Recent infection test calibration

29 April 2017

This vignette covers the use of functions `mdrical` and `frrcal`.

Introduction

Incidence estimates from cross-sectional surveys using biomarkers for 'recent infection' require that the test for recent infection (usually an adapted diagnostic assay) be accurately characterised. The two critical parameters of test performance are the Mean Duration of Recent Infection (MDRI), denoted Ω_T, (with T the recency cutoff time), and False Recent Rate (FRR), denoted β_T. The explicit time cutoff T was introduced by Kassanjee et al. *Epidemiology*, 2012.[1] to differentiate between 'true recent' and 'false recent' results. Also see Kassanjee, McWalter, Welte. *AIDS Research and Human Retroviruses*, 2014.[2], which notes:

> To lead to an informative estimator, this cut-off, though theoretically arbitrary, must be chosen to reflect the temporal dynamic range of the test for recent infection; i.e. at a time T post infection, the overwhelming majority of infected people should no longer be testing "recent", and furthermore, T should not be larger than necessary to achieve this criterion.[3]

MDRI is defined as the average time alive and returning a 'recent' result, while infected for times less than T. FRR is defined as a cross sectional context specific proportion of subjects returning a 'recent' result while infected for longer than T.

Test performance may be context-specific, and therefore, where available, local data should be used to calibrate tests. Often cross-sectional incidence surveys use a multi-step Recent Infection Testing Algorithm (RITA) and then the entire RITA must be appropriately calibrated. This may involve adapting MDRI estimates to account for the sensitivity of screening tests, or adapting FRR estimates based on weighted estimates for subpopulations such as treated individuals. Calibration should be performed using the same set of biomarkers used in a RITA, such as a viral load threshold to reduce false recency.

Estimating MDRI using binomial regression

This package provides the function `mdrical` to estimate MDRI for a given biomarker or set of biomarkers from a dataset of based on the test being applied to well-characterised specimens and subjects. That is, time since 'infection' should be well-known, as well as test result(s). While 'infection time' can be variously defined as refering to the exposure event, date of first detectability on an RNA assay, Western Blot seroconversion, etc., it should be consistently used. If the reference event used in test calibration differs from test conversion on the screening assay or algorithm that is used define someone as HIV-positive in a RITA, then the MDRI needs to be appropriately adapted to cater for this difference.

Function `mdrical` estimates MDRI by fitting a model for the probability of testing 'recent' as a function of time since infection $P_R(t)$. As an option, one of two functional forms (with their associated link functions) can be selected by the user. Fitting is performed using a generalised linear model (as implemented in the *glm2* package) to estimate parameters.

The linear binomial regression model takes the following form, with $g()$ the link function

[1] Kassanjee, R., McWalter, T.A., Baernighausen, T. and Welte, A. "A new general biomarker-based incidence estimator." Epidemiology; 2012, 23(5): 721-728.

[2] Kassanjee, R., McWalter, T.A. and Welte, A. "Short Communication: Defining Optimality of a Test for Recent Infection for HIV Incidence Surveillance." AIDS Research and Human Retroviruses; 2014, 30(1): 45-49.

[3] Kassanjee, R., McWalter, T.A., Baernighausen, T. and Welte, A. "A new general biomarker-based incidence estimator." Epidemiology; 2012, 23(5): 721-728.

$$g(P_R(t)) = f(t) \tag{1}$$

If the argument `functional_forms` is specified with the value `"cloglog_linear"`, $g()$ is the complementary log-log link function and $\ln(t)$ as linear predictor of $P_R(t)$, so that:

$$\ln\left(-\ln(1 - P_R(t)\right) = \beta_0 + \beta_1 \ln(t) \tag{2}$$

If the argument `functional_forms` is specified with the value `"logit_cubic"`, $g()$ is the complementary log-log link function and the linear predictor of $P_R(t)$ is a cubic polynomial in t, so that:

$$\ln\left(\frac{P_R(t)}{1 - P_R(t)}\right) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 \tag{3}$$

In both cases, MDRI is the integral of $P_R(t)$ from 0 to T.

$$\Omega_T = \int_0^T P_R(t)dt \tag{4}$$

The default behaviour is to implement both model forms if the argument `functional_forms` is omitted.

Confidence intervals are computed by means of subject-level bootstrapping, as measurements from subjects with more than one measurement in the dataset cannot be considered indpendent observations. An MDRI estimate is then calculated using the resampled data. The number of bootstrap iterations is specified using the argument `n_bootstraps`. We recommend 10,000 for reproducible confidence intervals and standard errors. To support subject level resampling, the subject identifier in the dataset must be specified using the `subid_var` argument.

