2019 Spanish Stata Conference
Madrid | 17 October 2019

# bsvalidation

## comando postestimación para la validación interna de modelos predictivos de regresión logística

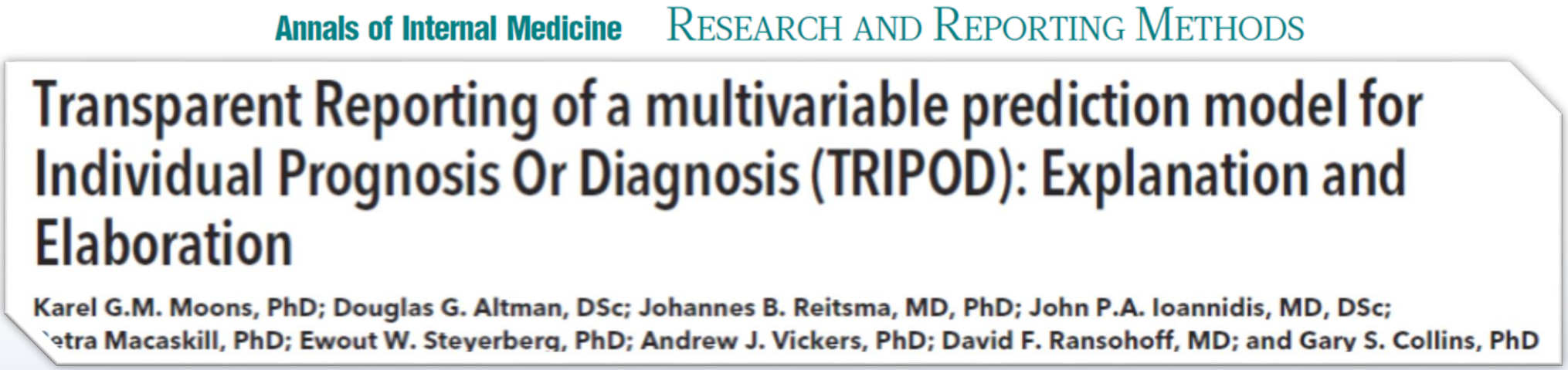
**Fernandez-Felix BM**, García-Esquinas E, Pérez T, Muriel A, Zamora J
Unidad de Bioestadística Clínica - Hospital Universitario Ramón y Cajal

# bsvalidation

En los modelos predictivos validar significa ver si el modelo predecirá bien la variable dependiente en nuevos individuos.

TRIPOD

TRIPOD Checklist: Prediction Model Development


Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

# bsvalidation

```
bsvalidation varlist [if], [dummy1()] [dummy2()] reps() [seed()] [pr()] [pe()] [graph] [group()]
```

| | |
|---|---|
| **varlist** | Lista de variables que se han evaluado pero no se han incluido en el modelo final. |
| **if** | Opción de selección de los datos sobre los que el comando trabajará. |
| **dummy1 y 2(#)** | Variables indicadoras o dummy. |
| **reps(#)** | Es la única opción obligatoria del comando. Indica el número de muestras bootstrap obtenidas desde la muestra original. Número máximo en la versión IC es de 800. |
| **seed(#)** | Semilla de aleatorización empleada en el proceso de muestreo bootstrap. |
| **pr(#)** | Nivel de significación para eliminar una variable del modelo. |
| **pe(#)** | Nivel de significación para añadir una variable al modelo. |
| **graph** | Genera los gráficos de calibración y discriminación del modelo. |
| **group(#)** | Número de grupos que se emplearán para generar los gráficos de calibración y en el test de Hosmer-Lemeshow. |

Sintaxis

# bsvalidation

- Rendimiento aparente
  - ***Discriminación***
    - AUC ROC
      `lroc`
  - ***Calibración***
    - Test de Hosmer-Lemeshow
      `estat gof`

- Rendimiento bootstrap
  - ***Discriminación***
    - AUC ROC ajustado por optimismo
  - ***Calibración***
    - Pendiente de calibración
    - Constante de calibración
  - ***Frecuencia de aparición de cada variable***
  - ***Gráficos***

# Ejemplo

**Base de datos:** nhanes2d.dta

**Objetivo:** Desarrollar y validar mediante técnicas bootstrap un modelo predictivo de heart attack.

**Variable dependiente:** heartatk

**Variables independientes:** height weight age female (black orace) diabetes highbp tcresult (region2 region3 region4) houssiz

**Estrategia de modelización:** Backward selection ($\alpha = 0.05$)

# bsvalidation

```
webuse nhanes2d

* Modelo máximo
logistic heartatk height weight age female (black orace) diabetes highbp
tcresult (region2 region3 region4) houssiz

* Estrategia backward selection (α = 0.05)
...
...

* Modelo final
logistic heartatk age female diabetes tcresult (region2 region3 region4)
```

```
* Validación bootstrap
bsvalidation height weight highbp houssiz, dummy1(black orace) reps(200)
seed(1485) pr(0.05) graph group(10)
```

