Anthony Orso

STAT 411

December 8, 2023

Time-to-Event Prediction Using Mixture Cure Models

The field of survival analysis is predicated on a baseline assumption: all survival curves converge to zero. In other words, everyone will experience the event. Thus, a censored observation only means it has not happened yet, as evidenced by multiplying the likelihood function by S(x) at that given point. This assumption provides the foundation for survival analysis, as we can easily cycle between the cumulative distribution, hazard, cumulative hazard, and probability density functions. While this enables statisticians to engage in complex and mathematically sound methods, a crack exists in this foundation. What happens when we know some people will *never* experience the event?

This caveat is essential for advanced practitioners to be trained in to better tackle problems. In cancer research, not everyone who has cancer will die during the study (right-censored), but some will never die at all because they are cured. In healthcare and epidemiological research, there is a wealth of information on mortality rates across demographics, and this provides us an opportunity to incorporate real-time baseline hazard rates. The article *Time to default in credit scoring using survival analysis: a benchmark study* (Dirick, Claeskens, & Baesens, 2016) introduces advanced methods to model credit defaulting using single- and multiple-event mixture cure models. In doing this, the authors also model profits and losses for lenders. Due to significant complexity and rarity of data on this issue, this paper will conclude with an original analysis on cancer data sets using the methods outlined in the article.

Article Summary

Before diving into the method and theory, the authors described two methods for censoring used in their analysis that require different modeling approaches. The first definition of censoring use is the most complex and requires highly sophisticated methods that can't be reproduced for this project. There are multiple levels of event types, which include default, early repayment, and maturation, which is when a loan is fully paid off. Maturation would be considered a cure, as the expected loan value was paid in full. This complicated definition of event types requires multi-event mixture models that involve multinomial logistic regression. The second definition is in line with typical survival analysis. Only loan defaults are listed as events, and everything else is a censored observation. The second definition allows for typical survival analysis where $\delta = 1$ for default and $\delta = 0$ for all other observations.

To evaluate the ideal model for predicting defaults and losses, the authors cover the exponential, Weibull, and log-logistic accelerated failure time (AFT); Cox proportional hazards (Cox PH); Cox PH with splines; mixture cure; and multi-event mixture cure models. This paper will do a deep dive into the mixture models since they are novel and unrelated to course content. However, the original analysis will look at both semiparametric and parametric mixture cure models to assess performance.

Mixture cure models

The development of curative models began with the medical profession's desire to incorporate the fact that some patients will never experience an event. However, typical censoring does not account for that. A censored observation simply means the event did not happen *yet* during the study. Thus, the survival curves generated from these data sets converge to zero. In essence, it is assumed everyone will die. But this is not the case with certain diseases. Some people who are diagnosed with cancer will achieve remission and never have cancer again. Existing techniques, such as the Cox PH and AFT models, do not account for this small but crucial detail.

To account for this, mixture cure models incorporate a logistic regression component to predict non-susceptibility and a survival model to describe survival behavior of the susceptible cases. This novel approach to modeling survival is important because most loans do not default. As such, it may be unwise to use a traditional model that assumes convergence to zero. The unconditional survival function for this model is given below.

$$S(t\,,\,x)\;=\;\pi(x)S(t,\,Y=1,\,x)+1-\pi(x)$$

Y is the susceptibility indicator (*Y* = 1 for susceptible to death, *Y* = 0 if not), and ***x*** is the covariate vector. In this model, we have a simple binomial logit since we are only modeling two event types: will default or won't default. This lines up with the author's second definition of censoring mentioned on page 2. The binomial logit probability is below.

$$\pi(x)\;=\;P(Y=1,\,x)\;=\;\frac{e^{\beta' x}}{1+e^{\beta' x}}$$

The authors then use a Cox PH model to develop the conditional survival given susceptibility.

$$S(t,\,Y=1,\,x)\;=\;[S_o(x,\,Y=1)]^{e^{\beta' x}}$$

In the author's analysis, the baseline survival is calculated using the Breslow estimator and Cox PH model. However, real-life estimates can be used in certain medical cases. In my original analysis, I will upload a raretable that provides cure estimates based on age, gender, and year for colon cancer patients. In these instances, we use a highly reliable baseline estimate. In all cases, the unconditional survival will plateau at a positive value of $1-\pi(x)$.