In addition to specifying the value of T (using the argument `recency_cutoff_time`), an `inclusion_time_threshold` is required, to *exclude* data points beyond a certain time (post infection). This is to prevent falsely recent measurements from unduly affecting the fit between 0 and T. This should typically be a value somewhat (but not too much) larger than T.

To specify recency status, one can either supply a list of variables and thresholds (indicating in whether a result above or below the thresholds signifies recency) or specify the `recency_rule` as `"binary_data"`, in which case a 1 indicates recency.

Example of `mdrical` using the complementary log-log functional form and pre-classified data

Load the package in order to use it

```
library(inctools)
```

The dataset `excalibdata` contains example data from an evaluation of an assay measuring recency of infection. At an assay result of <10, the specimen is considered to be recently infected. It further contains viral load data, which is commonly used to reduce false recency. For example, when recency is defined as assay result <10 and viral load > 1000, the FRR is substantially lower (but the MDRI is also reduced).

The first example provides a variable that has pre-classified results, and uses only the complementary log-log functional form.

Note: To keep compute time reasonable during execution of the example code, only 1000 bootstraps are performed. To obtain reasonable standard errors and confidence intervals, 10,000 bootstraps are recommended.

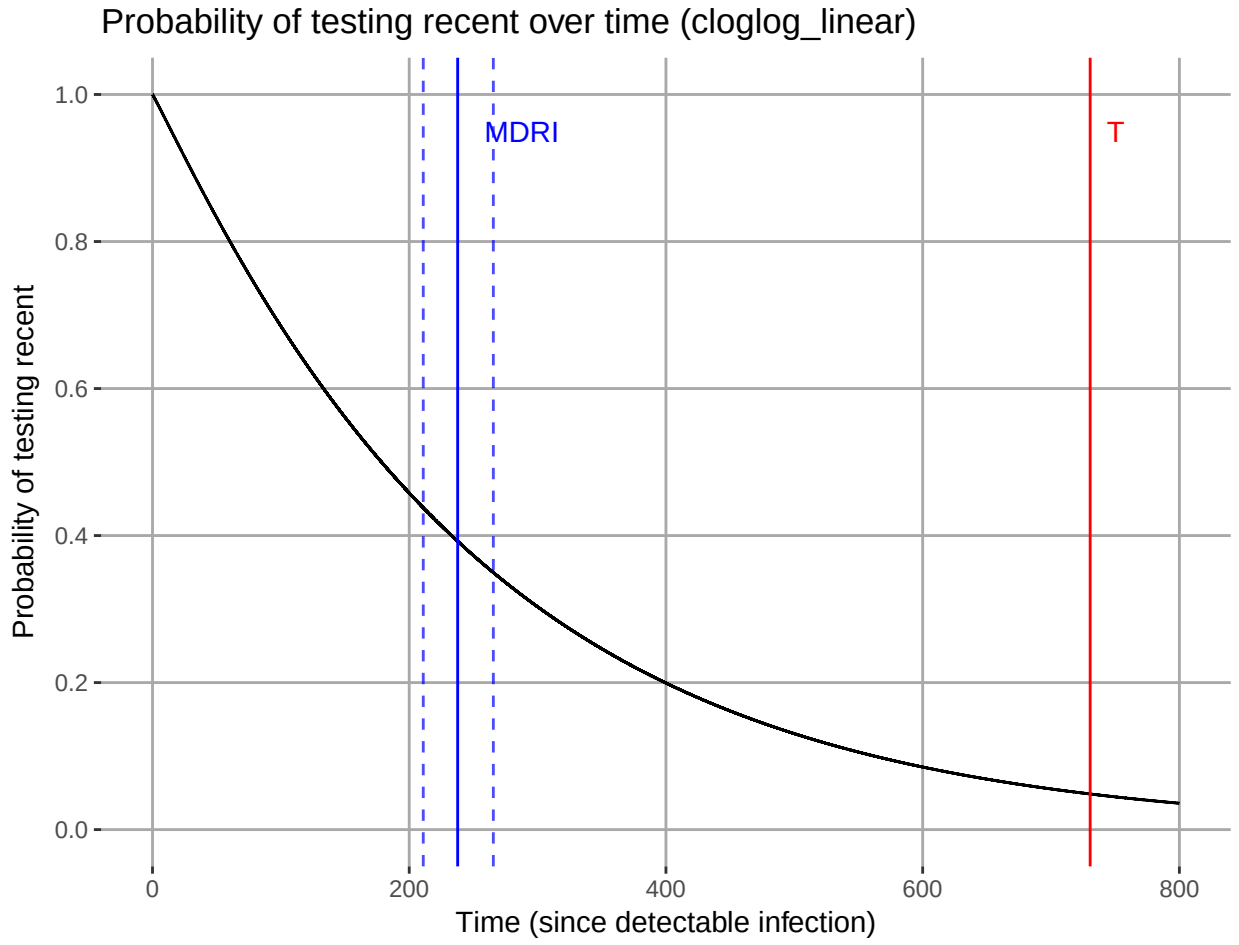
```r
mdri <- mdrical(data=excalibdata,
                subid_var = "SubjectID",
                time_var = "DaysSinceEDDI",
                recency_cutoff_time = 730.5,
                inclusion_time_threshold = 800,
                functional_forms = c("cloglog_linear"),
                recency_rule = "binary_data",
                recency_vars = "Recent",
                n_bootstraps = 1000,
                alpha = 0.05,
                plot = TRUE)
```

```r
print(mdri)
```

```
## $MDRI
##                       PE    CI_LB    CI_UB      SE n_recent n_subjects
## cloglog_linear 237.7463 210.9329 265.5173 13.5198      270        304
##                n_observations
## cloglog_linear            708
##
## $Plots
## $Plots$cloglog_linear
```



```
##
##
## $Models
## $Models$cloglog_linear
##
## Call:  glm2::glm2(formula = (1 - recency_status) ~ 1 + I(log(time_since_eddi)),
##     family = stats::binomial(link = "cloglog"), data = data,
##     control = stats::glm.control(epsilon = tolerance, maxit = maxit,
##         trace = FALSE))
##
## Coefficients:
##             (Intercept)  I(log(time_since_eddi))
##                  -5.786                   1.045
##
## Degrees of Freedom: 707 Total (i.e. Null);  706 Residual
## Null Deviance:       941.2
## Residual Deviance: 714   AIC: 718
```



