Survival Analysis Methods Applied to Establishing Covid-19 Vaccine Life

Jorge Luis Romeu, Ph.D.
https://www.researchgate.net/profile/Jorge_Romeu
http://web.cortland.edu/romeu/
Email: romeu@cortland.edu

1.0 Introduction

We apply *Survival analysis methodology* to *establishing the length of effectiveness of Covid-19 vaccines*. It is *assumed* that such *vaccine has already been developed and* our *reader is familiar* with our article *Some Statistical Methods to Accelerate Covid-19 Vaccine Testing*[1]. This work is part of *our pro-bono collaboration to the American struggle against Covid-19*, whereby retired professionals contribute based on their work experience. Read such *Proposal for Covid-19* in https://www.researchgate.net/publication/341282217_A_Proposal_for_Fighting_Covid-19_and_its_Economic_Fallout

Our previous work includes developing statistical methods to help accelerate vaccine testing: https://www.researchgate.net/publication/344193195_Some_Statistical_Methods_to_Accelerate_Covid-19_Vaccine_Testing and a Markov model to study problems of reopening college: https://www.researchgate.net/publication/343825461_A_Markov_Model_to_Study_College_Re-opening_Under_Covid-19 and another Markov model on the effects of Herd Immunization: https://www.researchgate.net/publication/343345908_A_Markov_Model_to_Study_Covid-19_Herd_Immunization?channel=doi&linkId=5f244905458515b729f78487&showFulltext=true and a discussion of socio-economic and racial issues affected by the Covid-19 Pandemic: https://www.researchgate.net/publication/343700072_A_Digression_About_Race_Ethnicity_Class_and_Covid-19 and developing *A Markov Chain Model for Covid-19 Survival Analysis:* https://www.researchgate.net/publication/343021113_A_Markov_Chain_Model_for_Covid-19_Survival_Analysis and *An Example of Survival Analysis Applied to analyzing Covid-19 Data*: https://www.researchgate.net/publication/342583500_An_Example_of_Survival_Analysis_Data_Applied_to_Covid-19, and *Multivariate Statistics in the Analysis of Covid-19 Data*, and *More on Applying Multivariate Statistics to Covid-19 Data*, both of which can also be found in: https://www.researchgate.net/publication/341385856_Multivariate_Stats_PC_Discrimination_in_the_Analysis_of_Covid-19, and the implementation of multivariate analyses methods such as: https://www.researchgate.net/publication/342154667_More_on_Applying_Principal_Components_Discrimination_Analysis_to_Covid-19 *Design of Experiments to the Assessment of Covid-19*: https://www.researchgate.net/publication/341532612_Example_of_a_DOE_Application_to_Coronavarius_Data_Analysis Offshoring: https://www.researchgate.net/publication/341685776_Off-Shoring_Taxpayers_and_the_Coronavarus_Pandemic and *reliability methods in ICU* assessment: https://www.researchgate.net/publication/342449617_Example_of_the_Design_and_Operation_of_an_ICU_using_Reliability_Principles and Quality Control methods for monitoring Covid-19: https://web.cortland.edu/matresearch/AplicatSPCtoCovid19MFE2020.pdf

[1] https://www.researchgate.net/publication/344193195_Some_Statistical_Methods_to_Accelerate_Covid-19_Vaccine_Testing in ResearchGate, or in https://web.cortland.edu/matresearch/DigressAccelVacTestCovid19.pdf

2.0 Problem Statement

Once new vaccine(s) have been established (i.e. they have been tested using Phases 1 through 3), found to be efficient (they immunize a large percent of the population), and safe (do no harm to most people), we need to explore for how long (and on which groups) such immunization works.

In the present paper we assume that a group of vaccinated individuals are followed for the most part of a calendar year (355 days). Some of these individuals may become *infected* with Covid-19 after a period of time. Others will remain *uninfected* until this experiment ends (*censored*).

All participants in the experiment provide *personal data*, to be used in establishing whether such characteristics actually modify (or not) the length of the Vaccine immunization period. Using these *covariates*, as well as the time in days to either (1) becoming infected, or (2) remaining as uninfected (censored), we implement survival, discrimination and regression analyses.

Survival analyses provides probabilities of time to becoming infected (event of interest), and its hazard rates, under such different covariates. This approach allows the establishment of *Vaccine Life Length*, in a manner *akin to* the establishment of *Warranty periods* in industry.

Discrimination analysis helps identify *which covariates impact Vaccine Life Length*, comparing (1) individuals that go through a long immunization period, without becoming infected, to (2) those other individuals who have become infected sometime after their Vaccination.

Finally, *regression analysis* helps assess those concomitant variables that do increase or decrease the Vaccine immunization Life Length (time of efficacy).

3.0 The Data

We have tried, unsuccessfully, to obtain Covid-19 patient and research data from organizations. Since we believe *it is important to show the use of survival analysis* techniques and to illustrate the power of their results, using appropriate data, *we have created a data set* to do so (Table #1).

We reused a data set, built from an example we had previously modified and used in our earlier papers. Such modifications included *adding several concomitant variables*. We created for each individual, using our judgment and experience, new data on *gender and socio-economic status*.

Table 1: Survival Analysis Data for Vaccine Life Length Analysis

Days	SocioEcon	Age	Comorb	Gender	Censor	LogDays
301	1	45	1	0	0	5.70711
337	0	50	1	1	0	5.820083
283	0	40	0	1	0	5.645447
283	0	43	0	0	0	5.645447
265	0	35	0	1	0	5.57973
274	0	38	0	0	0	5.613128

238	0	44	0	0	0	5.472271
355	1	47	1	1	0	5.872118
229	0	33	0	1	0	5.433722
310	0	48	1	0	0	5.736572
265	0	41	0	1	0	5.57973
346	1	49	1	0	0	5.846439
301	0	42	0	0	0	5.70711
229	0	45	0	1	0	5.433722
301	0	36	0	1	0	5.70711
301	0	39	0	0	0	5.70711
337	0	46	1	0	0	5.820083
229	0	32	0	1	0	5.433722
238	0	41	0	1	0	5.472271
301	1	44	1	1	0	5.70711
292	0	42	1	1	0	5.676754
256	0	39	0	0	0	5.545177
301	0	51	1	0	0	5.70711
301	0	49	0	0	0	5.70711
274	0	42	0	0	0	5.613128
346	0	55	1	0	0	5.846439
247	0	30	0	1	0	5.509388
265	1	36	1	0	0	5.57973
265	0	38	1	0	0	5.57973
256	0	34	0	0	0	5.545177
274	1	39	1	1	0	5.613128
256	0	36	0	0	0	5.545177
301	0	45	1	0	0	5.70711
337	0	47	1	1	0	5.820083
283	0	42	0	1	0	5.645447
283	1	44	1	0	0	5.645447
247	0	36	0	1	0	5.509388
310	0	47	1	1	0	5.736572
292	0	41	0	1	0	5.676754
238	0	73	1	1	1	5.472271
238	1	58	2	0	1	5.472271
274	0	60	1	1	1	5.613128
265	0	52	0	0	1	5.57973
355	0	65	2	0	1	5.872118
229	0	72	1	0	1	5.433722
256	0	66	1	0	1	5.545177
292	0	61	1	1	1	5.676754
229	1	55	2	0	1	5.433722
229	0	63	2	0	1	5.433722

229	0	78	1	1	1	5.433722
265	0	73	2	1	1	5.57973
283	0	77	1	1	1	5.645447
256	0	79	1	0	1	5.545177
238	0	82	1	0	1	5.472271
265	0	73	1	0	1	5.57973
238	0	78	2	0	1	5.472271
265	0	74	1	0	1	5.57973
283	0	68	1	1	1	5.645447
337	0	66	1	1	1	5.820083
247	0	69	2	0	1	5.509388
292	0	77	0	1	1	5.676754
229	0	85	2	0	1	5.433722
319	0	55	1	0	1	5.765191
283	1	45	2	1	1	5.645447
265	0	49	2	0	1	5.57973
328	0	57	1	1	1	5.793014
301	0	51	1	0	1	5.70711
274	0	66	2	1	1	5.613128
256	0	69	2	1	1	5.545177
283	0	59	1	1	1	5.645447
301	1	55	2	1	1	5.70711
247	0	67	2	0	1	5.509388
319	0	59	1	1	1	5.765191
265	0	68	2	0	1	5.57973
256	0	72	1	1	1	5.545177
283	0	77	1	1	1	5.645447
274	0	73	1	1	1	5.613128
301	0	70	0	1	1	5.70711
256	0	79	1	0	1	5.545177
247	0	80	2	0	1	5.509388
247	0	82	2	1	1	5.509388
283	0	81	0	0	1	5.645447
256	0	84	1	1	1	5.545177
256	0	85	2	1	1	5.545177
283	0	72	1	1	1	5.645447
247	1	66	2	1	1	5.509388
265	0	69	2	1	1	5.57973
247	0	77	2	1	1	5.509388
265	0	79	0	0	1	5.57973
247	0	84	0	0	1	5.509388
355	**1**	**85**	**2**	**1**	**1**	**5.872118**
229	**0**	**30**	**0**	**0**	**0**	**5.433722**

Description of the concomitant data provided by each individual participant, is shown below:

SocioEconomic	0.High	1.Low
Gender:	0.Male	1.Female
Censoring:	0.Suspended;1.Not	
Co-morbidities:	Number	

Days correspond to the *length of time* each participant is *in the experiment*, which was terminated short of a year (355 days). The experiment can continue at a later date, when more data can be collected, and summarized. This may be repeated as many times as needed, and obtain estimates.

