TP1 Bioestdística

Irisarri - Landa

# TP 1: Validación de datos

**Alumnos:** Malena Irisarri, Román Landa

UNIVERSIDAD NACIONAL
DE ROSARIO
FACULTAD DE CIENCIAS
ECONÓMICAS Y ESTADÍSTICA


UNR

Rosario, Argentina

17 de Abril de 2025

# Introducción

Se llevó a cabo un estudio de investigación clínica multicéntrico con el fin de implementar estándares mundiales de crecimiento fetal que faciliten la detección temprana de alteraciones en el desarrollo del feto dentro del últero, y de esta manera reducir la morbi-mortalidad perinatal asociada con el crecimiento.

Durante el período de reclutamiento, las mujeres admitidas en alguna de las clínicas de salud participantes cursando su primer trimestre de embarazo fueron invitadas a formar parte del estudio. Aquellas que cumplieron con los criterios de elegibilidad y dieron su consentimiento fueron seguidas según un esquema de visitas programado a las 14, 18, 24, 28, 32, 36 y 40 semanas de gestación. En cada visita, se tomaron medidas antropométricas del feto por medio de un ultrasonido.

La información necesaria para llevar a cabo el análisis se recolectó a lo largo de 17 formularios en papel. Particularmente, en el formulario de admisión se registraron algunas características de la mujeres al momento de ingresar en el estudio. A las mujeres que cumplieron todos los criterios les fue asignado un código identificador único (Subject Number) compuesto por su código de páis, código del médico responsable, y el orden de ingreso. Sólo se entrevistaron mujeres mayores de edad (18 años o más al momento de la entrevista).

# Datos

Contamos con una base de 1000 pacientes, con las siguentes variables:

- countrycode: Código del país (tipo: numérico)
- Patientid: Identificación única del paciente (tipo: carácter)
- interview: Fecha de la entrevista (tipo: fecha)
- ethnicgroup: Grupo étnico (tipo: entero)
- scr: Elegibilidad según el formulario CLIN-SCR (tipo: entero) Este campo indica si la mujer es elegible de acuerdo con los criterios establecidos en el formulario CLIN-SCR. Un valor de 1 indica elegibilidad, mientras que un valor de 0 indica no elegibilidad.
- usscr: Elegibilidad según el formulario US-SCR (tipo: entero) Similar al anterior, este campo indica la elegibilidad de la mujer según el formulario US-SCR. Los valores son 1 para elegible y 0 para no elegible.
- consent: Consentimiento (tipo: entero) Este campo refleja si la mujer aceptó participar y escribió el formulario de consentimiento. Un valor de 1 indica que firmó el consentimiento, mientras que un valor de 0 indica que no lo hizo.
- subjectnumber : Número de sujeto (tipo: carácter) Este campo representa un número único para identificar al sujeto en el estudio, expresado como texto.

Los primeros registros son:

```
## # A tibble: 6 x 10
##   countrycode patientid interview  ethnicgroup   scr usscr consent subjectnumber
##         <int> <chr>     <date>           <int> <int> <int>   <int> <chr>
## 1          14 15/03/19~ 2014-02-03           3     2     2       2 014003017
## 2          31 19/03/19~ 2011-03-29           1     2     2       2 031001011
## 3          23 26/11/19~ 2014-06-25           1     2     2       2 023009024
## 4          11 18/04/19~ 2014-03-18           3     2     2       2 011002135
## 5          11 01/01/19~ 2014-02-25           3     2     2       2 011002120
## 6          31 13/12/19~ 2011-04-01           1     2     2       2 031002012
## # i 2 more variables: fechaid <date>, inicialesid <chr>
```

## Reglas

Se realizaron diversas validaciones para asegurar la integridad y calidad de los datos de los pacientes. A continuación, se detallan las comprobaciones efectuadas:

1.Identificador del Paciente: Se comprobó la existencia del identificador único asignado a cada registro (no valores faltantes).

2.Código de País: Se comprobó la existencia del código de país para cada paciente (no valores faltantes).

