Dynamic prediction of survival in cystic fibrosis: A landmarking analysis using patient registry data

Example R code for obtaining estimated survival probabilities

Ruth H Keogh, Shaun R Seaman, Jessica K Barrett, David Taylor-Robinson, Rhonda Szczesniak

This document provides R code that can be used to obtain estimated survival probabilities for adults with cystic fibrosis (CF) in the UK. The code gives estimated probabilities of survival up to 10 years from ages 18 to 50, given values for a set of 16 predictor variables and conditional on survival to the current age. The results come from a dynamic prediction model developed using data from the UK Cystic Fibrosis Registry, as described in the following manuscript:

"Dynamic prediction of survival in cystic fibrosis: A landmarking analysis using UK patient registry data"
Ruth Keogh, Shaun Seaman, Jessica Barrett, David Taylor-Robinson, Rhonda Szczesniak.
Epidemiology 2018. In Press.

The abstract is given below.

ABSTRACT

Background

Cystic fibrosis (CF) is an inherited, chronic, progressive condition affecting around 10,000 individuals in the UK and over 70,000 worldwide. Survival in CF has improved considerably over recent decades and it is important to provide up to date information on patient prognosis.

Methods

The UK Cystic Fibrosis Registry is a secure centralized database, which collects annual data on almost all CF patients in the UK. Data from 43,592 annual records from 2005-2015 on 6181 individuals were used to develop a dynamic survival prediction model that provides personalised estimates of survival probabilities given a patient's current health status using 16 predictors. The model was developed using the landmarking approach, giving predicted survival curves up to 10 years from ages 18 to 50. Several models were compared using cross-validation.

Results

The final model has good discrimination (C-indexes 0.873, 0.843, 0.804 for 2-, 5-, 10-year survival prediction) and low prediction error (Brier scores 0.036, 0.076, 0.133). It identifies individuals at low and high risk of short- and long-term mortality based on their current status. For patients aged 20 during 2013-2015, for example, over 80% had a greater than 95% probability of 2-year survival and 40% were predicted to survive 10 years or more.

Conclusions

Dynamic personalised prediction models can guide treatment decisions and provide personalised information for patients. Our application illustrates the utility of the landmarking approach for making the best use of longitudinal and survival data and shows how models can be defined and compared in terms of predictive performance.

STEP 1

Enter values for predictor variables.

Please enter these carefully and read the notes about the values that are allowed for each predictor.

Example values are given.

```
# Current age in years (Range: 18 to 50. This must be a whole number.)
landmark.age = 30

# Sex: Male ('M') or female ('F')
sex = "M"

# Current FEV1 percent predicted (Range >0). This is assumed to be a value
# taken at an annual review visit when the patient is 'well'.
fev1.percent.predicted = 57

# Current FVC percent predicted (Range >0). This is assumed to be a value
# taken at an annual review visit when the patient is 'well'.
fvc.percent.predicted = 77

# Genotype: 'F508del-homozygous' (2 copies of F508del),
# 'F508del-heterozygous' (1 copy of F508del), 'other' (0 copies of F508del)
genotype = "F508del-homozygous"

# age of diagnosis in years (Range 0-50. This does not have to be a whole
# number, e.g. you can enter 0.5.)
age.of.diagnosis = 0.54

# Pseudomonas aeruginosa infection in the past year
pseudomonas.aeruginosa.infection = "yes"

# Burkholderia cepacia infection in the past year
burkholderia.cepacia.infection = "no"

# Staphylococcus aureus infection in the past year
staph.aureus.infection = "no"

# MRSA (Methicillin-resistant Staphylococcus aureus) infection in the past
# year
MRSA.infection = "no"

# Has the patient pancreatic insufficient: 'yes' or 'no'
pancreatic.insufficient = "yes"

# Has the patient been diagnosed with CF-related diabetes: 'yes' or 'no'
cfrd = "yes"
```

```r
# Current weight in kilograms (kg) (Range >0)
weight = 64

# height in centimetres (cm) (Range >0)
height = 167

# The number of days the patient has been in hospital to receive IV
# antibiotics in the past year#
#'0 days', '1-14 days', '15-28 days', '29+ days'
hospital.IVs = "0 days"

# The number of days the patient has been at home receiving IV antibiotics
# in the past year#
#'0 days', '1-14 days', '15-28 days', '29+ days'
home.IVs = "0 days"

# In the past year, has the patient been hospitalised for reasons other than
# receiving IVs: 'yes' or 'no'
hospital.nonIV = "no"
```

STEP 2

Give the survival time of interest (in years). You can enter any time from 0 to 10.