Example of `mdrical` using the logit cubic functional form and two independent thresholds on biomarkers

This example also specifies a vector of variables and a vector of paramaters to define recency, using the assay result and the viral load. The paramaters in the vector `c(10,0,1000,1)` mean that recency is defined as an assay biomarker reading below 10 and a viral load reading above 1000.

Note: To keep compute time reasonable during execution of the example code, only 1000 bootstraps are performed. To obtain reasonable standard errors and confidence intervals, 10,000 bootstraps are recommended.

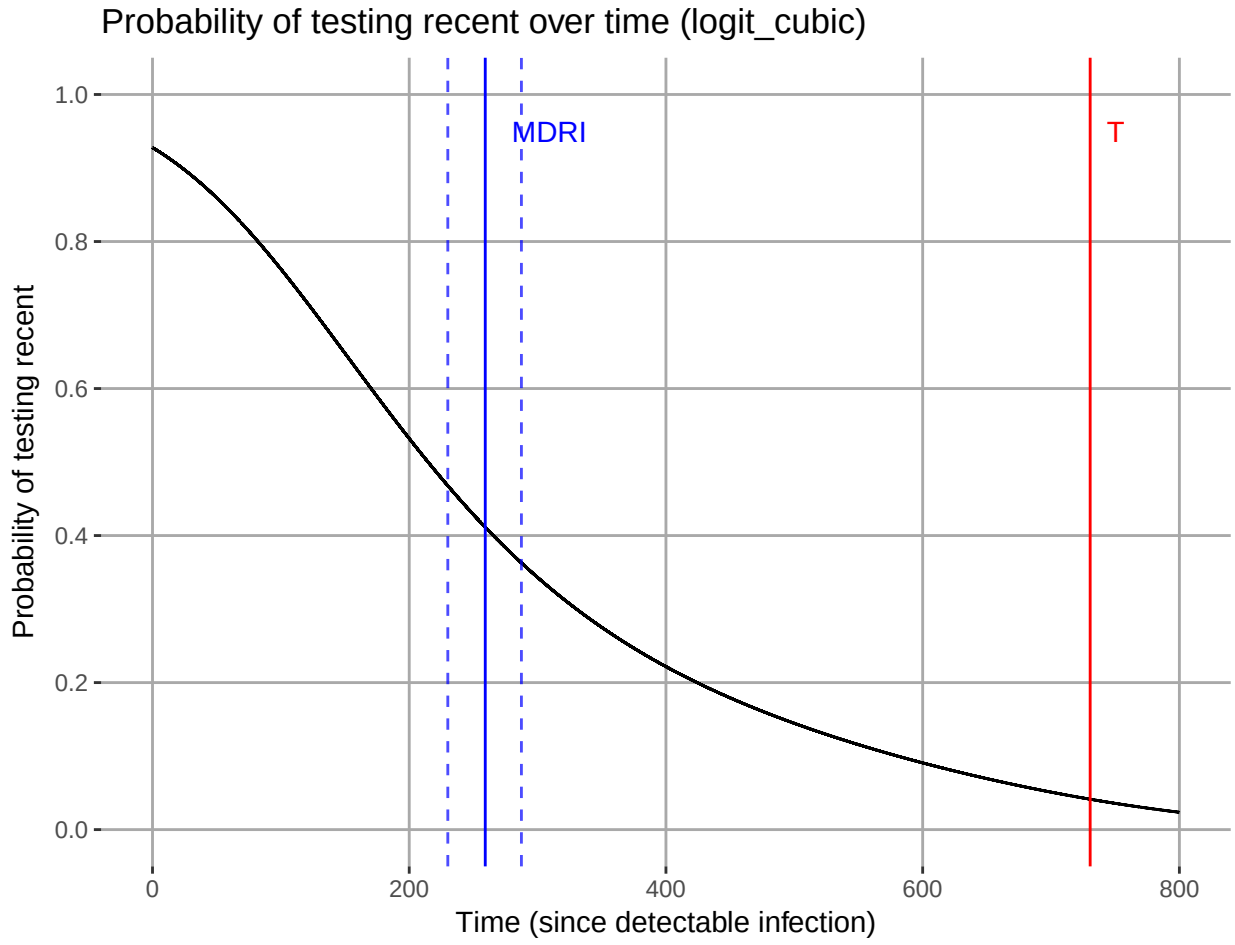
```
mdri <- mdrical(data=excalibdata,
                subid_var = "SubjectID",
                time_var = "DaysSinceEDDI",
                recency_cutoff_time = 730.5,
                inclusion_time_threshold = 800,
                functional_forms = c("logit_cubic"),
                recency_rule = "independent_thresholds",
                recency_vars = c("Result","VL"),
                recency_params = c(10,0,1000,1),
                n_bootstraps = 1000,
                alpha = 0.05,
                plot = TRUE)
```

```r
print(mdri)
```

```
## $MDRI
##                   PE    CI_LB    CI_UB      SE n_recent n_subjects
## logit_cubic 259.2234 230.0616 287.4033 14.5534      270        295
##             n_observations
## logit_cubic            644
##
## $Plots
## $Plots$logit_cubic
```



```
##
##
## $Models
## $Models$logit_cubic
##
## Call:  glm2::glm2(formula = recency_status ~ 1 + I(time_since_eddi) +
##     I(time_since_eddi^2) + I(time_since_eddi^3), family = stats::binomial(link = "logit"),
##     data = data, control = stats::glm.control(epsilon = tolerance,
##         maxit = maxit, trace = FALSE))
##
## Coefficients:
##          (Intercept)    I(time_since_eddi)  I(time_since_eddi^2)
##            2.554e+00            -1.591e-02             2.184e-05
## I(time_since_eddi^3)
##           -1.471e-08
##
## Degrees of Freedom: 643 Total (i.e. Null);  640 Residual
## Null Deviance:       875.9
## Residual Deviance: 635.3     AIC: 643.3
```



