Landmarking: An R package for analysis using landmark models

Isobel Barrott [1] Jessica Barrett [1]

[1]MRC Biostatistics Unit, University of Cambridge

What is the landmark model used for?

Landmark models are used to **dynamically predict** the risk of an event for an individual. This means that an individual has a personalised risk prediction which is updated as new information is collected about them. Landmark models are particularly applicable in a medical setting where electronic health records (EHRs), such as GP records, are continually being updated with new data.

What is the landmark model?

The key idea behind landmark models is the concept of predefined **landmark times**. A landmark time is a time point that we wish to make a risk prediction at. A model is built at the landmark time and this model is developed using only the data of individuals at risk (i.e. not censored or having experienced an event) at the landmark time. As we want to predict the risk of an event for an individual at a selection of time points, we define multiple landmark times and develop a model for each of these.

To form a model at each landmark age, landmark models use a 2-stage approach that is comprised of a **longitudinal submodel** and **survival submodel**. The typical form of the longitudinal submodel and survival submodel is the last observation carried forward model (LOCF) and the Cox proportional hazards model respectively, although there are extensions and these are discussed in the next section. Figure 1 shows the time intervals of the repeat measures data and time-to-event data that is included when using the model.

What are the advantages of the landmark model?

The other main statistical framework for this dynamic prediction problem is joint modelling. In comparison to joint modelling, the landmarking model is less computationally intensive and makes a weaker assumption about proportional hazards. Specifically this assumption only applies to the period between the landmark time and time horizon when using the landmark model.

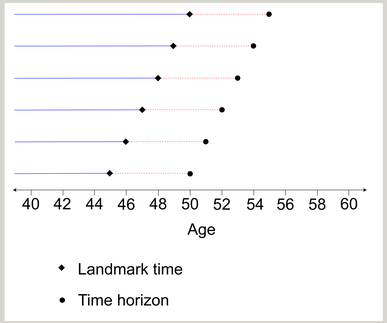

Figure 1. Illustration of a landmark model with six landmark times (ages 45 to 50 years). The blue solid line represents the time interval of repeat measures data to be inputted into the longitudinal submodel. The red dashed line represents the time interval of the time-to-event data to be inputted into the survival submodel. Events after the time horizon should be censored.

Features of the R package Landmarking

The package **Landmarking** offers benefits over existing packages *dynpred* and *landpred* for landmark model analysis. In particular, three features of the package are outlined below.

1. Linear mixed effects modelling (LME)

The LME model is an alternative to using the LOCF model for the longitudinal submodel. Using LME modelling means that the entire history of a risk factor is used to evaluate predicted risk, as opposed to the LOCF model where only one (the most recent) measurement is used. Figure 2 illustrates the use of these two types of models to predict systolic blood pressure at the landmark time. The LOCF model simply uses the most recently observed value of a risk factor. On the other hand, the LME model uses the entire history of the values of the risk factor from the individual, in combination with the trend in the risk factor from individuals used to develop the model. Moreover, using the Landmarking package, it is possible include a covariance structure in the random effects meaning that other risk factors can inform predictions. This improves the accuracy of the estimated value of the risk factor at the landmark time value compared to the LOCF model.

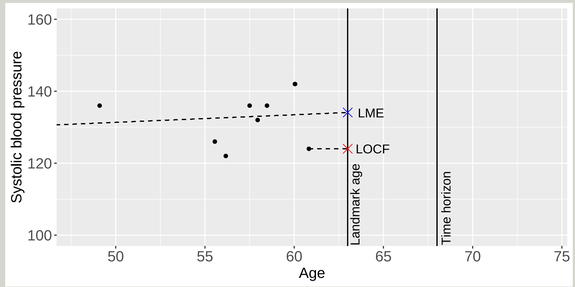

Figure 2. Comparison of the LOCF and LME model to predict systolic blood pressure at a landmark age for an individual with 8 repeat measures.

