Vaccine Effectiveness in Catalonia

Johannes Markus Burr and Erik Jansson

30 12 2021

Introduction

In this work data from the COVID19 pandemic in Catalonia from 2021 will be analyzed. Specifically, the relationship between the disease severity and vaccine status will be investigated. Approximately in the summer of 2021 European governments started to vaccinate persons with higher risk of a severe COVID19 infection. Since this data is purely observatory, this constitutes a study of vaccine effectiveness compared to vaccine efficacy, which was done in clearly designed randomized-control studies done for clinical admission.

Data Preparation

Since this is a real data set it naturally does not come precisely in the format our analysis expects it to be. Therefore, we first have to prepare and also translate it.

```
data = read.csv("Impacte_del_COVID.csv", encoding="UTF-8")
df = data
colnames(df) = c("sex", "age", "date", "event", "vaccinated", "count")
df$sex = as.factor(df$sex)               # male, female, NaN
df$age = as.factor(df$age)               # discretizied for some reason
df$event = as.factor(df$event)           # positive, hospitalized, critical (ICU)
df$date = as.Date(df$date,"%d/%m/%Y")    # now in format Year,month,day
df$vaccinated = as.factor(df$vaccinated) # none, partial, full
df$count = as.numeric(df$count)  # this means #count people with these characteristics have been observ

# the task asks us to only analyze people older than 30
df = subset(df, age=="30 a 39" | age == "40 a 49" | age == "50 a 59" |
         age == "60 a 69" | age == "70 a 79" | age == "80 o més")
# also only between 1st March 2020 and 12. December 2021
df = subset(df, date>"2021-03-01" | date > "2021-12-12")
dim(df)
```

```
## [1] 17059     6
```

The dataset captures these four event types.

```
table(df$event) # dependent variable
```

```
##
##             Cas         Crítics        Defunció Hospitalització
##            8042            2351            1565            5101
```

And shows the relationship to vaccine status.

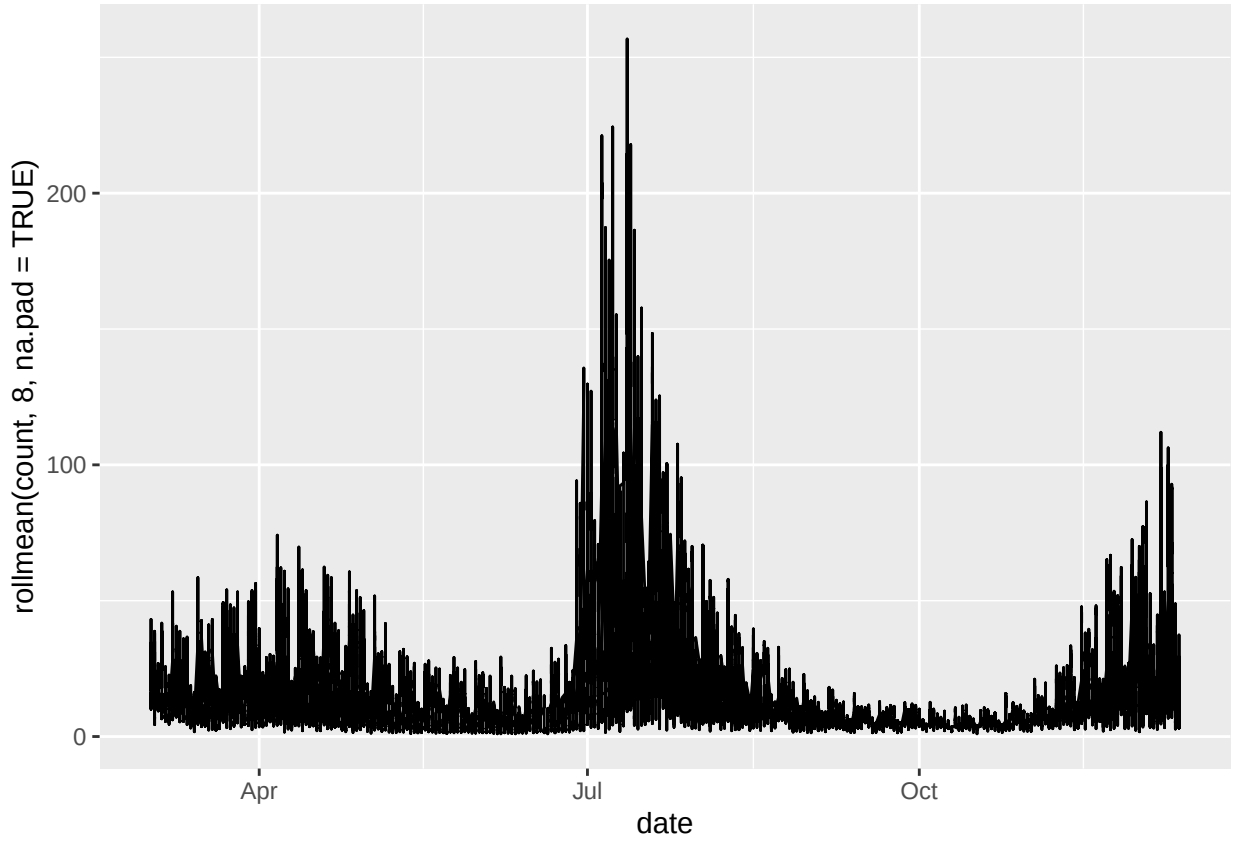
```
table(df$vaccinated)
```

```
##
##    Completa No iniciada     Parcial
##        5614        7981        3464
```