Multiple-event mixture cure models

Observed events in credit risk, however, are not as simple as default or no default. For example, cure is directly observed when a loan matures. In other words, the loan is fully paid off on the day the last payment is due. In addition, the basic mixture cure model doesn't account for competing risks, such as early repayment. Although this is not a default, early repayments are still a loss for the lender. The authors define three indicators for their analysis:

- Y_m : loan matures and is fully paid off
- Y_d : loan defaults
- Y_e : early repayment occurs

This analysis assumes loans with a fixed start and end date. As such, revolving credit card debt is not suitable for this type of modeling. The unconditional survival function is below.

$$S(t,\, x) \;=\; \pi_e(x)S_e(t,\, Y_e = 1,\, x) \;+\; \pi_d(x)S_d(t,\, Y_d = 1,\, x) \;+\; \big(1-\pi_e(x)-\pi_d(x)\big)$$

π_e and S_e correspond with Y_e (early repayment), and π_d and S_d correspond with Y_d (loan defaults). Each susceptibility probability is calculated in the typical way of multinomial logistic regression.

$$\pi_{i\in(d,e)}(x) \;=\; \frac{e^{\beta_i' x}}{1+e^{\beta_d' x}+e^{\beta_e' x}}$$

The probability of cure $(1 - \pi_e(x) - \pi_d(x))$ corresponds with the loan maturing. The survival probabilities attached to each susceptibility probability are derived from a Cox PH model.

Data analysis

The authors used several data sets that tracked loan defaulting across numerous banks and lending organizations in the U.K. Because survival models cannot cope with missing values, the authors use median imputation for continuous variables where ≤ 25% of values are missing, and rows were dropped if > 25% were missing. For categorical variables, an "NA" class was created if more than 15% of the column had a missing value. Otherwise, the rows were dropped from analysis. The two definitions of censoring yield dramatically different proportions of censoring in the final data set. The first definition that is used for multiple-event mixture cure models is less restrictive in defining an event, so the number of censored observations (loan maturing) ranged from 20 to 85% in each data set. However, the second definition that defines $\delta = 1$ as default and $\delta = 0$ as anything else generates no less than 95% of observations being censored.

The authors used ROC curves and annuity theory to evaluate model performance. Area Under the Curve is a run-of-the-mill model evaluation metric, but the use of annuity theory is novel. These data sets provide the final value of the loan, so the authors used survival and logistic regression estimates of each individual loan to project its expected final value. After using the outputs of the mixture cure model to forecast the loan value, the authors then compare the projected number to its actual value using Mean Squared Error.

There are existing formulae on how early payment, mature payment, and loan default yield differing loan values. We can then multiply all of these formulae by the probability of observing that event type to generate an overall Expected Future Value.With the second definition of censoring this is what the Expected Future Value looks like.

$$EFV_{s,m} = PD_s \cdot R_s\left(\sum_{j=1}^{n} S(j)_{s,m}^{d}(1 + i)^{n-j}\right) + (1 - PD_s) \cdot R_s\frac{(1+i)^n - 1}{i}$$

PD_s corresponds with the probability of default and R_s is the monthly payment. The function multiplied by PD_s is used to calculate the future value of the loan using survival analysis *without* a cure proportion. It helps us understand how much has been paid on the loan assuming the creditor has not defaulted up until that point. The probability of not defaulting is then multiplied by the expected future value of a fully matured loan, which is intuitive because that is what the loan will be worth if it never defaults. Adding these two products together then projects the value of the loan. However, this method becomes much more complex when using the first definition of censoring, which involves defaults, early repayments, and loan maturation.

$$EFV_{s,m} = PD_s \cdot R_s\left(\sum_{j=1}^{n} S(j)_{s,m}^{d}(1+i)^{n-j}\right) + PM_s \cdot R_s\frac{(1+i)^n - 1}{i}$$

$$+PE_s \cdot \left(R_s\left(\sum_{j=1}^{n} S(j)_{s,m}^{e}(1+i)^{n-j}\right) + \sum_{j=1}^{n-1}\left(\left(S(j-1)_{s,m}^{e} - S(j)_{s,m}^{e}\right)\right)L_{s,j}(1+i)^{n-j-1}\right)$$

In this formula, PD_s is the probability of default, which is multiplied by the monthly payment and the formula for the value of a loan if it defaults. PM_s corresponds with probability of a loan maturing (cured), multiplied by the monthly payment and value of the loan if it is paid in full.

