

Vaccine Adverse Event Reporting System (VAERS)

With Multilogistic
Regression and Gradient
Tree Boosting



Sabirah Shuaybi

Introduction to VAERS

- Created by FDA and CDC to receive reports about vaccine related adverse events
- Monitoring and recording of these reports in order to determine whether any vaccines or vaccine lots have a higher rate of adverse effects.
- Although a small percentage of population experiences an adverse reaction to vaccination, this number of vaccine injury sufferers is not small.

Problems

“The quality of VAERS data has been questioned. Because reports are submitted from a variety of sources, some inexperienced in completing medical data forms, reports may omit important data and contain obvious errors. Assessment is further complicated by the administration of multiple vaccines at the same time, following currently recommended vaccine schedules, because there may be no conclusive way to determine which vaccine or combination of vaccines caused the specific adverse event.”

“Former FDA commissioner David A. Kessler has estimated that VAERS reports currently represent **only a fraction** of the serious adverse events.”

Electronic Support for Public Health VAERS (ESP:VAERS)

HHS gave Harvard Medical School a \$1 million dollar grant to track VAERS reporting at Harvard Pilgrim Healthcare for 3 years and to create an automated reporting system which would revolutionize the VAERS reporting system-transforming it from “**passive**” to “**active**.”

Scope of the project was, “To create a generalizable system to facilitate detection and clinician reporting of vaccine adverse events, in order to **improve** the **safety** of national **vaccination** programs.”



Results of the ESP:VAERS Project

“Adverse events from drugs and vaccines are common, but underreported. Fewer than 1% of vaccine adverse events are reported. Low reporting rates preclude or slow the identification of ‘problem’ drugs and vaccines that endanger public health. New surveillance methods for drug and vaccine adverse effects are needed.”

“Barriers to reporting include a lack of clinician awareness, uncertainty about when and what to report, as well as the burdens of reporting: reporting is not part of the clinician’s usual workflow, takes time, and is duplicative.”



Motivation

- ❑ How can we reduce the number of adverse reactions suffered because of mandatory vaccination?
- ❑ How can we compensate those who experience these adverse reactions?



Data Merging

Vaers_DATA_2018

Vaers_VAX_2018

Vaers_SYMPTOMS_2018

vaers

Header	Type	Description of Contents
VAERS_ID	Num(6)	VAERS Identification Number
RECVDATE	Date	Date report was received
STATE	Char(2)	Box 1: State
AGE_YRS	Num(xxx.xx)	Box 4: Age in Years
CAGE_YR	Num(xxx)	Age of patient in years calculated by (vax_date-birthdate)*
CAGE_MO	Num(.x)	Age of patient in months calculated by (vax_date-birthdate).* The values for this variable range from (.0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1)
SEX	Char(1)	Box 5: Sex
RPT_DATE	Date	Box 6: Date Form Completed

Header	Type	Description of Contents
VAERS_ID	Num(6)	VAERS Identification Number
VAX_TYPE	Char(15)	Administered Vaccine Type
VAX_MANU	Char(40)	Vaccine Manufacturer
VAX_LOT	Char(15)	Manufacturer's Vaccine Lot
VAX_DOSE_SERIES	Char(2)	Dose number in series
VAX_ROUTE	Char(4)	Vaccination Route
VAX_SITE	Char(3)	Vaccination Site
VAX_NAME	Char(100)	Vaccination Name

Data Pre-Processing

```
# Subsetting Data (only keeping useful predictors in)
vaers <- vaers %>%
  select(AGE_YRS, SEX, DIED, ER_VISIT, HOSPDAYS, RECOVD, V_ADMINBY, VAX_MANU, VAX_ROUTE, VAX_SITE, VAX_TYPE)
```

