Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

STAT 222 PROJECT 2

A Simulation study of the Correlates of Future Violence in People Being Treated for Schizophrenia

Kenneth Broadhead*

[1]Department of Statistics, UC Davis, California, USA

Correspondence
*Kenneth Broadhead Email: kcbroadhead@ucdavis.edu

Abstract

In this project, we perform a simple simulation study based on the paper "Correlates of Future Violence in People Being Treated for Schizophrenia" by Buchanan, et al. In this paper, Buchanan, et al analyze survival data for patients being treated for schizophrenia, focusing on time to first injurious violence, in order to attain better understand the correlates for future violence. We simulate the data set using values presented in the paper, and reproduce major analysis and results. Our results agree with many major conclusions of the original paper; however, they our results often differ in the estimated effect size. The reason for this discrepancy is discussed, and suggestions for more accurate simulations are given.

KEYWORDS:
survival analysis, Cox ph model, simulation

1 | INTRODUCTION

The National Institute of Mental Health's (NIMH) CATIE study is a is a nationwide public health clinical trial aimed at comparing the effectiveness of older and newer medications designed to treat schizophrenia. In their paper, "Correlates of Future Violence in People Being Treated for Schizophrenia" from the September 2019 issue of the American Journal of Psychiatry, Buchanan, et al use a sample of 1435 patients participating in the CATIE study to examine the covariates of future violence in those being treated for schizophrenia. Baseline covariates were assessed upon entrance to the study; treatments were then assigned, and patients were followed for up to 18 months. Baseline covariates assessed included baseline violent behavior as well as other demographic variables, clinical conditions, current living conditions, and childhood risk factors. Dependent variables were injurious violence committed during the follow up time. Kaplan-Meier curves and Cox proportional hazards models of time to first injurious violence were used to assess the correlates of future violence. Patients were censored after 18 months, or when they stopped providing follow up data. No truncation was reported in the paper.

Overall, Buchanan, et al found 5 significant correlates of future violence: Baseline violent activity, recent violent victimization, childhood sexual abuse, drug use severity and medication non-adherence. The first three of these correlates are categorical, while the latter two were treated as continuous. All variables were found to increase the risk of injurious violent behavior, with baseline violent behavior being the most important predictor (having the highest hazard ratio).

Using these hazard ratios, and summary statistics of baseline covariates, we simulate the data set used by Buchanan, et al by employing a proportional hazard model to generate survival times. Additionally, uniform censoring is used to determine censoring status.

The rest of this report is structured as follows. In section 2, we briefly review the notation to be used throughout, the primary methods of analysis we will rely on when replicating results from the paper by Buchanan, et al, and the methods of simulating the data set used to replicate results. In section 3, we describe the results of our analysis on our simulated data, and compare them with the results of Buchanan, et al. We end with a discussion of the discrepancy between the two analyses, and suggestions for a more accurate simulation study.

2 | METHODS

2.1 | Notation

In this section we describe the notation used in describing the analytical methods for time to event data used in the following sections.

For a given sample of time to event data, suppose there are D distinct event times: $t_1, ..., t_D$. Denote the number of subjects at risk for the event at time t_i by Y_i, and let the number of those who experience the event at time t_i be denoted by d_i. Furthermore, let X_j, C_j and $T_j = min(T_j, C_j)$ denote the survival time (time to the event of interest), censoring time, and observed survival time for a subject in a given sample. Lastly, denote by $h(t|Z)$ the hazard function of a population conditioned on a set of covariates Z. Let $Z_{(i)}$ denote the covariate set for an individual with event time t_i. Denote an arbitrary baseline hazard function by $h_0(t)$. Finally, let $R(t_i)$ denote the risk set at time t_i; that is, the set of all individuals who have yet to experience the event just prior to time t_i.

2.2 | Analysis

We consider first to models below: the Kaplan Meier (product limit) estimator, and the cox proportional hazard model.

The Kaplan Meier estimator is a non-parametric estimator of the survival function for a population. For event times, and d_i, Y_i as described above, the quantity $\frac{d_i}{Y_i}$ gives an estimate of the conditional probability of an individual who survives just prior to time t_i experiences the event of interest at time t_i. It then follows that $1 - \frac{d_i}{Y_i}$ gives the proportion remaining event free at time t_i. The estimator is then given by:

$$\hat{S(t)} = \prod_{t_i \leq t} [1 - \frac{d_i}{Y_i}] \qquad t \geq t_1$$
$$\hat{S(t)} = 1 \qquad t < t_1$$