PE_s corresponds with the probability of early payment, multiplied by the monthly payment and formula that determines the value of a loan if it is paid early.

Results

In general, simple AFT models often performed the worst in terms of accuracy of projected loan value and consequent mean squared error. The single-event mixture cure model was in the top three models in terms of proximity to the actual loan value in eight out of 10 data sets, but Cox PH models (including ones with splines) accounted for the lowest MSE in six of the 10 data sets. Single-event mixture cure models accounted for the lowest MSE for two out of the 10. Multiple-event mixture models never produced the lowest MSE.

Critique

A major limitation of these models is that they are measurably more computationally expensive than typical survival modeling. Because maximum likelihood estimation is needed to model susceptibility, computational power can increase significantly. The authors used the *smcure* package with a Cox PH model for the survival proportion. Using this same package and model choice, it took hours to run on a colon cancer data set of 13,000 rows with only one covariate for regression. As such, using less flexible methods of survival modeling are more pragmatic unless someone has access to significant computational power. Although AFT models are prescriptive and perhaps less instantaneously precise than a Cox PH model, the ease in running them saves time, allows for more covariates in the regression equation, and prevents computer crashing. Another package (which I used in my original analysis) is *cuRe*, which allows the user to specify a parametric distribution for modeling. This ran quickly and produced easily interpretable regression coefficients.

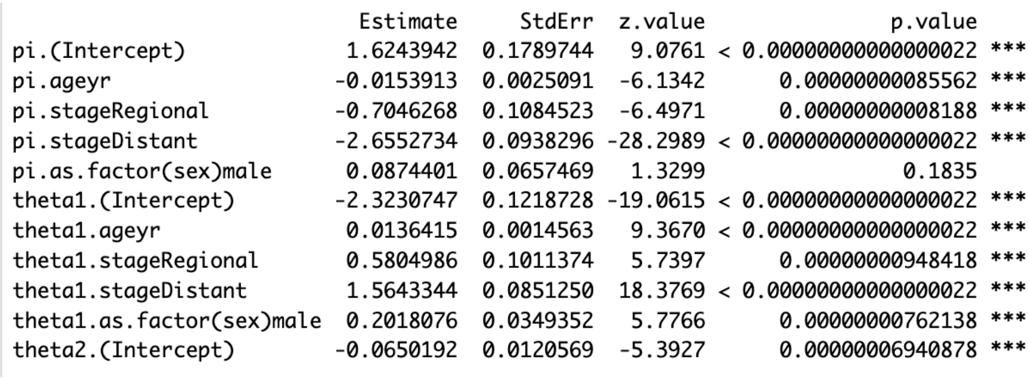
Regardless of how the survival probability is modeled, there is strong value in single-event mixture cure models when the presence of cure is an indisputable fact. Although Cox PH performed best overall in MSE, that metric was used for evaluating the accuracy of loan value rather than the accuracy in correctly predicting population cure rate. Furthermore, single-event mixture cure models were consistently ranked as a top model alongside Cox PH in terms of AUC and MSE.

However, the performance results of the multiple-event cure models show the bias-variance trade-off. The most complex, all-inclusive model is often not the best choice, and this was reflected in single-event mixture cure models always performing better. More parameters are not always needed and, in fact, can harm model performance. Thus, I would advise people to rely heavily on goodness-of-fit outputs to determine whether a multi-event model is needed over a simpler single-event model.

Original Analysis

The usage of credit defaulting data sets to project loan values using mixture cure modeling is a highly novel approach. As such, it was not feasible to do an original analysis on the exact same area of research due to the lack of data. Instead, I used a colon cancer data set from the *cuRe* package in R to model survival and cure rates. In the *cuRe* package, $\pi(x)$ is defined as the probability of cure, unlike the authors' definition of $P(Y = 1)$. Thus, my interpretations of the parameter estimates will change in direction. R provides a raretable that maps age, sex, and diagnosis year to a concrete baseline hazard rate that can be used in place of a manually calculated Breslow estimator. These real-life estimates come from SEER, which is an institution that provides real-time hazard and survival rates for diseases based on demographics. I also created an indicator variable that was 1 if distant stage cancer (metastatic) and 0 otherwise.