Descriptive Analysis

Firstly, let's plot the sum of all captured events as a function of time.

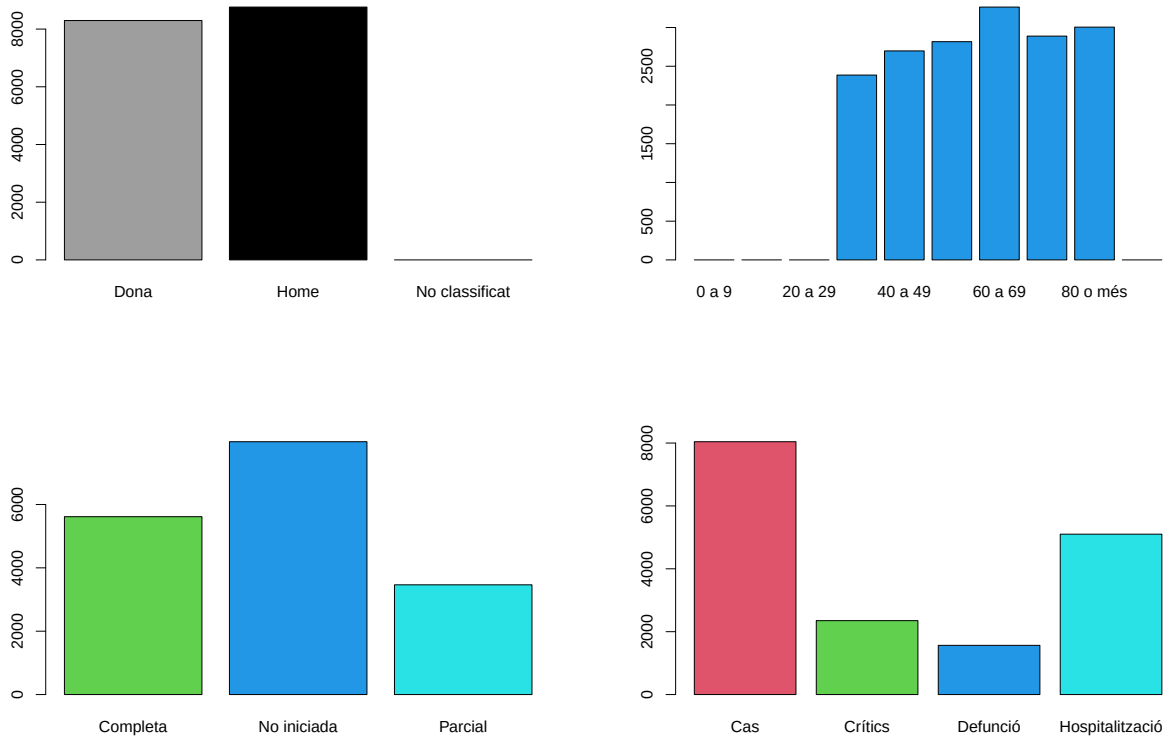
```
ggplot( data = df, aes( date, rollmean(count,8,na.pad=TRUE), vaccinated,
                        ylab="Number of Events")) + geom_line()
```



We have restricted the data to the time between March and December 2021, which includes the end of the Spring's wave but especially the big wave of infections in July as well as the beginning of the winter wave starting in November.

Exploratory Plots

```
plot(df$sex,col=8:9)
plot(df$age,col=4)
plot(df$vaccinated,col=3:5)
plot(df$event,col=2:6)
```



Slightly more male persons were observed than females. Since we restricted the data only to individuals older than 30, the age groups are quite balanced, with older people occurring slightly more often. People with a COVID related event were most of the time not vaccinated at all. The least common case was being partially vaccinated, which makes sense, since this status only held for around two months. Events captured in the dataset occurred by their severity: quarantine, hospitalization, critical (ICU), and death.