# Modelo máximo

Model maximum (The maximum model)

Logistic regression — Number of obs = 1,463; LR chi2(13) = 512.70; Prob > chi2 = 0.0000; Log likelihood = -666.57439; Pseudo R2 = 0.2778

| heartatk | Odds Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | 1.098593 | .0078285 | 13.20 | 0.000 | 1.083356 | 1.114044 |
| female | .3286807 | .0640385 | -5.71 | 0.000 | .2243521 | .4815242 |
| diabetes | 1.651728 | .3732161 | 2.22 | 0.026 | 1.060731 | 2.572004 |
| tcresult | 1.00264 | .0013895 | 1.90 | 0.057 | .9999201 | 1.005367 |
| region2 | 1.756602 | .3635485 | 2.72 | 0.006 | 1.170868 | 2.635355 |
| region3 | 1.321478 | .2658668 | 1.39 | 0.166 | .8908563 | 1.960254 |
| region4 | 1.902475 | .3910589 | 3.13 | 0.002 | 1.271605 | 2.846333 |
| weight | 1.005773 | .0053801 | 1.08 | 0.282 | .9952836 | 1.016374 |
| highbp | 1.13974 | .1613023 | 0.92 | 0.355 | .8636526 | 1.504086 |
| height | .9923349 | .0111999 | -0.68 | 0.495 | .9706246 | 1.014531 |
| houssiz | 1.022361 | .0526476 | 0.43 | 0.668 | .92421 | 1.130936 |
| black | 1.051649 | .2415631 | 0.22 | 0.826 | .6704254 | 1.649648 |
| orace | 1.178708 | .7974356 | 0.24 | 0.808 | .3129964 | 4.438873 |
| _cons | .0027971 | .0056289 | -2.92 | 0.003 | .0000542 | .1444294 |

Note: _cons estimates baseline odds.

# Modelo final

Model final (The final model)

Logistic regression — Number of obs = 1,463; LR chi2(7) = 509.50; Prob > chi2 = 0.0000; Log likelihood = -668.17141; Pseudo R2 = 0.2760

| heartatk | Odds Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | 1.098219 | .0069421 | 14.82 | 0.000 | 1.084697 | 1.11191 |
| female | .3398259 | .0471142 | -7.78 | 0.000 | .2589668 | .4459323 |
| diabetes | 1.726381 | .3863128 | 2.44 | 0.015 | 1.11343 | 2.676765 |
| tcresult | 1.002816 | .0013809 | 2.04 | 0.041 | 1.000113 | 1.005526 |
| region2 | 1.711613 | .3505669 | 2.62 | 0.009 | 1.145686 | 2.557089 |
| region3 | 1.28342 | .2527963 | 1.27 | 0.205 | .872382 | 1.888124 |
| region4 | 1.849009 | .3718546 | 3.06 | 0.002 | 1.246677 | 2.742359 |
| _cons | .0013269 | .000667 | -13.18 | 0.000 | .0004954 | .0035537 |

Note: _cons estimates baseline odds.

# Rendimiento del modelo

```
Apparent performance
------------------------------------------------------------------------------
                  ROC area =   0.834     95%CI (0.813-0.854)
 Hosmer-Lemeshow chi2(8) =  13.444    Prob > chi2 = 0.097
              Brier score =   0.154
------------------------------------------------------------------------------

Bootstrap performance
Number of replications: 200
------------------------------------------------------------------------------
                  Optimism =   0.007
      ROC area (adjusted) =   0.827     95%CI (0.807-0.847)
        Calibration slope =   0.962
    Calibration intercept =  -0.013
------------------------------------------------------------------------------
```