Socio economic corresponds to participant's background. It has been discussed how individuals of *lower socio-economic* levels are *more susceptible* to become infected with Covid-19. We want to test whether such covariate affects, in any way, Vaccine effectiveness in such cohorts.

The same considerations apply to a participant's *gender, age and Number of Co-morbidities*. These are all critical factors that we need to assess and test, with respect to Vaccine Life Length.

Censoring pertains to those participants that have remained uninfected, up to the end of this experiment, versus those who have become infected, during the experiment development.

The last two entries in Table #1, in red, correspond to *column Maximums and Minimums*.

4.0 Non Parametric Survival Analysis:

A *key objective* of this paper is to *demonstrate how survival analysis techniques* can contribute to *establish the Life Length of* the new *Covid-19 Vaccine*, defined as the period of time where it will cover an acceptable percent (e.g. 80%) of those vaccinated, against contracting this virus. This is in many ways *akin to establishing the warranty period* of a new industrial product.

In industry, we want the longest possible warranty period (to beat the competition), that will not bankrupt the producer (as the product will be fixed for free). In public health we want the longest possible vaccines immunization period (as to not re-vaccinate unnecessarily), without risking the health of the population (i.e. of not being properly immunized because the vaccine has expired).

4.1 Survival Analysis First Case: not considering any covariates

Assume we conduct a *study during 355 days on 90 vaccinated patients* and compute the Time to either (1) becoming infected with Covid-19, or (2) terminating the study, uninfected. At the time such study is suspended, *39 patients remain uninfected* -the other 51 are infected[2]. We *consider* the 39 *uninfected as Censored*, since we do not know if or when they would be infected, had the study continued beyond its 355 days -for the Vaccine will eventually lose its protective power.

[2] Participants may enter the study at different times. They are all censored at the time of termination of the study.

Initially, we will *not consider any of the concomitant variables collected*. Thence, the initial Vaccine Life Length will be of general character (i.e. for all individuals). Later on, we will.

We use the *Kaplan-Meier* Distribution Free survival procedure, as we *do not want to commit to any specific statistical distribution*. Analysis Results are given below:

Distribution (Survival) Analysis, in Days (All data; No Comorbidities):

Variable: Days (to become infected, or to leave the experiment)

Censoring Information:	*Count*
Uncensored value	51
Right censored value	39

Censoring value: Censor = 0

Nonparametric Estimates **(Event of Interest: Time to Infection)**

Mean(MTTF)	Standard Error	95.0% Normal CI Lower	Upper
295.713	5.18332	285.554	305.872

Median=283; IQR=81; Q1=256; Q3=337; (**days to censoring or infection)**

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	Upper
229	90	5	0.944444	0.0241452	0.897121	0.991768
238	82	4	0.898374	0.0321285	0.835403	0.961345
247	**76**	**7**	**0.815629**	**0.0416997**	**0.733899**	**0.897359**
256	67	7	0.730414	0.0482026	0.635939	0.824889
265	57	8	0.627900	0.0533509	0.523334	0.732466
274	**45**	**3**	**0.586040**	**0.0549965**	**0.478249**	**0.693831**
283	39	7	0.480853	0.0577339	0.367697	0.594009
292	28	2	0.446507	0.0584957	0.331857	0.561156
301	**24**	**3**	**0.390693**	**0.0594000**	**0.274271**	**0.507115**
319	11	2	0.319658	0.0665299	0.189262	0.450054
328	9	1	0.284141	0.0679602	0.150941	0.417340
337	8	1	0.248623	0.0681170	0.115116	0.382130
355	2	1	0.124311	0.0942690	0.000000	0.309075

An example of interpretation of the above Reliability/Survivability table is as follows: At the end of 229 days, there were 90 individuals *at risk of infection* of which five did become *infected*. The *Probability of Surviving uninfected* is 0.94 or 94% (of infection is: 1-0.94=0.06 or 6%). A *95% CI* (confidence interval) for such *surviving* probability is: 89.7% to 99.2%. There were 90 – 5 = 85 individuals *surviving (uninfected).* But only *82 were at risk at the end of Day 238, because three individuals were Censored (stopped being monitored when the experiment was suspended)*.

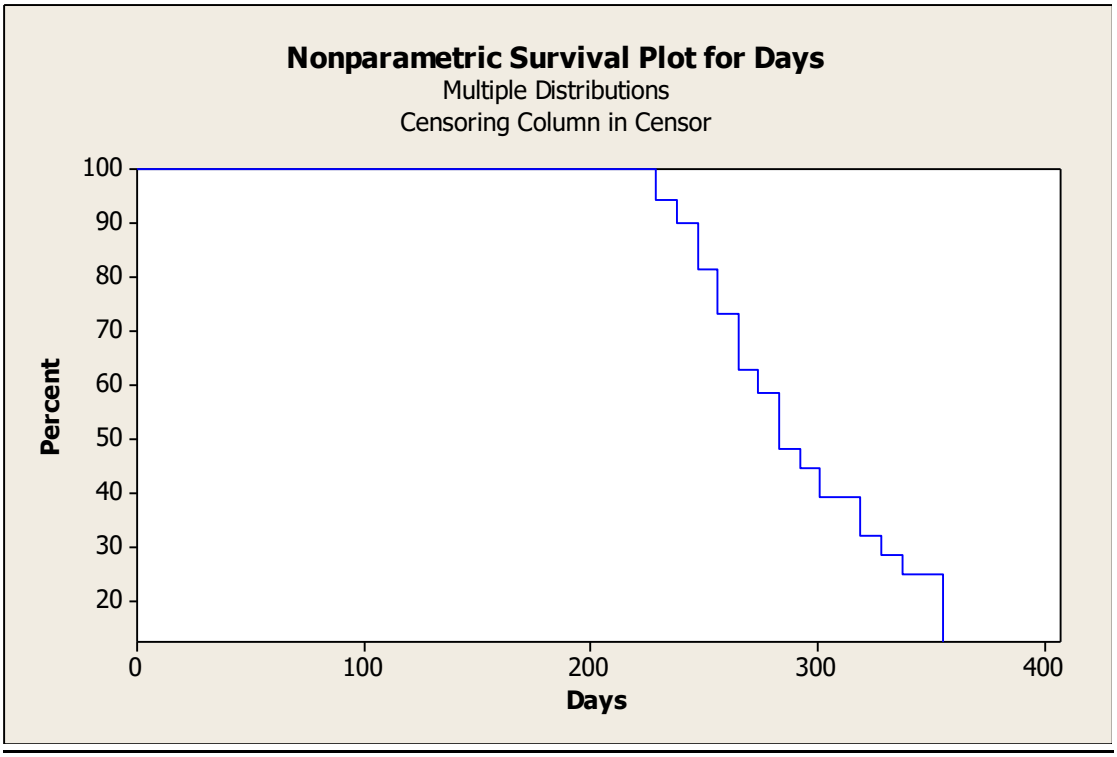
The *objective* of this research is *to establish an efficient Life Length for the Vaccine*. We will **use as thresh-hold 247 days** which yields a *Survival Probability of 0.816 (close to 80%)*, with 95% Confidence Interval (CI) lower (pessimistic) bound 0.734, and upper (optimistic) bound 0.897.

In practical terms this means that, *for 247 days, 81.6% of those vaccinated will be appropriately protected* from becoming infected with Covid-19. And this percentage would be as low as 73.4% or as high as 89.7%, both of which are well *over the minimum emergency value of 50%.* For, it is stated that "*in emergencies, regulators can authorize a vaccine's use based on interim analysis if it meets a minimum standard (in this case protection of half those who are vaccinated.*"[3]

However, if 50% population minimum coverage were preferred, *Life Length* could be *increased to 274 days*. Since, in this case the 95% CI lower bound would be 47.8%, meeting requirements.