My initial model included the distant, sex, age (in years), and cancer stage variables. However, this created numerical issues in which error messages showed up and the distant variable failed to achieve significance in both the logistic regression and survival estimation, so this variable was removed. I also experimented with both exponential and Weibull distributions, but the Weibull distribution reduced the AIC the most and led to the greatest number of statistically significant parameters. The R summary for the Weibull model is below.



```
                            Estimate     StdErr  z.value              p.value
pi.(Intercept)             1.6243942  0.1789744   9.0761 < 0.0000000000000000022 ***
pi.ageyr                  -0.0153913  0.0025091  -6.1342      0.00000000085562 ***
pi.stageRegional          -0.7046268  0.1084523  -6.4971      0.00000000008188 ***
pi.stageDistant           -2.6552734  0.0938296 -28.2989 < 0.0000000000000000022 ***
pi.as.factor(sex)male      0.0874401  0.0657469   1.3299                0.1835    
theta1.(Intercept)        -2.3230747  0.1218728 -19.0615 < 0.0000000000000000022 ***
theta1.ageyr               0.0136415  0.0014563   9.3670 < 0.0000000000000000022 ***
theta1.stageRegional       0.5804986  0.1011374   5.7397      0.00000000948418 ***
theta1.stageDistant        1.5643344  0.0851250  18.3769 < 0.0000000000000000022 ***
theta1.as.factor(sex)male  0.2018076  0.0349352   5.7766      0.00000000762138 ***
theta2.(Intercept)        -0.0650192  0.0120569  -5.3927      0.00000006940878 ***
```

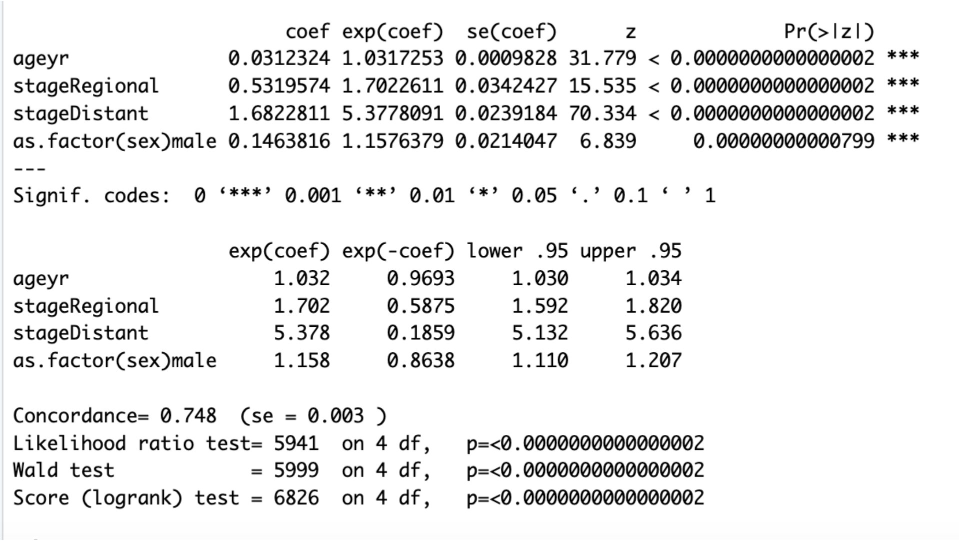
Based on these estimates, a one-unit increase in age decreases the odds of cure by 1.5%. Also, regional cancer decreases the odds of cure by 50.6% compared to localized cancer, and distant cancer decreases the odds of cure by 93% compared to localized cancer. These numbers align with common knowledge about cancer: older age and further-stage cancer are associated with higher mortality, lower remission, and lower cure rates. The sex parameter was statistically insignificant in the logistic regression portion of the model.

The parameters for the survival analysis must be interpreted differently. These parameters impact the survival estimate by being directly plugged into the Weibull survival function, whose parameters were fitted through maximum likelihood estimation of an accelerated failure time model. It looks like this:

$$S(t) = e^{\left(-(-2.32+0.013\cdot ageyr+0.58\cdot regional+1.56\cdot metastatic+0.20\cdot male)\cdot t^{-0.065}\right)}$$

Thus, a one-unit increase in age decreases survival by 1.4% ($1 - e^{-0.013}$). Regional cancer decreases survival by 44% compared to localized cancer ($1 - e^{-0.58}$), and metastatic cancer decreases survival by 79% compared to localized cancer ($1 - e^{-1.56}$). Male sex lowers survival by 18% ($1 - e^{-0.20}$) compared to female sex.