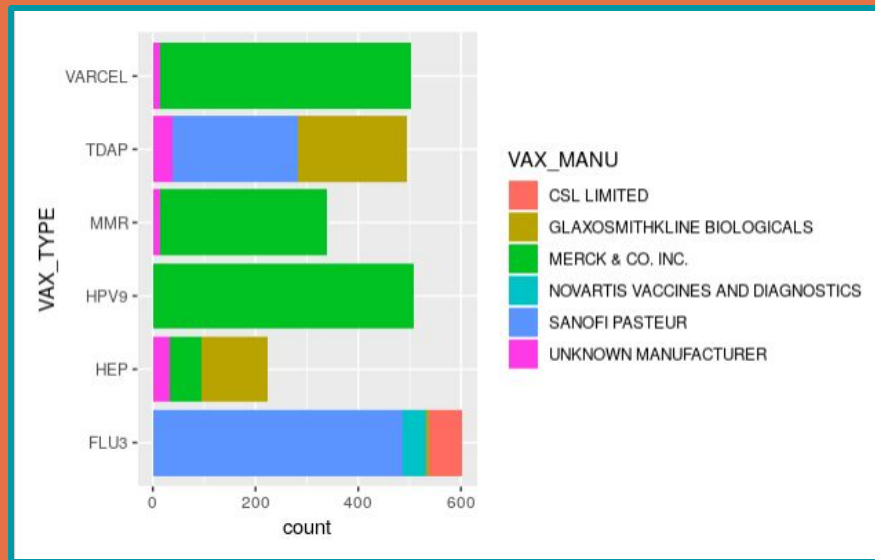
```
# Imputation of Missing Age Data (using median)
impute_missing_median <- function(x) {
  x[is.na(x)] <- median(x, na.rm = TRUE)
  return(x)
}
vaers <- vaers %>% mutate_at("AGE_YRS", impute_missing_median)
```

```
vaers <- vaers %>% filter(VAX_TYPE == c("FLU3", "MMR", "TDAP", "VARCEL", "HPV9", "HEP"))
```

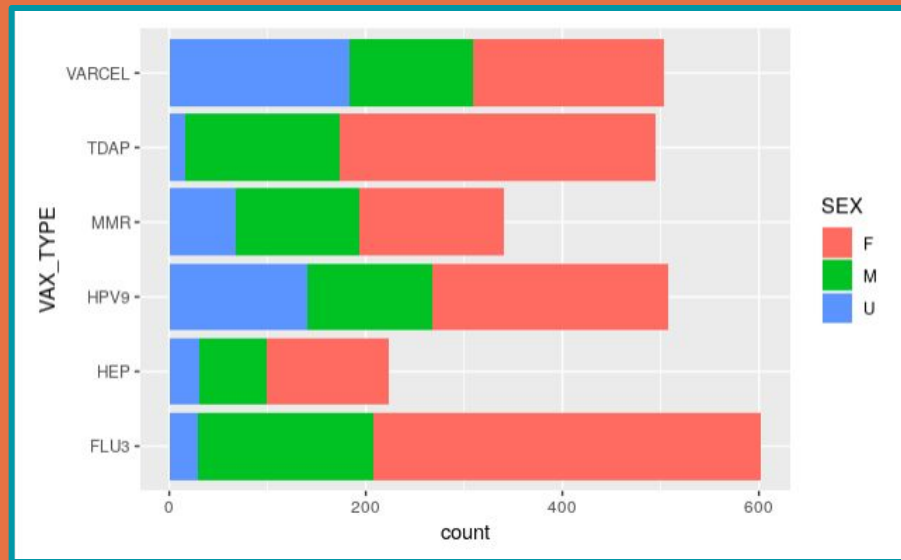
```
set.seed(723)

# Train/test split
tt_inds <- caret::createDataPartition(vaers$VAX_TYPE, p = 0.7)
train_set <- vaers %>% slice(tt_inds[[1]])
test_set <- vaers %>% slice(-tt_inds[[1]])
```