Critical Patients over time

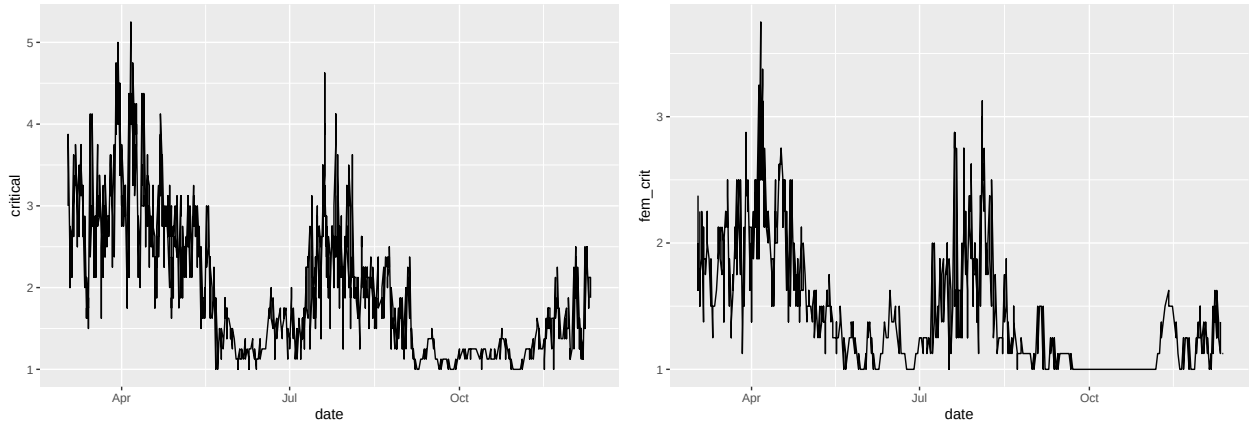
```
# males
df_critial_male = df[df$sex=="Home" & df$event=="Crítics",]
critical = rollmean(df_critial_male$count,8,na.pad=TRUE)
ggplot( data = df_critial_male, aes( date, critical, vaccinated )) + geom_line()
```

```
## Warning: Removed 4 row(s) containing missing values (geom_path).
```

```
#females
df_critial_female = df[df$sex=="Dona" & df$event=="Crítics",]
```

```
fem_crit = rollmean(df_critial_female$count,8,na.pad=TRUE)
ggplot( data = df_critial_female, aes( date, fem_crit, vaccinated )) + geom_line()
```

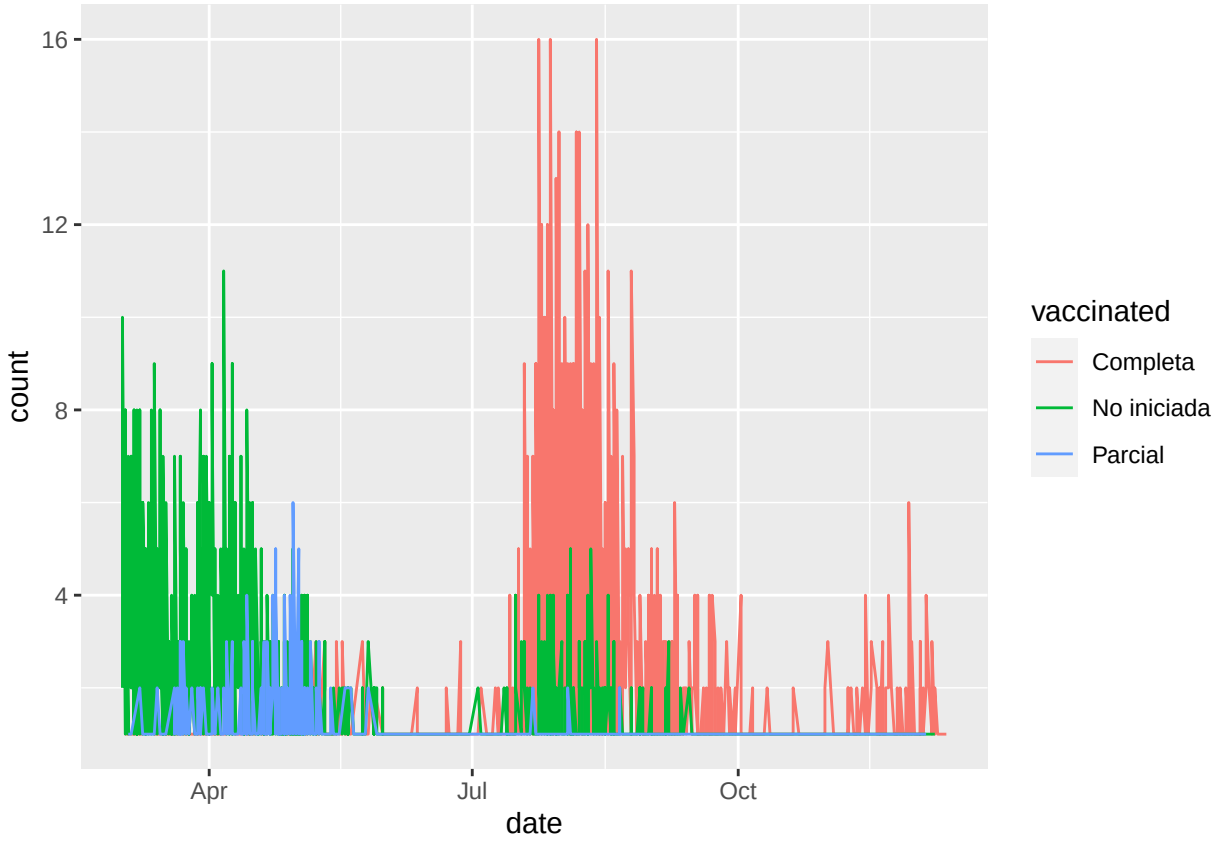
```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



On the left we see the development over time of male persons with a critical case, and on the right females on the ICU. To make the plot smoother amoving average over 8 days was calculated to tackle day-of-the-week effects. The three waves are clearly visible.

In the following plot we can see the development in dependency on the vaccine status.

