LICENCIATURA EN ESTADÍSTICA

“Validación de datos”

Trabajo Práctico 1

Autores: Tomás Anderson - Alejo Vaschetti

Docentes: Nora Arnesi - Gino Bartolelli - Cristina Cuesta

20/04/2025

Tabla de contenidos

Introducción

Es difícil saber el desarrollo esperado de un feto a cierta semana de gestación y por esto se decidió realizar un estudio de investigación clínica multicéntrico internacional para poder estimar e implementar estándares mundiales de crecimiento fetal. Con esta información, los médicos de todo el mundo podrían reducir la morbi-mortalidad perinatal asociada con el crecimiento. En el mismo se reclutan mujeres mayores de edad al momento del estudio que estén cursando su primer trimestre de embarazo, las cuales son seguidas con un esquema de visitas programadas 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.

Para hacerlo moral y éticamente correcto, se realiza una exhaustiva recolección de información de las mujeres, llevada a cabo a lo largo de 17 formularios en papel. Dos de estos aseguran que la mujer sea elegible de acuerdo con los formularios CLIN-SCR y US-SCR, además de contar con el consentimiento del paciente. Particularmente, se tiene especial interés en analizar que el formulario de admisión al estudio no contenga errores. Los datos del formulario son los siguientes:

- Código de país: Número de 3 cifras que representa un país.
- ID del paciente: Se compone de la fecha de nacimiento y las iniciales del paciente en el formato "dd/mm/yyyy-AA".
- Fecha de entrevista: Con formato "dd/mm/yyyy".
- Etnia: Número que representa una etnia, donde 1 es "caucásica", 2 es "asiática", 3 es "africana" y 4 es "otros".
- Elegibilidad CLIN-SCR: 2 si cumple con el criterio y 1 si no.
- Elegibilidad US-SCR: 2 si cumple con el criterio y 1 si no.
- Consentimiento dado: 2 si consiente y 1 si no.
- Número de sujeto: Número de 9 cifras compuesto por el código de país, el código del médico y el orden de ingreso. Este campo se llena únicamente si cumple con los criterios de elegibilidad y consiente.

Objetivos

- Definir reglas que permitan identificar errores en los datos cargados de las pacientes.
- Evaluar dichas reglas en el conjunto de datos.
- Calcular medidas sobre los tipos de errores.
- Obtener el número de participantes con formularios completados válidamente.

Resolución

Reglas

Las reglas se pueden clasificar en tres tipos de errores no mutuamente excluyentes:

- Existencia: El valor es faltante o existe cuando no debería serlo.
- Rango: Toma un valor distinto a los válidos.
- Consistencia: El valor no es congruente respecto del valor de otro/s campo/s.

Con todo esto en consideración se plantean las siguientas reglas con el objetivo de identificar todo tipo de error posible en la carga de datos:

Tabla 1: Definición de reglas

Nombre de la regla	Campos involucrados	Descripción	Representación lógica	Clasificación
A1	Código de país (a)	El código de país fue cargado	is.na(a)	Existencia
A2	Código de país (a)	El código de país es un número entre los posibles	!is.na(a) & is.numeric(a) & a {4, 11, 13, 23, 31, 48, 54, 65, 72, 97}	Rango
B1	Id de paciente (b)	El id de paciente fue cargado	is.na(b)	Existencia
B2	Id de paciente (b)	El formato del id de paciente es correcto	Ver anexo*	Rango
B3	Id de paciente (b)	La fecha de nacimiento del paciente es válida	Ver anexo*	Rango

Nombre de la regla	Campos involucrados	Descripción	Representación lógica	Clasificación
C1	Fecha de la entrevista (c)	La fecha fue cargada	is.na(c)	Existencia
C2	Fecha de la entrevista (c)	El formato de la fecha cargada es correcto	Ver anexo*	Rango
C3	Fecha de la entrevista (c)	La fecha de la entrevista es válida	Ver anexo*	Rango
BC1	Id de paciente (b) Fecha de la entrevista (c)	La paciente es mayor de edad y está en una edad fértil (menor a 51 años)	18 <= year(b - c) <= 50	Consistencia
D1	Grupo étnico (d)	El grupo étnico fue cargado	is.na(d)	Existencia
D2	Grupo étnico (d)	El grupo étnico es un número entre los posibles	!is.na(d) & is.numeric(d) & d {1, 2, 3, 4}	Rango
E1	Formulario CLIN-SCR (e)	El formulario fue cargado	is.na(e)	Existencia
E2	Formulario CLIN-SCR (e)	La elegibilidad por el formulario es un número entre los posibles	!is.na(e) & is.numeric(e) & x {1, 2}	Rango
F1	Formulario US-SCR (f)	El formulario fue cargado	is.na(f)	Existencia