2. Competing risks modelling

The Landmarking R package has the ability to fit competing risks models for the survival submodel. There are two types of competing risks model that can be fit using this package, the Fine Gray model and the cause-specific model, using the parameter *survival_submodel* in the function *fit_LOCF_landmark_model* or *fit_LME_landmark_model*. The cause-specific model involves regression on the cause-specific hazard function. A covariate with a hazard rate greater than 1 is associated with an increased rate of the event of interest and less than 1 with a decreased rate of the event. Conversely, the Fine Gray model involves regression on the cumulative incidence function (CIF). In this case, a covariate with a hazard rate greater than 1 is associated with an increased incidence of the event of interest and less than 1 with a decreased incidence of the event. Their differences has led to the proposal that the cause-specific model is preferred when the aim is to study the aetiology of a disease, and the Fine Gray model is preferred when the aim is prediction of the risk of an event.

3. k-fold cross validation

Another feature of this package is that is makes it easy to perform k-fold cross validation. The function *add_cv_number* assigns a fold to each individual in a dataset. This can be inputted as a parameters within the functions *fit_LOCF_landmark_model* or *fit_LME_landmark_model*, which ensure the same folds are used for both the longitudinal and survival submodels.

The Landmarking R package - Example

The package can be found at:
https://github.com/isobelbarrott/Landmarking

To show how to use the package we use a simple example which demonstrates its steps.

```
#Load package and dataset
library(Landmarking)
data(data_repeat_outcomes)
#
#Select individuals in the risk set at the landmark time
#and censor individuals at the horizon time.
data_landmark <- create_landmark_dataset(data = data_repeat_outcomes, x_L = 60, x_hor
= 65, assessment_time = "response_time_sbp_stnd", patient_id = "id", event_time =
"event_time", event_status = "event_status")
#
#Remove individuals with no LOCF value at landmark time
data_landmark <- return_ids_with_LOCF( data = data_landmark, patient_id = "id", covariates
= c("ethnicity", "smoking", "diabetes", "sbp_stnd", "tchdl_stnd"), covariates_time =
c(rep("response_time_sbp_stnd", 4), "response_time_tchdl_stnd"), x_L = 60)
#
#Add 10-fold cross validation number
data_landmark_cv <- add_cv_number(data = data_landmark, patient_id = "id", k = 10)
#
#Fit the landmark model with LOCF longitudinal model
#and cause-specific survival model
data_model_landmark_LOCF <- fit_LOCF_landmark_model( data = data_landmark_cv, x_L = 60,
x_hor = 65, covariates = c("ethnicity", "smoking", "diabetes", "sbp_stnd", "tchdl_stnd"),
covariates_time = c(rep("response_time_sbp_stnd", 4), "response_time_tchdl_stnd"), cv_name
= "cross_validation_number", patient_id = "id", event_time = "event_time", event_status =
"event_status", survival_submodel = "cause_specific" )
data_landmark_LOCF <- data_model_landmark_LOCF$data
#
#Calculate c-index and Brier score
#and cause-specific survival model
model_assessment_LOCF <- get_model_assessment( data = data_landmark_LOCF, patient_id =
"id", event_prediction = "event_prediction", event_status = "event_status", event_time
= "event_time", x_hor = 65, return_c_index = TRUE, return_brier_score = TRUE, b = 50,
standard_error = TRUE )
```

These steps need to be repeated for each landmark time that has been chosen.

Please see the package vignette for more details about using this package.

References

[1] Ruth H Keogh, Shaun R Seaman, Jessica K Barrett, David Taylor-Robinson, and Rhonda Szczesniak. Dynamic prediction of survival in cystic fibrosis: a landmarking analysis using uk patient registry data. *Epidemiology (Cambridge, Mass.)*, 30(1):29, 2019.

[2] Ellie Paige, Jessica Barrett, David Stevens, Ruth H Keogh, Michael J Sweeting, Irwin Nazareth, Irene Petersen, and Angela M Wood. Landmark models for optimizing the use of repeated measurements of risk factors in electronic health records to predict future disease risk. *American journal of epidemiology*, 187(7):1530–1538, 2018.

[3] Hans C Van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

[4] Marcel Wolbers, Michael T Koller, Vianda S Stel, Beat Schaer, Kitty J Jager, Karen Leffondre, and Georg Heinze. Competing risks analyses: objectives and approaches. *European heart journal*, 35(42):2936–2941, 2014.

isobel.barrott@mrc-bsu.cam.ac.uk