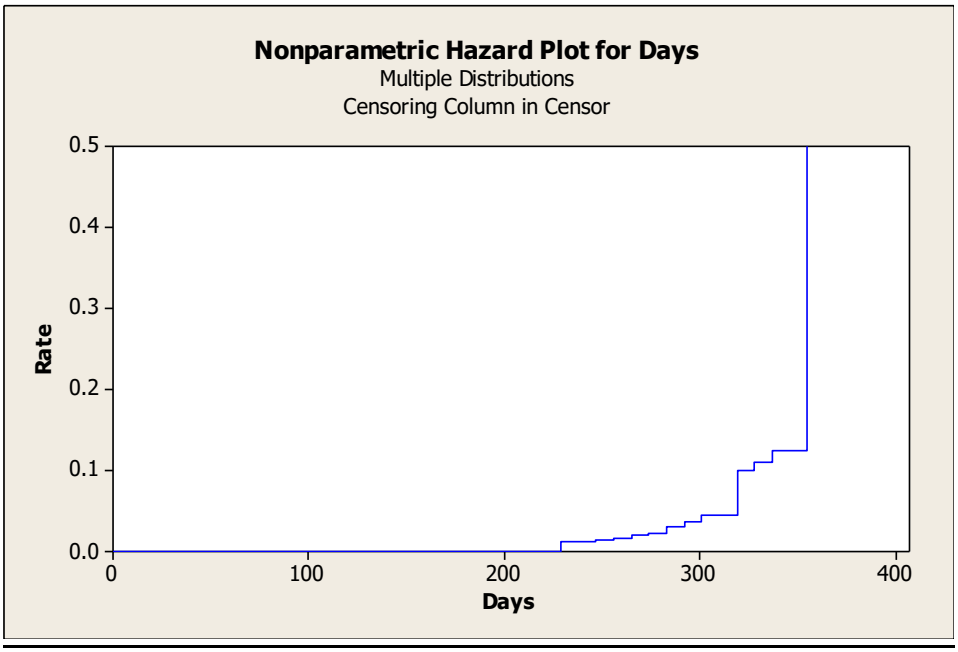
Survival results can also be *used for testing hypotheses about Life Length*. Assume Vaccine Life is *preset at 300 days*. The *95% CI at 301 days* (0.274, 0.507) *covers 0.5*. If *50%* is an acceptable coverage for the vaccine, *we can accept 300 days as Life Length*. *However if* acceptable Vaccine *performance implies covering 80%* of vaccinated individuals, *data would reject the hypothesis* that *300 days* is such *Vaccine Life*, as *0.8 is not within the 90% CI* for Survival, at 300 days.

The *survival plot for general survival probability* is shown below. We also present the graph of the corresponding general Failure or Hazard Rate (instantaneous failure probability).

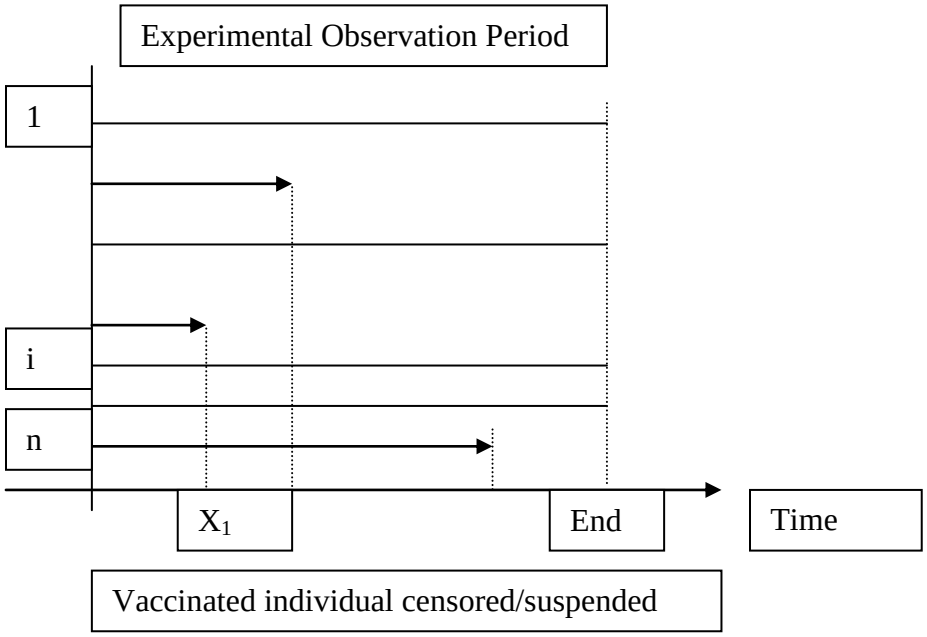


For example, a *vaccine survival (protection) probability* (for Covid-19 infection) for 300 days is about 50%. In practice, this means that after 300 days, only 50% of all the vaccinated individuals will be protected. Therefore, *300 days*, barely the minimum *50% thresh-hold*, could still be used.

[3] *Briefing the Covid-19 Pandemic*. The Economist, September 26th, 2020. Page 24.



Notice, in the general Hazard Plot above, how the instantaneous failure rate increases gradually, up to somewhat over 300 days, when there is an abrupt increase in the instantaneous failure rate.

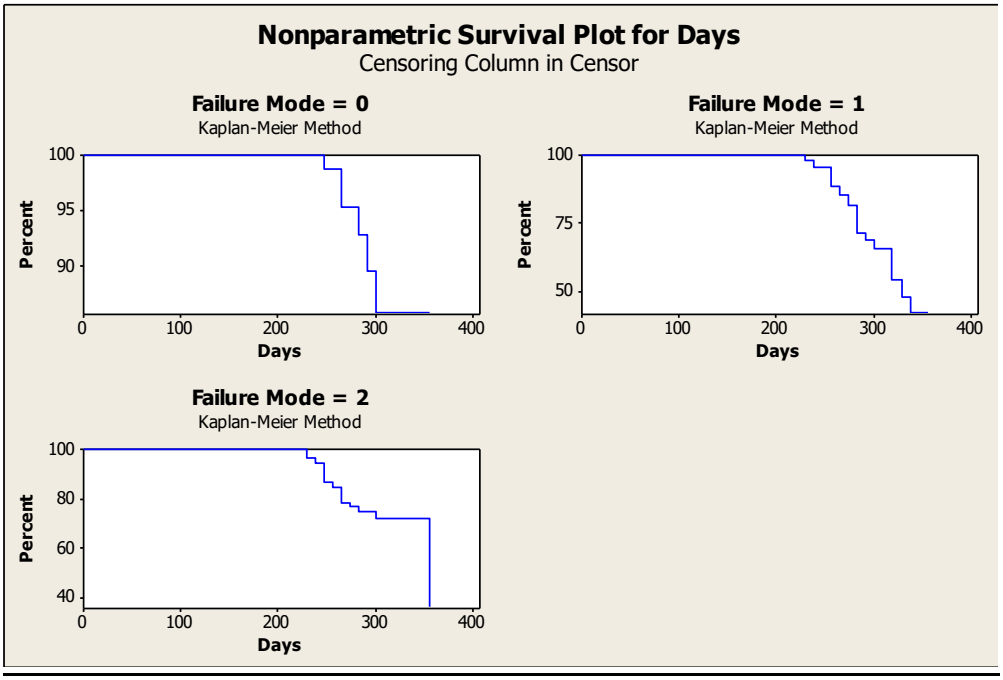


The graph above illustrates this experiment. We placed "n" vaccinated individuals under our observation. Some individuals "fail" (e.g. become infected) while others remain uninfected (i.e. they are "censored"), at the time this experiment is *suspended* (in our case, at 355 days).

4.2 Survival Analysis Second Case: considering all covariates as Failure Modes

Now assume *we conduct the same study described above*. But this time we will also consider the impact (if any) of the covariates collected. We will assess one covariate set at a time. We again apply the *Kaplan-Meier* Distribution Free survival method, for the same reasons given before: we *do not want to commit to any specific distribution*[4]. Analysis Results are given next.

Analysis for Failure Mode: Co-morbidities = 0, 1, 2 (No. of co-morbidities)



The above plots show very different *survival probabilities for Vaccinated individuals having, respectively, none, one or two Co-morbidities*. For example, the probability of an individual with *No Co-morbidities*, for 300 days without infection is 85%. If the individual has *one Co-morbidity* this probability is 70%. With *two Co-morbidities*, this probability is about 65%. The *Vaccine Life Length* against Covid-19 infection *decreases, as* the number of *Co-morbidities increases*.