In proportional hazards (PH) models, the effect of covariates on the hazard function are modeled, and assumed to act in a multiplicative fashion. The general model is:

$$h(t|Z) = h_0(t) f(\beta^\top Z),$$

where h_0 is an (unspecified) baseline hazard function, and f is some known function. In the Cox Proportional Hazard model, the function f is taken to be the exponential function, leading to the model

$$h(t|Z) = h_0(t) \exp(\beta^\top Z).$$

The parameters β in the Cox PH model are usually estimated by the method of maximum (partial) likelihood. Here we provide a brief derivation of the (partial) likelihood for the Cox Model in the absence of ties. We have for a given individual at time t_i, the probability of experiencing the event at t_i given there is only one death at time t_i is equal to:

$$\frac{P(event\ at\ t_i | no\ event\ to\ time\ t_i)}{P(one\ event\ at\ t_i | no\ event\ up\ to\ t_i)}$$
$$= \frac{h(t_i | Z_{(i)})}{\sum_{j \in R(t_i)} h(t_j | Z_j)}$$
$$= \frac{h_0(t_i) \exp \beta^\top Z_{(i)}}{\sum_{j \in R(t_i)} h_0(t_j) \exp \beta^\top Z_j}$$
$$\frac{\exp \beta^\top Z_{(i)}}{\sum_{j \in R(t_i)} \exp \beta^\top Z_j}$$
$$= L_i$$

The partial likelihood is then given by the product of each probability above for all D distinct event times:

$$L(\beta) = \prod_{i=1}^{D} L_i = \prod_{i=1}^{D} \frac{\exp \beta^\top Z_{(i)}}{\sum_{j \in R(t_i)} \exp \beta^\top Z_j}$$

For all analysis, modeling, and estimation in this report, we use SAS. Code is provided in the appendices.

2.3 | Simulation

To simulate the data set, we require the following: values for all five significant covariates, time to event data, and censoring indicators. Our simulations were carried out in R (see code in the appendices) and were structured as follows:

First, values for each of the covariates were simulated. Baseline violent behavior assumed one of three values: “no violence,” “non-injurious violence,” and “injurious violence.” To simulate this, a multinomial probability distribution was used, with probabilities computed from the sample information from the original paper. The variables tracking childhood sexual abuse and recent violent victimization (within past 6 months) were binary “yes” or “no” variables. A Bernoulli distribution was thus used, with probability of “yes” identical to the sample probability in the paper. The drug use severity of a patient was rated using a standard five-point (one to five) scale; in the paper’s analysis, it was treated as a continuous variable. It was simulated by randomly sampling an integer from one to five using probabilities the closely matched the sample mean and standard deviation reported in the paper. It is likely that the distribution of scores was heavily right skewed; this was also used to inform the probabilities used in simulation. Finally, medication adherence was rated using a number of methods from pill counts, to professional opinion and self-reported use. However, little information was given as to how these ratings were pooled into a single measure. As such, we simply simulated this score according to a normal distribution with mean and standard deviation matching those reported in the study. Overall sample proportions were used for categorical variables. As overall mean and standard deviation were not reported in the original study (variable summaries were stratified by outcome, which are not known in advance at the time of simulating covariates), the mean and standard deviation for the “no violence at follow up” group was used, as this group constituted a vast majority of the sample (85%), and the means and standard deviations did not vary greatly from group to group.

Next, to simulate event times, a (Cox) proportional hazard model was used. A baseline exponential distribution (Weibull with shape parameter equal to one) was used, with rate parameter modified according to the values of the covariates. The parameters (hazard ratios) reported by Buchanan, et al were used when determining the strength of the influence for each covariate. A uniform censoring distribution was used; event times were then censored if they were greater than the censoring time or greater than 18 months.

Covariates were simulated independently. Summary statistics and analysis for the simulated data in comparison to the original data set are provided in in the next section.

3 | RESULTS

3.1 | Summary of Simulation

Summary statistics for the simulated covariates can be found in Table 1. Note that similarity of the simulated data to the data set being simulated, in that categorical variables have similar proportions in both data sets, and continuous variables have similar first and second moments. Some comparisons for survival times between the two data sets can be made by examining the survival curves for both real and simulated data sets in figures 1 and two respectively. Here the difference is somewhat larger than for the values of the covariates. See below for a more in depth analysis of discrepancy’s between the analytical results and limitations of this simulation. Overall, however, this similarity suggests, prima facie, the appropriateness of attempting to replicate original analytical results using this simulated data.

3.2 | Analysis

First we present a comparison of summary statistics, stratified by event occurrence, for the real and simulated data sets, in tables 2 and 3 respectively. As the original paper reports summary statistics by violence reported at follow up (none, non-injurious, or injurious) we provide approximate summary statistics for the continuous covariates in the ’no injurious violence’ category.