Example of `mdrical` in which bootstraps are run in parallel

As above, but asking for both functional forms and parellelising the bootstrapping. In this case, the job is split over four cores.

Note: To make sure this vignette builds in a reasonable time, the example code, but not the output, is shown.

```
mdrical(data=excalibdata,
                  subid_var = "SubjectID",
                  time_var = "DaysSinceEDDI",
                  recency_cutoff_time = 730.5,
                  inclusion_time_threshold = 800,
                  functional_forms = c("logit_cubic","cloglog_linear"),
                  recency_rule = "independent_thresholds",
                  recency_vars = c("Result","VL"),
                  recency_params = c(10,0,1000,1),
                  n_bootstraps = 10000,
                  alpha = 0.05,
                  plot = TRUE,
                  parallel = TRUE,
                  cores=4)
```

Estimating FRR using binomial proportions with `frrcal`

FRR is simply the binomially estimated probability of a *subject's* measurements post-T being 'recent' on the recency test. A binomial exact test is performed using `binom.test`. All of a subject's measurements post-T are evaluated and if the majority are recent, the subject is considered to have measured falsely recent. Inversely, if a majority are non-recent, the subject contributes a 'true recent' result. Each subject represents one trial. In the case that exactly half of a subject's measurements are recent, they contribute 0.5 to the outcomes (which are rounded up to the nearest integer over all subjects).

This example calculates a false-recent rate, treating the data at subject level:

```
frrcal(data=excalibdata,
             subid_var = "SubjectID",
             time_var = "DaysSinceEDDI",
             recency_cutoff_time = 730.5,
             recency_rule = "independent_thresholds",
             recency_vars = c("Result","VL"),
             recency_params = c(10,0,1000,1),
             alpha = 0.05)
```

```
##  FRRest     LB     UB alpha n_recent n_subjects n_observations
##  0.0301 0.0131 0.0584  0.05        8        266            732
```