You can also specify would you would like to see a full survivor curve up to 10 years

---

```r
# If you are interested in 2-year survival enter 'time.for.survival=2'.
# This will be combined with the landmark age specified above.  So, if you
# entered landmark.age=18 and time.for.survival=2, the code below will
# provide an estimate of survival to age 20 for an individual currently aged
# 18 and with the values for the predictors entered above.
time.for.survival = 10

# Specify whether you would also like to see a survival curve.  This gives a
# plot of the estimated probability of survival to up to 10 years from the
# current age.
survival.curve = "yes"  #enter 'yes' or 'no'
```

STEP 3

FROM THIS POINT ONWARDS, YOU DO NOT NEED TO ENTER ANY VALUES OR MAKE ANY CHANGES TO THE CODE.

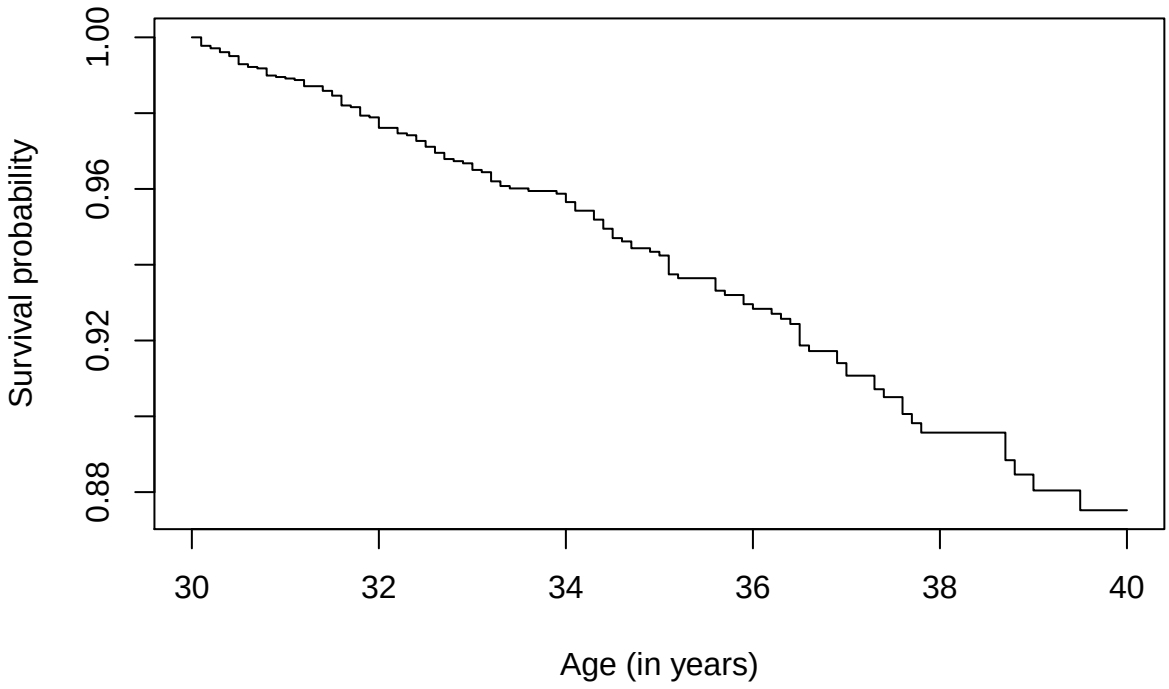
RUN THE CODE BELOW TO OBTAIN ESTIMATED SURVIVAL PROBABILITIES AND AN ESTIMATED SURVIVOR CURVE, GIVEN THE DETAILS ENTERED ABOVE.

```
#--------------------------------------
# read in the estimated baseline cumulative hazards and estimated log hazard
# ratios
#--------------------------------------

load(file = "./baseline_cumulative_hazards.RData")  #baseline.cumulative.hazards

load(file = "./times.RData")  #times

load(file = "./log_hazard_ratios.RData")  #log.hazard.ratios

#--------------------------------------
# obtain estimates of the probability of survival for a given number of
# years (time.for.survival: up to 10 years from the current age)
#--------------------------------------

if (!is.null(time.for.survival)) {
    risk.score = log.hazard.ratios["sex"] * (sex == "F") + log.hazard.ratios["fev1"] *
        fev1.percent.predicted + log.hazard.ratios["fvc"] * fvc.percent.predicted +
        log.hazard.ratios["genotype.1"] * (genotype == "F508del-heterozygous") +
        log.hazard.ratios["genotype.0"] * (genotype == "other") + log.hazard.ratios["age.diagnosis"] *
        age.of.diagnosis + log.hazard.ratios["p.aeruginosa"] * (pseudomonas.aeruginosa.infection ==
        "yes") + log.hazard.ratios["b.cepacia"] * (burkholderia.cepacia.infection ==
        "yes") + log.hazard.ratios["s.aureus"] * (staph.aureus.infection ==
        "yes") + log.hazard.ratios["mrsa"] * (MRSA.infection == "yes") + log.hazard.ratios["panc.insuff
        (pancreatic.insufficient == "yes") + log.hazard.ratios["cfrd"] * (cfrd ==
        "yes") + log.hazard.ratios["weight"] * weight + log.hazard.ratios["height"] *
        height + log.hazard.ratios["year"] * 10 + log.hazard.ratios["hospital.nonIV"] *
        (hospital.nonIV == "yes") + log.hazard.ratios["hospital.IVs:1-14"] *
        (hospital.IVs == "1-14 days") + log.hazard.ratios["hospital.IVs:15-28"] *
        (hospital.IVs == "15-28 days") + log.hazard.ratios["hospital.IVs:29+"] *
        (hospital.IVs == "29+ days") + log.hazard.ratios["home.IVs:1-14"] *
        (home.IVs == "1-14 days") + log.hazard.ratios["home.IVs:15-28"] * (home.IVs ==
        "15-28 days") + log.hazard.ratios["home.IVs:29+"] * (home.IVs == "29+ days")

    baseline.cumulative.hazard.landmark = na.omit(baseline.cumulative.hazards[,
        paste0("lm", landmark.age)])