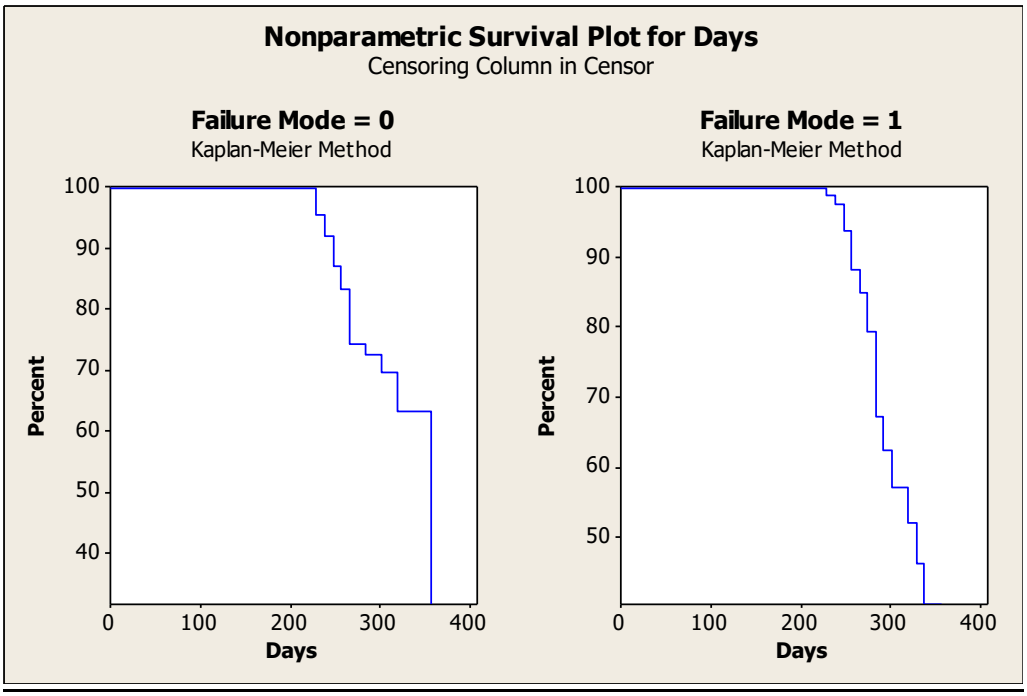
Summarizing: the comparison of survival plots show *an effect of Comorbidities on Life Length* of Covid-19 Vaccines. *Having more than one type of Vaccines may provide alternatives* that deal better with some situations, than others. This effect will be further explored in the next section.

Analysis of Gender (as Failure Mode): 0-Males; 1-Females

We now consider whether Gender (*0 for Males; 1 for Females*), which had an effect in the number of Covid-19 initial deaths, will also have an effect in the Life Length of the Vaccine

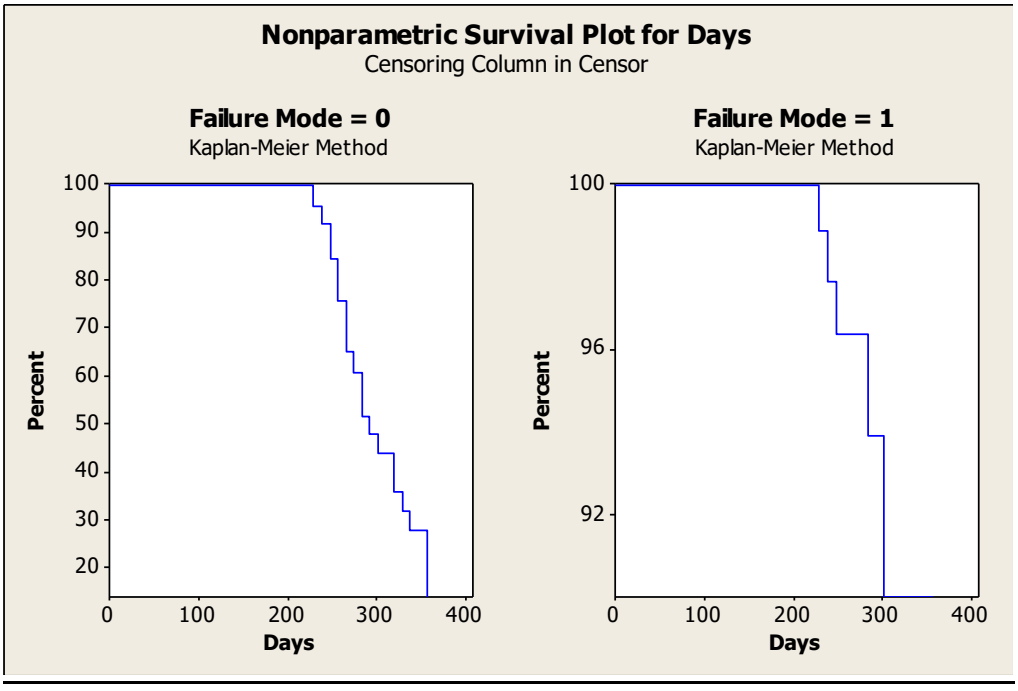
We can observe, from the survival plots below, how there seems to be minor gender effect in the Life Length of the Vaccine. This effect will be explored quantitatively in the next sections, using both, Discriminant Analysis and regression methods.

[4] There are ways to investigate and test for specific distributions, if desired. See related papers, in the Bibliography.



Analysis of Socio-Economic status (as Failure Mode):

Data on each individual was *collected*, regarding their socio-economic status. A Lower socio-economic status (i.e. Unit) corresponds to the working poor, the unemployed, etc. who tend to live in smaller, more crowded dwellings and neighborhoods, who work in *front-line* occupations (public transportation, etc,), lack effective health insurance and tend to have, thence, a higher number of Co-morbidities. All other socio-economic levels are denotes as Zero.



We observe in the survival plots above some socio-economic effects in the Life Length of the Vaccine, which had also been observed in the number of Covid-19 initial deaths[5]. Such effects will be explored quantitatively in the next sections, using Discriminant and Regression methods.

5.0 Discriminant Analysis

In this section we use Discriminant Analysis to *assess whether factors identified* in survival plots (Gender, Age, SocioEc, Co-morbidities) *have a significant impact in Vaccine Life Length*. We *first* develop a Discriminant Function using *Fisher's Regression* approach. We *then* develop, using the same data, a Discriminant approach using *Minitab procedure*. Then, we compare both.

To derive said Discriminant Functions we *define two groups*: *un-infected* (i.e. *Censored*), *and infected* participants. We select, among the Un-infected participants, only those who attained the longest times in the experiment (i.e. greater than 280 days) without becoming infected.

Fisher Regression Discrimination Analysis: DscGrps v. Gender, Age, Comorb

The regression equation is:

DscGrps_2 = - 2.84 + 0.001 SocioEcon_1 + 0.0462 Age_3 + 0.298 Comorb_3 + 0.088 Gender_2

Predictor	Coef	SE Coef	T	P	
Constant	-2.8405	0.3216	-8.83	0.000	
SocioEcon_1	0.0007	0.2133	0.00	0.998	NON SIGNIFICANT
Age_3	0.046227	0.005274	8.76	0.000	SIGNIFICANT
Comorb_3	0.2982	0.1065	2.80	0.007	SIGNIFICANT
Gender_2	0.0883	0.1297	0.68	0.498	NON SIGNIFICANT

S = 0.553807 R-Sq = 66.1% R-Sq(adj) = 64.1%

We see, from the regression equation above, how *factors Age and Co-morbidities* are *statistically significant*, while factors *Socio-economic and Gender are not* (i.e. these latter factors do not help discriminate/separate the two groups, because they have small impact on Vaccine Life Length).

Unusual Observations

Obs	SocioEcon_1	DscGrps_2	Fit	SE Fit	Residual	St Resid
26	0.00	1.0000	-0.4367	0.1388	1.4367	2.68R
47	1.00	1.0000	-0.0748	0.1999	1.0748	2.08R
50	0.00	1.0000	-0.1847	0.1159	1.1847	2.19R

Data Display

Row	Days_2	SocioEcon_1	Age_3	Comorb_3	Gender_2	DscGrps_2
26	265	0	52	0	0	1
47	283	1	45	2	1	1
50	301	0	51	1	0	1

Notice also how the three observations identified as unusual (Nos. 26, 47 and 50) correspond to infected, middle aged participants, with existing co-morbidities, or low infection times (Days).