```
df_deaths = df[df$event=="Defunció",]
ggplot(df_deaths,aes(x=date,y=count, color=vaccinated)) + geom_line()
```



Data Analysis

Contingency Table

A contingency table holds a count variable dependent on two (multilevel) factors. To build it we loop over the dataset, check the variables of interest and sum up the count variable and save it in a new dataframe.

Vaccine Status and Infection Gravity

Firstly, we investigate the relationship between vaccine status and the gravity of a coronavirus infection.

```
# init dataframe for count sums
table_vaccine = data.frame(matrix(0,ncol=3,nrow=4))
colnames(table_vaccine) = c("none","partial","full")  # vaccine status
rownames(table_vaccine) = c("casa","hospital","critical","dead") # event type
for (i in 1:dim(df)[1]){
  content_i = df[i,]

  age   = content_i[2]
  vacc  = content_i[5]
  event = content_i[4]
  count = content_i[6]
```

```
  #  define row and col variable, where count should be increased
  if (event=="Cas") row=1
  if (event=="Hospitalització") row = 2
  if (event=="Crítics") row = 3
  if (event=="Defunció") row = 4

  if (vacc=="No iniciada") col =1
  if (vacc=="Parcial") col =2
  if (vacc=="Completa") col =3

  table_vaccine[row,col] = count + table_vaccine[row,col]

}
table_vaccine
```

```
##            none partial  full
## casa     138487   24429 88167
## hospital  14323    1842  6243
## critical   3228     342   827
## dead       1513     336  1391
```

Now that we have the contingency table for event type by vaccine status, we can perform a χ^2 test testing the independence of both factors.