TABLE 1 Comparison of Simulated and Reported Summary Statistics

	Paper (n = 1435)		**Simulated** (n = 1435)	
Covariate	**N**	%	**N**	%
No Violence	1215	84.67	1215	84.67
Non. Inj. Viol	135	9.41	134	9.34
Inj. Viol	85	5.92	86	5.99
Childhood Sex Abuse	289	20.14	296	20.63
Violent Victimization	109	7.60	130	9.06
	Mean	**SD**	**Mean**	**SD**
Drug Use	1.34	0.70	1.33	0.69
Medication Non-adherence	1.29	0.71	1.29	0.71

TABLE 2 Reported Summary Statistics by Event Occurrence

	No Injurious Violence (n = 1358)		**Injurious Violence** (n = 77)	
Covariate	**N**	%	**N**	%
No Violence	1171	86.2	44	57.1
Non. Inj. Viol	120	8.8	15	19.5
Inj. Viol	67	4.9	18	23.4
Childhood Sex Abuse	260	19.1	29	37.7
Violent Victimization	26	1.9	8	10.5
	Mean	**SD**	**Mean**	**SD**
Drug Use	1.34	0.70	1.66	0.95
Medication Non-adherence	1.27	0. 71	1.38	0.81

The first primary fining of Buchanan, et al was that baseline violence was an important predictor (see below for Hazard Ratios) of time to first injurious violence, as seen in the Kaplan Meier curves shown in Figure 1. The curves confirms that individuals with violent behavior at baseline had much higher rates of future injurious violence, followed by those with individuals with non-injurious violence and no violence at baseline respectively. Similar analysis of Kalplan Meier curves conducted on the simulated data set once again confirm that individuals with violent behavior at baseline have the highest rates of future injurious violence, followed by those with non-injurious and no violence at baseline respectively. However, KM curves generated using the simulated data set noticeably underestimated these rates for individuals with violent behavior at baseline.

The second primary result of Buchanan, et al was the results of the Cox PH model run on the data. Table 2 shows the basic results of fitting such a model, along with the results from fitting the same model on the simulated data. We note that the results for Baseline violence are quite similar, with baseline injurious violence had a larger hazard ratio baseline non-injurious violence. Indeed, Baseline injurious violence had the largest hazard ratio overall in both model cases. Furthermore, the results for recent violent victimization, drug use, and medication non-adherence all having similar hazard ratios. The only covariate that differed between the two models is child hood sexual abuse. In the simulation study the covariate is not flagged as significant in the cox regression model, and the hazard ratio is much smaller than reported in the analysis of the real data set.

There are two likely reasons for the discrepancies in the Kaplan Meier Curves and the results of the proportional hazards model. The discrepancies in the KM Curves are likely due to the method of simulation, namely, that the values of the covariates were simulated independently. In reality, the covariates are likely to show some moderate to high dependency amongst one

TABLE 3 Simulated Summary Statistics by Event Occurrence

	No Injurious Violence (n = 1374)		Injurious violence (n = 61)	
Covariate	**N**	**%**	**N**	**%**
No Violence	1175	85.5	40	65.6
Non. Inj. Viol	125	9.1	9	14.8
Inj. Viol	74	5.4	12	19.7
Childhood Sex Abuse	284	20.7	12	19.7
Violent Victimization	116	8.4	14	23.0
	Mean	**SD**	**Mean**	**SD**
Drug Use	1.28	0.60	2.36	1.44
Medication Non-adherence	1.29	.71	1.35	0.71