[5] https://www.researchgate.net/publication/343700072_A_Digression_About_Race_Ethnicity_Class_and_Covid-19

II) Discriminant (Regression) Analysis: DscGrps_2 versus Age_3, Comorb_3

We repeat above Discriminant procedure with the two significant factors (Age, Co-morbidities):

The regression equation is:

DscGrps_2 = - 2.80 + 0.0463 Age_3 + 0.299 Comorb_3

Predictor	Coef	SE Coef	T	P	
Constant	-2.8000	0.2880	-9.72	0.000	
Age_3	**0.046281**	**0.004742**	**9.76**	**0.000**	**SIGNIFICANT**
Comorb_3	**0.29913**	**0.09808**	**3.05**	**0.003**	**SIGNIFICANT**

S = 0.547697 R-Sq = 65.8% R-Sq(adj) = 64.9%

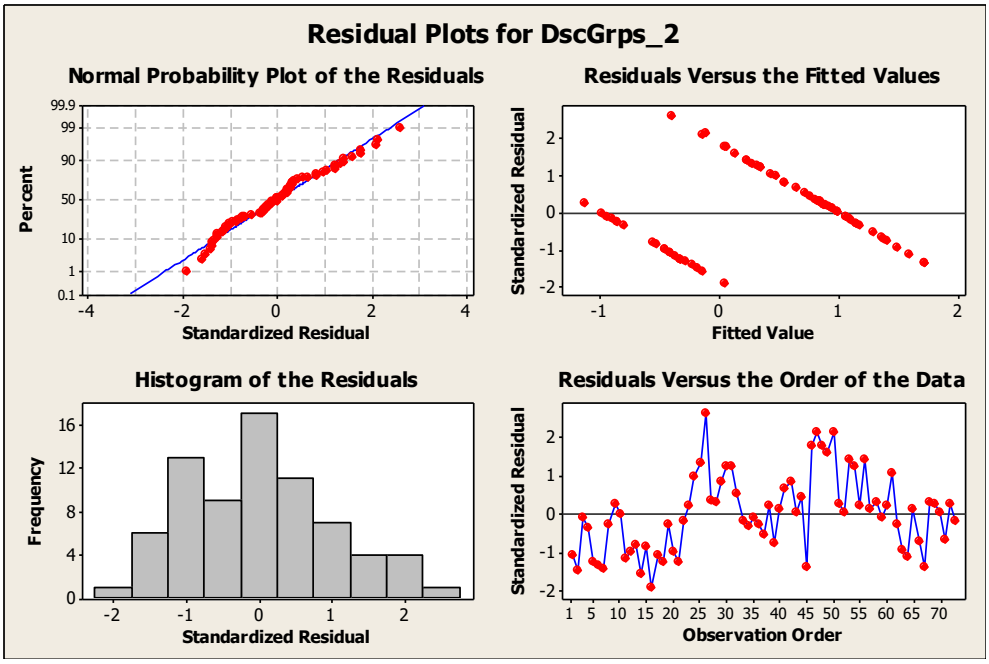
Analysis of Variance

Source	DF	SS	MS	F	P	
Regression	**2**	**40.481**	**20.241**	**67.48**	**0.000**	**SIGNIFICANT**
Residual Error	70	20.998	0.300			
Total	72	61.479				

Unusual Observations

Obs	Age_3	DscGrps_2	Fit	SE Fit	Residual	St Resid
26	52.0	1.0000	-0.3934	0.1218	1.3934	2.61R
47	45.0	1.0000	-0.1191	0.1510	1.1191	2.13R
50	51.0	1.0000	-0.1405	0.0814	1.1405	2.11R

Again, the same three observations identified as unusual (Nos. 26, 47 and 50) correspond to infected, middle aged participants, with existing co-morbidities, or with low infection times.



Minitab SW Discriminant Analysis: DscGrps_2 versus Age_3, Comorb_3

We repeat the Discriminant Analysis, but now using the Minitab Software procedure:

Linear Method for Response: DscGrps_2

Predictors: Age_3, Comorb_3

Group	-1	1
Count	22	51

Summary of classification

	True Group	
Put into Group	-1	1
-1	22	5
1	0	46
Total N	22	51
N correct	22	46
Proportion	**1.000**	**0.902**

N = 73 N Correct = 68 **Proportion Correct = 0.932**

Notice how the *Discrimination* procedure *misclassifies five entries*: from the Un-infected into the Infected group. All participants from Infected group were correctly classified. *Misclassification Probability is: 1 – 0.932 = 0.068 or 6.8%.*

Squared (Mahalanobis) Distance Between the Two Groups (-1, 1):

	-1	1
-1	0.00000	8.90565
1	8.90565	0.00000

Linear Discriminant Function for Groups

	-1	1
Constant	-13.683	-33.535
Age_3	0.575	0.888
Comorb_3	2.249	4.272

Summary of the Five Misclassified Observations

Observation	True Group	Pred Group	Group	Squared Distance	Probability
26**	1	-1	-1	1.526	0.973
			1	8.681	0.027
46**	1	-1	-1	1.672	0.649
			1	2.904	0.351
47**	1	-1	-1	4.793	0.848
			1	8.239	0.152
48**	1	-1	-1	5.132	0.616
			1	6.074	0.384
50**	1	-1	-1	0.8377	0.866
			1	4.5734	0.134

The *participant characteristics* for the five misclassified observations (which include the three misclassified observations from the Fisher Regression approach) *are given below*:

Data Display

Row	Days_2	SocioEcon_1	Age_3	Comorb_3	Gender_2	DscGrps_2
26	**265**	**0**	**52**	**0**	**0**	**1**
46	319	0	**55**	**1**	0	1
47	**283**	**1**	**45**	**2**	**1**	**1**
48	**265**	0	**49**	**2**	0	1
49	328	0	**57**	**1**	1	1
50	**301**	**0**	**51**	**1**	**0**	**1**

Notice how, again, *misclassified observations* (from Group1 to Group -1) correspond to infected, middle aged participants, with existing co-morbidities, or with low infection times.

Mahalanobis Distance for Fisher Regression Discrimination Function:

Fisher Discrimination Function explains **65%** of the problem and is able to correctly classify most cases in their respective group. The *Mahalanobis Distance* that *separates* these two *Groups of Infected (1), and Uninfected (-1), participants*, can be obtained in the following way:

For: n1= 51 (1); n2 = 22 (-1);
Lambda ^{2} = n1*n2/(n1+n2) = 51*22/73 = 1122/73 = 15.37
Dp^2 = [(n1+n2-2)/Lambda ^{2}]*[R^2 / (1- R^2)] = ((73-2)/15.4)*(0.65/0.35) = 8.58
Dp = Sqrt (Dp^2) = 2.93 => Prob (- ½ Dp) = Prob (-2.93/2) = 0.072

Comparison of results:

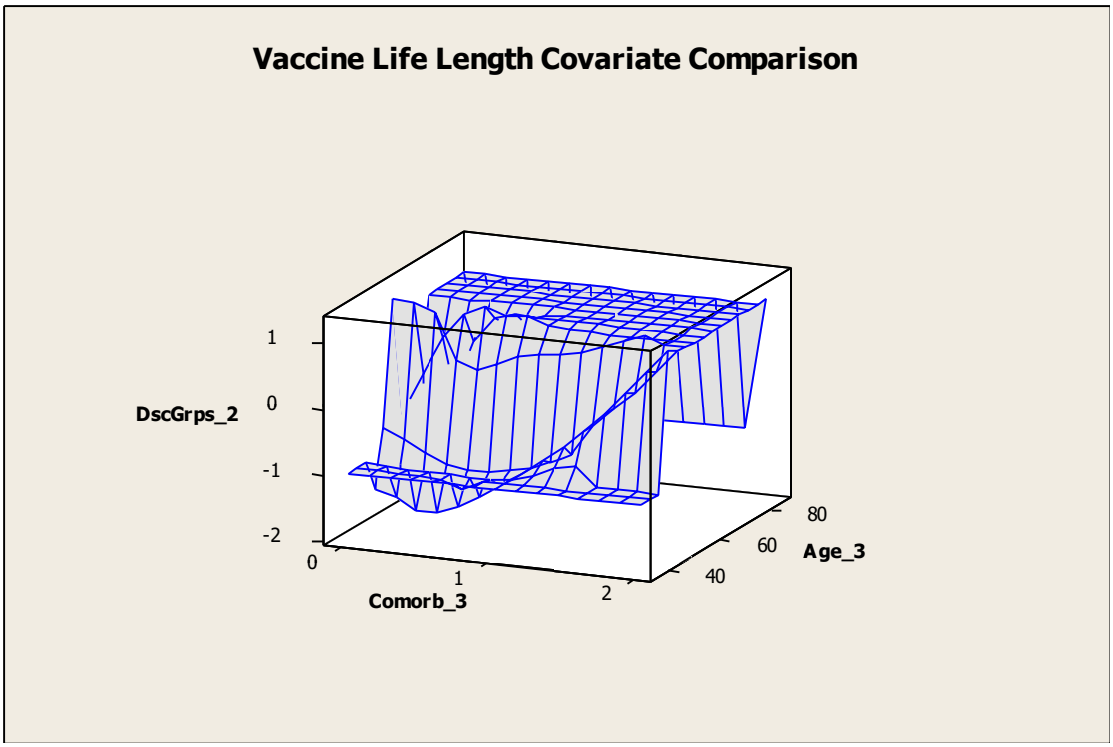
Disc. Function	Mahalanobis Dist.	Prob. Misclassific.	Factors
MINITAB	**8.90**	0.068	**Age, Co-morb**
REGRESSION	**8.58**	0.072	**Age, Co-morb**

Notice how Mahalanobis Distance and Probability of Miss-Classification, between two groups, established using *both Minitab and Fisher Discrimination* Functions, are *similar*. Thence:

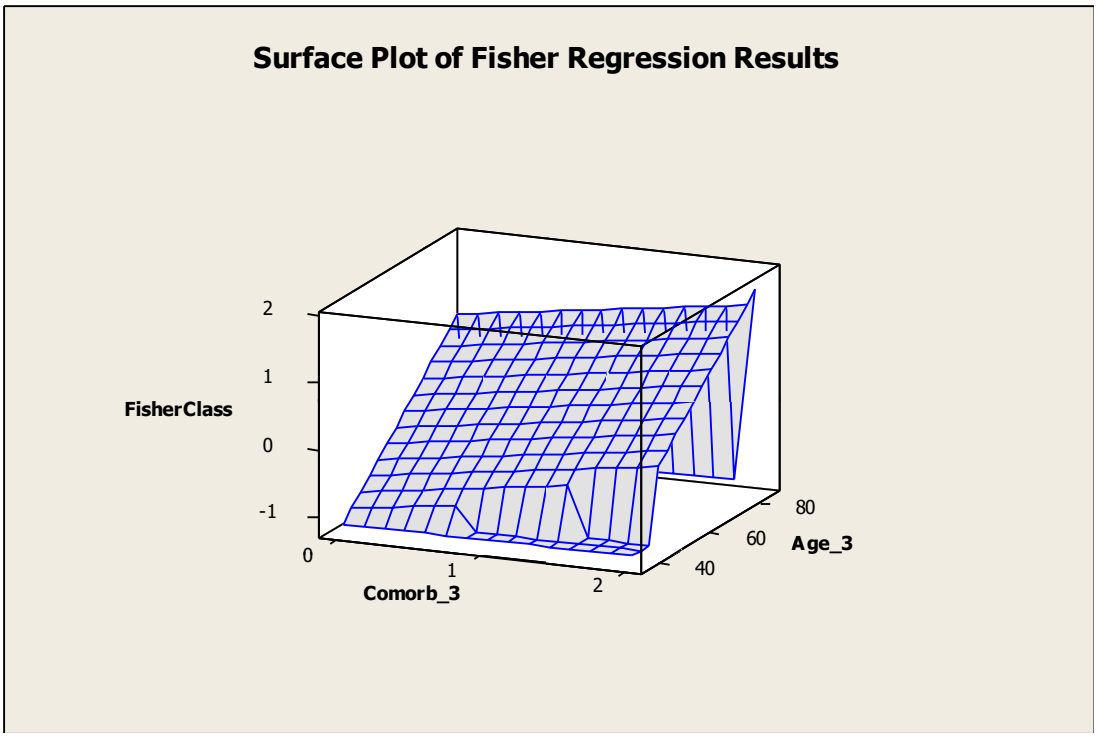
1. *Factors age and No. Co-morbidities* are good metrics to *differentiate the two groups*
2. *Both* Discrimination *Functions can be used* to establish an incoming Participant's group

If the analyst has access to a built-in Discrimination Function (such as the Minitab one, that we have used) it can be implemented directly on the data. However, if such built-in function is not available, the analyst can still use any of the many existing regression algorithms, and implement it as demonstrated in this paper. Both results are equivalent (when correctly done).

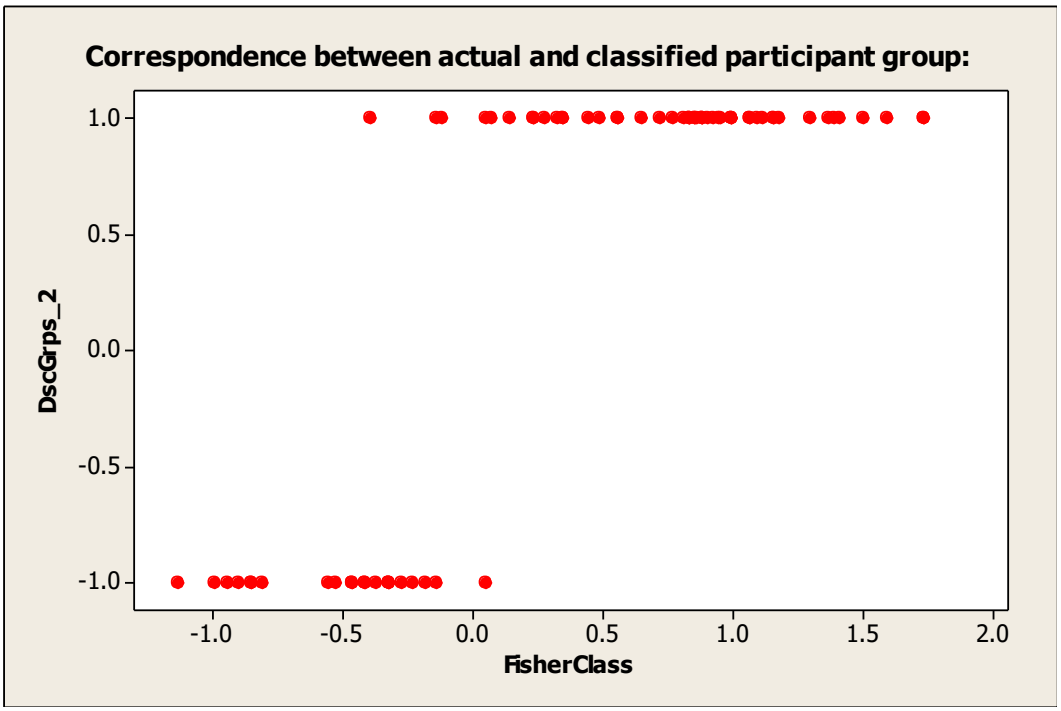
Below, we show *a surface plot of the two groups* Uninfected and Infected (-1, 1), separated using said Discrimination Functions and *factors Age and Number of Co-morbidities*. Notice how it is mostly smooth, except in the lower values of both factors, Age and Co-morbidities.



Alternatively, we show several Plots for Fisher Discrimination Function results. The surface plot below shows *Fisher's Discrimination function values by pa*rticipant, as a function of Age and the Number of Co-morbidities. Based upon Fisher's function and the two mentioned covariates, the *Values* corresponding to participants classified as *Uninfected are below the plane built at height Zero*; *values* corresponding to participants classified as *Infected*, *are above plane built at Zero*.



The Graph below shows how the *Fisher classification and actual participants group* correlate. *Values overlapping Zero*, in either group (-1 or +1), are *misclassification errors*.



6.0 Regression Analysis

We implement *regressions of Log(Days to Infection, or Censoring)* as functions of a participant *Age and Number of Co-morbidities*. We use the regression approach[6] whereby *Hazard function* $\Lambda(*)$ can be written as a *function of Time* to Failure t, *and* a set of *covariates Z*, here represented by *Patient Age and Number of Co-morbidities*. Vector β of the covariate coefficients of Z, will be estimated from the regression work. Per the theory, said variables are defined:

$$\mathbf{Z = (z_1,z_2) = (Age;\ No.\ Comorbidities);\ and\ \beta = (\beta_1, \beta_2)\ coefficients}$$

The *Hazard Rate function* is then: $\Lambda(t; Z) = \text{Exp}(\lambda_o + Z\beta)$, and the *Likelihood function* is:

$$\text{Ln}\{\Lambda(t; Z)\} = \text{Ln}\{\text{Exp}(\lambda_o)\text{Exp}(Z\beta)\} = \lambda_o + Z\beta = \lambda_o + z_1\beta_1 + z_2\beta_2$$

We use Ln(Time to Event[7]) as Response in the regression v. Age and No. Co-morbidities.

We implement the regression analysis using two Groups: Infected and Un-Infected

Regression Analysis is Reduced to 72 participants: uninfected, with longest censoring time:

LogDays v. SocioEconomic, No. Co-morbidities, Age, Gender

[6] For more on the Proportional Hazards model see Kalbfleisch and Prentice book, in the Bibliography section.
[7] Events of interest are, as before, time to Infection or to Censoring (by termination of the experiment w/o infection)

We will first consider all four factors. The regression equation is:

LogDays_2 = 5.93 - 0.0175 SocioEcon_1 - 0.00456 Age_3 - 0.0282 Comorb_3 + 0.0374 Gender_2

Predictor	Coef	SE Coef	T	P	
Constant	5.92736	0.05391	109.94	0.000	
SocioEcon_1	**-0.01754**	**0.03576**	**-0.49**	**0.625**	**NON SIGNIFICANT**
Age_3	**-0.0045588**	**0.0008843**	**-5.16**	**0.000**	**SIGNIFICANT**
Comorb_3	**-0.02823**	**0.01786**	**-1.58**	**0.119**	**NON SIGNIFICANT**
Gender_2	**0.03743**	**0.02174**	**1.72**	**0.090**	**NON SIGNIFICANT**