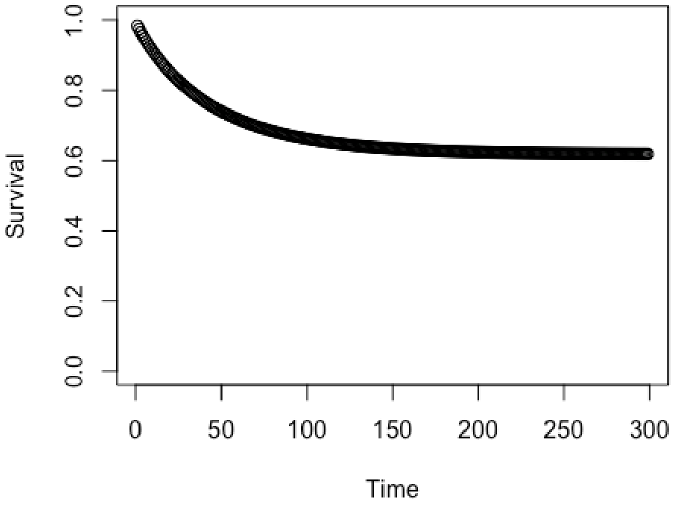
Like with the logistic regression component, these coefficients track with our intuitive knowledge of outcomes with cancer patients. The results of this model are powerful, but I want to establish a baseline model for a likelihood ratio test. To test the model's value, I built a Cox PH model with the same covariates as the mixture cure. It produced the below output.

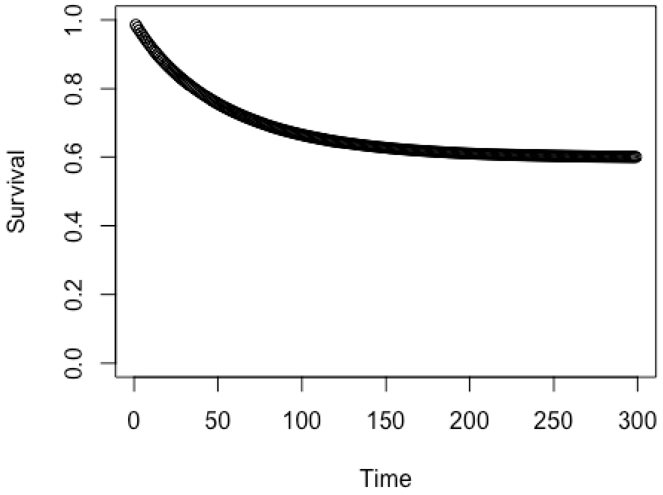


```
                        coef exp(coef)  se(coef)      z            Pr(>|z|)    
ageyr              0.0312324 1.0317253 0.0009828 31.779 < 0.0000000000000002 ***
stageRegional      0.5319574 1.7022611 0.0342427 15.535 < 0.0000000000000002 ***
stageDistant       1.6822811 5.3778091 0.0239184 70.334 < 0.0000000000000002 ***
as.factor(sex)male 0.1463816 1.1576379 0.0214047  6.839     0.00000000000799 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
ageyr                  1.032     0.9693     1.030     1.034
stageRegional          1.702     0.5875     1.592     1.820
stageDistant           5.378     0.1859     5.132     5.636
as.factor(sex)male     1.158     0.8638     1.110     1.207

Concordance= 0.748  (se = 0.003 )
Likelihood ratio test= 5941  on 4 df,   p=<0.0000000000000002
Wald test            = 5999  on 4 df,   p=<0.0000000000000002
Score (logrank) test = 6826  on 4 df,   p=<0.0000000000000002
```