    times.landmark = na.omit(times[, paste0("lm", landmark.age)])
    baseline.hazard.landmark = c(baseline.cumulative.hazard.landmark[1], diff(baseline.cumulative.hazar
        lag = 1))
```

```
    survival.probability = exp(-exp(risk.score) * sum(baseline.hazard.landmark[which(times.landmark <=
        landmark.age + time.for.survival)]))

    return(unlist(list(paste("Given the patient is now aged", landmark.age,
        ", and given the values provided for the predictors"), paste(" the patient's probability of surv
        landmark.age + time.for.survival, "is", round(survival.probability,
            3)))))

}
```

```
## [1] "Given the patient is now aged 30 , and given the values provided for the predictors"
## [2] " the patient's probability of survival to age 40 is 0.875"
```

```
#-------------------------------------
# obtain an estimated surivor curve up to 10 years from the current age
#-------------------------------------

if (survival.curve == "yes") {

    grid.times = seq(0, 10, 0.1)

    risk.score = log.hazard.ratios["sex"] * (sex == "F") + log.hazard.ratios["fev1"] *
        fev1.percent.predicted + log.hazard.ratios["fvc"] * fvc.percent.predicted +
        log.hazard.ratios["genotype.1"] * (genotype == "F508del-heterozygous") +
        log.hazard.ratios["genotype.0"] * (genotype == "other") + log.hazard.ratios["age.diagnosis"] *
        age.of.diagnosis + log.hazard.ratios["p.aeruginosa"] * (pseudomonas.aeruginosa.infection ==
        "yes") + log.hazard.ratios["b.cepacia"] * (burkholderia.cepacia.infection ==
        "yes") + log.hazard.ratios["s.aureus"] * (staph.aureus.infection ==
        "yes") + log.hazard.ratios["mrsa"] * (MRSA.infection == "yes") + log.hazard.ratios["panc.insuff
        (pancreatic.insufficient == "yes") + log.hazard.ratios["cfrd"] * (cfrd ==
        "yes") + log.hazard.ratios["weight"] * weight + log.hazard.ratios["height"] *
        height + log.hazard.ratios["year"] * 10 + log.hazard.ratios["hospital.nonIV"] *
        (hospital.nonIV == "yes") + log.hazard.ratios["hospital.IVs:1-14"] *
        (hospital.IVs == "1-14 days") + log.hazard.ratios["hospital.IVs:15-28"] *
        (hospital.IVs == "15-28 days") + log.hazard.ratios["hospital.IVs:29+"] *
        (hospital.IVs == "29+ days") + log.hazard.ratios["home.IVs:1-14"] *
        (home.IVs == "1-14 days") + log.hazard.ratios["home.IVs:15-28"] * (home.IVs ==
        "15-28 days") + log.hazard.ratios["home.IVs:29+"] * (home.IVs == "29+ days")

    baseline.cumulative.hazard.landmark = na.omit(baseline.cumulative.hazards[,
        paste0("lm", landmark.age)])
    times.landmark = na.omit(times[, paste0("lm", landmark.age)])
    baseline.hazard.landmark = c(baseline.cumulative.hazard.landmark[1], diff(baseline.cumulative.hazar
        lag = 1))

    survival.probability = exp(-exp(risk.score) * sapply(grid.times, FUN = function(x) {
        sum(baseline.hazard.landmark[which(times.landmark <= landmark.age +
            x)])
    }))

    plot(landmark.age + grid.times, survival.probability, type = "s", xlab = "Age (in years)",
        ylab = "Survival probability")

}
```


Survival probability
1.00
0.96
0.92
0.88
30
32
34
36
38
40
Age (in years)