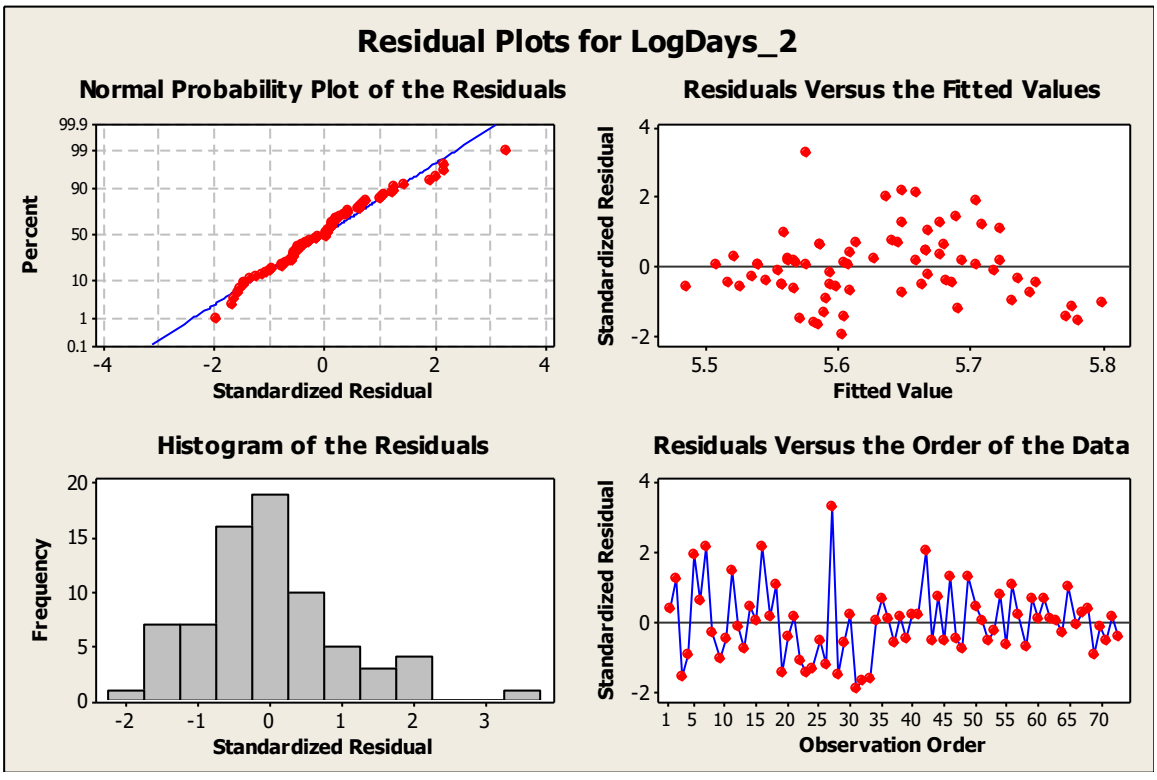
S = 0.0928495 R-Sq = 40.4% R-Sq(adj) = 36.9%

As expected, the *regression with all four covariates is weak*, as three of them are non-significant factors (only Age is). The overall regression is significant, because there is one significant factor.

Analysis of Variance

Source	DF	SS	MS	F	P	
Regression	**4**	**0.397241**	**0.099310**	**11.52**	**0.000**	**SIGNIFICANT**
Residual Error	68	0.586230	0.008621			
Total	72	0.983471				

Residual Plots for LogDays_2



We repeat the regression analysis below, using only the two factors (Age and Number of Co-morbidities) that were *significant in the Discriminant Function.*

Regression Analysis (Again reduced, as above, to 72 participants):

LogDays versus Age, Number of Co-morbidities

The regression equation is:

LogDays_2 = 5.93 - 0.00435 Age_3 - 0.0311 Comorb_3

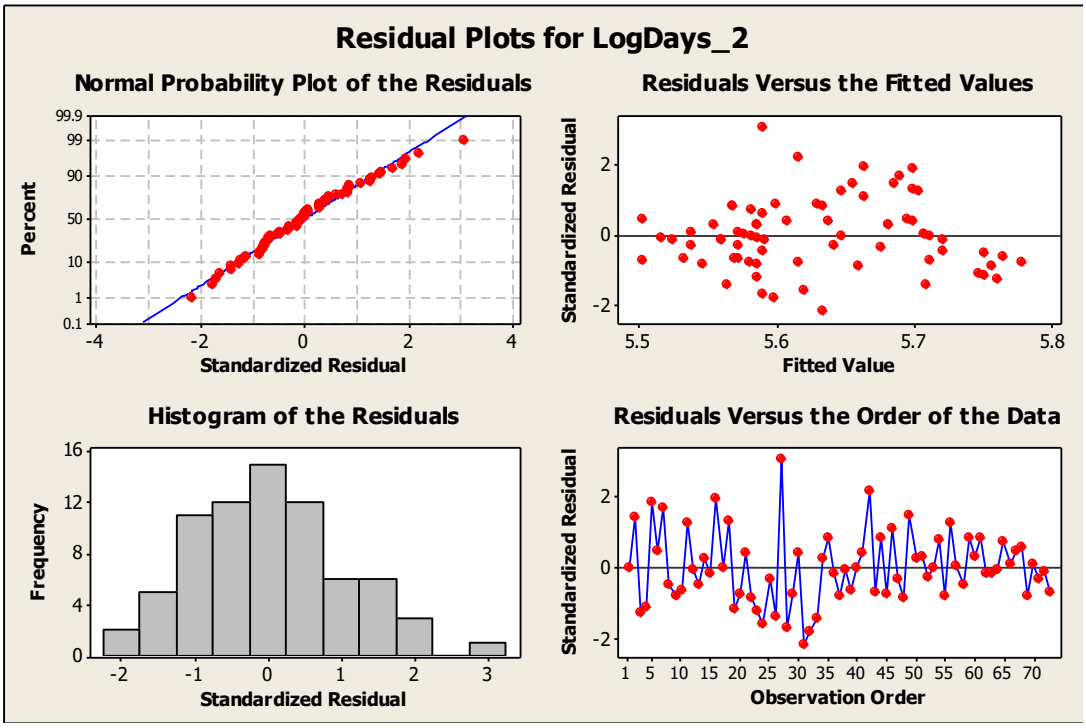
Predictor	Coef	SE Coef	T	P	
Constant	5.93428	0.04925	120.49	0.000	
Age_3	**-0.0043525**	**0.0008108**	**-5.37**	**0.000**	**SIGNIFICANT**
Comorb_3	**-0.03109**	**0.01677**	**-1.85**	**0.068**	**SIGNIFICANT**

S = 0.0936472 R-Sq = 37.6% R-Sq(adj) = 35.8%

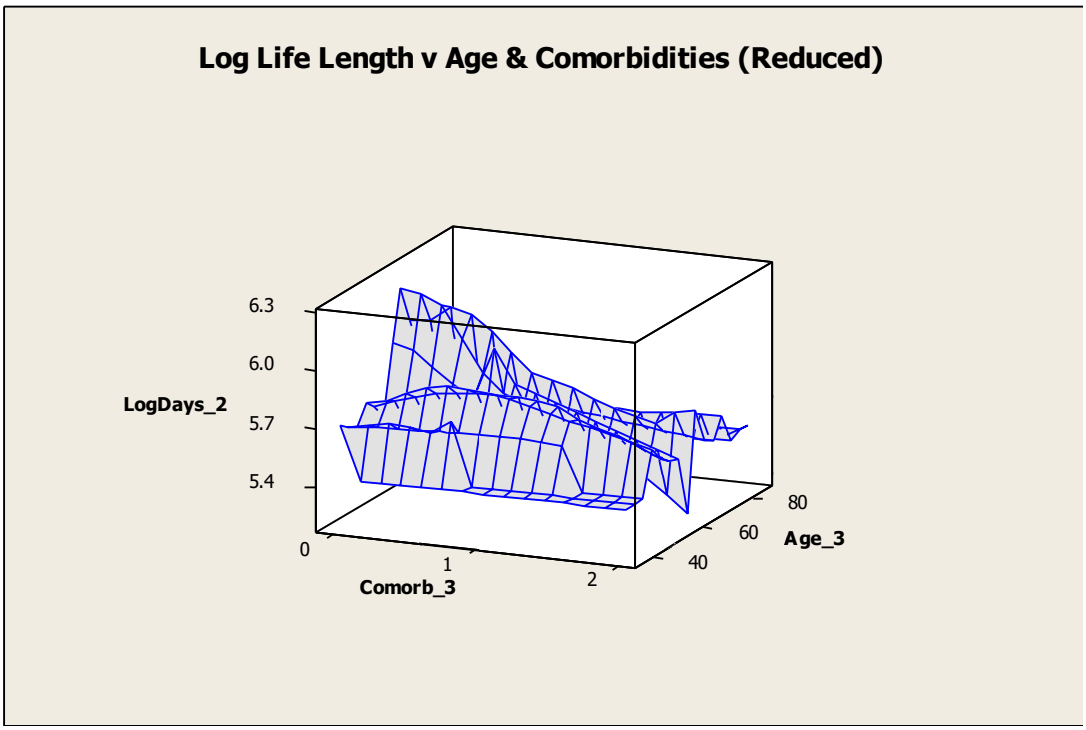
This *regression* is significant, and *explains only 36% of the problem*. B*oth factors considered (Age and Number of Co-morbidities) are statistically significant* and *negative.* This means that, *as these two factors increase, the effective Vaccine Life Length (response) decreases.*