```
(chisq_vac = chisq.test(table_vaccine) )
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_vaccine
## X-squared = 1332.2, df = 6, p-value < 2.2e-16
```

The null hypothesis that the type of event and vaccine status are independent can be rejected.

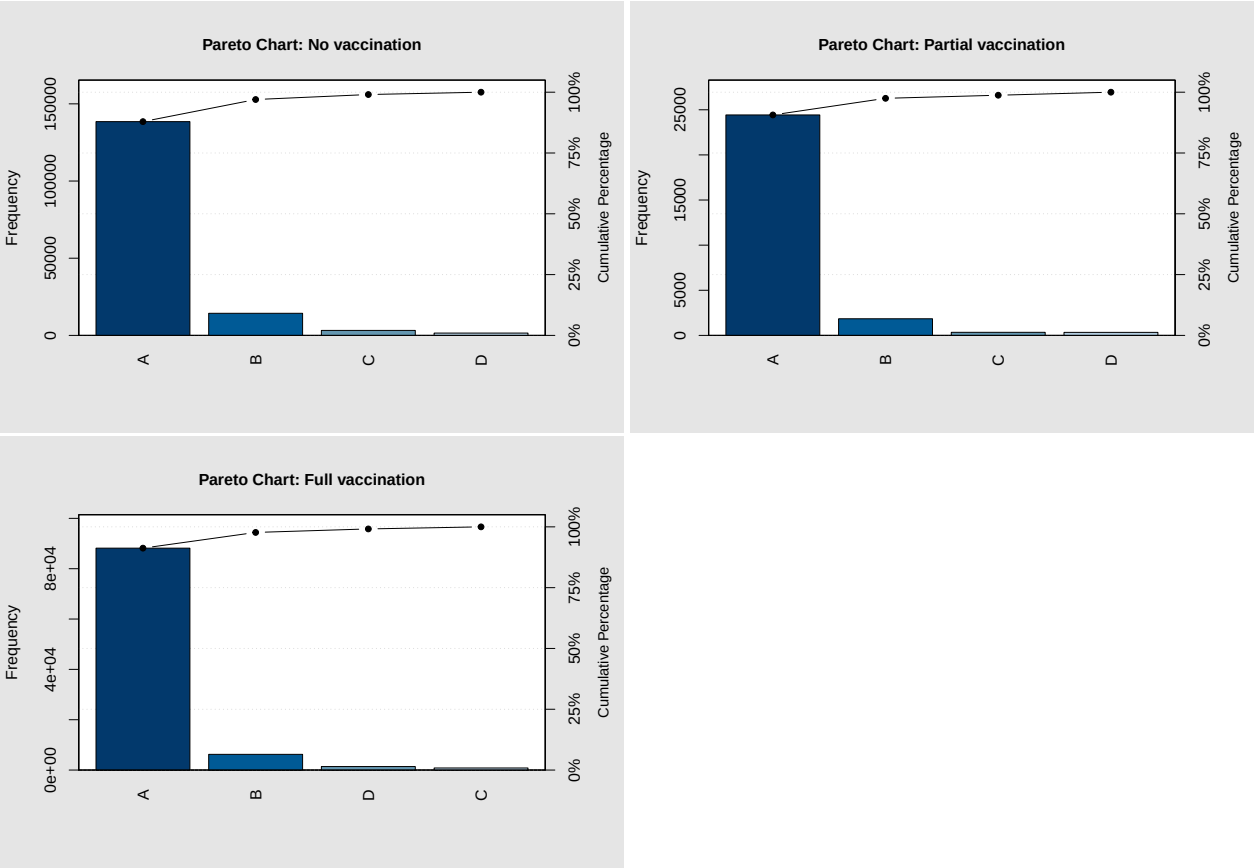
We can observe the difference between the observed contingency table and the table under the null-hypothesis.

```
table_vaccine - chisq_vac$expected
```

```
##                none    partial       full
## casa     -2226.0479  360.12117  1865.9267
## hospital  1765.0093 -306.03645 -1458.9729
## critical   763.8138  -79.49751  -684.3163
## dead      -302.7752   25.41279   277.3624
```

Looking at the "casa" row, we observe more "casa" cases than under the null-hypothesis for the fully vaccinated population rather than for the non-vaccinated. The reversed behavior is seen in the hospital and critical cases. By this observation one would expect vaccination to decrease the severity in case of covid. One unexpected observation is found in the "dead" row, where we observe more deaths among the fully vaccinated population than under the null-hypothesis. This could possibly be due to that the older population gets vaccinated first and/or is vaccinated to a greater extent, but still suffer a big risk of dying.

Pareto charts



On the x-axis we have A,B,C and D, which corresponds to the different cases “casa”, “hospital”, “critical” and “dead”.

The different pareto charts looks pretty similar for all vaccination statuses, with most observed cases being “casa”. We can see the fully vaccinated having a slightly larger percentage of “casa” cases compared to non-vaccinated, which is expected. For the fully vaccinated, we also have more deadly cases than critical. Both frequencies are similarly small so it this could be due to randomness, or that very old people that have a big risk of dying was probably vaccinated first.

Age and Infection Gravity

```
table_age = data.frame(matrix(0,ncol=6,nrow=4))
rownames(table_age) = c("casa","hospital","critical","dead")
dd = df
dd$age = as.numeric(dd$age)
for (i in 1:dim(df)[1]){
  row_i = dd[i,]

  age   = row_i[[2]]
  vacc  = row_i[5]
  event = row_i[[4]]
  count = row_i[6]
```

```
  if (event=="Cas") row=1
  if (event=="Hospitalització") row = 2
  if (event=="Crítics") row = 3
  if (event=="Defunció") row = 4


  table_age[row,age-3] = count + table_age[row,age-3]
  # the factor starts to count at 4 for some reason

}
table_age
```

```
##             X1    X2    X3    X4    X5    X6
## casa     74392 70147 43811 31794 17442 13497
## hospital  2449  3318  3766  4489  4041  4345
## critical   408   683   873  1237   967   229
## dead        13    36   125   383   701  1982
```

```
(chisq_age = chisq.test(table_age) )
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_age
## X-squared = 32235, df = 15, p-value < 2.2e-16
```

Again, the null hypothesis that age and severity of COVID19 infection are independent can be rejected.

We can observe the difference between observed counts and expected counts under null hypothesis.