# Frecuencia de variables

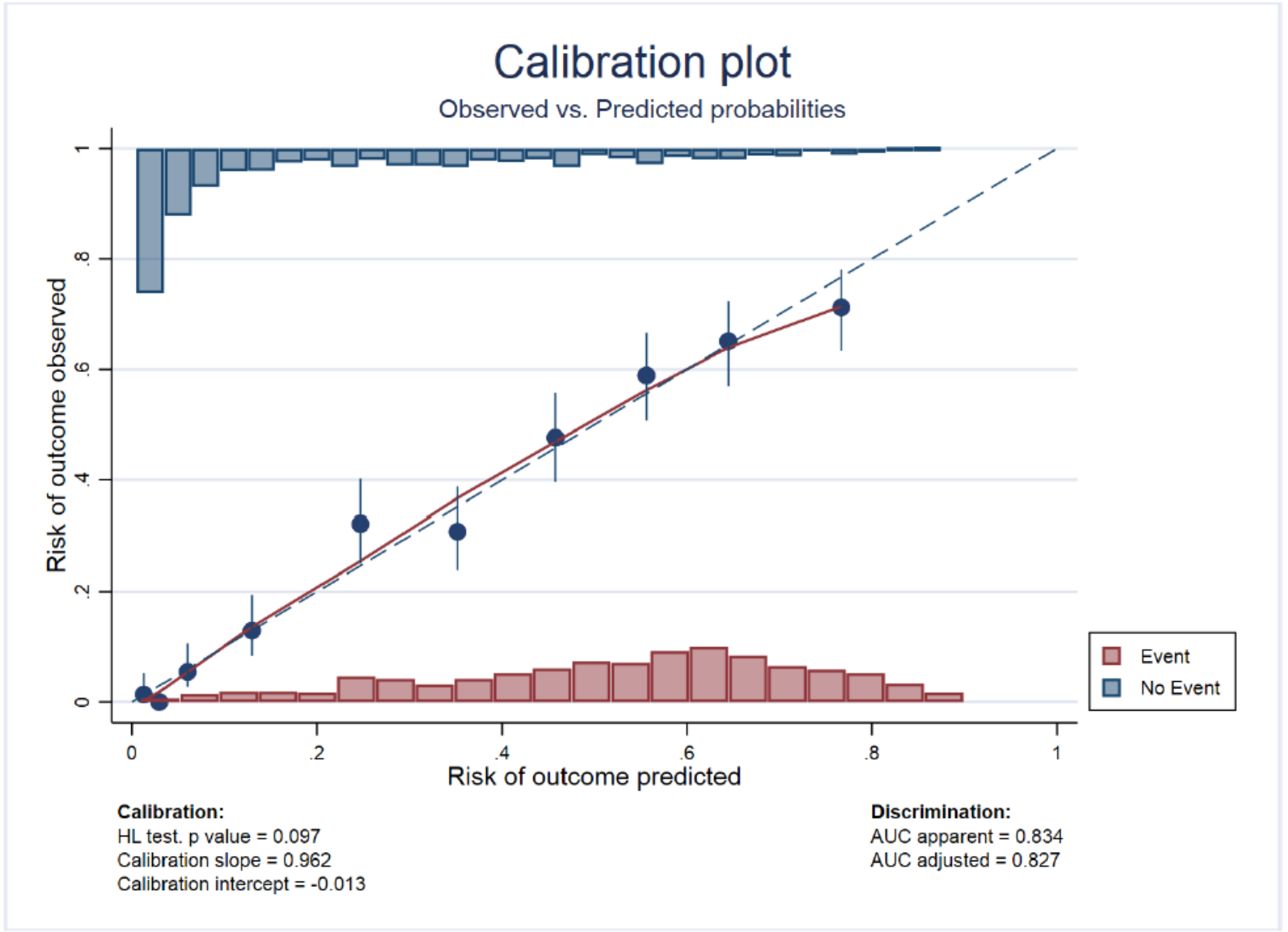
**Number of times each variable is selected**

| | Freq | % |
|---|---|---|
| age: | 200 | 100% |
| female: | 200 | 100% |
| diabetes: | 132 | 66% |
| tcresult: | 92 | 46% |
| region2: | 169 | 85% |
| region3: | 169 | 85% |
| region4: | 169 | 85% |
| weight: | 52 | 26% |
| highbp: | 45 | 23% |
| height: | 26 | 13% |
| houssiz: | 16 | 8% |
| black: | 8 | 4% |
| orace: | 8 | 4% |

**Number of variables in each model**

| | Freq | % |
|---|---|---|
| 2: | 6 | 3% |
| 3: | 7 | 4% |
| 4: | 14 | 7% |
| 5: | 16 | 8% |
| 6: | 42 | 21% |
| 7: | 67 | 34% |
| 8: | 37 | 19% |
| 9: | 10 | 5% |
| 10: | 1 | 1% |

# Gráfico


Calibration plot
Observed vs. Predicted probabilities
Risk of outcome observed
0
.2
.4
.6
.8
1
Risk of outcome predicted
Event
No Event
Calibration:
HL test. p value = 0.097
Calibration slope = 0.962
Calibration intercept = -0.013
Discrimination:
AUC apparent = 0.834
AUC adjusted = 0.827

# bsvalidation

## Fortalezas

- ***Aplicación sencilla***
- ***Automatización***
- ***Gráficos***
- ***Mejora del reporte***

## Debilidades

- ***Automatización***
- ***Código***
  - Eficiencia (optimizable)
  - Variables indicadoras
  - Predictores fijos
- ***Solo regresión logística***
- ***No estrategias complejas***
  - Ej. LASSO

# bsvalidation

## Futuro

- ***Mejorar el código***
- ***+ modelos de regresión***
- ***+ estrategias de selección de variables***
- ***Subir al repositorio de Stata***

# Muchas gracias

**Fernandez-Felix BM**, García-Esquinas E, Pérez T, Muriel A, Zamora J
Unidad de Bioestadística Clínica - Hospital Universitario Ramón y Cajal