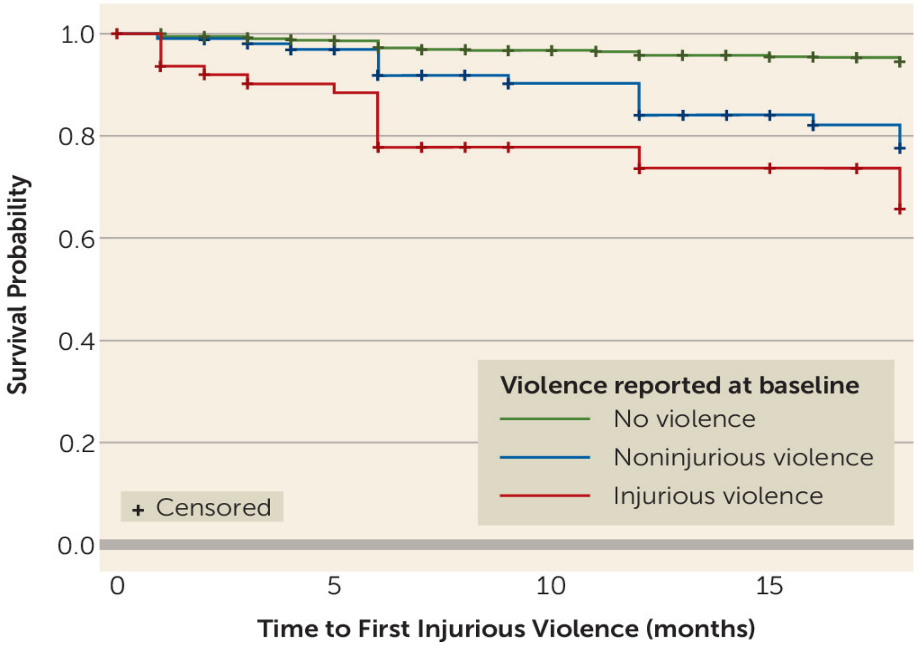

FIGURE 1 Kaplan Meier Curves for Real Data.

another (with the possible exception of childhood sexual abuse). For example, an individual who had higher drug use at baseline is more likely to have had higher medication non-adherence, have made more risky decisions, and in turn be have victim of violent victimization due to aggressive (or violent) behavior. Thus, the covariates are more likely to have either all (or mostly) been all moderate to higher, or all low. Since all covariates had positive hazard ratio, this would have the effect of "dragging down" the KM curves for baseline injurious and non-injurious violence. This would lead give an even greater similarity between the KM curves for simulated and real data.

The discrepancies in the results of the cox model are likely due to the high censoring rate. The incredibly high censoring rate is likely to make hazard estimates in the cox model unstable, thus leading to potential discrepancy between the results of the two analyses.

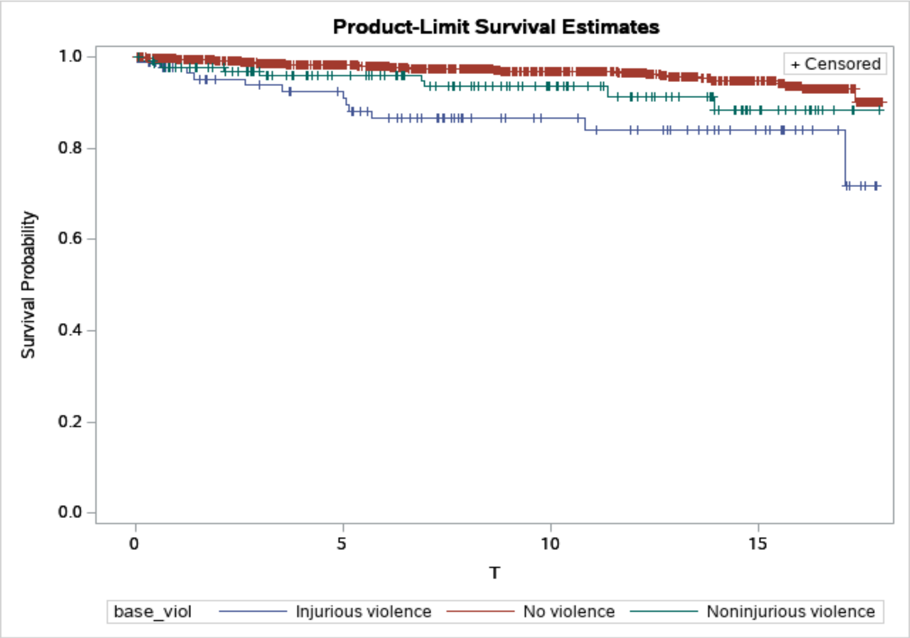

FIGURE 2 Kaplan Meier Curves for Simulated Data.

TABLE 4 Results of fitted Cox PH Model

	Paper			**Simulated**		
Covariate	**HR**	**95% Wald CI**	**p-value**	HR	**95%Wald CI**	**p-value**
Non. Inj. Viol	2.72	1.45, 5.09	<0.01	2.40	1.14, 5.08	0.0218
Inj. Viol	4.02	2.12, 7.60	<0.01	5.20	2.68, 10.09	<.0001
Childhood Sex Abuse	1.85	1.12, 3.05	<0.01	1.11	0.58, 2.11	0.7525
Violent Victimization	3.52	1.62, 7.64	<0.01	4.21	2.26, 7.83	<.0001
Drug Use	2.93	1.65, 5,18	<0.01	2.96	2.44, 3.58	<.0001
Medication Non-adherence	1.39	1.04, 1.86	<0.01	1.58	1.12, 2.24	0.0101

4 | CONCLUSIONS

In this report, we simulated and reanalyzed a data set from the paper “Correlates of Future Violence in People Being Treated for Schizophrenia” by Buchanan, et al. The results of our analysis on the simulated data set were similar to those obtained by Buchanan, et al. All major conclusions remain the same, though the magnitude of some effects differed somewhat significantly. These discrepancies were likely due to the methods of simulation, and the fact that, even amongst schizophrenics, violence is a rare event. To improve upon the results of this simulation study, future simulations studies of this nature should incorporate dependencies between covariates into the simulation process.