```
table_age - chisq_age$expected
```

```
##                  X1         X2        X3          X4         X5         X6
## casa      5387.2237  3891.2685  427.3572 -2058.19170 -3234.7826 -4412.8752
## hospital -3709.3581 -2595.0185 -105.7901  1467.84798  2195.6925  2746.6262
## critical  -800.4211  -477.2795  113.2597   644.17577   604.9053   -84.6402
## dead      -877.4445  -818.9705 -434.8268   -53.83205   434.1847  1750.8892
```

In the table, column X1 is the first interval of ages, 30-40 years. X2 is 40-50 years and so on up to X6 which is 80 years old and above. If we look at the "casa" row, we observe more "casa" cases among the younger people than we expect under the null-hypothesis. For the older population we notice the opposite, less "casa" cases than expected. We can also make an analogous but reversed observation for the "hospital" and "dead" row. These observations supports that older people seem to have a more severe case of covid rather than young. We have one unexpected result in the table that is less observed critical cases among age of 80 and above. This could be explained by that people in this age are weak and do not survive critical cases, which increases the number of deaths for this group of age.

Statistical Inference

Parametric Analysis

Poisson Regression

To further investigate the relationship between vaccine status and event type, we fit a poisson model on the count variable from the contingency table. The model predicts the count variable through the event type and the vaccine status assuming a poisson distribution. We use poisson's canonical link function.

```
table_vaccine
```

```
##            none partial  full
## casa     138487   24429 88167
## hospital  14323    1842  6243
## critical   3228     342   827
## dead       1513     336  1391
```

```
y = unname(unlist(table_vaccine))
print(y)            # y is a vector of the count variable
```

```
##  [1] 138487  14323   3228   1513  24429   1842    342    336  88167   6243
## [11]    827   1391
```

```
vaccine_status = as.factor(c(rep(1,4),rep(2,4),rep(3,4)))
event_type = as.factor(c(rep(c(1,2,3,4),3)))
poiss_glm = glm(y~vaccine_status+event_type, family="poisson")
summary(poiss_glm)
```

```
##
## Call:
## glm(formula = y ~ vaccine_status + event_type, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -19.271   -6.907   -1.291    6.746   15.401
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     11.854478   0.002602  4555.1   <2e-16 ***
## vaccine_status2 -1.765803   0.006592  -267.9   <2e-16 ***
## vaccine_status3 -0.488881   0.004086  -119.6   <2e-16 ***
## event_type2     -2.416366   0.006972  -346.6   <2e-16 ***
## event_type3     -4.044861   0.015212  -265.9   <2e-16 ***
## event_type4     -4.350210   0.017681  -246.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 647701  on 11  degrees of freedom
```

```
## Residual deviance:   1382  on  6  degrees of freedom
## AIC: 1516.2
##
## Number of Fisher Scoring iterations: 4
```

All parameters are significantly different from zero. Here we can better interpret the vaccination effect because it is estimated in dependency on the event type. Generally, being partially vaccinated is linked to fewer events compared to the baseline (not vaccinated), because the parameter vaccine_status2 has a negative sign. The estimated parameter for fully vaccinated people is also negative, but smaller. However, this does not mean that being partially vaccinated is advantageous. Instead, it is probably due to the fact, that people were only partially vaccinated for a very short time, where there was less chance to be infected.

```
anova(poiss_glm)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev
## NULL                              11     647701
## vaccine_status  2   102469         9     545233
## event_type      3   543851         6       1382
```

Multinomial Model

Another option for this data with a categorical would be a multinomial model. Because the model interpretation is difficult and it was not covered in class, it is not further discussed.

```
library(nnet)
multinom_model  = multinom(event~vaccinated,data=df)
```

```
## # weights:  16 (9 variable)
## initial  value 23648.795506
## iter  10 value 20325.467825
## final  value 20285.735343
## converged
```

```
summary(multinom_model)
```

```
## Call:
## multinom(formula = event ~ vaccinated, data = df)
##
## Coefficients:
##                 (Intercept) vaccinatedNo iniciada vaccinatedParcial
## Crítics          -1.5905962           0.85955748        -0.4719066
## Defunció         -1.5458369           0.08055976        -0.6105049
```

```
## Hospitalització  -0.5151402            0.32111808        -0.4002018
##
## Std. Errors:
##                 (Intercept) vaccinatedNo iniciada vaccinatedParcial
## Crítics          0.04603660            0.05565248        0.07953033
## Defunció         0.04519186            0.06111938        0.08133331
## Hospitalització  0.03097434            0.04077959        0.05117251
##
## Residual Deviance: 40571.47
## AIC: 40589.47
```

Logistic Regression

In order to fit a simpler, easier to interpret model, we create a binary variable "having a severe infection". Some people might argue, that being infected with the coronavirus is itself not a bad thing. What we are worried about, is having a severe disease, which is consideradly worse than the common flu. Therefore we define a severe case as anything worse than being quarantined at home: Being hospitalized, on the ICU, or dying.