Data Exploration and Visualization

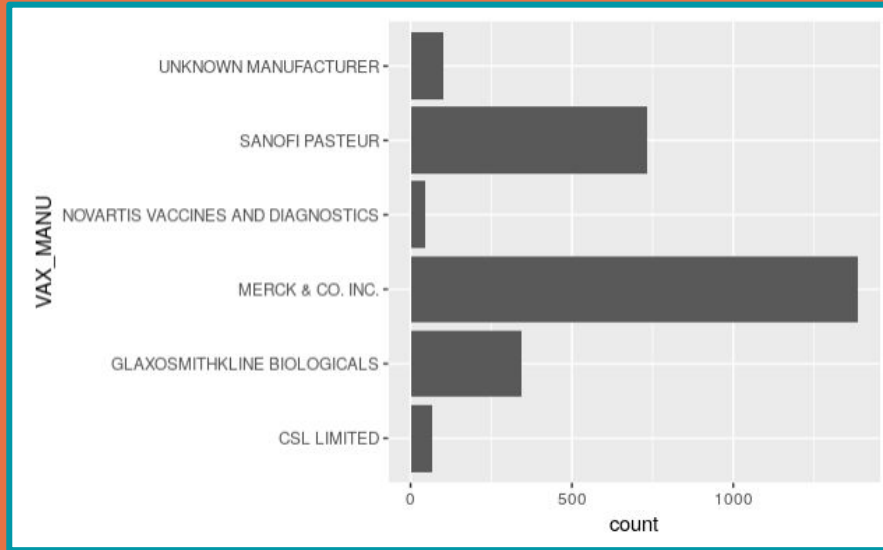


Vaccine Type and Vaccine Manufacturers

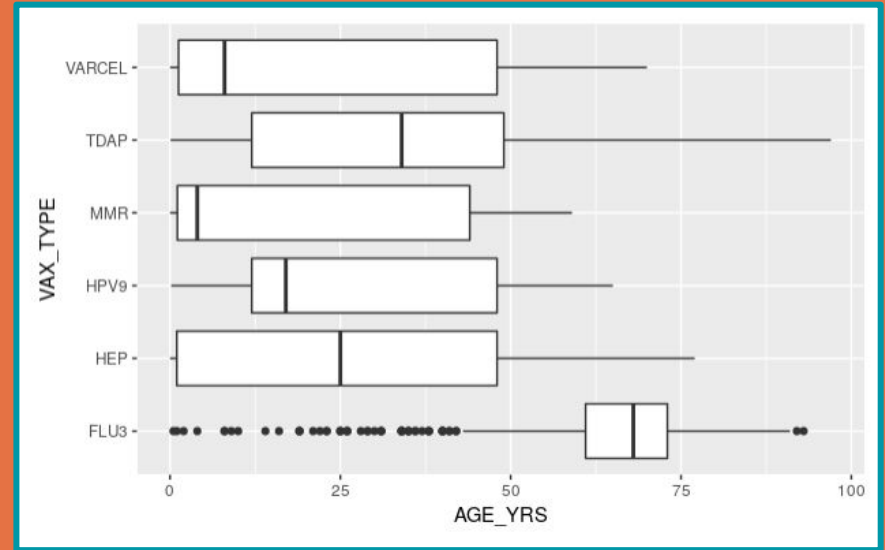


Vaccine Type and Gender

Data Exploration and Visualization

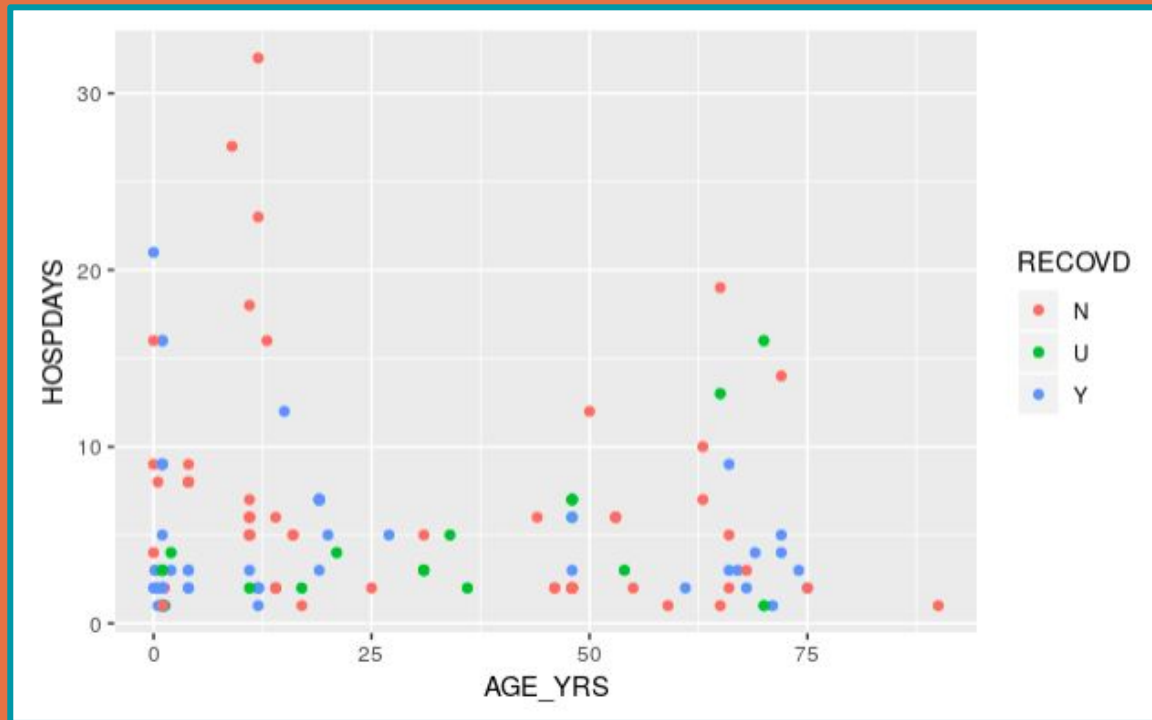


Vaccine Manufacturers



Vaccine Type and Age (in years)

Data Exploration and Visualization



Hospital Days and Age (in years)

Multilogistic Regression with $K > 2$

Define $Y_i = k$ if observation i is in class k

$$Y_i^{*(k)} = \begin{cases} 1 & \text{if } Y_i = k \\ 0 & \text{otherwise} \end{cases}$$

The probability that Y_i is in class k is

$$\frac{\exp(\beta_0^{(k)} + \beta_1^{(k)} x_{i1} + \cdots + \beta_p^{(k)} x_{ip})}{\exp(\beta_0^{(1)} + \beta_1^{(1)} x_{i1} + \cdots + \beta_p^{(1)} x_{ip}) + \cdots + \exp(\beta_0^{(k)} + \beta_1^{(k)} x_{i1} + \cdots + \beta_p^{(k)} x_{ip})}$$

Fitting a Multinomial Logistic Model on VAERS

```
multilogistic_fit <- train(
  VAX_TYPE ~ .,
  data = train_set,
  trace = FALSE,
  method = "multinom",
  trControl = trainControl(
    method = "cv",
    number = 10,
    returnResamp = "all",
    savePredictions = TRUE,
  ),
  tuneGrid = data.frame(decay = seq(from = 0, to = 0.2, length = 30))
)

# Pick tuning parameter values yielding highest cross-validated accuracy
multilogistic_fit$results %>% filter(Accuracy == max(Accuracy))
```
```

# Assessing Test Set Performance