Nombre de la regla	Campos involucrados	Descripción	Representación lógica	Clasificación
F2	Formulario US-SCR (f)	La elegibilidad por el formulario es un número entre los posibles	!is.na(f) & is.numeric(f) & x {1, 2}	Rango
G1	Formulario de consentimiento (g)	El formulario fue cargado	is.na(f)	Existencia
G2	Formulario de consentimiento (g)	El consentimiento por el formulario es un número entre los posibles	!is.na(f) & is.numeric(f) & x {1, 2}	Rango
H1	Número de sujeto (h) CLIN-SCR (e) US-SCR (f) Consentimiento (g)	El número de sujeto debería ser cargado pero no lo fue	is.na(h) & c(e, f, g) == c(2, 2, 2)	Existencia y consistencia
H2	Número de sujeto (h) CLIN-SCR (e) US-SCR (f) Consentimiento (g)	El número de sujeto no debería ser asignado y cargado pero lo fue	!is.na(h) & c(e, f, g) != c(2, 2, 2)	Existencia y consistencia
H3	Número de sujeto (h)	El número de sujeto está conformado por 9 números	!is.na(h) & is.numeric(h) & floor(log10(h)) == 8	Rango
H4	Número de sujeto (h)	El número de sujeto es único	!is.na(h) & sum(h == H) == 1	Consistencia

Nombre de la regla	Campos involucrados	Descripción	Representación lógica	Clasificación
H5	Número de sujeto (h) Código de país (a)	Los primeros tres dígitos del número del sujeto corresponden con el código de país cargado de la paciente	Ver anexo*	Consistencia

Evaluación de las reglas

Se crea un algoritmo para evaluar el cumplimiento de las reglas anteriormente planteadas, identificar a los individuos con errores y clasificar qué tipo de error se está cometiendo. Se tiene la información del formulario de admisión de 1000 pacientes con las que se procede a utilizar dicho algoritmo.

Tabla 2: Cantidad de errores por regla

A1	A2	B1	B2	B3	C1	C2	C3	BC1	D1	D2	E1	E2	F1	F2	G1	G2	H1	H2	H3	H4	H5
0	0	0	0	0	0	0	25	0	26	51	0	0	0	0	1	0	97	56	0	0	41

El 59% de las reglas planteadas no fueron incumplidas, en cambio, para las que sí lo fueron se registra un alto número de errores. El error más común se encuentra en la falta del *número de sujeto* del paciente cuando este cumplía todos los rquisitos de elegibilidad y consentimiento. Este fallo puede ser dado por error en no asignar un *número de sujeto* cuando se debería o que algunos de los requisitos anteriores no se cumplía y por lo tanto fueron mal cargados.

Sumando los errores de cada campo y separándolos por el tipo de error se obtiene:

Tabla 3: Errores clasificados

Campo	Existencia	Rango	Consistencia
Código de país	0	0	41
Id del paciente	0	0	0
Fecha de la entrevista	0	25	0
Grupo étnico	26	51	-
CLIN-SCR	0	0	107
US-SCR	0	0	115
Consentimiento	1	0	127
Número de sujeto	153	0	194

Todos estos fallos llevan a 336 pacientes a tener algún tipo de inconsistencias en los datos cargados del formulario de amisión, dejando un total de 664 pacientes con sus datos en orden. El campo con más errores resulta ser el *número de sujeto* con un total de 194. Como estos están relacionados con el valor en otros campos, la cantidad de errores en estos otros se ve aumentada.

Conclusiones

Un 33,6% de las pacientes presentaron al menos un error en los datos cargados del formualario de admisión. Esto representa un gran problema ya que puede ser una cuestión sistemática. Dado los errores más comunes, se recomienda hacer un esfuerzo mayor en asignar y cargar bien el grupo étnico de las pacientes, prestar mayor atención al cargar el año de la entrevista y hacer mayor incapié al asignar el número del sujeto únicamente en los casos requeridos, por lo tanto debería tenerse mayor cuidado con la certificación de los formularios de elegibilidad.

Anexo

Hay representaciones lógicas que son complicadas de describir y dependen del software que se use por lo tanto el código usado para la realización de este trabajo se encuentra en los archivos codigo.py y TPBIO.qmd en el siguiente link:

https://github.com/AlejoVaschetti/TP_Bio