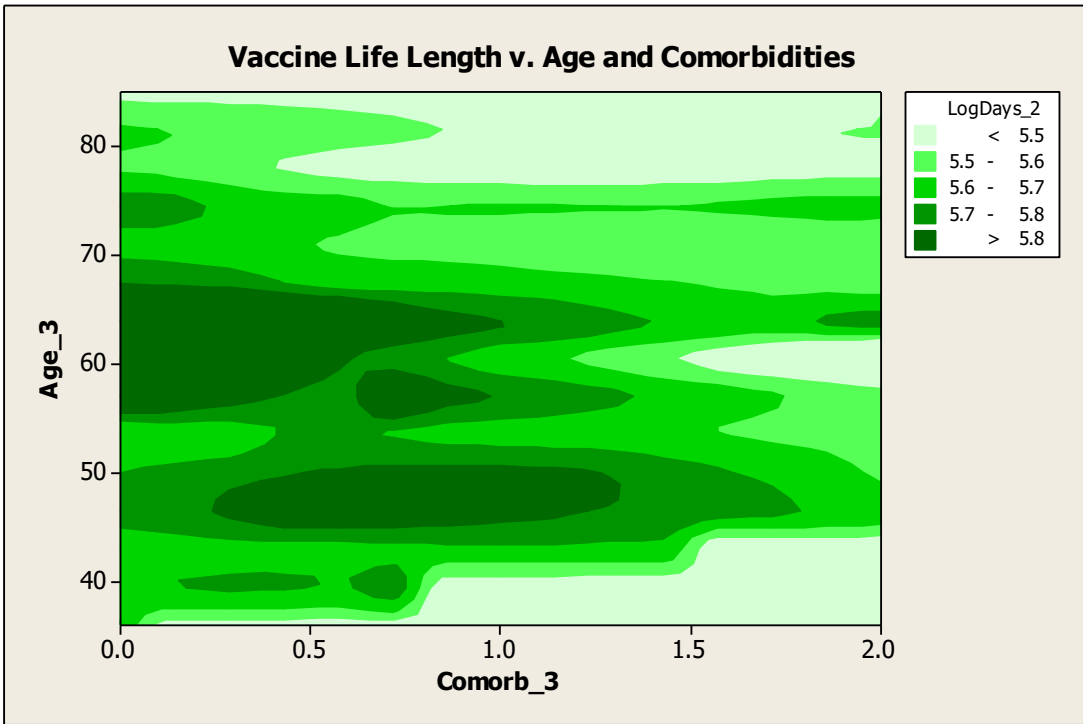
Analysis of Variance

Source	DF	SS	MS	F	P	
Regression	**2**	**0.36959**	**0.18479**	**21.07**	**0.000**	**SIGNIFICANT**
Residual Error	70	0.61389	0.00877			
Total	72	0.98347				



The residual plot shows how the regression, using Patient Age and Number of Co-morbidities v. LogDays, is acceptable under regression assumptions. We next present *response surface plots*.





Vaccine Life surface plot assesses the *regions* for Patient Age and Number of Co-morbidities where said LogDays regression responses are greater (or smaller). We can see how the longest Vaccine Lives occur when Co-morbidities are 0 or at most 1, and Ages correspond to mid-life (in this example almost all participants were between the ages of 40 and 80).

7.0 Discussion

This research has followed *three logical steps*. *First,* a *survival analysis* has been implemented to assess the general decay of a Vaccine Life Length, as time elapses in order to establish its *Useful Life*. Then, we have studied how such Life Length is modified by different population factors such as age, socio-economic conditions, gender, number of co-morbidities (others can be added).

Once this has been established, we have implemented a *Discrimination Analysis* between the two vaccinated groups (those that have become *infected* with the Covid-19 virus, *and* those that have *not contracted the virus* by the time the study has ended, and are thus considered as *Censored*). The significant variables in this analysis provide an indication of *which factors impact* and which ones do not, the *Vaccine Life Length*. It was established that *Age and Number of Co-morbidities* had a (negative) impact in the Vaccine Life Length.

Finally, a *regression analysis* was implemented *to quantify such impact*. It was determined that, *as Age and Number of Co-morbidities increases*, the length of time of the effect of the Vaccine (i.e. *Life Length) diminishes*. Such indication may be *useful in determining the number of doses* that each patient needs (and its timing), to remain safe.

The *approach* followed in this research allows the *analysis* to be *on-going*. *Initial censoring time* (or experimental length) can be established, in order *to have a first, working, estimation* of the *Vaccine Life Length*. The *study can continue further*, and data from the *additional time* can then be added and *reanalyzed,* and used to *adjust and refine the initial estimations at a later date*.

8. Conclusions

The main objective of the present analysis is *to provide* a detailed *example of the use of Survival Analysis, to* the study and assessment of the *Life Length* of *Covid-19 Vaccines*. The data used in our analysis has been built by this researcher, using our experience and information. The specific results obtained here have only an illustrative value.

The main use of this work is as *a tutorial example*, whereby Public Health, drug *developers* and medical researchers *can follow our statistical procedures as a guide using their own data*, or as a brain storming exercise, *to generate additional analysis ideas*.

This framework can *include more factors* (metrics), *as they become available* to researchers. For example, instead of (or in addition to) Number of Co-morbidities it can include diseases at their levels (e.g. if considering Cancer, give its Phase: I, II, III and IV).

Finally, all Covid-19 experimental and public health research has *one objective*: to contribute to *defeat this scourge*. *Individual results* from specific cases are *not as important*; the possibility of *extrapolating said results* to the general population is! *In statistics*, this is known as *Inference.*

Statisticians, supporting the excellent work of Public Health professionals, can achieve this goal.

Bibliography

Ebeling, C. E. An Introduction to Reliability and Maintainability Engineering. Waveland Press, Long Grove, Ill. 1997.

Kalbfleisch, J. D. and R. L. Prentice. Statistical Analysis of Failure Time Data. John Wiley & Sons. New York. 1980.

O'Connor, P. T. Practical Reliability Engineering. John Wiley & Sons, NY. Fourth Ed. 2002

Reliability Toolkit. Reliability Analysis Center. RAC.

Romeu, J. L. *A Comparative Study of Goodness-of-Fit Tests for Multivariate Normality*. Journal of Multivariate Analysis. V. 46, No. 2. August 1993. 309--334.

Romeu, J. L. RAC START. Vol. 9, No. 6: *Empirical Assessment of Normal and Lognormal Distribution Assumptions*. https://web.cortland.edu/matresearch/NormAssumSTART.pdf

Romeu, J. L. RAC START: Vol. 8, No. 2: *Statistical Assumptions of an Exponential Distribution.* https://web.cortland.edu/matresearch/ExpAssumSTART.pdf

Romeu, J. L. RAC START. Vol. 10, No. 3: *Empirical Assessment of the Weibull Distribution*. https://web.cortland.edu/matresearch/WeibAssumpSTART.pdf

Romeu, J. L. RAC START. Vol. 10, No. 7: *Reliability Estimations for Exponential Life*. https://web.cortland.edu/matresearch/RExpLifeSTART.pdf

Romeu, J. L. RAC START: Vol. 8, No. 2: *About Censored Data*. https://web.cortland.edu/matresearch/CensorDatSTART.pdf .

About the Author:

Jorge Luis Romeu retired Emeritus from the State University of New York (SUNY). He was for sixteen years, a Research Professor at Syracuse University, where he is currently an Adjunct Professor of Statistics. Romeu worked for many years as a Senior Research Engineer at the Reliability Analysis Center (RAC), an Air Force Information and Analysis Center operated by IIT Research Institute (IITRI). Romeu received seven Fulbright assignments: in Mexico (3), the Dominican Republic (2), Ecuador, and Colombia. He holds a doctorate in Statistics/O.R., is a C. Stat. Fellow, of the Royal Statistical Society, a Senior Member of the American Society for Quality (ASQ) and Member of the American Statistical Association. Romeu is a Past ASQ Regional Director (currently Deputy Regional Director), and holds Reliability and Quality ASQ Professional Certifications. Romeu created and directs the Juarez Lincoln Marti International Ed. Project (JLM, https://web.cortland.edu/matresearch/), which supports (i) higher education in Ibero-America and (ii) maintains the Quality, Reliability and Continuous Improvement Institute (QR&CII, https://web.cortland.edu/matresearch/QR&CIInstPg.htm) technical web site.