3.Formato del Código de País: Se validó que el código de país se registrara como un número entero.

4.Validez del Código de País: Se comprobó que el código de país correspondiera a una lista predefinida de códigos válidos.

5.Fecha de Identificación: Se comprobó la existencia de la fecha de identificación para cada paciente (no valores faltantes).

6.Formato de la Fecha de Identificación: Se verificó que la fecha de identificación se registrara en un formato de fecha válido.

7.Iniciales del Paciente: Se comprobó la existencia de las iniciales del paciente en cada registro (no valores faltantes).

8.Formato de las Iniciales: Se validó que las iniciales del paciente se registraran como texto.

9.Formato Específico de las Iniciales: Se comprobó que las iniciales consistieran exactamente en dos letras mayúsculas.

10.Fecha de la Entrevista: Se comprobó la existencia de la fecha de la entrevista para cada paciente (no valores faltantes).

11.Formato de la Fecha de la Entrevista: Se validó que la fecha de la entrevista se registrara en un formato de fecha válido.

12.Grupo Étnico: Se comprobó la existencia de información sobre el grupo étnico del paciente (no valores faltantes).

13.Formato del Grupo Étnico: Se validó que el grupo étnico se registrara como un número entero.

14.Validez del Grupo Étnico: Se comprobó que el código del grupo étnico correspondiera a una lista predefinida de valores válidos.

15.Resultado del Screening (SCR): Se comprobó la existencia del resultado del screening (no valores faltantes).

16.Validez del Resultado del Screening (SCR): Se comprobó que el resultado del screening correspondiera a una lista predefinida de valores válidos.

17.Resultado del Ultra-Screening (USSCR): Se comprobó la existencia del resultado del ultra-screening (no valores faltantes).

18.Validez del Resultado del Ultra-Screening (USSCR): Se comprobó que el resultado del ultra-screening correspondiera a una lista predefinida de valores válidos.

19.Consentimiento Informado: Se comprobó la existencia de información sobre el consentimiento informado del paciente (no valores faltantes).

20.Validez del Consentimiento Informado: Se comprobó que la información del consentimiento informado correspondiera a una lista predefinida de valores válidos.

21.Número de Sujeto: Se comprobó la existencia del número de sujeto asignado a cada paciente (no valores faltantes).

22.Obligatoriedad del Número de Sujeto: Se validó que el número de sujeto estuviera presente en los casos donde el paciente pasó las etapas de screening y consentimiento.

23.Consistencia del Número de Sujeto con el Código de País: Se comprobó que los primeros tres dígitos del número de sujeto coincidieran con el código de país del paciente.

24.Ausencia Inesperada del Número de Sujeto (SCR=1): Se verificó que el número de sujeto no estuviera presente en los casos donde el resultado del screening fue negativo.

25.Ausencia Inesperada del Número de Sujeto (USSCR=1): Se verificó que el número de sujeto no estuviera presente en los casos donde el resultado del ultra-screening fue negativo.

26.Ausencia Inesperada del Número de Sujeto (Consent=1): Se verificó que el número de sujeto no estuviera presente en los casos donde el paciente no dio su consentimiento.

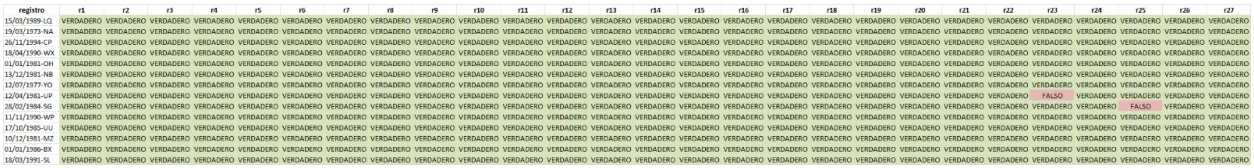
27.Mayoría de Edad del Paciente: Se validó que, para los casos donde se disponía de la fecha de identificación y la fecha de la entrevista, el paciente tuviera la edad legal establecida.