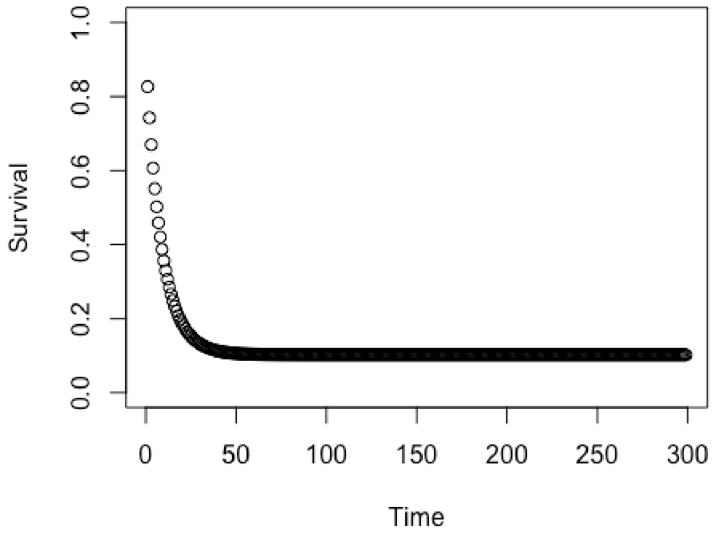
The parameters tell a similar story, but the model fit compared to the mixture cure model is worse. The AIC of the Cox PH model was 157,209, but the AIC for the mixture cure model was 32,707. This is an enormous difference. In addition, the likelihood ratio test p-value was so small that the computer rounded it down to 0. We confidently reject the null hypothesis that there is no difference between a mixture cure and Cox PH model and opt for using the mixture cure model.

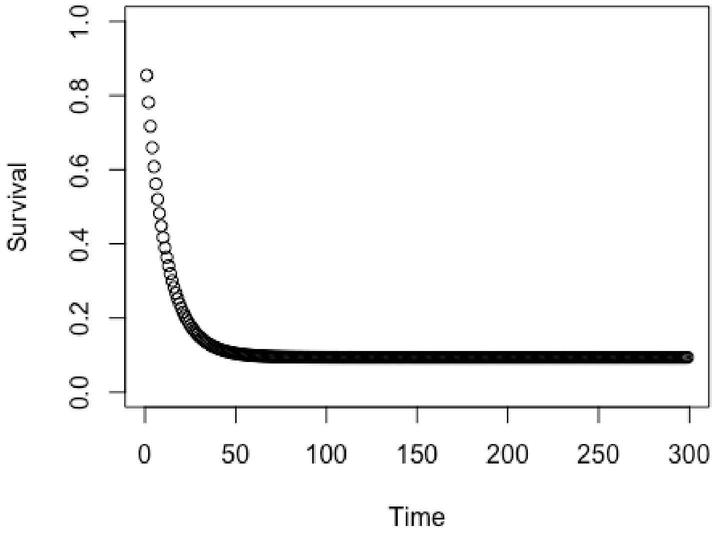
I then wanted to generate survival curves using the predict() function. Below are the graphs of 80-year-old, localized cancer patients. The first is for men, the second is for women.





A large proportion of people with localized cancer get cured, and the differences by gender were minimal despite a slight dip in death rate early on for men. I then wanted to see how these patients' survival rates would differ if they instead had metastatic cancer. The first is male, and the second is female.





Metastatic cancer has much dire outcomes, with the survival rate converging to around 9% and all deaths happening sooner than localized cancer. This tracks with the model coefficients and our knowledge of cancer outcomes.

Future Considerations

Although multiple-event mixture cure models had less power in the authors' study, there are still strong implications for their use in biomedical contexts. For example, there are meaningful differences between death, relapse, and remission. A simple $\delta \in (0,1)$ doesn't sufficiently capture the differences in these three event types. However, the issue of bias-variance trade-off is still a consideration when using multinomial logistic regression in the context of survival analysis.

In addition, the validity of using cure models could also be tested using strata in a Cox PH model. Logistic regression could first be run to test whether an event or no event will happen, and the results of that could be stored as a binary variable. This variable could then be entered into the Cox PH model as strata. Model fit measures would then determine if there were meaningful survival differences among the strata.

Overall, mixture cure models are a novel and promising alternative to traditional survival analysis. The wealth of information on baseline survival metrics in biomedical research makes curative models particularly powerful in analyzing patient health data. More research is needed to build out the efficacy, use cases, and understanding of multiple-event data sets.

