too numerous to exclude.

surgical_cpt code

We considered to control with a random effect for *surgical_ctp* code but 51183 cases do not have a *surgical_ctp* code defined, which are too many to exclude.

combined_cpt code

We considered to control with a random effect for *combined_cpt* code but 63688 cases do not have a *combined_cpt* code defined, which are too many to exclude.

reported_anesthesia_code

We considered to control with a random effect for *reported_anesthesia_code* code but 218213 cases do not have a *reported_anesthesia_code* code defined, which are too many to exclude.

primary_anesthesia_type

primary_anesthesia-type is recoded as a factor with 7 levels [General, Neuroaxial, Regional, MAC, Sedation, Local, Other]. We considered to control with a fixed or a effect for *primary_anesthesia_type* code but 5739 cases do not have a *primary_anesthesia_type* code defined, which may be too many to exclude.

We did exclude NA leaving us with 212628 unique cases.

procedure_status

It would make sense to try to control for *procedure_status*, (which indicates if the case was Emergency or Elective); but 179851 of the remaining cases do not have a *procedure_status* code defined, which obviously are too many to exclude.

case_type

It would make sense to try to control for *case_type*, (which indicates if the case was Non - OR, OB/GYN NON Surgical, OB/GYN Surgical, OR, OTHER . . . , but 46042 of the remaining cases do not have a *case_type* code defined, which obviously are too many to exclude.

asaps_imputed

asaps

It would make sense to try to control for *asaps* or *asaps_imputed*, (ASA Status, which indicates how sick a patient was, and only 67 of the remaining cases do not have an ASA status recorded; so we exclude them, leaving us with 212561 unique cases with also *asaps* as a predictor.

prov1

It would be great to control for individual provider behavior, to show variability among providers in their propensity to admister antiemetics contingent on insurance status. There are 607 different *prov1* levels, I believe they are coding for individual providers. 226 of the remaining cases do not have the *prov1* recorded; if we exclude them, it leaves us with 212335 unique cases with provider coded as *prov1* as predictor.

Save cleaned datasets

a dataset with the *Income_MedicanR* and individual provider as predictor is saved (after removing cases with *prov1* NA) in *prov1_Income_AQI_4_14.Rdata*

```
## 'data.frame':    212335 obs. of  13 variables:
##  $ ond      : Factor w/ 2 levels "no Ondan","Ondan": 2 1 1 2 2 2 1 2 2 2 ...
##  $ dex      : Factor w/ 2 levels "no Dex","Dex": 2 1 1 1 2 1 1 2 2 2 ...
##  $ drop     : Factor w/ 2 levels "no Drope","Drope": 1 1 1 1 1 1 1 1 1 1 ...
##  $ any      : Factor w/ 2 levels "neither","either": 2 1 1 2 2 2 1 2 2 2 ...
##  $ inc      : int  31000 44000 44000 68000 45000 44000 105000 107000 34000 83000 ...
##  $ age      : int  68 71 43 39 68 72 73 59 35 49 ...
##  $ age_group: Factor w/ 6 levels "19 - 49","50 - 64",..: 3 3 1 1 3 3 3 2 1 1 ...
##  $ sex      : Factor w/ 2 levels "female","male": 1 1 2 2 1 2 1 2 2 1 ...
##  $ ASA      : Factor w/ 6 levels "1","2","3","4",..: 3 4 5 3 2 3 4 2 3 2 ...
##  $ duration : int  205 359 153 235 355 431 444 195 295 246 ...
##  $ anes_type: Factor w/ 7 levels "General","Neuroaxial",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ prov     : Factor w/ 607 levels "5622","5623",..: 61 228 50 17 4 21 27 1 5 91 ...
##  $ practice : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
```