**NOTA:** se adjunta en la entrega un archvio *reglas.xlsx* con la especificaciín de todas las reglas.

# Resultados

Luego de aplicar las reglas de validación a nuestra base, compartimos los siguientes resultados:

- Matriz de validación de tipo Registro (filas) X Regla (columnas)



| registro | r1 | r2 | r3 | r4 | r5 | r6 | r7 | r8 | r9 | r10 | r11 | r12 | r13 | r14 | r15 | r16 | r17 | r18 | r19 | r20 | r21 | r22 | r23 | r24 | r25 | r26 | r27 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15/03/1989-LQ | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 19/03/1973-NA | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 26/11/1994-CP | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 18/04/1990-WX | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 01/01/1981-OH | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 13/12/1981-NB | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 12/07/1977-YO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 12/04/1981-UP | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | FALSO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 28/02/1984-SG | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | FALSO | FALSO | VERDADERO | VERDADERO |
| 11/11/1990-WP | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 17/10/1985-UU | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 10/12/1981-MZ | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 01/01/1996-BX | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |
| 18/03/1991-SL | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO | VERDADERO |

- Número de participantes limpios:

La cantidad de registros sin errores es 759, lo que equivale a un 75,9% de la base.

- Participantes con inconsistencias:

La lista de participantes con inconsistencias es

| Registro | Errores | Reglas fallidas | Descripción |
|---|---|---|---|
| 01/01/1989-KR | 2 | r21, r22 | |
| 01/04/1981-RT | 2 | r21, r22 | |
| 01/04/1981-ZT | 2 | r12, r14 | |
| 01/04/1987-CI | 1 | r25 | (subjectnumber) no debería estar si usscr=1 |
| 01/04/1987-GU | 1 | r26 | (subjectnumber) no debería estar si consent=1 |
| 01/05/1982-UX | 2 | r21, r22 | |
| 01/06/1977-HF | 2 | r21, r22 | |
| 01/06/1992-EJ | 1 | r24 | (subjectnumber) no debería estar si scr=1 |
| 01/09/1990-YH | 1 | r23 | (subjectnumber) coincide con countrycode |
| 01/09/1993-QZ | 2 | r12, r14 | |
| 01/10/1982-YA | 2 | r21, r22 | |
| 01/10/1990-AZ | 2 | r12, r14 | |
| 01/11/1983-EU | 2 | r21, r22 | |
| 02/02/1981-DT | 2 | r21, r22 | |
| 02/02/1982-FB | 1 | r25 | (subjectnumber) no debería estar si usscr=1 |

- Inconsistencias más frecuentes:

Las inconsistencias mas frecuentes son:

1. En primer lugar, la regla 21, que indica *subject number* como dato faltante. El campo cuenta con 100 valores perdidos en nuestra base.
2. Lugo, la regla 22, que verifica que *subject number* es vacio siendo que debia estar completo.
3. Por último, la regla 14, que evalua si el campo *ethnic group* es uno de los valores admitidos.

Presentamos el resto de las reglas con errores en las siguiente tabla:

| Regla | Errores | Descripción |
|---|---|---|
| r21 | 100 | (subjectnumber) es faltante |
| r22 | 97 | (subjectnumber) obligatorio si pasa screening |
| r14 | 51 | (ethnicgroup) fuera de rango |
| r23 | 41 | (subjectnumber) coincide con countrycode |
| r26 | 30 | (subjectnumber) no debería estar si consent=1 |
| r12 | 26 | (ethnicgroup) es faltante |
| r25 | 18 | (subjectnumber) no debería estar si usscr=1 |
| r24 | 10 | (subjectnumber) no debería estar si scr=1 |
| r27 | 3 | paciente mayor de edad |
| r19 | 1 | (consent) es faltante |
| r20 | 1 | (consent) dentro de rango |

- Campos con más errores:

El campo con mas errores de la base es *subject number* con 97 errores.