```
# define critical case as anything worse than being positive and quarantined ("Cas")
df$severe = as.numeric(! df$event=="Cas")
# now we would need to multiply the rows by count
df_enrolled = df[rep(row.names(df),df$count),]
```

In order to for a logistic Regression, we need to expand the dataset, so that each event constitutes its own row. We realize that by copying each row as often as it's count variable indicates via rep. The sanity check confirms that the resulting dataframe has as many rows as the sum of table_age in the old one.

```
print("sanity check: ")
```

```
## [1] "sanity check: "
```

```
      dim(df_enrolled)[1] == sum(df$count) # sanity check
```

```
## [1] TRUE
```

We want to model the probability of a person of varying age, vaccination status and sex to have a non-severe or a severe case of Covid19. We are using a logistic model for this problem, because our dependent variable is binary, and we also have multiple explanatory variables. We also consider the explanatory variables as independent. For age and sex this is obviously true, but it might not be true for vaccination status and age, since older people have been vaccinated first and because old and young people might get vaccinated to different extents.

```
logitReg = glm(severe ~ age + vaccinated + sex, data=df_enrolled, family=binomial(link='logit'))
summary(logitReg)
```

```
##
## Call:
## glm(formula = severe ~ age + vaccinated + sex, family = binomial(link = "logit"),
##     data = df_enrolled)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3018  -0.4767  -0.3305  -0.2504   2.9922
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.46513    0.02525 -176.81   <2e-16 ***
## age40 a 49             0.55041    0.02515   21.89   <2e-16 ***
## age50 a 59             1.23045    0.02465   49.92   <2e-16 ***
## age60 a 69             1.94236    0.02430   79.93   <2e-16 ***
## age70 a 79             2.53634    0.02532  100.17   <2e-16 ***
## age80 o més            3.16772    0.02616  121.07   <2e-16 ***
## vaccinatedNo iniciada  1.11729    0.01571   71.12   <2e-16 ***
## vaccinatedParcial      0.49110    0.02552   19.25   <2e-16 ***
## sexHome                0.46780    0.01315   35.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 191126  on 281127  degrees of freedom
## Residual deviance: 164713  on 281119  degrees of freedom
## AIC: 164731
##
## Number of Fisher Scoring iterations: 6
```

Again, all parameters are statistically significant having a p-value numerically equal to zero. We can interpret the parameters more easily when we take their exponent, which constitutes the change in the Odds Ratio:

Another meaningful link function for the binary regression is the probit link. This assumes a latent normal process.

```
probitReg = glm(severe ~ age + vaccinated + sex, data=df_enrolled, family=binomial(link='probit'))
summary(probitReg)
```

```
##
## Call:
## glm(formula = severe ~ age + vaccinated + sex, family = binomial(link = "probit"),
##     data = df_enrolled)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2505  -0.4832  -0.3408  -0.2330   3.1346
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.439663   0.012146 -200.87   <2e-16 ***
## age40 a 49             0.260994   0.011577   22.54   <2e-16 ***
## age50 a 59             0.599291   0.011808   50.75   <2e-16 ***
## age60 a 69             0.988519   0.011992   82.43   <2e-16 ***
## age70 a 79             1.329429   0.012957  102.60   <2e-16 ***
## age80 o més            1.692007   0.013548  124.89   <2e-16 ***
## vaccinatedNo iniciada  0.614790   0.008353   73.60   <2e-16 ***
```

```
## vaccinatedParcial      0.269894   0.013410   20.13   <2e-16 ***
## sexHome                0.239457   0.006912   34.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 191126  on 281127  degrees of freedom
## Residual deviance: 164497  on 281119  degrees of freedom
## AIC: 164515
##
## Number of Fisher Scoring iterations: 6
```

The last link function we try is the complementary log log link.

```
cllReg = glm(severe ~ age + vaccinated + sex, data=df_enrolled, family=binomial(link='cloglog'))
summary(cllReg)
```

```
##
## Call:
## glm(formula = severe ~ age + vaccinated + sex, family = binomial(link = "cloglog"),
##     data = df_enrolled)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3450  -0.4746  -0.3256  -0.2616   2.9406
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.31700    0.02357 -183.16   <2e-16 ***
## age40 a 49             0.52397    0.02446   21.42   <2e-16 ***
## age50 a 59             1.17156    0.02369   49.45   <2e-16 ***
## age60 a 69             1.81506    0.02296   79.07   <2e-16 ***
## age70 a 79             2.32566    0.02326   99.97   <2e-16 ***
## age80 o més            2.83725    0.02340  121.27   <2e-16 ***
## vaccinatedNo iniciada  0.96158    0.01378   69.80   <2e-16 ***
## vaccinatedParcial      0.42666    0.02299   18.56   <2e-16 ***
## sexHome                0.41778    0.01180   35.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 191126  on 281127  degrees of freedom
## Residual deviance: 164994  on 281119  degrees of freedom
## AIC: 165012
##
## Number of Fisher Scoring iterations: 6
```