```
mean(test_set$VAX_TYPE != predict(multilogistic_fit, test_set))
mean(test_set$VAX_TYPE == predict(multilogistic_fit, test_set))
```

Test Set **Error** Rate

[1] 0.35075

Test Set **Accuracy** Rate

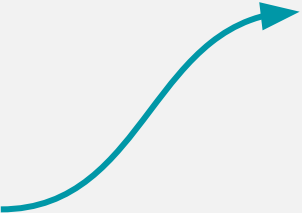
[1] 0.64925

# Gradient Tree Boosting

- An **ensemble** method; predictors selected *sequentially*
- **Conversion** of weak learners → **strong learners**
- After evaluating the first tree, we increase weights of observations that are difficult to classify and lower **weights** for those that are easy to classify
- Each new tree = modified
- New predictors are learning from mistakes committed by previous predictors → it takes **less time/iterations** to get close to final predictions

# Gradient Tree Boosting Fit on VAERS

```
xgb_fit <- train(
 VAX_TYPE ~ .,
 data = train_set,
 method = "xgbTree",
 trControl = trainControl(
 method = "cv",
 number = 10,
 returnResamp = "all",
 savePredictions = TRUE
),
 tuneGrid = expand.grid(
 nrounds = c(5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100),
 eta = c(0.5, 0.6, 0.7), # learning rate; 0.3 is the default
 gamma = 0, # minimum loss reduction to make a split; 0 is the default
 max_depth = 1:5, # how deep are our trees?
 subsample = c(0.5, 0.9, 1), # proportion of observations to use in growing each tree
 colsample_bytree = 1, # proportion of explanatory variables used in each tree
 min_child_weight = 1 # think of this as how many observations must be in each leaf node
)
)
```



**Tuning  
Parameters**



# Tuning Parameters

**Nrounds** → # of component models

Tried: 5, 10, 20 to 100

**Learning rate** → if too high, can result in overfitting.

Tried: 0.5, 0.6, and 0.7

**Subsample** → prop of obs used to grow each tree.

Tried: 0.5, 0.9, 1.

**Max depth** → deeper trees more complex, shorter pref.

Tried: values from 1 to 5.

# Results & Test Set Performance on XGB Fit

Predictions of final ensemble model = **weighted sum** of predictions made by previous tree models.

After experimenting with parameters, model with highest cross validation accuracy selected:

```
xgb_fit$results %>% filter(Accuracy == max(Accuracy))
```

```
mean(test_set$VAX_TYPE != predict(multilogistic_fit, test_set))
mean(test_set$VAX_TYPE == predict(multilogistic_fit, test_set))
```

Test Set **Error** Rate

Test Set **Accuracy** Rate

[1] 0.26866

[1] 0.73134

\* Better Test Set  
Performance than  
Multinomial Logistic  
Regression Model

# Conclusion & Limitations

- As science progresses, physicians and researchers will continue to establish connections between vaccines and certain adverse reactions.
- Constructing effective models like Multinomial Logistic Fit and Gradient Tree Boosting will help identify problematic areas in the vaccine schedule
- Provide efficient metrics to predict the likelihood of an adverse reaction based on predictors (gender, age, vaccine type, prior medications, allergies, etc.)



**Limitations** → analysis conducted on subsetting data. Future work would be to systematically expand this scope to obtain more scalable results

# References & Acknowledgements

<https://vaers.hhs.gov/>

[https://vaers.hhs.gov/docs/VAERSDataUseGuide\\_October2017.pdf](https://vaers.hhs.gov/docs/VAERSDataUseGuide_October2017.pdf)

<https://www.ncbi.nlm.nih.gov/pubmed/26060294>

[http://www.evanlray.com/stat340\\_f2018/materials/20181126\\_gradient\\_boosting\\_multi\\_class/20181126\\_gradient\\_boosting\\_multi\\_class.pdf](http://www.evanlray.com/stat340_f2018/materials/20181126_gradient_boosting_multi_class/20181126_gradient_boosting_multi_class.pdf)

<https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/>

Thank you!

Questions?