Sources

Dirick, L., Claeskens, G. & Baesens, B. Time to default in credit scoring using survival analysis: a benchmark study. J Oper Res Soc 68, 652–665 (2017).

https://link.springer.com/article/10.1057/s41274-016-0128-9

Rasmus Kuhr Jensen, Mark Clements, Lars Klingen Gjærde, Lasse Hjort Jakobsen. Fitting parametric cure models in R using the packages cuRe and rstpm2. Computer Methods and Programs in Biomedicine, Volume 226 (2022).

https://www.sciencedirect.com/science/article/pii/S0169260722005065?ref=pdf_download&fr=RR-2&rr=831f34c41d8c6201

R Code

```
library(relsurv)
library(cuRe)
library(rstpm2)
library(tidyverse)
library(KMsurv)
library(smcure)
data("colonDC")


## Filter out Unknown Cancer
colonDC = colonDC %>%
  filter(stage %in% c('Regional','Distant','Localised'))


options(scipen = 999)


#Create an indicator variable for metastatic cancer and turn stage into
```

```
#factor variable

colonDC = transform(colonDC,
                    stage = factor(stage),
                    distant = as.numeric(stage=='Distant'))
#Convert the age variable into days for easier interpretability
colonDC = colonDC %>%
  mutate(ageyr = agedays/365)

#Upload the SEER baseline hazard rates for modeling as a new column

colonDC$bhaz <- general.haz(time = "FU",
                     rmap = list(age = "agedays", sex = "sex", year= "dx"),
                     data = colonDC,
                     ratetable = survexp.dk)

#Experiment with various parameters and the Weibull distribution
```

```
fit1 <- fit.cure.model(formula = Surv(FUyear,status)~distant+ageyr+stage+as.factor(sex),
               formula.surv = list(~distant+ageyr+stage+as.factor(sex),~1),
               type='mixture',dist="weibull",link="logit",
               bhazard="bhaz",data=colonDC)
summary(fit1)
AIC(fit1)


#remove distant variable since it causes numerical issues


fit.wei <- fit.cure.model(formula = Surv(FUyear,status)~ageyr+stage+as.factor(sex),
                 formula.surv = list(~ageyr+stage+as.factor(sex),~1),
                 type='mixture',dist="weibull",link="logit",
                 bhazard="bhaz",data=colonDC)


summary(fit.wei)
AIC(fit.wei)


#Try an exponential model to see how it performs
```

```
fit.exp <- fit.cure.model(formula = Surv(FUyear,status)~ageyr+stage+as.factor(sex),
                          formula.surv = list(~ageyr+stage+as.factor(sex)),
                          type='mixture',dist="exponential",link="logit",
                          bhazard="bhaz",data=colonDC)
summary(fit.exp)
AIC(fit.exp)


#We will use Weibull model because of better outputs


#Build a cox baseline


cox <- coxph(Surv(FUyear,status)~ageyr+stage+as.factor(sex), data = colonDC)
summary(cox)
cox$loglik
AIC(cox)


#params for mixture cure is 11, params for cox PH is 4, df = 7 for lRT
```

```
test = -2*(-78600.55--16342.97)

1-pchisq(test,7)

#Tiny p-value. Reject null and use mixture model

#Use predict function to observe survival and cure rate of different types of patients


a=predict(fit.wei, data.frame(ageyr = 80,stage='Localised',sex='male'),type='surv')[[1]][1]
a1<-predict(fit.wei, data.frame(ageyr = 80,stage='Localised',sex='female'),type='surv')[[1]][1]
b=predict(fit.wei, data.frame(ageyr = 80,stage='Distant',sex='male'),type='surv')[[1]][1]
b1<-predict(fit.wei, data.frame(ageyr = 80,stage='Distant',sex='female'),type='surv')[[1]][1]

vec = c()
veca1 <- c()

for (as in a){
  vec=c(vec,as)
}

for (val in a1){
  veca1<-c(veca1,val)
}

vecb = c()
vecb1 = c()

for (bs in b){
  vecb=c(vecb,bs)
}

for (val in b1){
  vecb1<-c(vecb1,val)
}
```

```
vecb1


plot(1:299,vec,ylim=c(0,1),ylab='Survival',xlab='Time')

plot(1:299,veca1,ylim=c(0,1),ylab='Survival',xlab='Time')

plot(1:299,vecb,ylim=c(0,1),ylab='Survival',xlab='Time')

plot(1:299,vecb1,ylim=c(0,1),ylab='Survival',xlab='Time')


###




##


#Code continuously crashes if more than one covariate

#Code takes anywhere from 1 hour to 24 hours to run. Not suitable for analysis


pd0<-smcure(Surv(FUyear,status)~stage,

            cureform=~stage,data=colonDC,model="ph")
```

```
printsmcure(pd)
```