```
# function to calculate fisher statistic
fisherStat = function(RegM){
  return(sum(residuals(RegM,type="deviance")^2)/RegM$df.residual)
}
fisherStat(logitReg)
```

```
## [1] 0.58592
```

```
fisherStat(probitReg)
```

```
## [1] 0.5851513
```

```
fisherStat(cllReg)
```

```
## [1] 0.5869195
```

All three Fisher Statistics are very similar between 0.585-0.586 indicating Underdispersion.

Comparing the Models In order to compare the three binary regressions and to check which link function results in the best fit, we compare the log-Likelihood and the Akaike Information Coefficient. The Poisson Model cannot be compared with binary regressions because the predicted y is of different nature.

```
round(c(logitReg$aic,probitReg$aic,cllReg$aic),5)
```

```
## [1] 164731.2 164515.2 165012.2
```

With AIC smaller values are considered to show a better model fit.

```
logLik(logitReg)
```

```
## 'log Lik.' -82356.62 (df=9)
```

```
logLik(probitReg)
```

```
## 'log Lik.' -82248.58 (df=9)
```

```
logLik(cllReg)
```

```
## 'log Lik.' -82497.11 (df=9)
```

With the log-Likelihood bigger values indicate better model fit. So for both metrics the probit-Regression model has the best value.

Model Interpretation However, the values of the logistic Regression are easier to interpret because the estimated parameters can be seen as the log(Odds). Since the models are of similar quality, let's interpret these parameters. Taking the exponential of the parameter, gives the Odds Ratio relative to the baseline group.

```
round(exp(logitReg$coef),2)
```

```
##           (Intercept)            age40 a 49            age50 a 59
##                  0.01                  1.73                  3.42
##            age60 a 69            age70 a 79          age80 o més
##                  6.98                 12.63                 23.75
## vaccinatedNo iniciada     vaccinatedParcial               sexHome
##                  3.06                  1.63                  1.60
```

All age categories have a Odds Ratios (OR) bigger than 1. This means people of all ages have a higher risk to get a severe COVID19 infection than the baseline defined to be individuals between 30 and 40. Also the OR increases with age, which means that the Odds increase the older an individual is. The OR for persons being 80 or more is estimated as 23.75, which is the biggest value in this model. This indicates, that age played a bigger role than vaccine status (in 2021 in Catalonia). The Odds for male individuals are estimated to be 60% to suffer from a severe COVID19 disease. For vaccine status, the fully vaccinated people are taken as baseline group. Non-vaccinated people of the same age group and sex are estimated to have approximately 3 times the Odds of having a severe Coronavirus infection. Also being partially vaccinated increases the odds by around 63% given that other covariates remain the same.

Conclusion

This work studying vaccine effectiveness shows the issues of non-randomized data. Most of the issues, we had to wrap our heads around would naturally dissolve with a randomized control study, which is why it is the gold standard for pharmaceutical studies.

One issue present is that older people were vaccinated first, but still were at high risk for severe Covid infections. Therefore the age necessarily has to be included in the models. Another issue is the baseline fallacy, that at some point more people were vaccinated than not, which automatically shifts the number of events towards vaccinated individuals.

An additional issue is that we did not study interactions between age and vaccination status. This was because of the increased complexity in the model and the time constraints of the project. Though, it is likely that more old people are vaccinated than young, because they suffer a greater risk in case of covid and in general gets vaccinated first. An interaction variable between age and vaccination status could therefore have made our models more accurate. This is especially motivated by more deaths than expected under the null-hypothesis for the population 80 years old and above, as well as for the fully vaccinated group (remember the observations of the contingency tables). It is possible that this decreases the effectiveness vaccination in our model, and that in reality vaccination has a stronger effect.

Another observation was that all p-values displayed were strikingly small. This is because both the effect sizes and the sample size of this data set were very large, which naturally influence the statistical inference.

One conclusion, which we were not aware of before, is that male persons suffered of more severe infections. This might be due to statistical link between males and less healthy lifestyles like smoking resulting in cardiovascular diseases, which are a negative predictor for survival of COVID19.

Regarding the main analysis, we can confirm what others have found before:

- Age plays a major role for severe COVID19 diseases. Older people are at much higher risk. In this data the Odds for a severe infection were estimated to be 23 times higher for individuals older than 80 than those between 30 and 40.
- Being (fully) vaccinated clearly lowers the risk for a severe Coronavirus infection. Not being vaccinated is estimated to result in a three times increased risk for a severe COVID19 disease progression.