□

APPENDIX

A REFERENCES

Buchanan, et al. (2019). Correlates of Future Violence in People Being Treated for Schizophrenia. *The American Journal of Psychiatry, 179(9),694-701*. https://doi.org/10.1176/appi.ajp.2019.18080909

B CODE (SAS)

```
/* Import */
proc import datafile="/folders/myfolders/proj2/simulate.csv"
out=simulate dbms=csv replace;
datarow=2;
getnames=yes;
guessingrows=100;
run;
/* Note: censor=0 indicates a censored observation*/

proc means data=simulate
mean std sum;
run;

proc freq data=simulate;
table base_viol;
run;

PROC SORT Data=simulate Out=simulate_sort;
BY censor;
RUN;

proc means data=simulate_sort
mean std sum;
by censor;
run;

proc freq data=simulate_sort;
table base_viol;
by censor;
run;

proc lifetest data=simulate method=KM nelson plots=(s hazard(kernel=E));
time T*censor(0);
strata base_viol;
run;

proc phreg data=simulate plots=survival;
class base_viol(ref="No violence");
model T*censor(0) = base_viol sex_abuse viol_victim drug_use medication;
hazardratio base_viol / CL=WALD;
hazardratio sex_abuse/ CL=WALD;
hazardratio viol_victim / CL=WALD;
hazardratio drug_use / CL=WALD;
hazardratio medication / CL=WALD;
run;
```

C CODE (R)

```
## Simulation of Data
## Stat 222 project 2

# ------------------------------        Seed Set        ------------------------------

set.seed(1234)

# ------------------------------ Simulating Covariates ------------------------------
# Simulating Baseline violence covariates
N<-1435

probs<-c(1215/N,135/N,85/N)
base_viol<-t(rmultinom(n=1435,1,probs))
base_viol<-as.data.frame(base_viol)
colnames(base_viol)<-c("no_viol", "noninj_viol", "inju_viol")

# Simulating Childhood risk factor
sex_abuse<-as.data.frame(rbinom(N,1,289/N))
colnames(sex_abuse)<-"sex_abuse"
covars<-cbind(base_viol, sex_abuse)

# Simulating Current circumstances
viol_victim<-as.data.frame(rbinom(N,1,109/N))
colnames(viol_victim)<-"viol_victim"
covars<-cbind(covars, viol_victim)

# Simulating Clinical condition
drug_use<-as.data.frame(sample(1:5, size=N, replace=TRUE, prob=c(0.75,.20,.02,.02,.01)))
medication<-as.data.frame(rnorm(N,1.29,0.71))
colnames(drug_use)<-"drug_use"
colnames(medication)<-"medication"

covars<-cbind(covars, cbind(drug_use, medication))


# ------------------------------ Simulate Observed time ------------------------------
L_0<-0.000305 ## 0.000705
HR<-log(c(2.72,4.02,1.85,3.52,2.93,1.39))
X<-rexp(1435, rate=(L_0 * exp(as.matrix(covars[,-1])%*%HR)))

censor_time<-runif(1435,0,18)
T<-pmin(X, censor_time, 18)
censor <- as.numeric(!(X>censor_time | X>18))
```

```
##drop_out<- as.numeric( X>censor_time & X<18 )
##summary(drop_out)
##summary(censor)
# Note: a value of 0 indicates the observation is censored

# ———————————————— Final Formatting ————————————————
base_viol<-NULL
for(i in 1:length(covars[,1]))
{
        if(covars[i,1]==1){base_viol[i]<-'No_violence'}
        if(covars[i,2]==1){base_viol[i]<-'Noninjurious_violence'}
        if(covars[i,3]==1){base_viol[i]<-'Injurious_violence'}
}

covars<-cbind(base_viol,covars)

covars<-covars[,-c(2,3,4)]

simulate<-cbind(T,cbind(censor,covars))

write.csv(simulate,
          file="C:/Users/kenne/Documents/SASUniversityEdition/myfolders/proj2/simulate.csv")
```